IBM SPSS Modeler Text Analytics
Guide de l'utilisateur version 15

IBM®

Remarque : Avant d'utiliser ces informations et le produit concerné, lisez les informations générales de la rubrique Avis sur p. 387.

Cette version concerne IBM® SPSS® Modeler Text Analytics 15 et toutes les versions et modifications ultérieures sauf mention contraire dans les nouvelles versions.

Les captures d'écran de produits Adobe sont réimprimées avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran de produits Microsoft sont réimprimées avec l'autorisation de Microsoft Corporation.

Préface

IBM® SPSS® Modeler Text Analytics propose de puissantes fonctionnalités d'analyse de texte qui utilisent des technologies linguistiques avancées et le traitement du langage naturel (NLP - Natural Language Processing) pour traiter rapidement une grande variété de données textuelles non structurées et, à partir de ce texte, extraire et organiser les concepts clés. De plus, SPSS Modeler Text Analytics peut regrouper ces concepts par catégories.

Environ 80 % des données d'une société se présentent sous la forme de documents texte (rapports, pages Web, messages électroniques et notes de centre d'appel, etc.). Pour qu'une entreprise soit en mesure de mieux comprendre le comportement de ses clients, les données textuelles représentent un facteur essentiel. Les systèmes dotés du traitement du langage naturel extraient les concepts de manière intelligente, y compris les mots composés. En outre, grâce à la maîtrise du langage sous-jacent, ils classent les termes en groupes d'informations similaires (produits, entreprises ou personnes, par exemple), s'aidant du sens et du contexte. Par conséquent, vous pouvez rapidement savoir si les informations du document présentent un intérêt pour vous. Ces concepts extraits et les catégories peuvent ensuite être combinés aux données structurées existantes, telles que les données démographiques, et appliqués à la modélisation grâce à la gamme complète d'outils de Data mining de IBM® SPSS® Modeler, afin de favoriser une prise de décision précise et efficace.

Les systèmes linguistiques dépendent des connaissances disponibles : plus leurs dictionnaires contiennent d'informations, plus les résultats obtenus sont probants. SPSS Modeler Text Analytics est livré avec un ensemble de ressources linguistiques, comme les dictionnaires de termes et de synonymes, les bibliothèques et les modèles. Ce produit vous permet également d'approfondir ces ressources linguistiques et de les adapter à votre contexte. La mise au point des ressources linguistiques est souvent un processus itératif nécessaire pour assurer avec précision l'extraction et la catégorisation des concepts. Enfin, des modèles personnalisés, des bibliothèques et des dictionnaires spécialisés dans des domaines précis, tels que la gestion de la relation client et la génomique, sont fournis.

À propos d'IBM Business Analytics

Le logiciel IBM Business Analytics propose des informations complètes, cohérentes et précises que les preneurs de décisions peuvent utiliser en toute confiance pour améliorer les performances de leur entreprise. Un portefeuille complet de veille économique, d'analyses prédictives, de gestion des performances et de stratégie financière et d'applications analytiques qui propose des informations claires, immédiates et décisionnelles sur les performances actuelles permettant de prévoir les résultats à venir. Ce logiciel allie des solutions dédiées à l'industrie, des pratiques ayant fait leur preuve et des services professionnels afin que les organisations de toute taille puissent obtenir la meilleure productivité possible, automatiser leurs décisions en toute confiance et améliorer leurs résultats.

Ce portefeuille contient le logiciel IBM SPSS Predicitive Analytics qui aide les organisations à prévoir les événements à venir et à agir en fonction de ces informations pour améliorer leurs résultats. Les clients de l'industrie du commerce et des domaines gouvernementaux et de l'éducation du monde entier font confiance à la technologie IBM SPSS et reconnaissent son avantage concurrentiel pour attirer et fidéliser les clients, pour grossir la base de clientèle tout en réduisant la fraude et en limitant les risques. En intégrant le logiciel IBM SPSS à leurs

opérations quotidiennes, les organisations deviennent des entreprises prédictives, capables de diriger et d'automatiser leurs décisions afin de répondre à leurs objectifs et d'obtenir des avantages concurrentiels mesurables. Pour des informations supplémentaires ou pour joindre un représentant, consultez le site *http://www.ibm.com/spss*.

Assistance technique

L'assistance technique est disponible pour les clients de la maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour joindre l'assistance technique, consultez le site Web IBM Corp. à l'adresse *http://www.ibm.com/support*. Lorsque vous demandez une assistance, veillez à être en mesure de pouvoir vous identifier ainsi que votre organisation et votre contrat d'assistance.

Contenu

Partie II: Session interactive

8 Mode Session interactive 116

9 Extraction de concepts et de types 139

10 Catégorisation des données textuelles 162

Partie I:
Nœuds Text Mining

Chapitre

1

A propos de IBM SPSS Modeler Text Analytics

IBM® SPSS® Modeler Text Analytics propose de puissantes fonctionnalités d'analyse de texte qui utilisent des technologies linguistiques avancées et le traitement du langage naturel (NLP - Natural Language Processing) pour traiter rapidement une grande variété de données textuelles non structurées et, à partir de ce texte, extraire et organiser les concepts clés. De plus, SPSS Modeler Text Analytics peut regrouper ces concepts par catégories.

Environ 80 % des données d'une société se présentent sous la forme de documents texte (rapports, pages Web, messages électroniques et notes de centre d'appel, etc.). Pour qu'une entreprise soit en mesure de mieux comprendre le comportement de ses clients, les données textuelles représentent un facteur essentiel. Les systèmes dotés du traitement du langage naturel extraient les concepts de manière intelligente, y compris les mots composés. En outre, grâce à la maîtrise du langage sous-jacent, ils classent les termes en groupes d'informations similaires (produits, entreprises ou personnes, par exemple), s'aidant du sens et du contexte. Par conséquent, vous pouvez rapidement savoir si les informations du document présentent un intérêt pour vous. Ces concepts extraits et les catégories peuvent ensuite être combinés aux données structurées existantes, telles que les données démographiques, et appliqués à la modélisation grâce à la gamme complète d'outils de Data mining de IBM® SPSS® Modeler, afin de favoriser une prise de décision précise et efficace.

Les systèmes linguistiques dépendent des connaissances disponibles : plus leurs dictionnaires contiennent d'informations, plus les résultats obtenus sont probants. SPSS Modeler Text Analytics est livré avec un ensemble de ressources linguistiques, comme les dictionnaires de termes et de synonymes, les bibliothèques et les modèles. Ce produit vous permet également d'approfondir ces ressources linguistiques et de les adapter à votre contexte. La mise au point des ressources linguistiques est souvent un processus itératif nécessaire pour assurer avec précision l'extraction et la catégorisation des concepts. Enfin, des modèles personnalisés, des bibliothèques et des dictionnaires spécialisés dans des domaines précis, tels que la gestion de la relation client et la génomique, sont fournis.

Déploiement. Vous pouvez déployer les flux de text mining à l'aide de IBM® SPSS® Modeler Solution Publisher pour le scoring en temps réel des données non structurées. Cette possibilité permet de garantir une mise en oeuvre réussie d'opérations de Text Mining en boucles fermées. Par exemple, votre organisation peut désormais analyser de manière plus pertinente les notes consignées dans le Bloc-notes, issues des appelants entrants et sortants, en appliquant vos modèles prédictifs. Votre communication marketing en temps réel est ainsi mieux adaptée.

Remarque : Pour exécuter SPSS Modeler Text Analytics avec SPSS Modeler Solution Publisher, ajoutez le répertoire `<install_directory>/ext/bin/spss.TMWBServer` à la variable d'environnement `$LD_LIBRARY_PATH`.

Traduction automatique des langues prises en charge. Associé à Software as a Service (SaaS) de SDL, SPSS Modeler Text Analytics vous permet de traduire en anglais un texte rédigé dans l'une des langues prises en charge, parmi lesquelles l'arabe, le chinois et le persan. Vous pouvez ensuite effectuer l'analyse du texte traduit et transmettre les résultats obtenus aux utilisateurs qui n'auraient pas pu en comprendre le contenu dans la langue source. Les résultats du Text Mining étant automatiquement reliés au texte rédigé en langue étrangère correspondant, votre organisation peut alors attirer l'attention des locuteurs natifs sur les résultats les plus significatifs de l'analyse. SDL propose une fonction de traduction automatique basée sur des algorithmes de traduction statistiques, qui sont le fruit de 20 années de recherches approfondies dans le domaine de la traduction.

Mise à niveau vers la version 15 de IBM SPSS Modeler Text Analytics

Mise à niveau à partir des versions précédentes de PASW Text Analytics ou de Text Mining for Clementine

Avant d'installer IBM® SPSS® Modeler Text Analytics version 15, enregistrez et exportez tous les TAP, modèles et bibliothèques de votre version actuelle que vous souhaitez utiliser dans la nouvelle version. Nous vous recommandons d'enregistrer ces fichiers dans un répertoire qui ne sera pas effacé ou remplacé lors de l'installation de la nouvelle version.

Après avoir installé la dernière version de SPSS Modeler Text Analytics , vous pouvez charger le fichier TAP enregistré, ajouter toutes les bibliothèques enregistrées ou importer et charger les modèles enregistrés pour les utiliser dans la dernière version.

Important ! Si vous désinstallez votre version actuelle sans d'abord enregistrer ou exporter les fichiers nécessaires, tout le travail effectué sur les TAP, modèles et bibliothèques publiques dans la version précédente sera perdu et ne pourra pas être utilisé dans SPSS Modeler Text Analytics version 15.

A propos de Text Mining

De nos jours, de plus en plus d'informations sont stockées dans des formats non structurés et partiellement structurés (messages électroniques de clients, notes de centre d'appel, réponses ouvertes à des enquêtes, actualités, formulaires Web, etc.). Ce flot d'informations pose problème à de nombreuses organisations qui souhaitent trouver la méthode leur permettant de collecter, d'étudier et d'exploiter ces informations.

Le processus de **Text Mining** consiste à analyser des ensembles de documents textuels afin de capturer les concepts et thèmes-clés, et de découvrir les relations et les tendances cachées. Il ne nécessite pas que vous connaissiez les mots ou les termes précis utilisés par les auteurs pour exprimer ces concepts. Bien qu'il s'agisse de processus très différents, le Text Mining est parfois confondu avec la récupération d'informations. Si l'extraction et le stockage précis des informations représentent un défi considérable, l'extraction et la gestion efficaces du contenu, de la terminologie et des relations compris dans ces informations jouent un rôle vital.

Text Mining et Data Mining

Pour chaque élément du texte, le système de Text Mining linguistique renvoie un index de concepts, ainsi que des informations à propos de ces concepts. Ces informations simplifiées et structurées peuvent être combinées à d'autres sources de données afin de répondre aux questions du type :

- Quels concepts sont associés ?
- A quel autre élément sont-ils liés ?
- Quelles sont les catégories de niveau supérieur pouvant découler des informations extraites ?
- Quels résultats les catégories ou les concepts permettent-ils de prédire ?
- De quelle façon les catégories ou les concepts prédisent-ils les comportements ?

Par une utilisation conjointe de Text Mining et de Data mining, vous obtenez des résultats plus probants que sur la base des données structurées ou non structurées seules. Ce processus comprend généralement les étapes suivantes :

1. **Identification du texte à explorer.** Préparation du texte avant exploration. Si le texte apparaît dans plusieurs fichiers, enregistrez-les tous au même endroit. Dans le cas de bases de données, déterminez le champ contenant le texte.
2. **Exploration du texte et extraction des données structurées.** Appliquez les algorithmes de Text Mining au texte source.
3. **Création des modèles de concepts et de catégories.** Identifiez les principaux concepts et/ou créez des catégories. Généralement, le système renvoie de nombreux concepts à partir de données non structurées. Identifiez les meilleurs concepts et catégories en vue de scoring des catégories.
4. **Analyse des données structurées.** Utilisez les techniques standard du Data mining (comme le clustering, la classification et la modélisation prédictive) pour connaître les relations unissant les concepts. Fusionnez les concepts extraits avec d'autres données structurées afin de prévoir le comportement sur la base des concepts.

Analyse de texte et catégorisation

L'analyse de texte, sorte d'analyse qualitative, est l'extraction d'informations utiles d'un texte, de manière à regrouper les principaux concepts ou idées qui figurent dans ce texte dans un nombre approprié de catégories. Vous pouvez effectuer une analyse de texte sur tout type et toute longueur de texte, bien que l'approche analytique varie quelque peu.

Etant donné que les enregistrements ou les documents courts sont moins complexes et contiennent généralement moins de mots et de réponses ambigus, leur catégorisation est plus simple. Par exemple, si nous posons des questions ouvertes et courtes au cours d'une enquête sur les trois activités préférées des personnes interrogées lorsqu'elles sont en vacances, leurs réponses seront pour la plupart courtes : *aller à la plage*, *visiter des parcs nationaux* ou *ne rien faire*. Des réponses ouvertes plus longues risquent, par contre, d'être plutôt complexes et démesurées, en particulier si les personnes interrogées sont instruites, motivées et qu'elles disposent de suffisamment de temps pour remplir un questionnaire. Si nous interrogeons des personnes sur leurs opinions politiques dans le cadre d'une enquête ou si nous mettons au point un

blog concernant la politique, nous nous attendons à recevoir de très longs commentaires sur une grande variété de problèmes et de prises de position.

La possibilité d'extraire les principaux concepts et de créer des catégories avec pertinence à partir de ces longues sources textuelles en très peu de temps est un avantage-clé de l'utilisation de IBM® SPSS® Modeler Text Analytics . Pour obtenir les résultats les plus fiables à chacune des étapes du processus d'analyse de texte, des techniques statistiques et linguistiques automatiques sont associées.

Traitement linguistique et traitement du langage naturel

Le principal problème lié à la gestion de ces données textuelles non structurées est l'absence de règles standard de rédaction permettant aux ordinateurs de comprendre les textes. La langue, et par conséquent le sens des mots, varie d'un document à l'autre et même au sein d'un même document. Pour pouvoir récupérer et organiser efficacement ces données non structurées, vous devez analyser la langue et découvrir la signification du texte. Il existe plusieurs méthodes automatisées permettant l'extraction des concepts d'informations non structurées. Ces méthodes peuvent être réparties en deux types : linguistiques et non linguistiques.

Certaines entreprises ont tenté d'employer des solutions non linguistiques automatisées basées sur des statistiques et des réseaux de neurones. Grâce aux technologies informatiques, ces solutions permettent d'analyser et de catégoriser les principaux concepts plus rapidement qu'un être humain. Le degré de précision de ces solutions est malheureusement relativement faible. La plupart des systèmes basés sur les statistiques comptent simplement le nombre d'occurrences des mots et calculent leur proximité statistique vis-à-vis des concepts associés. Ils produisent un grand nombre de résultats non pertinents (« bruit ») et passent à côté de ceux qu'ils doivent trouver. On parle alors de « silence ».

Pour compenser leur précision limitée, certaines solutions intègrent des règles non linguistiques complexes permettant de distinguer les résultats pertinents des résultats non pertinents. Cette technique est appelée *Text Mining basé sur des règles*.

Le *Text mining linguistique* applique quant à lui les principes du traitement du langage naturel (NLP)— l'analyse des langues humaines assistée par ordinateur—), analyse des langues assistée par ordinateur, à l'analyse des mots, des expressions et de la syntaxe, ou de la structure, du texte. Les systèmes dotés du traitement du langage naturel extraient les concepts de manière intelligente, y compris les mots composés. En outre, grâce à la maîtrise du langage sous-jacent, ils classent les concepts en groupes d'informations similaires (produits, organisations ou personnes, par exemple), s'aidant du sens et du contexte.

Le Text Mining linguistique recherche le sens du texte comme le font les êtres humains, en reconnaissant divers types de mot comme ayant un sens similaire et en analysant la structure de la phrase pour établir le cadre de compréhension du texte. Tout en garantissant la rapidité et la rentabilité des systèmes statistiques, cette méthode offre un degré de précision nettement supérieur et exige une intervention considérablement moindre de l'utilisateur.

Pour illustrer la différence entre la méthode statistique et la méthode linguistique pendant le processus d'extraction dans toutes les langues, à l'exception du japonais, examinons le mode d'action de chacune de ces méthodes dans le cadre d'une requête concernant l'expression `reproduction de documents`. La solution statistique et la solution linguistique doivent toutes les deux étendre le mot `reproduction` à ses synonymes (`copie` et `duplication`, par exemple). Sinon, des informations pertinentes risquent d'être ignorées. Si toutefois une solution

statistique tente d'appliquer ce type de synonymie (recherche d'autres termes possédant la même signification), elle inclura vraisemblablement le terme `naissance`, générant ainsi de nombreux résultats inappropriés. Comme la compréhension de la langue permet de lever toute ambiguïté dans le texte, l'exploration de texte linguistique reste par définition la méthode la plus fiable.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Pour du texte en japonais, la différence entre les approches basées sur les statistiques et celles basées sur la linguistique pendant l'extraction peut être illustrée à l'aide du mot 沈む comme exemple. A l'aide de ce mot, nous pouvons rechercher des expressions telles que 日が沈む, traduites par *le soleil se couche*, ou 気分が沈む, traduite par *avoir le blues*. Si vous utilisez uniquement les techniques de statistiques, 日 (traduit par *soleil*), 気分 (traduit par *ressentir*), et 沈む (traduit par *en bas*) sont extraits séparément. Mais si vous utilisez l'analyseur de sentiment, qui utilise des techniques linguistiques, 日, 気分, et 沈む sont extraits, mais 気分が沈む (traduit comme *avoir le blues*) est également extrait et affecté au type <悪い － 悲しみ全般>. L'utilisation de techniques basées sur la linguistique à l'aide de l'analyseur de sentiment permet d'extraire plus d'expressions significatives. L'analyse et la capture des émotions permettent de lever toute ambiguïté dans le texte, et fait de l'exploration de texte linguistique la méthode la plus fiable, par définition.

Si vous comprenez le fonctionnement du processus d'extraction, vous êtes plus à même de prendre les décisions-clés lorsque vous affinez vos ressources linguistiques (bibliothèques, types, synonymes, etc.). Les principales étapes du processus d'extraction sont les suivantes :

- Conversion des données source en un format standard
- Identification des termes susceptibles d'être extraits
- Identification des classes d'équivalence et intégration des synonymes
- Affectation d'un type
- Indexation et, si nécessaire, mise en correspondance de patrons avec un deuxième analyseur

Etape 1. Conversion des données source en un format standard

Au cours de cette première étape, les données que vous importez sont converties dans un format uniforme pouvant être utilisé pour effectuer d'autres analyses. Cette conversion, qui s'effectue en interne, ne modifie pas les données d'origine.

Etape 2. Identification des termes susceptibles d'être extraits

Il est important de comprendre le rôle des ressources linguistiques dans l'identification des termes susceptibles d'être extraits lors de l'extraction linguistique. Les ressources linguistiques sont utilisées lors de chaque exécution d'une extraction. Elles se présentent sous la forme de ressources compilées, de bibliothèques et de modèles. Les bibliothèques comportent des listes de mots, des relations et des informations complémentaires qui permettent de spécifier ou d'affiner l'extraction. Vous ne pouvez pas afficher ni éditer les ressources compilées. Toutefois, les autres ressources peuvent être modifiées dans l'Editeur de modèle ou, si vous êtes dans une session interactive, dans l'Editeur de ressources.

Les ressources compilées sont des composants internes essentiels du moteur du programme d'extraction de SPSS Modeler Text Analytics . Ces ressources comportent un dictionnaire général qui répertorie les formes de base avec un code concernant la catégorie grammaticale (nom, verbe, adjectif, etc.). Les ressources comprennent également des types intégrés et réservés qui permettent d'affecter de nombreux termes extraits aux types suivants : <地名>, <組織>, ou <

人名>. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais dans l'annexe A sur p. 378.*Remarque* : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Outre ces ressources compilées, plusieurs bibliothèques sont fournies avec le produit et peuvent être utilisées pour compléter les types et les définitions de concept figurant dans les ressources compilées, ainsi que pour proposer des synonymes. Ces bibliothèques —et toutes les bibliothèques personnalisées que vous créez— comprennent plusieurs dictionnaires : déclarations de types, dictionnaires de synonymes et dictionnaires d'exclusions. Pour plus d'informations, reportez-vous à la section Modification des ressources pour du texte en japonais dans l'annexe A sur p. 374.*Remarque* : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Une fois les données importées et converties, le moteur du programme d'extraction commence à identifier les termes susceptibles d'être extraits. Ces termes sont des mots ou des groupes de mots qui permettent d'identifier des concepts du texte. Pendant le traitement du texte, les mots uniques (**unitermes**) et les mots composés (**multitermes**) sont identifiés à l'aide d'extracteurs de patrons de catégorie grammaticale. Par exemple, l'expression multiterme 青森りんご, qui suit le patron de catégorie grammaticale <地名> + <名詞> possède deux composants. Par conséquent, les mots-clés de sentiment susceptibles d'être extraits sont identifiés à l'aide de l'analyse des liens du texte de sentiment.

Imaginons par exemple que vous ayez le texte suivant en japonais : 写真が新鮮で良かった. Dans ce cas, le moteur d'extraction affecte le type de sentiment 良い - 褒め・賞賛, après avoir mis en correspondance（品物） + が + 良い à l'aide des règles de liens du texte de sentiment.

Remarque : les termes qui figurent dans le dictionnaire général compilé susmentionné constituent la liste de tous les mots qui risquent de s'avérer inintéressants ou de présenter une ambiguïté linguistique en tant qu'unitermes. Ces mots sont exclus de l'extraction lorsque vous identifiez les unitermes. Ils font toutefois l'objet d'une réévaluation lorsque vous déterminez les catégories grammaticales ou que vous recherchez des mots composés (expressions multitermes) plus longs, susceptibles d'être extraits.

Etape 3. Identification des classes d'équivalence et intégration des synonymes

Une fois les expressions unitermes et multitermes susceptibles d'être extraites identifiées, le logiciel utilise un dictionnaire de normalisation afin d'identifier des classes d'équivalence. Une classe d'équivalence désigne la forme de base d'une phrase ou la forme unique de deux variantes d'une même phrase. L'affectation d'expression à des classes d'équivalence a pour objectif de veiller à ce que, par exemple, `effet secondaire` et 副作用 ne soient pas traités comme des concepts distincts. Pour déterminer quel concept utiliser pour la classe d'équivalence (c'est-à-dire si `effet secondaire` ou 副作用 est utilisé en tant que terme principal), le moteur du programme d'extraction applique les règles suivantes dans l'ordre indiqué ci-dessous :

- Forme définie par l'utilisateur dans une bibliothèque.
- La forme la plus fréquente, comme définie par les ressources précompilées.

Etape 4. Affectation d'un type

Des types sont ensuite affectés aux concepts extraits. Un type correspond à un regroupement sémantique de concepts. Les ressources compilées et les bibliothèques sont utilisées au cours de cette étape. Les types comprennent des éléments tels que des concepts de niveau supérieur, des

mots positifs et négatifs, des prénoms, des lieux, des organisations, etc. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

Les ressources en japonais contiennent un ensemble de types distinct. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais dans l'annexe A sur p. 378.*Remarque* : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Les performances des systèmes linguistiques dépendent des connaissances disponibles : plus leurs dictionnaires contiennent d'informations, plus les résultats obtenus sont probants. Modifier le contenu du dictionnaire, les définitions de synonyme par exemple, permet de simplifier les informations obtenues. Souvent itératif, ce processus est nécessaire pour obtenir une extraction précise des concepts. Le traitement du langage naturel est un élément fondamental de SPSS Modeler Text Analytics .

Fonctionnement de l'extraction

Lors de l'extraction des principaux concepts et idées de vos réponses, IBM® SPSS® Modeler Text Analytics s'appuie sur une analyse de texte linguistique. Cette approche a la même efficacité en temps et en argent que les systèmes statistiques. Mais elle offre un plus grand degré de précision tout en ne nécessitant que peu d'intervention humaine. L'analyse de texte linguistique se base sur un domaine d'étude appelé processus de langage naturel, également connu sous le nom de linguistique computationnelle.

Important ! Pour le texte en japonais, le processus d'extraction suit des étapes différentes. Pour plus d'informations, reportez-vous à la section Fonctionnement de l'extraction dans l'annexe A sur p. 368.*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Si vous comprenez le fonctionnement du processus d'extraction, vous êtes plus à même de prendre les décisions-clés lorsque vous affinez vos ressources linguistiques (bibliothèques, types, synonymes, etc.). Les principales étapes du processus d'extraction sont les suivantes :

- Conversion des données source en un format standard
- Identification des termes susceptibles d'être extraits
- Identification des classes d'équivalence et intégration des synonymes
- Affectation d'un type
- Indexation
- Mise en correspondance de l'extraction des patrons et des événements

Etape 1. Conversion des données source en un format standard

Au cours de cette première étape, les données que vous importez sont converties dans un format uniforme pouvant être utilisé pour effectuer d'autres analyses. Cette conversion, qui s'effectue en interne, ne modifie pas les données d'origine.

Etape 2. Identification des termes susceptibles d'être extraits

Il est important de comprendre le rôle des ressources linguistiques dans l'identification des termes susceptibles d'être extraits lors de l'extraction linguistique. Les ressources linguistiques sont utilisées lors de chaque exécution d'une extraction. Elles se présentent sous la forme de

ressources compilées, de bibliothèques et de modèles. Les bibliothèques comportent des listes de mots, des relations et des informations complémentaires qui permettent de spécifier ou d'affiner l'extraction. Vous ne pouvez pas afficher ni éditer les ressources compilées. Toutefois, les autres ressources (modèles) peuvent être modifiées dans l'Editeur de modèle ou, si vous êtes dans une session interactive, dans l'Editeur de ressources.

Les ressources compilées sont des composants internes essentiels du moteur du programme d'extraction de IBM® SPSS® Modeler Text Analytics . Ces ressources comportent un dictionnaire général qui liste les formes de base avec un code concernant la catégorie grammaticale (nom, verbe, adjectif, adverbe, participe, élément de coordination, déterminant ou préposition). Les ressources comprennent également des types intégrés et réservés qui permettent d'affecter de nombreux termes extraits aux types suivants : `<Location>`, `<Organization>`, ou `<Person>`. Pour plus d'informations, reportez-vous à la section Types intégrés dans le chapitre 17 sur p. 306.

Outre ces ressources compilées, plusieurs bibliothèques sont fournies avec le produit et peuvent être utilisées pour compléter les types et les définitions de concept figurant dans les ressources compilées, ainsi que pour proposer d'autres types et synonymes. Ces bibliothèques —et toutes les bibliothèques personnalisées que vous créez— comprennent plusieurs dictionnaires : déclarations de types, dictionnaires de substitutions (synonymes et éléments optionnels) et dictionnaires d'exclusions. Pour plus d'informations, reportez-vous à la section Utilisation des bibliothèques dans le chapitre 16 sur p. 290.

Une fois les données importées et converties, le moteur du programme d'extraction commence à identifier les termes susceptibles d'être extraits. Ces termes sont des mots ou des groupes de mots qui permettent d'identifier des concepts du texte. Lors du traitement du texte, les mots uniques (**unitermes**) qui ne figurent pas dans les ressources compilées sont considérés comme étant des termes susceptibles d'être extraits. Les mots composés (**multitermes**) susceptibles d'être extraits sont identifiés à l'aide d'extracteurs de patrons de catégorie grammaticale. Par exemple, l'expression multiterme `belle voiture`, qui suit le patron de catégorie grammaticale adjectif nom, possède deux composants. L'expression multiterme `belle petite voiture`, qui suit le patron de catégorie grammaticale "adjectif adjectif nom", possède trois composants.

Remarque : les termes qui figurent dans le dictionnaire général compilé susmentionné constituent la liste de tous les mots qui risquent de s'avérer inintéressants ou de présenter une ambiguïté linguistique en tant qu'unitermes. Ces mots sont exclus de l'extraction lorsque vous identifiez les unitermes. Ils font toutefois l'objet d'une réévaluation lorsque vous déterminez les catégories grammaticales ou que vous recherchez des mots composés (expressions multitermes) plus longs, susceptibles d'être extraits.

Enfin, un algorithme spécial est appliqué pour traiter les chaînes en majuscules (intitulés de postes, par exemple), de telle sorte que ces patrons puissent être extraits.

Etape 3. Identification des classes d'équivalence et intégration des synonymes

Une fois les expressions multitermes et les unitermes susceptibles d'être extraits identifiés, le logiciel utilise un ensemble d'algorithmes pour les comparer et identifier des classes d'équivalence. Une classe d'équivalence désigne la forme de base d'une phrase ou la forme unique de deux variantes d'une même phrase. L'affectation de phrases à des classes d'équivalence a pour objectif de veiller à ce que, par exemple, `président de l'entreprise` et `président d'entreprise` ne soient pas traités comme des concepts distincts. Pour déterminer le concept à utiliser pour la classe d'équivalence — (à savoir, si `président de l'entreprise` ou

`président d'entreprise` est utilisé en tant que terme principal), le moteur du programme d'extraction applique les règles suivantes dans l'ordre indiqué ci-dessous :

- Forme définie par l'utilisateur dans une bibliothèque.
- Forme la plus fréquente dans l'ensemble du corps du texte.
- Forme la plus courte dans l'ensemble du corps du texte (ce qui correspond généralement à la forme de base).

Etape 4. Affectation d'un type

Des types sont ensuite affectés aux concepts extraits. Un type correspond à un regroupement sémantique de concepts. Les ressources compilées et les bibliothèques sont utilisées au cours de cette étape. Les types comprennent des éléments tels que des concepts de niveau supérieur, des mots positifs et négatifs, des prénoms, des lieux, des organisations, etc. Vous pouvez définir d'autres types. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

Etape 5. Indexation

L'indexation de l'ensemble des documents ou des enregistrements s'effectue en définissant un pointeur entre une position de texte et le terme représentatif de chaque classe d'équivalence. Cela suppose que toutes les instances de forme fléchies d'un concept susceptible d'être extrait sont indexées en tant que forme de base susceptible d'être extraite. La fréquence globale est calculée pour chaque forme de base.

Etape 6. Mise en correspondance de l'extraction des patrons et des événements.

IBM SPSS Modeler Text Analytics peut non seulement détecter les types et les concepts, mais également les relations qui existent entre eux. Plusieurs algorithmes et bibliothèques sont fournis avec ce produit ; ils permettent d'extraire les patrons de relations entre les types et les concepts. Ils s'avèrent particulièrement utiles lorsque vous tentez de détecter des opinions spécifiques (relations entre des produits, par exemple) ou les liens relationnels entre des personnes ou des objets (liens entre des groupes politiques ou des génomes, par exemple).

Fonctionnement de la catégorisation

Lorsque vous créez des modèles de catégories dans IBM® SPSS® Modeler Text Analytics , vous disposez de plusieurs techniques de création de catégories. Etant donné que chaque ensemble de données est unique, le nombre de techniques et leur ordre d'application peuvent varier. Votre interprétation des résultats pouvant être différente de celle d'une autre personne, vous pouvez être amené à essayer plusieurs techniques de manière à déterminer celle qui donne les meilleurs résultats pour vos données textuelles. Dans SPSS Modeler Text Analytics , vous pouvez créer des modèles de catégories dans une session interactive qui vous permettra d'explorer et d'affiner vos catégories.

Dans ce manuel, la **création de catégories** fait référence à la génération de définitions et de classification de catégories à l'aide d'une ou de plusieurs techniques intégrées et la **catégorisation** fait référence au scoring, ou à l'étiquetage, processus par lequel des identificateurs uniques (nom/ID/valeur) sont affectés aux définitions de catégorie de chaque enregistrement ou de chaque document.

Pendant la création de catégories, les concepts et les types qui ont été extraits sont utilisés en tant que blocs de construction de vos catégories. Lorsque vous créez des catégories, les documents ou les enregistrements sont automatiquement affectés aux catégories s'ils contiennent du texte qui correspond à un élément d'une définition de catégorie.

SPSS Modeler Text Analytics vous propose plusieurs techniques automatisées de classification supervisée qui vous permettent de catégoriser rapidement vos documents ou vos enregistrements.

Techniques de regroupement

Chaque technique disponible convient à certains types de données et de situation. Cependant, il est souvent judicieux de combiner des techniques dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. Aussi est-il possible de voir un concept figurer dans plusieurs catégories ou de rencontrer des catégories redondantes.

Dérivation des racines de concept. Cette technique crée des catégories en sélectionnant un concept et en recherchant les autres concepts y étant associés. Pour cela, elle analyse si certains composants du concept sont liés d'un point de vue morphologique ou s'ils partagent des racines. Cette technique est très utile pour l'identification de concepts de mots composés synonymes, car les concepts de chaque catégorie générée sont synonymes ou leur signification est proche. Elle utilise des données de différentes longueurs et génère un nombre inférieur de catégories compactes. Par exemple, le concept `opportunités d'avancer` serait regroupé avec les concepts `opportunité d'avancement` et `opportunité d'un avancement`. Pour plus d'informations, reportez-vous à la section Dérivation des racines de concept dans le chapitre 10 sur p. 187.Vous ne pouvez pas utiliser cette option pour le texte en japonais.

Réseau sémantique. Cette technique commence en identifiant les sens possibles de chaque concept à partir de son index complet de relations existant entre les mots puis crée des catégories en regroupant les concepts associés. Cette technique est plus performante lorsque les concepts sont connus dans le réseau sémantique et qu'ils ne sont pas trop ambigus. Son efficacité est cependant amoindrie lorsque le texte contient des termes spécialisés dont le réseau n'a pas connaissance. Par exemple, le concept `pomme granny smith` pourrait être regroupé avec `pomme gala` et `pomme golden` car il s'agit de soeurs de la granny smith. Pour donner un autre exemple, le concept `animal` pourrait être regroupé avec `chat` et `kangourou` car il s'agit d'hyponymes d'`animal`. Cette technique est uniquement disponible pour les textes en anglais dans cette version. Pour plus d'informations, reportez-vous à la section Réseaux sémantiques dans le chapitre 10 sur p. 189.

Inclusion de concept. Cette technique crée des catégories en regroupant les concepts multitermes (mots composés) selon qu'ils contiennent ou non des mots qui sont des sous-ensembles ou des super-ensembles d'un mot dans l'autre. Par exemple, le concept `siège` serait regroupé avec `siège de sécurité`, `siège couchette` et `commande de siège éjectable`. Pour plus d'informations, reportez-vous à la section Inclusion de concepts dans le chapitre 10 sur p. 188.

Cooccurrence. Cette technique crée des catégories à partir des cooccurrences trouvées dans le texte. Ainsi, lorsque des concepts ou des patrons de concept apparaissent souvent ensemble dans des documents et des enregistrements, la cooccurrence reflète une relation sous-jacente qui a vraisemblablement de l'intérêt dans vos définitions de catégorie. Lorsque des mots font l'objet d'une cooccurrence de manière significative, une règle de cooccurrence est créée et peut être utilisée comme descripteur de catégorie pour une nouvelle sous-catégorie. Par exemple, si de nombreux enregistrements contiennent les mots `prix` et `disponibilité`, (mais que peu

d'enregistrements contiennent l'un sans l'autre) alors ces concepts peuvent être regroupés dans une règle de cooccurrence, (`prix &disponible`) et ils peuvent être affectés à une sous-catégorie de la catégorie `prix` par exemple.Pour plus d'informations, reportez-vous à la section Règles de cooccurrence dans le chapitre 10 sur p. 191.

- **Nombre minimum de documents.**. Pour aider à déterminer l'intérêt des coocurrences, déterminez le nombre minimum de documents ou d'enregistrements devant contenir une cooccurrence donnée pour être utilisés en tant que descripteurs dans une catégorie.

IBM SPSS Modeler Text Analytics Noeuds

Outre les nombreux noeuds standard fournis avec IBM® SPSS® Modeler, vous pouvez utiliser les noeuds de Text Mining pour intégrer la puissance de l'analyse de texte dans vos flux. IBM® SPSS® Modeler Text Analytics vous propose plusieurs noeuds de Text Mining à cet effet. Ces noeuds sont stockés dans l'onglet SPSS Modeler Text Analytics de la palette des noeuds.

Figure 1-1
Onglet IBM SPSS Modeler Text Analytics de la palette des noeuds

Figure 1-2
Effectuer le zoom sur tous les noeuds IBM SPSS Modeler Text Analytics

Les noeuds suivants sont inclus :

- Le **noeud source Liste fichiers** génère une liste de noms de documents utilisée comme entrée pour le processus de Text Mining. Cela s'avère utile lorsque le texte se trouve dans des documents externes, et non dans une base de données ou un fichier structuré. Le noeud génère un champ unique comportant un enregistrement pour chaque document ou dossier répertorié, pouvant servir d'entrée dans un noeud de Text Mining suivant. Pour plus d'informations, reportez-vous à la section Noeud Liste fichiers dans le chapitre 2 sur p. 14.
- Grâce au **noeud source Fil de nouvelles**, vous pouvez lire du texte issu de fils de nouvelles, tels que des blogs ou de l'actualité au format RSS ou HTML, et utiliser ces données dans le processus de Text Mining. Le noeud génère un ou plusieurs champs pour chaque enregistrement figurant dans les fils. Ces champs peuvent servir d'entrée dans un noeud de Text Mining suivant. Pour plus d'informations, reportez-vous à la section Noeud Fil de nouvelles dans le chapitre 2 sur p. 19.
- Le **noeud de Text Mining** applique des méthodes linguistiques pour extraire les principaux concepts du texte, permet de créer des catégories avec ces concepts et d'autres données, et offre la possibilité d'identifier les relations et les associations existant entre les concepts en fonction de patrons connus (analyse des liens du texte). Grâce à ce noeud, vous pouvez explorer le contenu de données textuelles ou générer soit un modèle de concepts, soit un modèle de catégories. Les concepts et les catégories peuvent être combinés avec les données structurées existantes, telles que des données démographiques, et appliqués à la modélisation.

Pour plus d'informations, reportez-vous à la section Nœud de modélisation Text Mining dans le chapitre 3 sur p. 32.

- Le **noeud Analyse des liens du texte** extrait des concepts et identifie également les relations existant entre les concepts en fonction de patrons connus dans le texte. L'extraction de patrons permet de révéler les relations existant entre vos concepts, ainsi que tout qualificatif ou opinion associé à ces concepts. Le noeud Analyse des liens du texte propose une méthode plus directe pour identifier les patrons, les extraire du texte et ajouter leurs résultats à l'ensemble de données figurant dans le flux. Cependant, vous pouvez également effectuer une analyse des liens du texte en lançant une session interactive dans le noeud de modélisation Text Mining. Pour plus d'informations, reportez-vous à la section Nœud Analyse des liens du texte dans le chapitre 4 sur p. 82.
- Vous pouvez utiliser le **noeud Traduire** pour traduire en anglais ou dans toute autre langue un texte rédigé dans l'une des langues prises en charge, parmi lesquelles l'arabe, le chinois et le persan, et ce, à des fins de modélisation. Il est ainsi possible d'explorer des documents rédigés dans une langue utilisant des ensembles de caractères à deux octets et qui, sinon, ne serait pas prise en charge ; cela permet aux analystes d'extraire les concepts de ces documents même s'ils ne parlent pas la langue concernée. La même fonctionnalité peut être appelée à partir de n'importe quel noeud de modélisation de texte ; toutefois, en utilisant un noeud Traduire distinct, vous pouvez mettre une traduction en mémoire cache et la réutiliser dans plusieurs noeuds. Pour plus d'informations, reportez-vous à la section Noeud Traduire dans le chapitre 5 sur p. 94.
- Lors de l'exploration du texte contenu dans des documents externes, vous pouvez utiliser le **noeud de sortie Text Mining** pour générer une page HTML qui comporte des liens pointant vers les documents desquels les concepts ont été extraits. Pour plus d'informations, reportez-vous à la section Noeud Afficheur de fichiers dans le chapitre 6 sur p. 102.

Applications

En règle générale, IBM® SPSS® Modeler Text Analytics profite à toutes les personnes amenées à rechercher régulièrement des éléments-clés dans de gros volumes de documents.

Voici quelques exemples d'applications :

- **Recherche scientifique et médicale.** Explorer des documents de recherche divers, tels que des rapports sur les brevets, des articles et des publications relatives aux protocoles. Identifier des associations jusque-là inconnues (par exemple, l'association d'un docteur à un produit particulier), ouvrant la voie à de nouvelles explorations. Réduire le délai nécessaire au processus de découverte de médicaments. Utiliser le programme dans le cadre de recherches en génomique.
- **Recherche dans le domaine des investissements.** Passer en revue les rapports d'analyse quotidiens et les articles des journaux afin d'identifier les changements de stratégie et les évolutions du marché. A partir de ces informations, il est possible d'analyser les tendances, et de détecter les problèmes et opportunités que rencontre une société ou un secteur sur une période donnée.
- **Détection des fraudes.** Dans le secteur bancaire et dans le domaine de la santé, ce logiciel peut servir à détecter les anomalies et les alertes dans de gros volumes de texte.

- **Etude de marché.** Dans le domaine de l'étude de marché, cette application permet d'identifier les rubriques essentielles contenues dans les réponses ouvertes formulées à l'occasion d'enquêtes.
- **Analyse de fils de nouvelles et de blogs.** Cette application permet de créer et d'explorer des modèles en s'appuyant sur les principales idées figurant dans l'actualité, les blogs, etc.
- **CRM.** Créer des modèles sur la base des données issues de l'ensemble des points de communication avec la clientèle (messages électroniques, transactions et enquêtes).

Chapitre
2

Lecture du texte source

Dans le Text Mining, les données utilisées peuvent se trouver dans tout format standard utilisé par IBM® SPSS® Modeler, notamment les bases de données et autres formats dits « rectangulaires », qui représentent les données sous forme de lignes et de colonnes, ainsi que dans les formats de document (par exemple, les formats Microsoft Word, Adobe PDF ou HTML), qui n'obéissent pas à cette structure.

- Pour lire du texte à partir de documents qui n'obéissent pas à la structure de données standard (Microsoft Word , Microsoft Excel et Microsoft PowerPoint, mais aussi les formats Adobe PDF, XML et HTML, entre autres), vous pouvez utiliser le noeud Liste fichiers afin de générer une liste de documents ou de dossiers qui servira d'entrée au processus de Text Mining. Pour plus d'informations, reportez-vous à la section Noeud Liste fichiers sur p. 14.
- Pour lire du texte à partir de fils de nouvelles, tels que des blogs ou des actualités au format RSS ou HTML, il est possible d'utiliser le noeud Fil de nouvelles afin de formater des données de fils de nouvelles pour les convertir en entrée dans le processus de Text Mining. Pour plus d'informations, reportez-vous à la section Noeud Fil de nouvelles sur p. 19.
- Pour lire un texte à partir d'un format de données standard utilisé par SPSS Modeler, tel qu'une base de données comprenant un ou plusieurs champs texte dédiés aux commentaires des clients, vous pouvez utiliser n'importe quel noeud source SPSS Modeler natif et standard. Pour plus d'informations, reportez-vous à la documentation du noeud SPSS Modeler.

Noeud Liste fichiers

Pour lire du texte à partir de documents non structurés, enregistrés dans des formats tels que Microsoft Word, Microsoft Excel et Microsoft PowerPoint, mais aussi Adobe PDF, XML et HTML (entre autres), vous pouvez utiliser le noeud Liste fichiers afin de générer une liste de documents ou de dossiers qui servira d'entrée au processus de Text Mining. Cette opération s'avère nécessaire car il est impossible de représenter les documents texte non structurés par des champs et des enregistrements (c'est-à-dire des lignes et des colonnes), contrairement aux autres données utilisées par IBM® SPSS® Modeler. Ce noeud se trouve dans la palette Text Mining.

Le nœud Liste fichiers fonctionne comme un nœud source, à une exception près toutefois : il ne lit pas les données proprement dites mais le nom des documents ou des répertoires figurant sous la racine indiquée, et génère ces noms sous la forme d'une liste. La sortie consiste en un champ unique comportant un enregistrement pour chaque fichier répertorié, pouvant servir d'entrée pour un noeud de Text Mining suivant.

Vous pouvez trouver ce noeud dans l'onglet IBM® SPSS® Modeler Text Analytics de la palette de noeuds en bas de la fenêtre SPSS Modeler. Pour plus d'informations, reportez-vous à la section IBM SPSS Modeler Text Analytics Noeuds dans le chapitre 1 sur p. 11.

Important ! Les noms de répertoires et de fichiers contenant des caractères non inclus dans l'encoding local de l'ordinateur ne sont pas pris en charge. Lorsque vous essayez d'exécuter un flux contenant un noeud Liste fichiers, les noms de fichiers ou de répertoires contenant ces

caractères feront échouer l'exécution du flux. Cela peut arriver avec des noms de répertoires ou de fichiers en langues étrangères, comme un nom de fichier japonais dans un paramètre local français.

Traitement des fichiers RTF. Pour traiter les fichiers RTF, un filtre est nécessaire. Un filtre RTF peut être téléchargé à partir du site de Microsoft et enregistré manuellement.

Adobe PDF Traitement. Pour extraire du texte de fichiers Adobe PDF, Adobe Reader version 9 doit être installé sur l'ordinateur où se trouvent SPSS Modeler Text Analytics et IBM® SPSS® Modeler Text Analytics Server .

- **Remarque** : ne mettez pas à niveau vers la version 10 de Adobe Reader ou une version ultérieure car elle ne contient pas le filtre nécessaire.
- La mise à niveau vers la version 9 de Adobe Reader vous permet d'éviter une fuite importante de la mémoire dans le filtre, pouvant causer des erreurs dans le traitement des fichiers, lorsque vous travaillez avec un grand nombre de documents Adobe PDF (aux alentours de 1000 ou plus) . Si vous prévoyez le traitement de documents Adobe PDF sur des systèmes d'exploitation Microsoft Windows 32 bits ou 64 bits, effectuez une mise à niveau vers Adobe Reader version 9.x pour les systèmes 32 bits ou vers Adobe PDF iFilter 9 pour les systèmes 64 bits. Ils sont tous deux disponibles sur le site Web d'Adobe.
- Adobe a modifié le logiciel de filtrage utilisé au démarrage de Adobe Reader 8.x. Les anciens fichiers Adobe PDF peuvent ne pas être lisibles ou contenir des caractères corrompus. Ceci est un problème lié au logiciel d'Adobe et n'est pas contrôlé par SPSS Modeler Text Analytics .
- Si des restrictions de sécurité d'un fichier Adobe PDF pour « Copie de contenu ou extraction » sont configurées sur « Non autorisé » dans l'onglet Sécurité de la boîte de dialogue des propriétés du document Adobe PDF, le document ne peut pas être filtré ni lu dans le produit.
- Les fichiers Adobe PDF ne peuvent pas être traités sur des plates-formes non Microsoft Windows.
- En raison des limitations d'Adobe, il est impossible d'extraire du texte à partir de fichiers Adobe PDF image.

Microsoft Office Traitement.

- Afin de traiter les nouveaux formats des documents Microsoft Word, Microsoft Excel et Microsoft PowerPoint introduits dans Microsoft Office 2007, Microsoft Office 2007 doit être installé sur l'ordinateur d'exécution de SPSS Modeler Text Analytics Server (en local ou à distance) ou vous pouvez installer le nouveau pack de filtre Microsoft Office 2007 (disponible sur le site Web de Microsoft).
- Les fichiers Microsoft Office ne peuvent pas être traités sur des plates-formes non Microsoft Windows.

Prise en charge des données locales. Si vous êtes connecté à un SPSS Modeler Text Analytics Server distant et que vous avez un flux avec un noeud Liste fichiers, les données doivent résider sur le même ordinateur que SPSS Modeler Text Analytics Server , ou assurez-vous que le serveur a accès au dossier dans lesquel les données source du noeud Liste fichiers sont stockées.

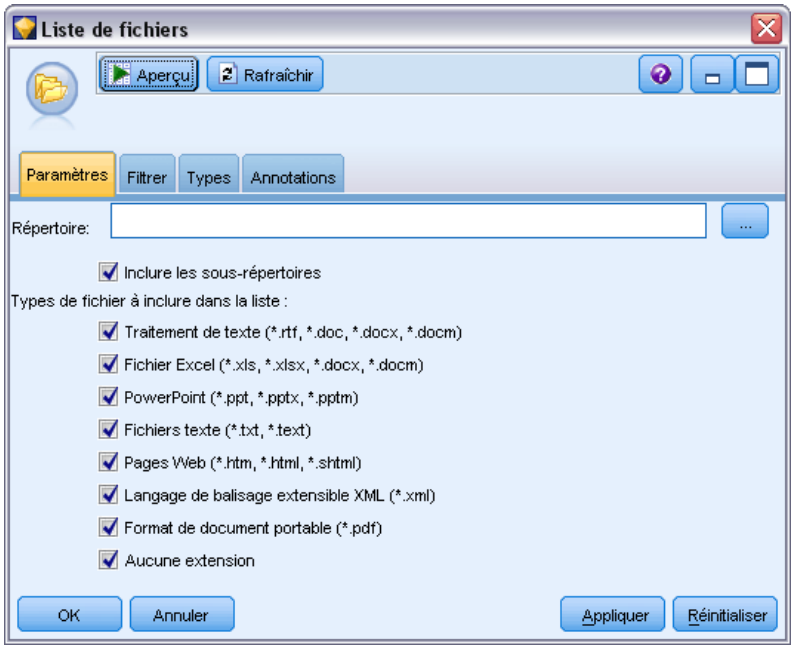
Nœud Liste fichiers : Onglet Paramètres

Dans cet onglet, vous pouvez définir les répertoires, les extensions de fichier et la sortie du noeud que vous souhaitez.

Remarque : l'extraction Text Mining ne peut pas traiter les fichiers Microsoft Office et Adobe PDF sur des plates-formes non Microsoft Windows. Toutefois, il est toujours possible de traiter les fichiers XML, HTML ou texte.

Les noms de répertoires et de fichiers contenant des caractères non inclus dans l'encoding local de l'ordinateur ne sont pas pris en charge. Lorsque vous essayez d'exécuter un flux contenant un noeud Liste fichiers, les noms de fichiers ou de répertoires contenant ces caractères feront échouer l'exécution du flux. Cela peut arriver avec des noms de répertoires ou de fichiers en langues étrangères, comme un nom de fichier japonais dans un paramètre local français.

Figure 2-1
Boîte de dialogue du noeud Liste fichiers : onglet Paramètres



Répertoire. Définit le dossier racine qui contient les documents à répertorier.

- **Inclure les sous-répertoires.** Indique que les sous-répertoires doivent également être analysés.

Types de fichier à inclure dans la liste : Vous pouvez sélectionner ou désélectionner les types et extensions de fichier à utiliser. Si une extension de fichier est désélectionnée, les fichiers présentant cette extension sont ignorés. Vous pouvez appliquer un filtre sur les extensions suivantes :

- *.rtf, .doc, .docx, .docm*
- *.xls, .xlsx, .xlsm*
- *.ppt, .pptx, .pptm*
- *.txt, .text*
- *.htm, .html, .shtml*
- *.xml*
- *.pdf*
- *.$*

Remarque : Pour plus d'informations, reportez-vous à la section Noeud Liste fichiers sur p. 14.

Important ! L'option « Liste des répertoires » n'est plus disponible à partir de la version 14 et le seul résultat sera une liste de fichiers.

Nœud Liste fichiers : autres onglets

L'onglet Types est un onglet standard dans les noeuds IBM® SPSS® Modeler, tout comme l'onglet Annotations.

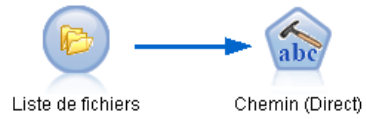
Utilisation du noeud Liste fichiers pour le Text Mining

Le noeud Liste fichiers est utilisé lorsque les données textuelles résident dans des documents non structurés externes dans des formats tels que Microsoft Word, Microsoft Excel et Microsoft PowerPoint, ainsi que Adobe PDF, XML et HTML (entre autres). Il permet de générer une liste de documents ou de dossiers qui servira d'entrée au processus de Text Mining (un noeud de Text Mining ou Analyse des liens du texte suivant).

Si vous utilisez le noeud Liste fichiers, veillez à spécifier que le champ Texte correspond au chemin d'accès des documents dans le noeud de Text Mining ou Analyse des liens du texte ; vous indiquez ainsi que, plutôt que de contenir le texte réel à explorer, le champ sélectionné contient les chemins vers les documents où est situé le texte.

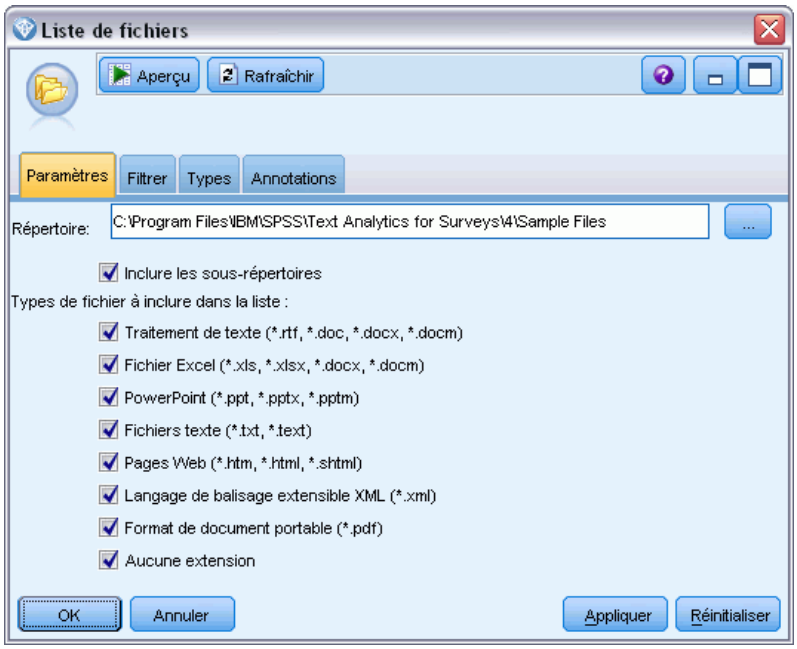
Dans l'exemple suivant, nous avons connecté un noeud Liste fichiers à un noeud de Text Mining de façon à fournir du texte qui réside dans des documents externes.

Figure 2-2
Exemple de flux : nœud Liste fichiers avec nœud de modélisation Text Mining



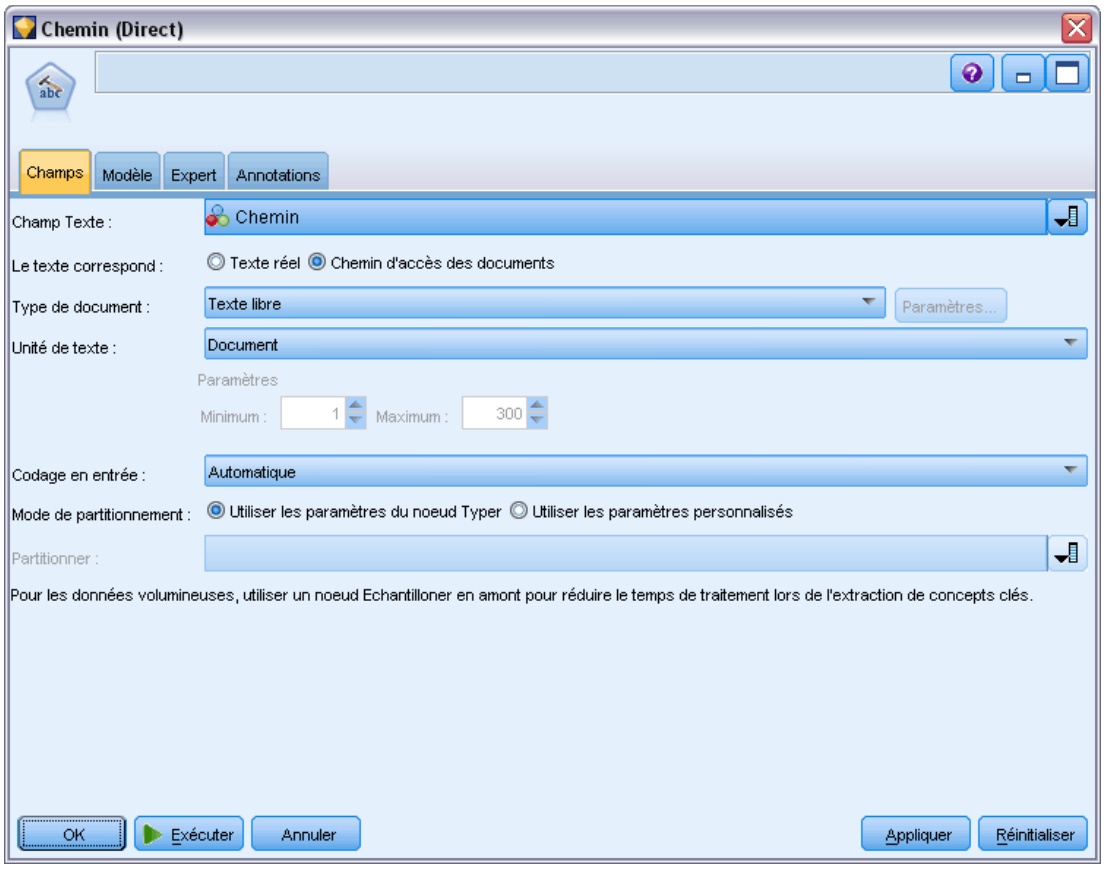
- **Noeud Liste fichiers (onglet Paramètres).** Nous avons tout d'abord ajouté ce nœud au flux pour indiquer l'emplacement de stockage des documents texte. Nous avons sélectionné le répertoire contenant tous les documents sur lesquels nous souhaitons effectuer le processus de Text Mining.

Figure 2-3
Boîte de dialogue du noeud Liste fichiers : onglet Paramètres



- **Noeud de Text Mining (onglet Champs).** Nous avons ensuite ajouté et connecté un noeud de Text Mining au noeud Liste fichiers. Dans ce noeud, nous avons défini le format d'entrée, le modèle de ressources et le format de sortie. Nous avons choisi le nom de champ créé à partir du noeud Liste fichiers et sélectionné l'option Le texte correspond au chemin d'accès des documents, ainsi que d'autres paramètres. Pour plus d'informations, reportez-vous à la section Utilisation du nœud Text Mining dans un flux dans le chapitre 3 sur p. 50.

Figure 2-4
Boîte de dialogue du noeud de Text Mining : onglet Champs



Pour plus d'informations sur l'utilisation du noeud de Text Mining, reportez-vous à le chapitre 3.

Noeud Fil de nouvelles

Vous pouvez utiliser le noeud Fil de nouvelles afin de préparer des données textuelles à partir de fils de nouvelles pour le processus de Text Mining. Ce nœud accepte les fils de nouvelles dans les deux formats suivants :

- **Format RSS.** RSS est un format normalisé simple, basé sur le langage XML et destiné au contenu Web. Pour ce format, l'URL pointe vers une page qui présente un ensemble de liens vers d'autres articles tels que les sources d'actualité et les blogs publiés. Le format RSS étant normalisé, chaque lien d'article est automatiquement identifié et traité comme un enregistrement séparé dans le flux de données obtenu. Aucune autre donnée n'est nécessaire pour que vous puissiez identifier les données textuelles et les enregistrements importants du fils de nouvelles sauf si vous souhaitez appliquer une technique de filtrage au texte.
- **Format HTML.** Vous pouvez définir une ou plusieurs URL vers des pages HTML dans l'onglet Entrée. Ensuite, dans l'onglet Enregistrements, définissez la balise de début d'enregistrement et identifiez également les balises qui délimitent le contenu cible et assignez ces balises aux champs de sortie de votre choix (description, titre, date modifiée, etc.) Pour plus

d'informations, reportez-vous à la section Nœud Fil de nouvelles : onglet Enregistrements sur p. 22.

Important ! Si vous essayez de récupérer des informations sur le Web avec un serveur proxy, vous devez activer ce serveur dans le fichier `net.properties` pour le serveur et le client IBM® SPSS® Modeler Text Analytics . Suivez les instuctions détaillées dans ce fichier. Cela s'applique lorsque vous accédez au Web avec le noeud Fil de nouvelles ou lorsque vous récupérez une licence SDL Software as a Service (SaaS) car ces connexions utilisent Java. Pour le client, l'emplacement du fichier est *C:\Program Files\IBM\SPSS\Modeler\15\jre\lib\net.properties* par défaut. Pour le client, l'emplacement du fichier est *C:\Program Files\IBM\SPSS\Modeler\15\ext\bin\spss.TMWBServer\jre\lib\net.properties* par défaut.

La sortie de ce noeud est un ensemble de champs utilisés pour décrire les enregistrements. Le champ Description est le plus fréquemment utilisé car il contient la plus grosse partie du contenu textuel. Toutefois, d'autres champs sont intéressants, tels que la description courte de l'enregistrement (champ Description courte) ou le titre de l'enregistrement (champ Titre). Il est possible de sélectionner n'importe quel champ de sortie en tant qu'entrée pour un noeud de Text Mining suivant.

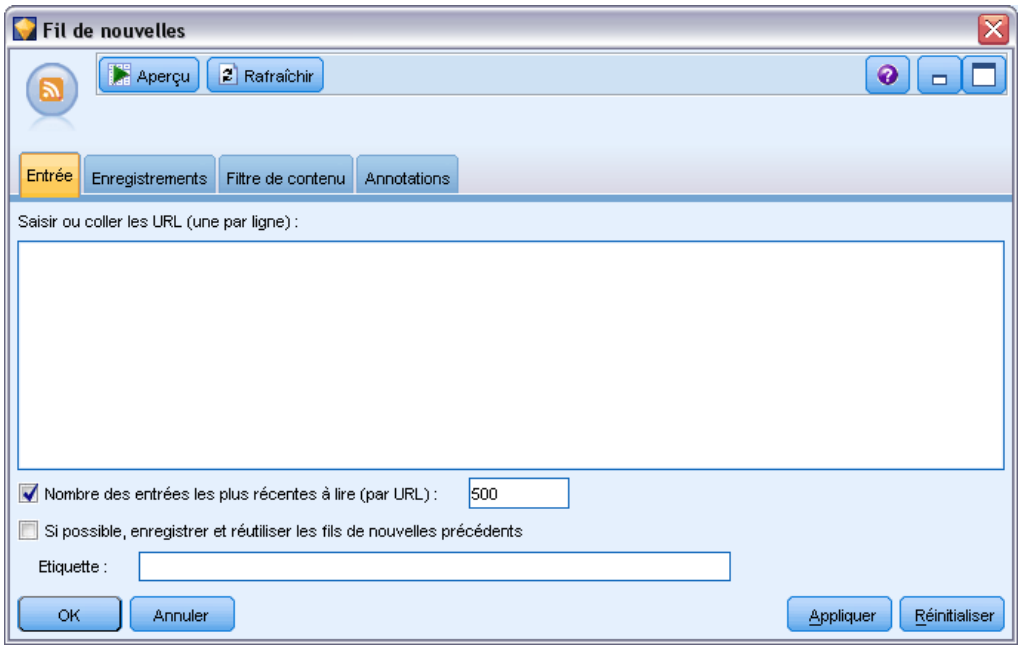
Vous pouvez trouver ce noeud dans l'onglet SPSS Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM® SPSS® Modeler. Pour plus d'informations, reportez-vous à la section IBM SPSS Modeler Text Analytics Noeuds dans le chapitre 1 sur p. 11.

Nœud Fil de nouvelles : onglet Entrée

L'onglet Entrée permet de spécifier une ou plusieurs adresses Web, ou URL, afin de capturer les donnés textuelles. Dans le contexte du processus de Text Mining, vous pouvez indiquer des URL pour des fils qui contiennent des données textuelles.

Important ! Lorsque vous travaillerez avec des données non RSS, vous préfèrerez peut-être utiliser un outil de récupération de données Web, tel que WebQL®, pour automatiser le rassemblement du contenu puis faire référence à la sortie de cet outil à l'aide d'un nœud source différent.

Figure 2-5
Boîte de dialogue du nœud Fil de nouvelles : onglet Entrée



Les paramètres pouvant être définis sont les suivants :

Saisir ou coller les URL. Dans ce champ, vous pouvez entrer ou coller une ou plusieurs URL. Dans le cas de plusieurs URL, entrez une seule URL par ligne et utilisez la touche Entrée/Retour pour séparer les lignes. Saisissez le chemin d'accès de l'URL vers le fichier. Ces URL peuvent correspondre à des fils apparaissant dans l'un des deux formats suivants :

- *Format RSS.* RSS est un format normalisé simple, basé sur le langage XML et destiné au contenu Web. Pour ce format, l'URL pointe vers une page qui présente un ensemble de liens vers d'autres articles tels que les sources d'actualité et les blogs publiés. Le format RSS étant normalisé, chaque lien d'article est automatiquement identifié et traité comme un enregistrement séparé dans le flux de données obtenu. Aucune autre donnée n'est nécessaire pour que vous puissiez identifier les données textuelles et les enregistrements importants du fils de nouvelles sauf si vous souhaitez appliquer une technique de filtrage au texte.
- *Format HTML.* Vous pouvez définir une ou plusieurs URL vers des pages HTML dans l'onglet Entrée. Ensuite, dans l'onglet Enregistrements, définissez la balise de début d'enregistrement et identifiez également les balises qui délimitent le contenu cible et assignez ces balises aux champs de sortie de votre choix (description, titre, date modifiée, etc.) Lorsque vous travaillerez avec des données non RSS, vous préfèrerez peut-être utiliser un outil de récupération de données Web, tel que WebQL®, pour automatiser le rassemblement du contenu puis faire référence à la sortie de cet outil à l'aide d'un nœud source différent. Pour plus d'informations, reportez-vous à la section Nœud Fil de nouvelles : onglet Enregistrements sur p. 22.

Nombre des entrées les plus récentes à lire (par URL). Ce champ spécifie le nombre maximal d'enregistrements à lire pour chaque URL répertoriée, en commençant par le premier enregistrement détecté dans le fil. La quantité de texte influe sur la vitesse de traitement pendant l'extraction dans un noeud Text Mining ou Analyse des liens du texte, en aval.

Si possible, enregistrer et réutiliser les fils de nouvelles précédents. Cette option permet d'analyser les fils de nouvelles et de mettre en cache les résultats traités. Puis, après plusieurs exécutions de flux, si le contenu d'un fil donné reste inchangé ou si le fil est inaccessible (interruption d'Internet, par exemple), la version mise en cache est utilisée pour accélérer le temps de traitement. Tout nouveau contenu détecté dans ces fils est également mis en cache pour la prochaine exécution du noeud.

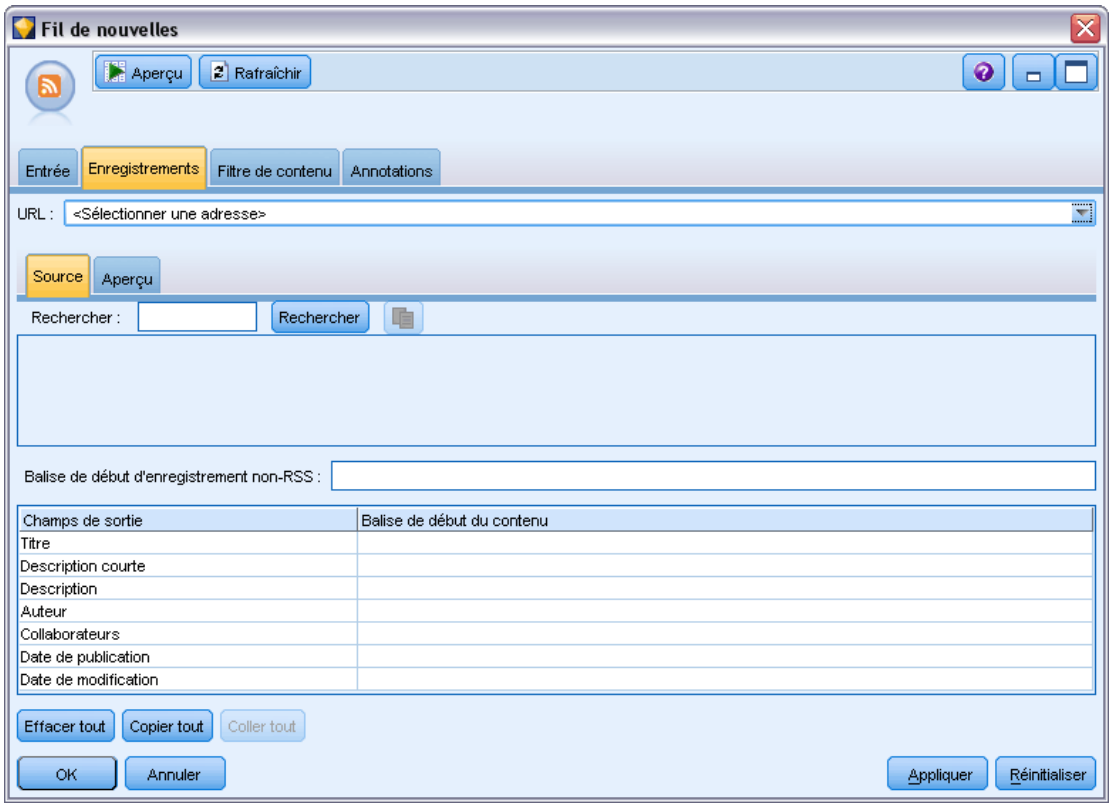
- **Etiquette**. Si vous sélectionnez l'option Si possible, enregistrer et réutiliser les fils de nouvelles précédents, vous devez spécifier un nom d'étiquette pour les résultats. Cette étiquette permet de décrire les fils mis en cache sur le serveur. Si aucune étiquette n'est spécifiée ou si l'étiquette n'est pas reconnue, la réutilisation n'est pas possible. Vous pouvez gérer ces caches de fils de nouvelles dans la table de session de IBM® SPSS® Text Analytics Administration Console . Pour plus d'informations, reportez-vous au guide utilisateur de SPSS Text Analytics Administration Console .

Nœud Fil de nouvelles : onglet Enregistrements

L'onglet Enregistrements permet de définir le contenu textuel de fils non-RSS en identifiant le début de chaque nouvel enregistrement, ainsi que d'autres informations intéressantes concernant chaque enregistrement. Si vous savez qu'un fil non-RSS (HTML) contient du texte se trouvant dans plusieurs enregistrements, vous devez identifier ici la balise de début d'enregistrement ou bien le texte sera traité comme enregistrement indépendant. Pendant la normalisation des fils RSS qui ne nécessitent aucune spécification de balise dans cet onglet, vous pouvez prévisualiser le contenu dans l'onglet Aperçu.

Important ! Lorsque vous travaillerez avec des données non RSS, vous préfèrerez peut-être utiliser un outil de récupération de données Web, tel que WebQL®, pour automatiser le rassemblement du contenu puis faire référence à la sortie de cet outil à l'aide d'un nœud source différent.

Figure 2-6
Boîte de dialogue du nœud Fil de nouvelles : Onglet Enregistrements



URL. Cette liste déroulante contient la liste des URL entrées dans l'onglet Entrée. Sont présents à la fois des fils au format HTML et RSS. Si l'adresse URL est trop longue pour la liste déroulante, elle est automatiquement coupée au milieu et le texte coupé est remplacé par des points de suspension, par exemple *http://www.ibm.com/exemple/début-de-l'adresse...reste-de-l'adresse/chemin.htm*.

- Avec les **fils au format HTML**, si le fil contient plusieurs enregistrements (ou entrées), vous pouvez déterminer les balises HTML qui contiennent les données correspondant au champ indiqué dans le tableau. Par exemple, vous pouvez définir la balise de début (qui indique le commencement d'un nouvel enregistrement), une balise de date modifiée ou un nom d'auteur.
- Avec les **fils au format RSS**, il ne vous est pas demandé d'entrer des balises car le format RSS est normalisé. Cependant, vous pouvez prévisualiser les exemples de résultats dans l'onglet Aperçu si nécessaire. Tous les fils RSS reconnus sont précédés de l'image du logo RSS.

Onglet Source. Dans cet onglet, vous pouvez visualiser le code source des fils HTML. Ce code n'est pas modifiable. Vous pouvez utiliser le champ Rechercher pour localiser des balises ou des informations spécifiques dans cette page ; vous pouvez ensuite copier et coller ces dernières dans le tableau situé en dessous. Le champ Rechercher ne distingue pas les majuscules des minuscules et fournit des correspondances partielles.

Onglet Aperçu. Dans cet onglet, vous pouvez prévisualiser la façon dont un enregistrement est lu par le noeud Fil de nouvelles. Ceci est particulièrement utile pour les fils HTML car vous pouvez modifier ces modalités de lecture en définissant des balises HTML dans le tableau situé sous l'onglet Aperçu.

Balise de début d'enregistrement non RSS. Cette option ne s'applique qu'aux fils non RSS. Si votre fil HTML contient beaucoup de texte que vous souhaitez séparer en plusieurs enregistrements, indiquez ici la balise HTML qui signalera le début d'un enregistrement (un article ou un billet de blog par exemple). Si vous ne définissez pas de balise de début pour un fil non RSS, la totalité de la page est traitée en tant qu'enregistrement unique, la totalité du contenu apparaît dans le champ Description et la date d'exécution du noeud est utilisée à la fois comme date de modification et date de publication.

Table des champs. Cette option ne s'applique qu'aux fils non RSS. Dans cette table, vous pouvez découper le contenu textuel en champs de sortie spécifiques en saisissant une balise de début pour chacun des champs de sortie prédéfinis. Entrez uniquement la balise de début. Toutes les correspondances sont obtenues via l'analyse du code HTML et la mise en correspondance du contenu de la table et des noms et attributs de balise détectés dans le code HTML. Vous pouvez utiliser les boutons situés sous la table pour copier les balises définies et les réutiliser pour d'autres fils.

Table 2-1
Champs de sortie possibles pour fils non RSS (formats HTML)

Nom du champ de sortie	Contenu de balise prévu
Titre	La balise délimitant le titre de l'enregistrement. (facultatif)
Description courte	La balise délimitant la description courte ou l'étiquette. (facultatif)
Description	La balise délimitant le texte principal. Si ce champ n'est pas renseigné, il inclura le contenu de la balise `<body>` (s'il existe un enregistrement unique) ou le contenu détecté au sein de l'enregistrement actuel (lorsqu'un séparateur d'enregistrement a été spécifié).
Author	La balise délimitant l'auteur du texte. (facultatif)
Collaborateurs	La balise délimitant les noms des collaborateurs. (facultatif)
Date de publication	La balise délimitant la date de publication du texte. Si ce champ n'est pas renseigné, il contiendra la date de lecture des données par le noeud.
Date de modification	La balise délimitant la date de modification du texte. Si ce champ n'est pas renseigné, il contiendra la date de lecture des données par le noeud.

Lorsque vous entrez une balise dans le tableau, le fil est analysé à l'aide de cette balise, en vue d'obtenir une correspondance minimale plutôt qu'une correspondance exacte. En d'autres termes, si vous avez entré `<div>` comme champ Titre, toutes les balises `<div>` du fil sont renvoyées, y compris celles présentant des attributs spécifiques, (telles que `<div class="post three">`), ainsi `<div>` est égale à la balise racine (`<div>`) ainsi qu'à tout dérivé incluant un attribut et qui utilise ce contenu pour le champ de sortie Titre. Si vous entrez une balise racine, tous les autres attributs sont également inclus.

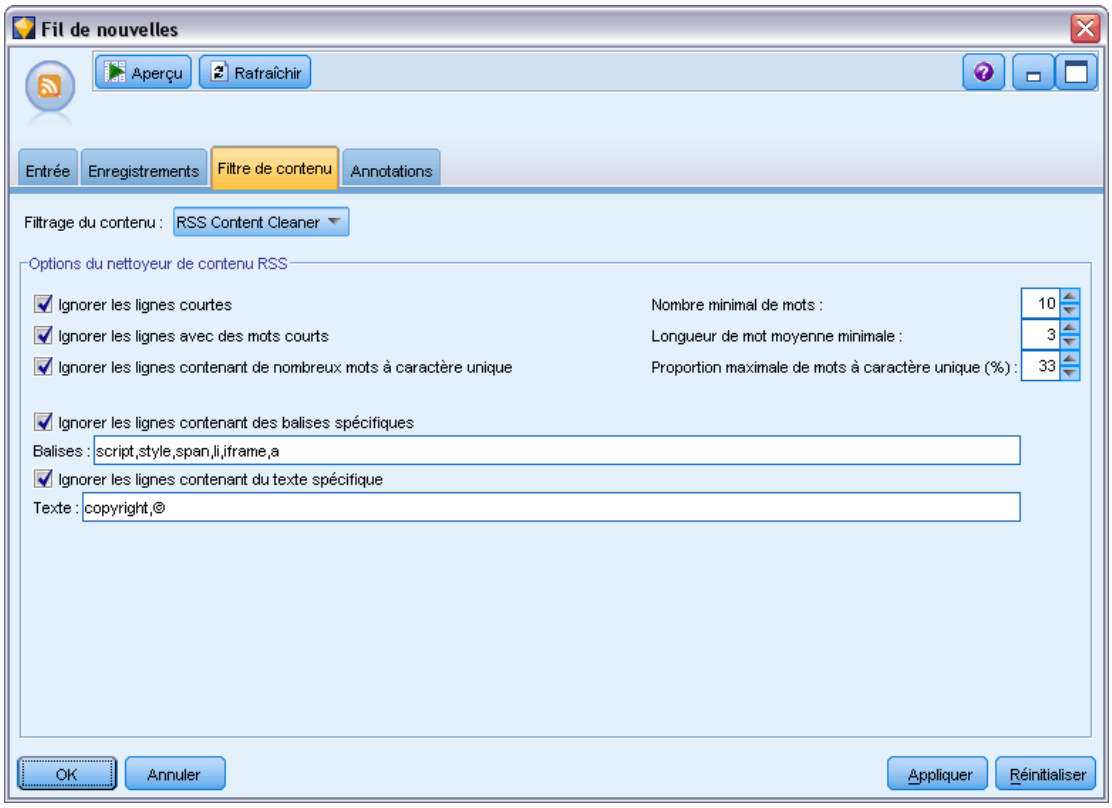
Table 2-2
Les exemples de balises HTML utilisés identifient le texte des champs de sortie.

Si vous entrez :	Cela renvoie :	Et également :	Mais cela ne renvoie pas :
`<div>`	`<div>`	`<div class="post">`	toutes les autres balises
`<p class="auth">`	`<p class="auth">`	`<p color="black" class="auth" id="85643">`	`<p color="black">`

Nœud Fil de nouvelles : Onglet Filtrer le contenu

L'onglet Filtrer le contenu permet d'appliquer une technique de filtre au contenu des fils RSS. Cet onglet ne s'applique pas aux fils HTML. Le filtrage peut être utile lorsque le fil contient beaucoup de texte sous forme d'en-têtes, de bas de page, de menus, de publicités, etc. Vous pouvez utiliser cet onglet pour supprimer les balises HTML, JavaScript et des lignes ou des mots courts du contenu.

Figure 2-7
Boîte de dialogue du nœud Fil de nouvelles : Onglet Filtrer le contenu



Filtrage de contenu. Si vous ne souhaitez pas appliquer de technique de nettoyage, sélectionnez Aucun. Sinon, sélectionnez Nettoyeur de contenu RSS.

Options du nettoyeur de contenu RSS. Si vous sélectionnez Nettoyeur de contenu RSS, vous pouvez choisir d'ignorer des lignes basées sur certains critères. Une ligne est délimitée par une balise HTML telle que `<p>` et `<li>` mais pas par des balises incorporées telles que `<span>`, `<b>` et `<font>`. Veuillez noter que les balises `<br>` sont exécutées comme des sauts de ligne.

- **Ignorer les lignes courtes**. Cette option ignore les lignes qui ne contiennent pas le nombre de mots minimum défini ici.
- **Ignorer les lignes avec des mots courts**. Cette option ignore les lignes qui contiennent plus que la longueur de mot minimum moyenne définie ici.
- **Ignorer les lignes avec beaucoup de mots à caractère unique**. Cette option ignore les lignes qui contiennent plus d'une certaine proportion de mots à caractère unique.
- **Ignorer les lignes contenant des balises spécifiques**. Cette option ignore le texte des lignes contenant une des balises spécifiées dans ce champ.
- **Ignorer les lignes contenant du texte spécifique**. Cette option ignore le texte des lignes contenant une partie du texte spécifié dans ce champ.

Utilisation du noeud Fil de nouvelles dans le processus de Text Mining

Vous pouvez utiliser le noeud Fil de nouvelles afin de préparer des données textuelles à partir de fils de nouvelles Internet pour le processus de Text Mining. Ce nœud accepte les fils de nouvelles au format HTML ou RSS. Ces fils servent d'entrée au processus de Text Mining (un noeud de Text Mining ou Analyse des liens du texte suivant).

Si vous utilisez le nœud Fil de nouvelles, veillez à spécifier que Le texte correspond au texte réel dans le nœud de Text Mining ou Analyse des liens du texte pour indiquer que ces fils conduisent directement à chaque article ou billet de blog.

Important ! Si vous essayez de récupérer des informations sur le Web avec un serveur proxy, vous devez activer ce serveur dans le fichier `net.properties` pour le serveur et le client IBM® SPSS® Modeler Text Analytics . Suivez les instuctions détaillées dans ce fichier. Cela s'applique lorsque vous accédez au Web avec le noeud Fil de nouvelles ou lorsque vous récupérez une licence SDL Software as a Service (SaaS) car ces connexions utilisent Java. Pour le client, l'emplacement du fichier est *C:\Program Files\IBM\SPSS\Modeler\15\jre\lib\net.properties* par défaut. Pour le client, l'emplacement du fichier est *C:\Program Files\IBM\SPSS\Modeler\15\ext\bin\spss.TMWBServer\jre\lib\net.properties* par défaut.

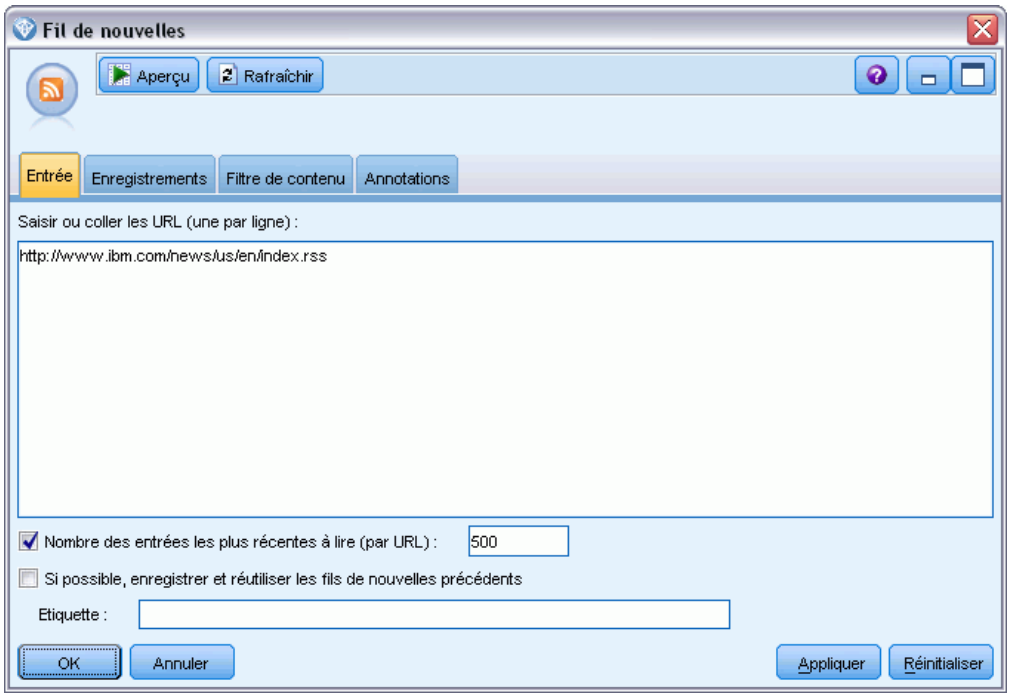
Exemple : Noeud de fil de nouvelles (fil RSS) avec noeud de modélisation Text Mining

Dans l'exemple suivant, nous connectons un noeud Fil de nouvelles à un noeud de Text Mining de façon à fournir des données textuelles provenant d'un fil RSS dans le processus de Text Mining.

Figure 2-8
Exemple de flux : Nœud de fil de nouvelles avec nœud de Text Mining

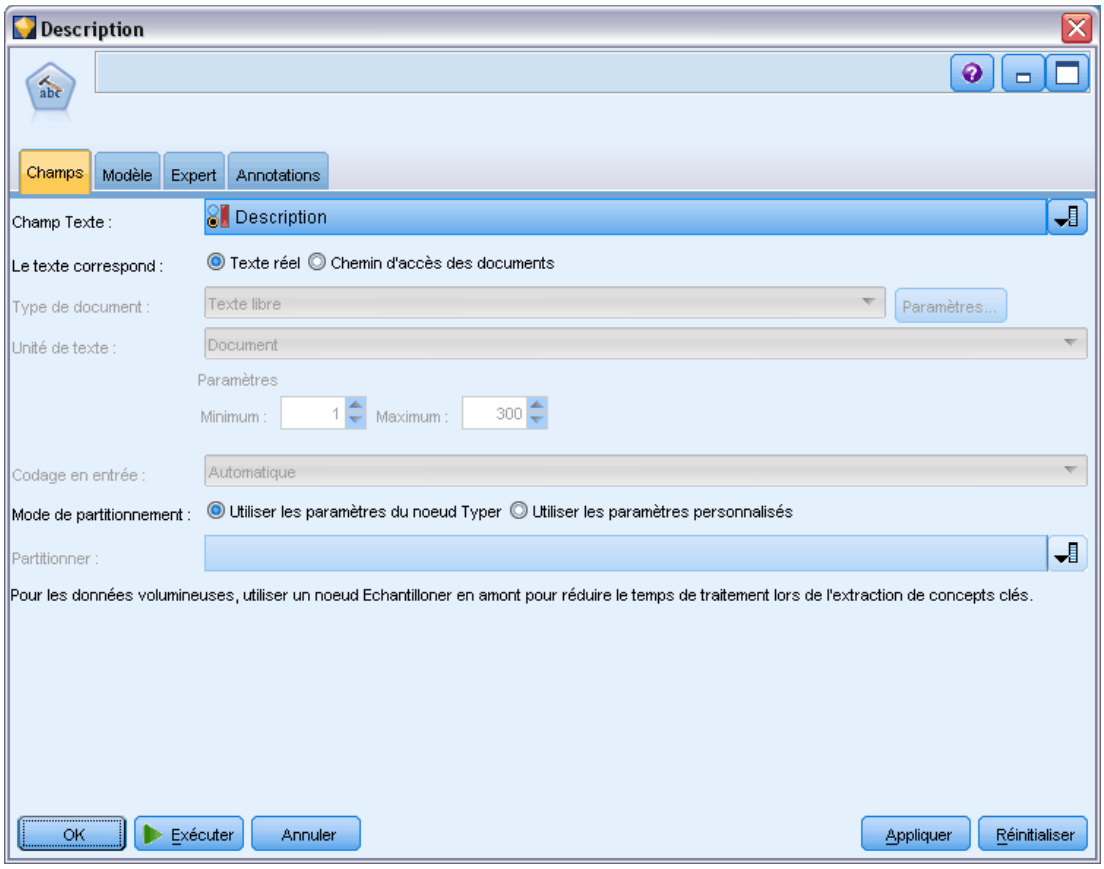
- **Nœud de fil de nouvelles (onglet Entrée).** Nous avons tout d'abord ajouté ce nœud au flux afin d'indiquer l'emplacement du contenu du fil et de vérifier la structure du contenu. Dans le premier onglet, nous avons fourni l'URL d'un fil RSS. Étant donné que notre exemple concerne un fil RSS, le formatage est déjà défini et il n'est pas nécessaire d'apporter des modifications dans l'onglet Enregistrements. Un algorithme de filtrage de contenu facultatif est disponible pour les fils RSS mais n'a pas été utilisé dans cet exemple.

Figure 2-9
Boîte de dialogue du nœud Fil de nouvelles : onglet Entrée



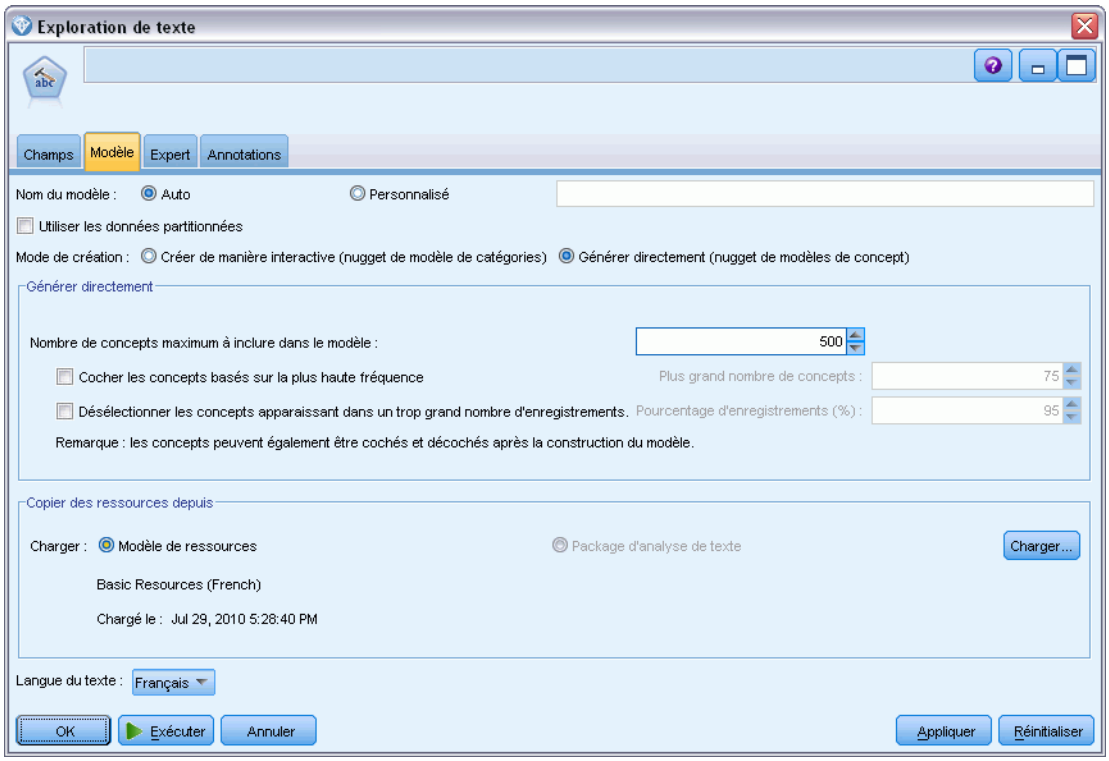
- **Noeud de Text Mining (onglet Champs).** Ensuite, nous avons ajouté et connecté un nœud de Text Mining au nœud Fil de nouvelles. Dans cet onglet, nous avons défini la sortie du champ de texte par un noeud de fil de nouvelles. Dans ce cas, nous voulions utiliser le champ Description. Nous avons également sélectionné l'option Le texte correspond au texte réel, ainsi que d'autres paramètres.

Figure 2-10
Boîte de dialogue du noeud de Text Mining : onglet Champs



- **Noeud de Text Mining (onglet Modèle).** Ensuite, dans l'onglet Modèle, nous avons choisi le mode et les ressources de création. Dans cet exemple, nous avons choisi de créer un modèle de concepts directement à partir de ce noeud à l'aide du modèle de ressources par défaut.

Figure 2-11
Nœud de Text Mining : onglet Modèle



Pour plus d'informations sur l'utilisation du noeud de Text Mining, reportez-vous à le chapitre 3.

Chapitre
3

Text Mining pour les concepts et les catégories

Le nœud de modélisation Text Mining est utilisé pour générer l'un des deux nuggets de modèle Text Mining :

- *Les nuggets de modèles de concepts* explorent et extraient les concepts fondamentaux de données textuelles, structurées ou non.
- *Les nuggets de modèles de catégories* scorent et attribuent des documents et des enregistrements à des catégories, qui sont constitués des concepts (et des patrons) extraits.

Les concepts, les patrons et les catégories extraits de vos nuggets de modèle peuvent tous être combinés aux données structurées existantes, telles que les données démographiques, et appliqués grâce à la gamme complète d'outils de IBM® SPSS® Modeler, afin de favoriser une prise de décision plus précise et plus efficace. Par exemple, si des clients désignent souvent les problèmes de connexion comme le principal obstacle à la réalisation des tâches de gestion des comptes en ligne, vous pouvez intégrer les "problèmes de connexion" dans vos modèles.

En outre, le nœud de modélisation Text Mining est totalement intégré dans SPSS Modeler. Vous pouvez ainsi déployer des flux de Text Mining via IBM® SPSS® Modeler Solution Publisher pour l'affectation des données non structurées dans les catégories en temps réel dans des applications comme PredictiveCallCenter. Cette possibilité permet de garantir une mise en œuvre réussie d'opérations de Text Mining en boucles fermées. Par exemple, votre organisation peut désormais analyser de manière plus pertinente les notes consignées dans le Bloc-notes, issues des appelants entrants et sortants, en appliquant vos modèles prédictifs. Votre communication marketing en temps réel est ainsi mieux adaptée. L'utilisation des résultats des modèles Text Mining dans les flux améliore la précision des modèles de données prédictifs.

Remarque : Pour exécuter IBM® SPSS® Modeler Text Analytics avec SPSS Modeler Solution Publisher, ajoutez le répertoire `<install_directory>/ext/bin/spss.TMWBServer` à la variable d'environnement `$LD_LIBRARY_PATH`.

Dans SPSS Modeler Text Analytics , il est souvent fait référence aux concepts et aux catégories extraits. Il est important de comprendre la signification des concepts et des catégories car ils peuvent faciliter la prise de décisions plus avisées lors du travail exploratoire et de la création de modèles.

Concepts et nuggets de modèles de concepts

Lors du processus d'extraction, les données textuelles sont analysées pour découvrir des mots isolés intéressants ou pertinents (par exemple, `élection` ou `paix`) et des groupes de mots (par exemple, `élection présidentielle`, `élection du président` ou `traités de paix`). Ces mots et groupes de mots sont appelés des *termes*. En utilisant les ressources linguistiques,

les termes pertinents sont extraits et les termes similaires sont regroupés sous un terme principal appelé **concept**.

Ainsi, un concept peut représenter plusieurs termes sous-jacents en fonction de votre texte et des ressources linguistiques utilisées. Par exemple, prenons une enquête de satisfaction destinée à des employés et supposons que le concept `salaire` a été extrait. Supposons également que lorsque vous avez examiné les enregistrements associés à `salaire`, vous avez noté que `salaire` n'est pas toujours présent dans le texte mais qu'au lieu de cela certains enregistrements contiennent des termes similaires, tels que `revenu`, `revenus`, et `salaires`. Ces termes sont regroupés sous `salaire` car le moteur du programme d'extraction a déterminé qu'ils étaient similaires ou qu'il s'agissait de synonymes en fonction des règles de traitement ou des ressources linguistiques. Dans ce cas, les documents ou les enregistrements contenant ces termes sont traités de la même manière que s'ils contenaient le mot `salaire`.

Si vous souhaitez prendre connaissance des termes regroupés sous un concept, vous pouvez explorer le concept dans une session interactive ou consulter les synonymes indiqués dans le modèle de concepts. Pour plus d'informations, reportez-vous à la section Termes sous-jacents dans les modèles de concepts sur p. 62.

Un **nugget de modèle de concepts** contient un ensemble de concepts pouvant être utilisés afin d'identifier des enregistrements ou des documents qui contiennent également le concept (notamment ses synonymes ou des groupes de termes). Un modèle de concepts peut être utilisé de deux manières. La première consiste à explorer et à analyser les concepts rencontrés dans le texte source d'origine ou à identifier rapidement les documents intéressants. La seconde consiste à appliquer ce modèle aux nouveaux enregistrements ou documents texte afin d'identifier rapidement les concepts-clés similaires contenus dans ceux-ci (par exemple, la recherche en temps réel de concepts-clés dans les notes d'un centre d'appel).

Pour plus d'informations, reportez-vous à la section Nugget Text Mining : Modèle de concepts sur p. 57.

Catégories et nuggets de modèles de catégories

Vous pouvez créer des **catégories** représentant, essentiellement, des concepts de niveau supérieur ou des rubriques pour capturer les principales idées, les connaissances et les attitudes exprimées dans le texte. Les catégories sont constituées d'un ensemble de descripteurs, tels que des *concepts*, des *types* et des *règles*. Ensemble, ces descripteurs permettent d'identifier si un enregistrement ou un document appartient ou non à une catégorie. Un document ou un enregistrement peut être analysé afin de déterminer si un texte qu'il contient correspond à un descripteur. Si une correspondance est détectée, le document/l'enregistrement est attribué à cette catégorie. Ce processus est appelé **catégorisation**.

Les catégories peuvent être créées automatiquement à l'aide des techniques fiables et automatisées du produit, ou manuellement, en utilisant les informations supplémentaires dont vous disposez concernant les données, ou à l'aide d'une combinaison des deux. Vous pouvez aussi charger un ensemble de catégories prédéfinies à partir d'un package d'analyse de textes grâce à l'onglet Modèle de ce nœud. La création manuelle de catégories ou l'affinement de catégories ne peuvent être effectuées que par l'intermédiaire de la session interactive. Pour plus d'informations, reportez-vous à la section Nœud de Text Mining : onglet Modèle sur p. 37.

Un **nugget de modèle de catégories** contient un ensemble de catégories et les descripteurs associés. Le modèle peut être utilisé afin de regrouper en catégories un ensemble de documents ou d'enregistrements en fonction du texte contenu dans chaque document/enregistrement. Chaque document ou enregistrement est lu et affecté à chaque catégorie pour laquelle une correspondance de descripteur a été identifiée. De cette manière, il est possible d'attribuer un document ou un enregistrement à plus d'une catégorie. Vous pouvez utiliser des nuggets de modèles de catégories pour visualiser les idées essentielles contenues dans des réponses ouvertes à des enquêtes ou dans un ensemble d'entrées de blog, par exemple.

Pour plus d'informations, reportez-vous à la section Nugget Text Mining : Modèle de catégories sur p. 72.

Nœud de modélisation Text Mining

Le nœud de Text Mining applique des techniques linguistiques et de fréquence pour extraire les principaux concepts du texte et créer des catégories avec ces concepts et d'autres données. Grâce à ce nœud, vous pouvez explorer le contenu de données textuelles ou générer soit un nugget de modèle de concepts, soit un nugget de modèle de catégories. Lorsque vous exécutez ce nœud de modélisation, un moteur d'extraction linguistique interne extrait et organise les concepts, les patrons et/ou les catégories à l'aide de méthodes de traitement du langage naturel.

Vous pouvez exécuter le nœud Text Mining et générer automatiquement un nugget de modèle de concepts ou de catégories grâce à l'option Générer directement. Sinon, vous pouvez aussi utiliser une approche plus pragmatique et exploratoire grâce au mode Créer de manière interactive dans lequel vous pouvez non seulement extraire des concepts, créer des catégories et affiner vos ressources linguistiques mais aussi procéder à une analyse des liens du texte et explorer des classes. Pour plus d'informations, reportez-vous à la section Nœud de Text Mining : onglet Modèle sur p. 37.

Vous pouvez trouver ce noeud dans l'onglet IBM® SPSS® Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM® SPSS® Modeler. Pour plus d'informations, reportez-vous à la section IBM SPSS Modeler Text Analytics Noeuds dans le chapitre 1 sur p. 11.

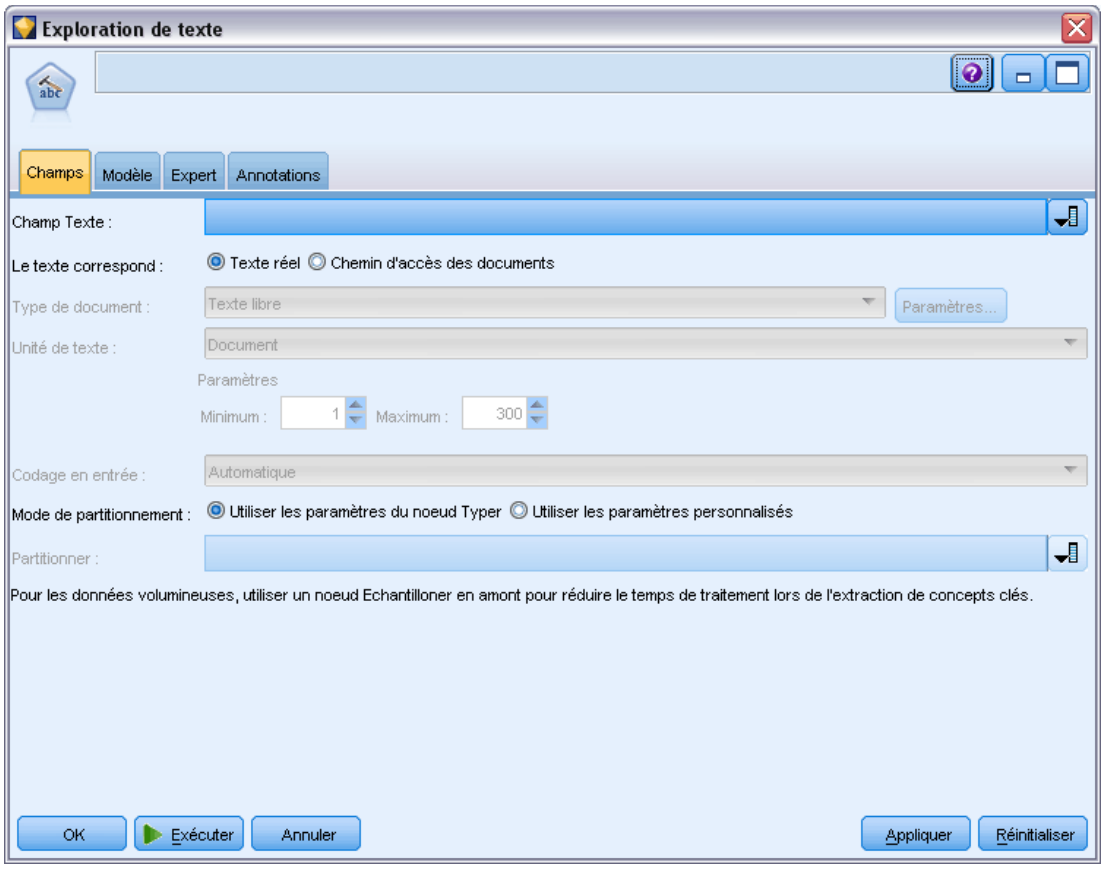
Conditions requises. Les nœuds de modélisation Text Mining acceptent des données textuelles d'un nœud Fil de nouvelles, d'un nœud Liste fichiers ou de nœuds source standard. Le nœud est installé avec SPSS Modeler Text Analytics et est accessible sur la palette SPSS Modeler Text Analytics .

Remarque : Ce nœud remplace le nœud Extraction de texte pour tous les utilisateurs et l'ancien nœud Text Mining pour les utilisateurs japonais, qui était proposé dans des versions précédentes de Text Mining for Clementine. Si vous disposez de flux plus anciens qui utilisent ces nœuds ou des nuggets de modèle, vous devez recréer vos flux à l'aide du nouveau nœud de Text Mining. *Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Nœud de Text Mining : onglet Champs

L'onglet Champs sert à indiquer les paramètres de champ des données dont vous allez extraire les concepts. Pour accélérer la durée du traitement, pensez à utiliser un nœud Echantillonner en amont de ce nœud lorsque vous travaillez avec de grands ensembles de données. Pour plus d'informations, reportez-vous à la section Echantillonnage en amont pour gagner du temps sur p. 50.

Figure 3-1
Boîte de dialogue du nœud de modélisation Text Mining : onglet Champs



Les paramètres pouvant être définis sont les suivants :

Champ Texte. Sélectionnez le champ contenant le texte à explorer, le chemin d'accès au document ou le chemin d'accès au répertoire contenant les documents. Ce champ dépend de la source de données.

Le texte correspond au. Indique ce que contient le champ de texte spécifié dans le paramètre précédent. Les différents choix sont :

- **Texte réel.** Sélectionnez cette option si le champ contient le texte exact à partir duquel les concepts doivent être extraits.
- **Chemin d'accès des documents.** Sélectionnez cette option si le champ contient un ou plusieurs noms de chemin d'accès pour les emplacements des documents texte.

Type de document. Cette option n'est disponible que si vous indiquez que le champ texte correspond au Chemin d'accès des documents. Le type de document indique la structure du texte. Sélectionnez l'un des types suivants :

- **Texte libre.** S'utilise pour la plupart des documents ou des sources textuelles. L'ensemble de texte entier est analysé pour l'extraction. Contrairement aux autres options, il n'existe pas de paramètres supplémentaires pour cette option.

- **Texte structuré.** Utilisez ce format pour les bibliographies, les brevets et les fichiers contenant des structures standard qui peuvent être identifiées et analysées. Ce type de document est utilisé pour éviter tout ou une partie du processus d'extraction. Il permet de définir des séparateurs de termes, d'affecter des types et d'imposer une valeur de fréquence minimale. Si vous sélectionnez cette option, vous devez cliquer sur le bouton Paramètres et entrer les séparateurs de texte dans la zone Formatage de texte structuré de la boîte de dialogue Paramètres du document. Pour plus d'informations, reportez-vous à la section Paramètres de document de l'onglet Champs sur p. 35.
- **Texte XML.** Permet de signaler les balises XML qui contiennent le texte à extraire. Toutes les autres balises sont ignorées. Si vous sélectionnez cette option, vous devez cliquer sur le bouton Paramètres et indiquer de manière explicite les éléments XML contenant le texte à lire au cours du processus d'extraction dans la zone Formatage de texte XML de la boîte de dialogue Paramètres du document. Pour plus d'informations, reportez-vous à la section Paramètres de document de l'onglet Champs sur p. 35.

Unité de texte. Cette option n'est disponible que si vous indiquez que Le texte correspond au chemin d'accès des documents et que vous sélectionnez le type de document Texte libre. Sélectionnez le mode d'extraction parmi les choix suivants :

- **Document.** Utilisez ce mode pour les documents courts et homogènes d'un point de vue sémantique (par exemple, dans le cas d'articles issus d'agences de presse).
- **Paragraphe.** Utilisez ce format pour les pages Web et les documents sans balise. Le processus d'extraction divise les documents en unités sémantiques, sur la base de certaines caractéristiques, comme des balises internes et des éléments syntaxiques. Si ce mode est sélectionné, le scoring est appliquée paragraphe par paragraphe. Par conséquent, la règle `pomme & orange` est vraie uniquement si `pomme` et `orange` se trouvent dans le même paragraphe, par exemple.

Paramètres de Paragraphe. Cette option n'est disponible que si vous avez indiqué que Le texte correspond au chemin d'accès des documents et défini l'option d'unité de texte sur Paragraphe. Indiquez les nombres maximal et minimal de caractères à utiliser dans les extractions. La taille employée est en fait arrondie de manière à inclure les caractères compris avant le point le plus proche (qu'il se situe avant ou après la limite fixée). Pour vous assurer que les associations de mots obtenues à partir du texte du groupe de documents sont représentatives, n'indiquez pas de taille d'extraction trop petite.

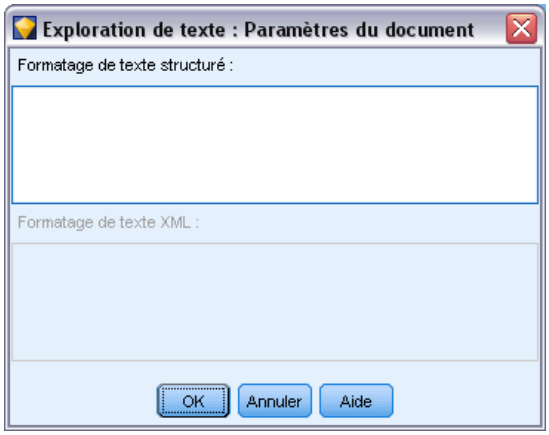
- **Minimum.** Indiquez le nombre minimum de caractères à utiliser dans les extractions.
- **Maximum.** Indiquez le nombre maximal de caractères à utiliser dans les extractions.

Encoding du texte source. Cette option n'est disponible que si vous indiquez que Le texte correspond au chemin d'accès des documents. Elle indique l'encoding de texte par défaut. Dans toutes les langues, à l'exception du japonais, l'encoding spécifié ou reconnu est converti en un encoding ISO-8859-1. Ainsi, même si vous indiquez un autre encoding, le moteur d'extraction le remplace par l'encoding ISO-8859-1 avant de traiter le texte. Les caractères qui ne figurent pas dans la définition de l'encoding ISO-8859-1 sont convertis en espaces. Pour le texte en japonais, vous pouvez choisir l'une des options de codage suivantes : SHIFT_JIS, EUC_JP, UTF-8, ou ISO-2022-JP.

Partitionnement. Le partitionnement permet d'opter pour un partitionnement basé sur les paramètres du nœud Typer ou pour un autre type de partitionnement. La partition répartit les données en échantillons d'apprentissage et de test.

Paramètres de document de l'onglet Champs

Figure 3-2
Boîte de dialogue Paramètres du document



Formatage de texte structuré

Si vous souhaitez éviter une partie ou tout le processus d'extraction parce que vous avez des données structurées ou que vous souhaitez imposer des règles sur la façon de traiter le texte, utilisez l'option de type de document Texte structuré et déclarez les champs ou les balises contenant le texte dans la section Formatage de texte structuré de la boîte de dialogue Paramètres de document. Les termes extraits sont calculés à partir du texte contenu entre les champs ou les balises déclarés (et entre leurs balises enfant). Tout champ ou balise non déclaré sera ignoré.

Dans certains contextes, le traitement linguistique n'est pas obligatoire et le moteur d'extraction linguistique peut être remplacé par des déclarations explicites. Dans un fichier bibliographique où les champs de mot-clé sont séparés par des séparateurs, comme un point-virgule (;) ou une virgule (,), il suffit d'extraire la chaîne comprise entre deux séparateurs. C'est pourquoi il est possible de suspendre le processus d'extraction complet et de définir des règles de traitement spécifiques pour déclarer les séparateurs de termes, affecter les types au texte extrait, ou imposer une fréquence minimale pour l'extraction.

Utilisez les règles suivantes pour déclarer des éléments de texte structuré :

- Un seul champ, balise ou élément peut être déclaré par ligne. Il n'est pas nécessaire de les faire figurer dans les données.
- Les déclarations font la distinction entre majuscules et minuscules.
- Si vous déclarez une balise comportant des attributs, par exemple `<title id="1234">`, et que vous souhaitez inclure toutes les variations, ou dans le cas présent tous les ID, ajoutez la balise sans l'attribut ou le signe « supérieur à » (>), comme `<title`

- Ajoutez le signe deux-points après le nom du champ ou de la balise pour indiquer qu'il s'agit de texte structuré. Ajoutez ces deux points directement après le champ ou la balise mais avant les séparateurs, les types ou les valeurs de fréquence, comme `auteur:` ou `<emplacement>:`.
- Pour indiquer que plusieurs termes sont contenus dans le champ ou la balise et qu'un séparateur est utilisé pour désigner les termes individuels, déclarez le séparateur après le signe deux-points, comme `auteur:,` ou `<section>:;`.
- Pour assigner un type au contenu trouvé dans la balise, déclarez le nom du type après le signe deux-points et un séparateur comme `author:,Person` ou `<place>:;Location`. Déclarez le type à l'aide des noms tels qu'ils apparaissent dans l'éditeur de ressources.
- Pour définir une fréquence minimale pour un champ ou une balise, déclarez un chiffre à la fin de la ligne, comme `author:,Person1` ou `<place>:;Location5`. Où `n` est la fréquence définie : les termes trouvés dans le champ ou la balise doivent se produire au moins `n` fois dans l'ensemble entier de documents ou d'enregistrements à extraire. Il vous faut également définir un séparateur.
- Si vous avez une balise qui contient un signe deux-points, une barre oblique inverse doit précéder ces deux points afin que la déclaration ne soit pas ignorée. Par exemple, si vous avez un champ nommé `<topic:source>`, saisissez-le sous la forme `<topic\:source>`.

Pour illustrer la syntaxe, imaginons les champs bibliographiques récurrents suivants :

```
author:Morel, Kawashima
abstract:cet article décrit la façon dont les champs sont déclarés.
publication:Documentation de Text Mining
datepub:Mars 2010
```

Dans cet exemple, si nous souhaitions que le processus d'extraction considère l'auteur et l'extrait mais ignore le reste du contenu, nous déclarerions uniquement les champs suivants :

```
author:,Person1
abstract:
```

Dans cet exemple, la déclaration de champ `author:,Person1` indique que le traitement linguistique du contenu des champs a été suspendu. Au lieu de cela, il spécifie que le champ auteur contient plus d'un nom, lequel est séparé du suivant par le séparateur virgule, et que les noms doivent être attribués au type Personne et que si le nom apparaît au moins une fois dans l'ensemble complet des documents ou des enregistrements, il doit être extrait. Parce que le champ `abstract:` est affiché sans autre déclaration, le champ sera analysé pendant le processus d'extraction et le traitement linguistique standard et la définition de typage seront appliqués.

Formatage de texte XML

Si vous souhaitez limiter le processus d'extraction au texte contenu entre des balises XML spécifiques uniquement, utilisez l'option de type de document texte XML pour déclarer les balises contenant le texte dans la section formatage de texte XML de la boîte de dialogue Paramètres du document. Les termes extraits sont calculés à partir du texte contenu entre ces balises ou entre leurs balises enfant.

Important ! Si vous souhaitez éviter le processus d'extraction et imposer des règles sur les séparateurs de termes, affecter des types au texte extrait ou imposer une fréquence aux termes extraits, utilisez l'option texte structuré suivante.

Lorsque vous déclarez des balises pour le formatage de texte XML, utilisez les règles suivantes :

- Une seule balise XML peut être déclarée par ligne.
- Les éléments de balise respectent la casse.
- Si une balise comporte des attributs, par exemple `<title id="1234">`, et que vous souhaitez inclure toutes les variations, ou dans le cas présent tous les ID, ajoutez la balise sans l'attribut ou le signe « supérieur à » (>), comme `<title`

Pour illustrer la syntaxe, imaginons le document XML suivant :

```
<section>Rules of the Road
     <title id="01234">Traffic Signals</title>
     <p>Road signs are helpful.</p>
</section>
<p>Learning the rules is important.</p>
```

Pour cet exemple, déclarons les balises suivantes :

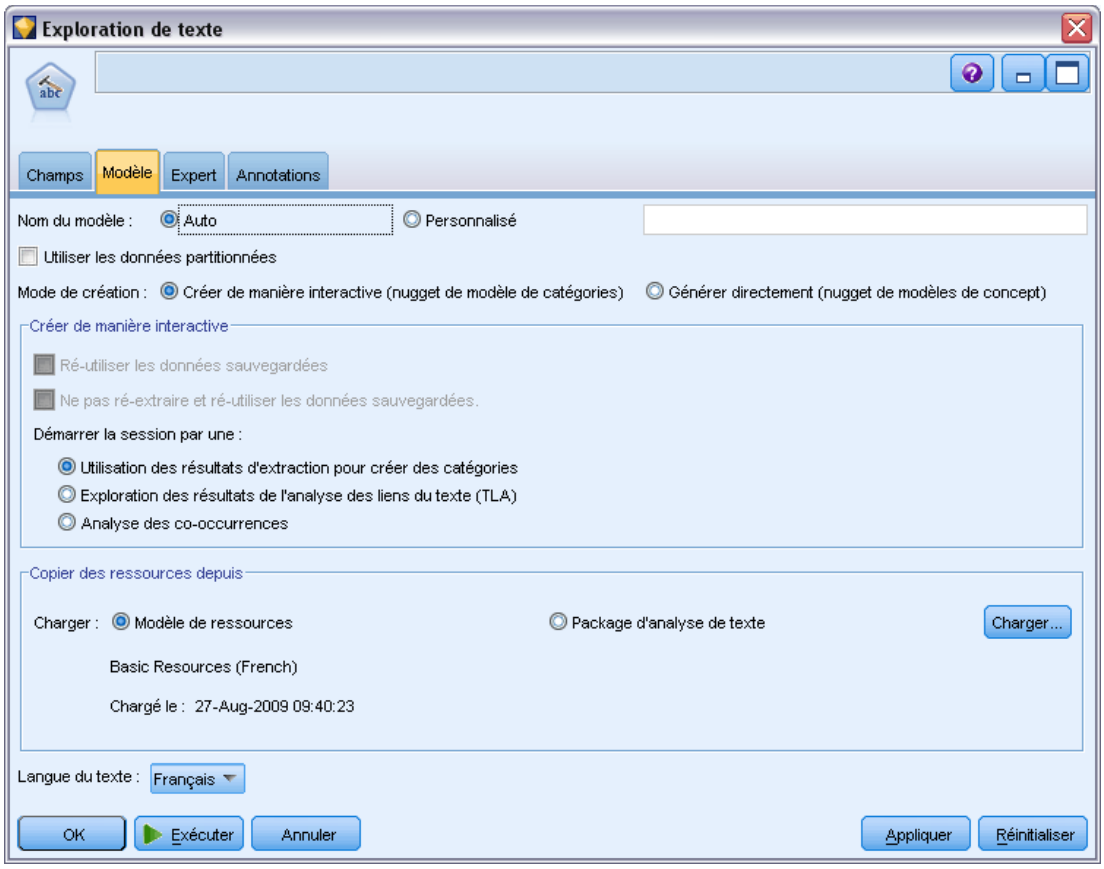
```
<section>
<title
```

Dans cet exemple, parce que vous avez déclaré la balise `<section>`, le texte à l'intérieur de cette balise et de ses balises imbriquées, `Traffic Signals` et `Road signs are helpful`, est analysé pendant le processus d'extraction. Mais, `Learning the rules is important` est ignoré, car cette balise`<p>` n'a pas été explicitement déclarée et qu'elle n'a pas été imbriquée dans une balise déclarée.

Nœud de Text Mining : onglet Modèle

L'onglet Modèle est utilisé pour indiquer la méthode de création et les paramètres de modèle généraux pour la sortie du nœud.

Figure 3-3
Boîte de dialogue du nœud de Text Mining : onglet Modèle



Les paramètres pouvant être définis sont les suivants :

Nom du modèle. Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

Utiliser les données partitionnées. Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle.

Mode Création. Indique la façon dont les nuggets de modèle sont générés lorsqu'un flux contenant ce nœud de Text Mining sera exécuté. Sinon, vous pouvez aussi utiliser une approche plus pragmatique et exploratoire grâce au mode Créer de manière interactive dans lequel vous pouvez non seulement extraire des concepts, créer des catégories et affiner vos ressources linguistiques mais aussi procéder à une analyse des liens du texte et explorer des clusters.

- **Créer de manière interactive.** Lorsqu'un flux est exécuté, cette option lance une interface interactive dans laquelle vous pouvez extraire des concepts et des patrons, explorer et affiner les résultats extraits, créer et affiner des catégories, affiner les ressources linguistiques

(modèles, synonymes, types, bibliothèques, etc.) et créer des nuggets de modèle de catégories. Pour plus d'informations, reportez-vous à la section Créer de manière interactive sur p. 39.

- **Générer directement.** Cette option indique, que lorsque le flux est exécuté, un modèle doit être automatiquement créé et ajouté à la palette Modèles. À la différence de la session interactive, aucune manipulation supplémentaire n'est nécessaire de votre part au moment de l'exécution outre les paramètres définis dans le nœud. Si vous sélectionnez cette option, des options propres au modèle apparaissent et vous permettent de définir le type de modèle que vous souhaitez produire. Pour plus d'informations, reportez-vous à la section Générer directement sur p. 41.

Copiez les ressources de. Lors de l'exploration de texte, l'extraction est basée sur les paramètres de l'onglet Expert mais également sur les ressources linguistiques. Ces ressources servent de base au traitement et à l'exécution du texte pendant l'extraction afin d'obtenir des concepts, des types et parfois des patrons. Vous pouvez copier les ressources dans ce nœud à partir d'un modèle de ressources ou d'un package d'analyse de texte. Sélectionnez-en un puis cliquez sur Charger pour définir le package ou le modèle depuis lequel les ressources seront copiées. Au moment du chargement, une copie des ressources est stockée dans le nœud. Par conséquent, si vous souhaitez utiliser un modèle mis à jour ou un TAP (package d'analyse de texte), vous devez le recharger ici ou dans une session interactive. Pour faciliter votre travail, la date et l'heure de copie et de chargement des ressources sont indiquées dans le nœud. Pour plus d'informations, reportez-vous à la section Copie des ressources à partir de modèles et de TAP sur p. 42.

Langue du texte. Identifie la langue du texte exploré. Les ressources copiées dans le noeud contrôlent les options de langue présentées. Vous pouvez sélectionner la langue correspondant aux ressources ou choisir l'option TOUTES. Nous vous recommandons vivement de spécifier la langue exacte des données textuelles ; cependant, si vous avez un doute, vous pouvez choisir l'option TOUTES. TOUTES ne s'applique pas au texte en japonais. L'option TOUTES allonge la durée d'exécution car la reconnaissance automatique de la langue est utilisée pour analyser tous les documents et enregistrements de façon à identifier d'abord la langue du texte. Avec cette option, tous les enregistrements ou documents dont la langue est prise en charge et fait l'objet d'une licence sont lus par le moteur d'extraction à l'aide des dictionnaires internes propres aux langues. Pour plus d'informations, reportez-vous à la section Identificateur de langue dans le chapitre 18 sur p. 336. Si vous souhaitez acquérir la licence d'une langue prise en charge à laquelle vous n'avez pas accès actuellement, contactez votre représentant commercial.

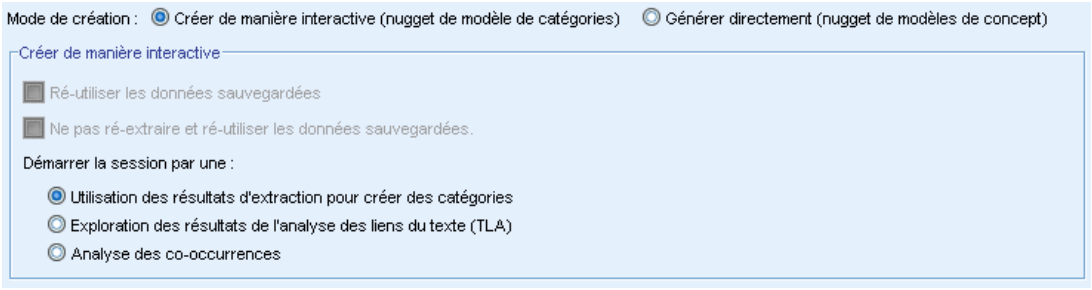
Créer de manière interactive

Dans l'onglet Modèle du nœud de modélisation de Text Mining, vous pouvez choisir un mode de création pour vos nuggets de modèle. Si vous sélectionnez Créer de manière interactive, une interface interactive s'ouvre lorsque vous exécutez le flux. Dans cette session interactive, vous pouvez effectuer les opérations suivantes :

- Extraire et explorer les résultats de l'extraction, y compris des concepts et typages pour découvrir les idées fondamentales dans vos données de texte.
- Utilisez différentes méthodes pour créer et étendre des catégories provenant de concepts, de types, de patrons TLA et de règles et pouvoir scorer vos documents et enregistrements dans ces catégories.

- Affinez vos ressources linguistiques (modèles de ressources, bibliothèques, dictionnaires, synonymes, etc.) pour pouvoir améliorer vos résultats grâce à un processus itératif dans lequel les concepts sont extraits, examinés et affinés.
- Procéder à une analyse des liens du texte (TLA) et utiliser les patrons TLA découverts afin de créer de meilleurs nuggets de modèle de catégories. Le nœud Analyse des liens du texte ne propose pas les mêmes options exploratoires ni les mêmes capacités de modélisation.
- Générez des clusters pour découvrir de nouvelles relations ou explorer des relations entre les concepts, les types, les patrons et les catégories dans le panneau Visualisation.
- Générez des nuggets de modèles de catégories affinés dans la palette Modèles de IBM® SPSS® Modeler et utilisez-les dans d'autres flux.

Figure 3-4
Options de l'onglet Modèle pour créer de manière interactive



Ré-utiliser les données sauvegardées. Lorsque vous travaillez dans une session interactive, vous pouvez mettre à jour le nœud avec les données de la session (paramètres d'extraction, ressources, définitions des catégories, etc.). L'option Utiliser la session interactive permet de redémarrer la session interactive en utilisant les données de la session enregistrées. Cette option est désactivée la première que vous utilisez ce nœud, car aucune donnée de session n'a pu être enregistrée. Pour apprendre à mettre à jour le nœud avec les données de session et utiliser cette option, consultez Mise à jour des nœuds de modélisation et enregistrement sur p. 136.

Si vous lancez une session *avec* cette option, alors les paramètres d'extraction, les catégories, les ressources et tout autre travail effectué lors de la dernière mise à jour du nœud depuis une session interactive seront disponibles au prochain lancement de session. Les données de session enregistrées étant utilisées avec cette option, certains contenus, comme les ressources copiées dans le modèle ci-dessous et d'autres onglets, sont désactivés et ignorés. Mais si vous lancez une session *sans* cette option, seuls les contenus du nœud comme définis actuellement sont utilisés, ce qui signifie que tout travail précédent effectué dans l'utilitaire ne sera pas disponible.

Remarque : Si vous modifiez le nœud source de votre flux après la mise en cache des résultats d'extraction à l'aide de l'option Utiliser le travail d'une session..., vous devrez exécuter une nouvelle extraction une fois la session interactive démarrée, pour obtenir les résultats d'extraction mis à jour.

Ne pas ré-extraire et ré-utiliser les données cachées et les résultats. Vous pouvez ré-utiliser les résultats et les données cachés de l'extraction dans la session interactive. Cette option est particulièrement utile lorsque vous souhaitez gagner du temps et ré-utiliser les résultats de l'extraction plutôt que d'attendre l'exécution d'une toute nouvelle extraction lors du lancement de la session. Afin d'utiliser cette option, vous devez d'abord avoir mis à jour ce nœud depuis une session interactive et avoir choisi l'option de Conserver la session interactive et les données de texte cachées avec les résultats de l'extraction pour une utilisation ultérieure. Pour apprendre à mettre à

jour le nœud avec les données de session et utiliser cette option, consultez Mise à jour des nœuds de modélisation et enregistrement sur p. 136

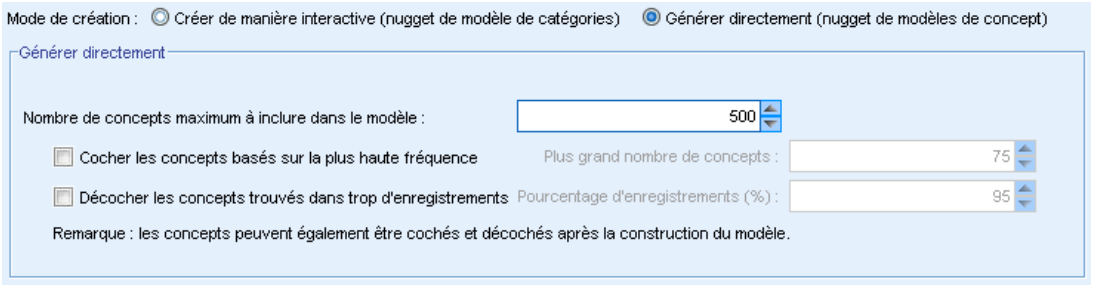
Démarrer la session par. Sélectionner l'option indiquant la vue à afficher et l'action à effectuer en premier lors du lancement de la session interactive. Quelle que soit la vue dans laquelle vous commencez, une fois la session ouverte, vous pouvez choisir n'importe quelle vue.

- **Utilisation des résultats de l'extraction pour créer des catégories.** Cette option lance la session interactive dans la vue Catégories et concepts et, le cas échéant, effectue une extraction. Dans cette vue, il est possible de créer des catégories et de générer un modèle de catégories. Vous pouvez également afficher une autre vue. Pour plus d'informations, reportez-vous à la section Mode Session interactive dans le chapitre 8 sur p. 116.
- **Exploration des résultats de l'analyse des liens du texte (TLA).** Au démarrage, cette option commence par extraire et identifier les relations entre les concepts contenus dans le texte, comme les opinions ou les autres liens de la vue Analyse des liens du texte. Vous devez sélectionner un modèle ou un package d'analyse de texte qui contient des règles de patrons TLA pour utiliser cette option et obtenir des résultats. Si vous travaillez avec des ensembles de données plus importants, l'extraction TLA peut prendre du temps. Dans ce cas, vous pouvez envisager d'utiliser un nœud Echantillonner en amont. Pour plus d'informations, reportez-vous à la section Exploration de l'analyse des liens du texte dans le chapitre 12 sur p. 245.
- **Analyse des co-occurrences.** Cette option est lancée dans la vue Clusters et met à jour les résultats d'extraction obsolètes. Dans cette vue, vous pouvez effectuer une analyse des classes de co-termes, ce qui produit un ensemble de classes. La classification non supervisée des co-termes est un processus qui commence par évaluer la force de la valeur du lien entre deux concepts d'après leur cooccurrence dans un enregistrement ou un document spécifique et se termine par le regroupement des concepts fortement liés dans des clusters. Pour plus d'informations, reportez-vous à la section Mode Session interactive dans le chapitre 8 sur p. 116.

Générer directement

Dans l'onglet Modèle du nœud de modélisation de Text Mining, vous pouvez choisir un mode de création pour vos nuggets de modèle. Si vous sélectionnez Générer directement, vous pouvez définir les options dans le nœud puis simplement exécuter votre flux. La sortie est un nugget de modèle de concepts qui a été directement placé dans la palette Modèles. À la différence de la session interactive, aucune manipulation supplémentaire n'est nécessaire de votre part au moment de l'exécution outre les paramètres de fréquence définis pour cette option dans le nœud.

Figure 3-5
Options de l'onglet Modèle pour générer directement un modèle



Nombre maximum de concepts à inclure dans le modèle. Cette option, valable uniquement lorsque vous créez un modèle automatiquement (non interactif), indique que vous souhaitez créer un modèle de concepts. Elle indique également que ce modèle ne doit pas contenir plus que le nombre indiqué de concepts.

- **Activer les concepts en fonction de la fréquence la plus élevée. Plus grand nombre de concepts.** Il s'agit du nombre de concepts qui seront cochés, en partant de celui dont la fréquence est la plus élevée. Le terme Fréquence fait ici référence au nombre d'occurrences d'un concept (et de tous ses termes sous-jacents) dans l'ensemble des documents/enregistrements. Il est parfois supérieur aux effectifs des enregistrements car un concept peut figurer plusieurs fois dans un même enregistrement.
- **Désactiver les concepts apparaissant dans trop d'enregistrements. Pourcentage d'enregistrements.** Désactiver les concepts apparaissant dans un pourcentage supérieur à celui indiqué pour le nombre d'enregistrements. Cette option permet d'exclure les concepts qui figurent fréquemment dans le texte ou les enregistrements, mais qui ne présentent pas d'intérêt pour l'analyse.

Copie des ressources à partir de modèles et de TAP

Lors de l'exploration de texte, l'extraction est basée sur les paramètres de l'onglet Expert mais également sur les ressources linguistiques. Ces ressources servent de base au traitement et à l'exécution du texte pendant l'extraction afin d'obtenir des concepts, des types et parfois des patrons. Vous pouvez copier des ressources dans ce nœud depuis un *modèle de ressources*, et si vous êtes dans le nœud Text Mining, vous pouvez également sélectionner un *package d'analyse de texte* (TAP).

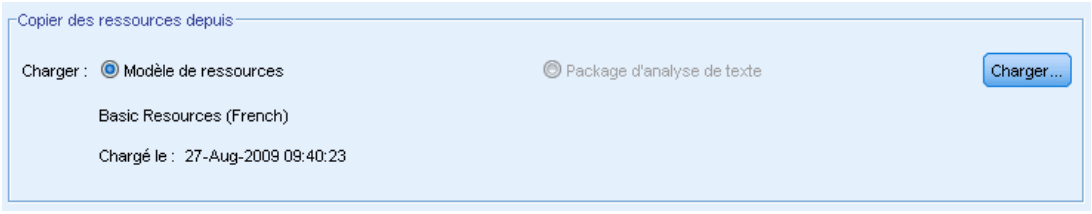
Par défaut, les ressources sont copiées dans le nœud du modèle de base de la langue dont vous possédez la licence pour votre produit lorsque vous ajoutez le nœud à l'espace de travail. Si vous possédez des licences pour plusieurs langues, la première langue sélectionnée est utilisée pour déterminer le modèle à charger automatiquement.

Au moment du chargement, une copie des ressources sélectionnées est stockée dans le nœud. Seul le contenu du modèle ou du TAP est copié, mais le modèle ou TAP n'est pas lui-même lié au nœud. Par conséquent, si ce modèle ou TAP est par la suite mis à jour, ces mises à jour ne sont pas automatiquement disponibles dans le nœud. Pour résumer, les ressources chargées dans le nœud sont toujours utilisées sauf si vous rechargez une copie d'un modèle ou d'un TAP ou si vous mettez à jour un nœud Text Mining et sélectionnez l'option Utiliser le travail d'une session. Pour des informations sur l'option Utiliser le travail d'une session, consultez la rubrique suivante.

Lorsque vous sélectionnez un modèle ou TAP, choisissez-en un ayant le même langage que vos données texte. Vous ne pouvez utiliser que les modèles ou TAP définis dans les langues pour lesquelles vous détenez une licence. Pour effectuer une analyse des liens du texte, vous devez sélectionner un modèle contenant des patrons TLA. Si un modèle contient des patrons TLA, une icône s'affiche dans la colonne TLA de la boîte de dialogue Charger le modèle de ressources.

Remarque : Vous ne pouvez pas charger les TAP dans le nœud d'analyse des liens du texte.

Figure 3-6
Nœud de Text Mining, onglet Modèle : options pour copier les ressources dans le nœud

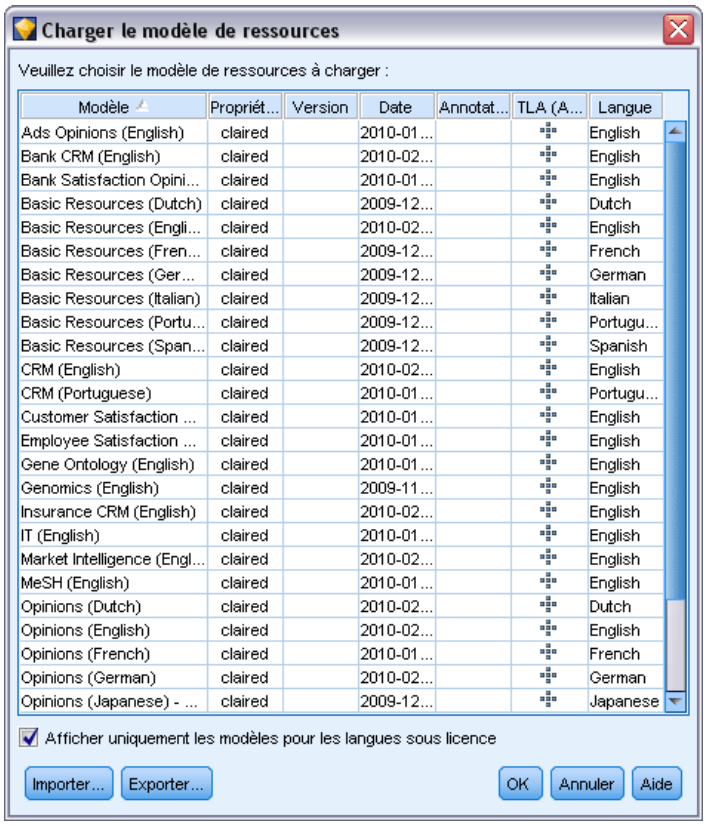


Modèles de ressources

Un modèle de ressources est un ensemble prédéfini de bibliothèques et de ressources linguistiques et non linguistiques avancées qui ont été affinées pour un domaine ou une utilisation spécifique. Dans le nœud de modélisation de text mining, une copie des ressources d'un modèle de base est déjà chargée dans le nœud lorsque vous ajoutez le nœud au flux, mais vous pouvez modifier les modèles ou charger un package d'analyse de texte en sélectionnant Modèle de ressources ou Package d'analyse de texte puis en cliquant sur Charger. Pour les modèles, vous pouvez ensuite sélectionner le modèle dans la boîte de dialogue Charger un modèle de ressources.

Remarque : Si le modèle désiré ne se trouve pas dans la liste mais que vous possédez une copie exportée sur votre ordinateur, vous pouvez alors l'importer. Vous pouvez également exporter depuis cette boîte de dialogue pour un partage avec d'autres utilisateurs. Pour plus d'informations, reportez-vous à la section Import et export des modèles dans le chapitre 15 sur p. 283.

Figure 3-7
Boîte de dialogue Charger le modèle de ressources



Packages d'analyse de texte (TAP)

Un package d'analyse de texte (TAP) est un ensemble prédéfini de bibliothèques et de ressources linguistiques et non linguistiques avancées associées à un ou à plusieurs ensembles de catégories prédéfinies. IBM® SPSS® Modeler Text Analytics propose désormais plusieurs TAP prédéfinis pour des textes en langue anglaise et également pour des textes en japonais, chacun étant affiné pour un domaine spécifique. Vous ne pouvez pas éditer ces TAP mais vous pouvez les utiliser pour commencer votre création de modèle de catégorie. Vous pouvez également créer vos propres TAP dans la session interactive. Pour plus d'informations, reportez-vous à la section Chargement des packages d'analyse de texte dans le chapitre 10 sur p. 227.*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Remarque : Vous ne pouvez pas charger les TAP dans le nœud d'analyse des liens du texte.

Utilisation de l'option « Utiliser le travail d'une session » (onglet Modèle)

Alors que les ressources sont copiées dans le nœud de l'onglet Modèle, vous pouvez également effectuer des modifications ultérieures sur les ressources dans une session interactive et mettre à jour le nœud de modélisation de text mining avec ces dernières modifications. Dans ce cas, vous

pouvez sélectionner l'option Utiliser le travail d'une session dans l'onglet Modèle du nœud de modélisation de text mining.

Si vous sélectionnez Utiliser le travail d'une session, le bouton Charger est désactivé dans le nœud pour indiquer que ces ressources qui provenaient de la session interactive seront utilisées à la place des ressources chargées précédemment.

Après avoir sélectionné l'option Utiliser le travail d'une session, vous pouvez modifier vos ressources directement dans la session interactive avec la vue Editeur de ressources. Pour plus d'informations, reportez-vous à la section Mise à jour des ressources d'un nœud après le chargement dans le chapitre 15 sur p. 281.

Nœud de Text Mining : Onglet Expert

L'onglet Expert contient des paramètres avancés ayant une incidence sur le mode d'extraction et de traitement du texte. Les paramètres de cette boîte de dialogue déterminent le fonctionnement de base du processus d'extraction, ainsi que quelques procédures avancées. Cependant, ils ne représentent qu'une partie des options disponibles. Il existe également un certain nombre de ressources linguistiques et d'options ayant une incidence sur les résultats de l'extraction, qui sont contrôlées par le modèle de ressources sélectionné dans l'onglet Modèle. Pour plus d'informations, reportez-vous à la section Nœud de Text Mining : onglet Modèle sur p. 37.

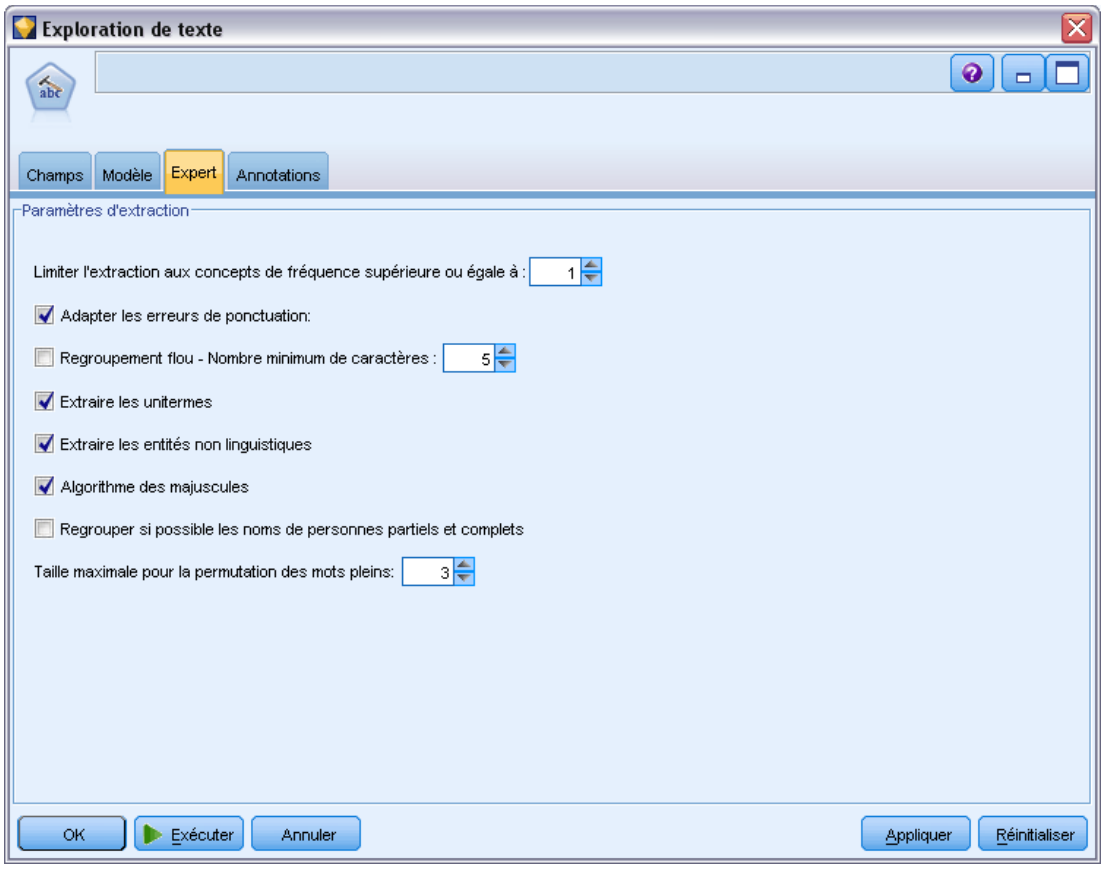
Remarque : Cet onglet est entièrement désactivé si vous avez sélectionné le mode Créer de manière interactive à l'aide des informations de la session interactive enregistrées dans l'onglet Modèle. Dans ce cas, les paramètres d'extraction utilisés sont ceux de la dernière session interactive enregistrée.

Pour le texte en néerlandais, en anglais, en français, en allemand, en italien, en portugais, et en espagnol

Vous pouvez définir les paramètres suivants à chaque extraction de langues autres que le japonais par exemple l'anglais, l'espagnol, le français, l'allemand, etc :

Remarque : Consultez la suite de cette rubrique pour obtenir des informations sur les paramètres Expert pour le texte en japonais. L'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Figure 3-8
Boîte de dialogue du nœud de Text Mining : Onglet Expert



Limiter l'extraction aux concepts de fréquence supérieure ou égale à [n]. Spécifie à partir de combien d'occurrences un mot ou un groupe de mots présent dans un texte doit être extrait. Ainsi, une valeur de 5 limite l'extraction aux mots ou groupes de mots figurant au moins cinq fois dans l'ensemble des enregistrements ou des documents.

Dans certains cas, modifier cette limite peut faire une grande différence dans les résultats d'extraction et par conséquent, dans les catégories. Imaginons que vous travaillez avec des données concernant un restaurant et que vous n'avez pas augmenté la limite au-dessus de 1 pour cette option. Dans ce cas, vos résultats d'extraction pourront contenir *pizza (1), pizza fine (2), pizza épinards (2)*, et *pizza préférée (2)*. Mais si l'extraction était limitée à une fréquence globale de 5 ou plus et que vous recommenciez l'extraction, trois de ces concepts ne seraient pas renvoyés. Vous obtiendriez *pizza (7)*, car *pizza* est la forme la plus simple et que ce mot existait déjà comme candidat possible. Et en fonction du reste du texte, vous pourriez obtenir une fréquence supérieure à 7, si le texte contient d'autres phrases avec le mot pizza. De plus, si *pizza épinards* était déjà un descripteur de catégorie, vous pourriez le remplacer par*pizza* pour pouvoir capturer tous les enregistrements. C'est pour cette raison que lorsque des catégories ont déjà été créées, la modification de cette limite doit être effectuée avec prudence.

Veuillez noter qu'il s'agit d'une fonctionnalité d'extraction uniquement ; si votre modèle contient des termes (ce qui est généralement le cas) et qu'un terme pour le modèle est trouvé dans le texte, alors le terme sera indexé quelle que soit sa fréquence.

Par exemple, imaginons que vous utilisez le modèle Ressources de base qui contient « los angeles » sous le type <Location> dans Core library ; si votre document contient Los Angeles une fois seulement, alors Los Angeles fera partie de la liste des concepts. Pour éviter cela, vous devrez définir un filtre pour afficher les concepts se produisant au moins le même nombre de fois que la valeur saisie dans le champ Limiter l'extraction aux concepts ayant une fréquence globale supérieure à [n].

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Traitement des fautes de frappe. Nombre de caractères minimum requis ([n]). Cette option applique une technique de regroupement flou qui permet de regrouper les mots mal orthographiés ou d'orthographe similaire sous un seul concept. L'algorithme de regroupement flou retire temporairement toutes les voyelles (à l'exception de la première) et les consonnes doubles/triples des mots extraits pour comparer ces derniers et voir s'ils sont identiques, afin par exemple que `modélisation` et `modelisation` soient regroupés. Néanmoins, si chaque type est affecté à un type différent et que vous excluez le type `<Unknown>`, le regroupement flou n'est pas appliqué.

Vous pouvez également définir le nombre minimum de caractères *racines* requis avant d'utiliser le regroupement flou. Le nombre de caractères racine d'un terme est calculé en ajoutant l'ensemble des caractères et en soustrayant les caractères formant des suffixes flexionnels et, dans le cas des termes apparaissant sous la forme de mots composés, les déterminants et les prépositions. Par exemple, le terme `exercices` serait déterminé comme comportant 8 caractères racine dans la forme « exercice », étant donné que la lettre *s* située à la fin du mot représente une flexion (forme plurielle). De même, `jus énergétique` comporte 14 caractères racine « jus énergétique » et `conception de dessins`, 16 caractères racine « conception dessin ». Cette méthode de calcul est uniquement utilisée pour vérifier si le regroupement flou doit être appliqué, mais n'influe pas sur la mise en correspondance des mots.

Remarque : Si, d'après vous, certains mots sont ensuite groupés de manière incorrecte, vous pouvez exclure des paires de mots de cette technique en les déclarant de manière explicite dans la section Regroupement flou : Exceptions de l'onglet de ressources avancées. Pour plus d'informations, reportez-vous à la section Regroupement flou dans le chapitre 18 sur p. 328.

Extraire les unitermes. Cette option extrait les mots uniques (unitermes) dans la mesure où le mot ne fait pas déjà partie d'un mot composé et si c'est un nom ou une catégorie grammaticale non reconnue.

Extraction des entités non linguistiques. Cette option extrait les entités non linguistiques, telles que les numéros de téléphone, numéros de sécurité sociale, heures, dates, devises, chiffres, pourcentages, adresses électroniques, adresses HTTP, etc. Vous pouvez inclure ou exclure certains types d'entités non linguistiques dans la section Entités non linguistiques : configuration de l'onglet de ressources avancées. Désactivez les entités dont vous n'avez pas besoin pour éviter au moteur d'extraction un temps de traitement inutile. Pour plus d'informations, reportez-vous à la section Configuration dans le chapitre 18 sur p. 332.

Algorithme des noms propres. Cette option extrait les termes simples et composés ne figurant pas dans les dictionnaires intégrés, si la première lettre est une majuscule. Cette option offre une bonne manière d'extraire la plupart des noms propres.

Regroupement éventuel des noms de personnes partiels et complets. Cette option regroupe les noms qui apparaissent différemment dans le texte. Cette fonction est utile car les noms sont souvent cités dans leur forme complète au début du texte puis uniquement en version abrégée. Cette option essaye de faire correspondre tout uniterme ayant le type `<Unknown>` avec le dernier mot de tout terme composé entré comme `<Person>`. Par exemple, si *martin* est trouvé et initialement saisi comme `<Unknown>`, le moteur d'extraction vérifie si un terme composé de type `<Person>` contient *martin* comme dernier mot, tel que *pierre martin*. Cette option ne s'applique pas aux prénoms, car la plupart ne sont jamais extraits comme unitermes.

Nombre de mots pleins soumis à une permutation pour le regroupement. Cette option indique le nombre maximal de mots utiles pouvant être présents lorsque s'applique la technique de permutation. Cette technique de permutation regroupe les expressions similaires qui ne diffèrent les unes des autres que par la présence de mots utiles (par exemple, de et la), quelle que soit leur flexion. Par exemple, si vous avez paramétré cette valeur sur au moins deux mots et si les deux expressions `représentants d'entreprise` et `représentants de l'entreprise` ont été extraites. Dans ce cas, les deux termes extraits sont regroupés dans la liste de concepts finale car les deux termes sont considérés comme étant les mêmes lorsque `de l'` est ignoré.

Remarque : Pour activer l'extraction des résultats de l'analyse des liens du texte, vous devez démarrer la session avec l'option Exploration des résultats de l'analyse des liens du texte et également choisir les ressources qui contiennent des définitions TLA. Vous pouvez toujours extraire les résultats TLA ultérieurement pendant une session interactive à partir de la boîte de dialogue Paramètres d'extraction. Pour plus d'informations, reportez-vous à la section Extraction de données dans le chapitre 9 sur p. 143.

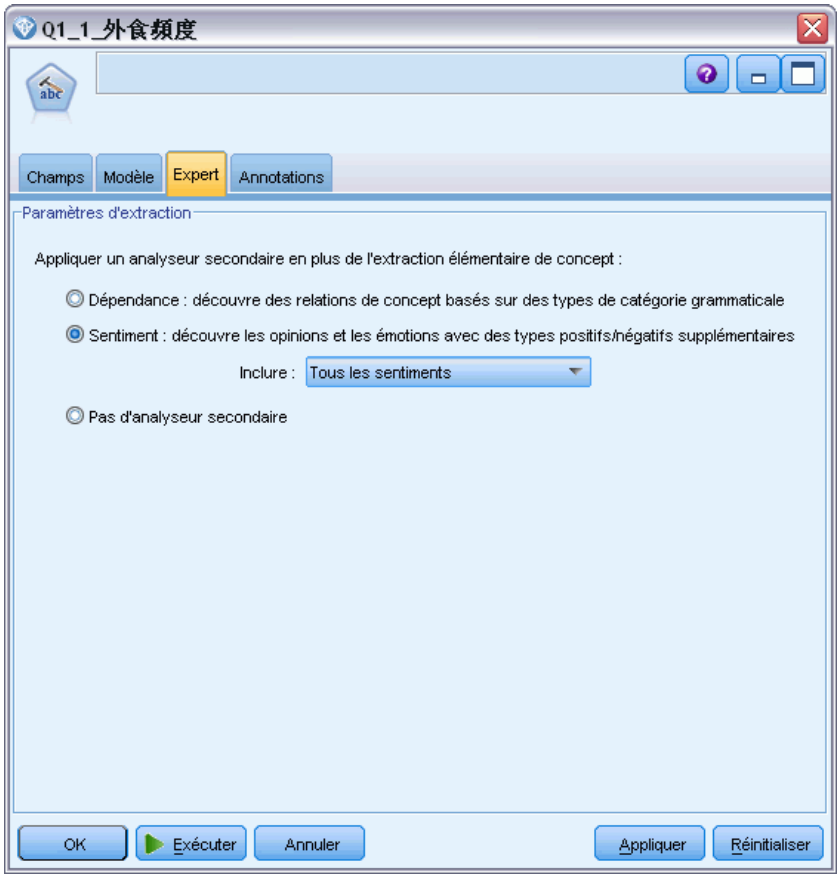
Pour les textes en japonais

Remarque : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

La boîte de dialogue contient des options différentes pour le texte en japonais car le processus d'extraction n'est pas exactement le même. Afin de pouvoir travailler avec des textes en japonais, vous devez également sélectionner un modèle ou un package d'analyse de texte adapté au japonais dans l'onglet Modèle de ce nœud. Pour plus d'informations, reportez-vous à la section Copie des ressources à partir de modèles et de TAP sur p. 42.

Les paramètres pouvant être définis sont les suivants :

Figure 3-9
Boîte de dialogue du nœud de Text Mining : Onglet Expert (texte en japonais)



Remarque : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Analyse secondaire. Lorsqu'une extraction est lancée, l'extraction des mots-clés de base est effectuée à l'aide de l'ensemble de types par défaut. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais dans l'annexe A sur p. 378. Mais lorsque vous sélectionnez un analyseur secondaire, vous pouvez obtenir des concepts plus nombreux et plus riches car l'extracteur inclura désormais des verbes à particules et des auxiliaires comme faisant partie du concept. Par exemple, supposons que nous avons une phrase 肩の荷が下りた, traduite par "*Ca m'a enlevé un gros poids*". Dans cet exemple, l'extraction des mots-clés de base peut extraire chaque concept séparément comme suit : 肩 (*poids*), 荷 (*gros*), 下りる (*a enlevé*), mais la relation entre ces mots n'est pas extraite. Cependant, si vous avez appliqué l'analyse de sentiment, vous pouvez extraire des concepts plus riches relatifs à un type de sentiment comme le concept =肩の荷が下りた, qui est traduit par "*avoir enlevé un gros poids*", affecté au type <良い-安心>. Dans le cas d'une analyse de sentiment, un grand nombre de types supplémentaires est également inclus. De plus, choisir un analyseur secondaire vous permet également de générer des résultats d'analyse des liens du texte.

Remarque : Lorsqu'un analyseur secondaire est appelé, le processus d'extraction nécessite plus de temps. Pour plus d'informations, reportez-vous à la section Fonctionnement de l'extraction secondaire dans l'annexe A sur p. 371.

- **Analyse des dépendances** Choisir cette option génère des particules étendues pour les concepts d'extraction du type de base et de l'extraction des mots-clés. Vous pouvez également obtenir des résultats de patrons plus riches avec l'analyse de dépendance des liens du texte (TLA).
- **Analyse des sentiments**. Choisir cet analyseur génère l'extraction de concepts supplémentaires et, le cas échéant, l'extraction de résultats de patrons TLA. En plus des types de base, vous bénéficiez également de plus de 80 types de sentiments, notamment 嬉しい, 吉報, 幸運, 安心, 幸福, etc. Ces types permettent de découvrir des concepts et des patrons dans le texte grâce à l'expression des émotions, des sentiments et des opinions. Ce sont trois options qui dictent la cible de l'analyse des sentiments : Tous les sentiments, Sentiment représentatif uniquement et Conclusions uniquement.
- **Pas d'analyseur secondaire**. Cette option désactive tous les analyseurs secondaires. Cette option est masquée si l'option Exploration des résultats de l'analyse des liens du texte (TLA) a été sélectionnée dans l'onglet Modèle car une analyse secondaire est nécessaire afin d'obtenir les résultats TLA. Si vous sélectionnez cette option mais choisissez ensuite l'option Exploration des résultats de l'analyse des liens du texte (TLA), une erreur surviendra pendant l'exécution du flux.

Echantillonnage en amont pour gagner du temps

La durée de traitement d'une grande quantité de données peut aller de plusieurs minutes à plusieurs heures, particulièrement lors de l'utilisation de la session interactive. Plus la taille des données est importante, plus la durée des processus d'extraction et de catégorisation est longue. Pour travailler plus efficacement, il est possible d'ajouter l'un des nœuds Echantillonner de IBM® SPSS® Modeler en amont de votre nœud Text Mining. Utilisez ce nœud Echantillonner pour prendre un échantillon aléatoire à l'aide d'un sous-ensemble de documents ou d'enregistrements moins important et d'effectuer les premiers transferts.

Un échantillon de taille moins importante est souvent parfaitement indiqué pour décider de la façon dont vous modifiez vos ressources et même créez la plupart, voire toutes, vos catégories. Une fois l'opération exécutée sur ce petit sous-ensemble de données et les résultats escomptés obtenus, vous pouvez appliquer la même technique pour créer des catégories dans l'ensemble de données entier. Vous pouvez ensuite rechercher des documents et des enregistrements qui ne font pas partie des catégories créées et faire les ajustements nécessaires.

Remarque : Le nœud Echantillonner est un nœud standard de SPSS Modeler.

Utilisation du nœud Text Mining dans un flux

Le nœud de modélisation Text Mining permet d'accéder aux données et d'extraire des concepts dans un flux. Vous pouvez utiliser n'importe quel nœud source pour accéder aux données, comme un nœud SGDB, Délimité, Fil de nouvelles ou Fixe. Un nœud Liste fichiers peut être utilisé pour les textes résidant dans des documents externes.

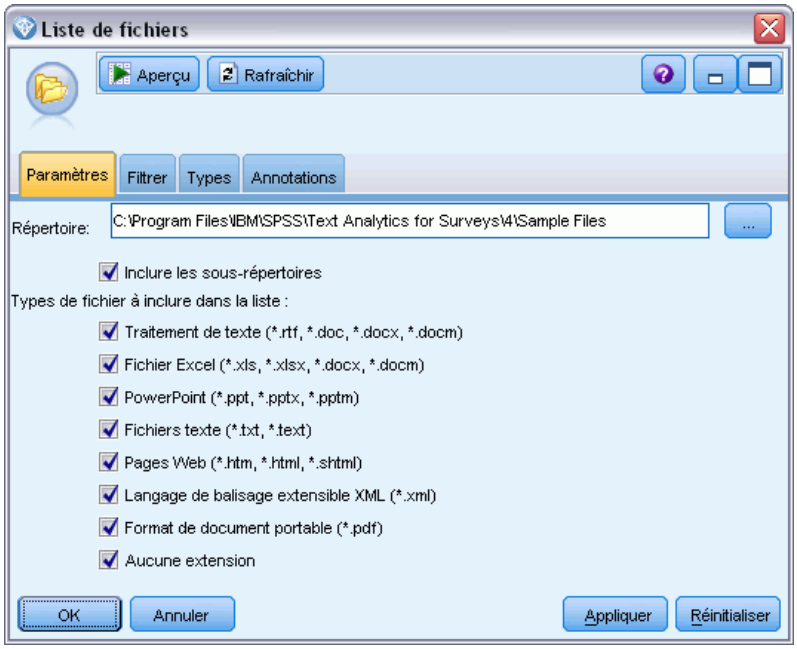
Exemple 1 : Nœud Liste fichiers et nœud de Text Mining pour créer un nugget de modèle de concepts directement

L'exemple suivant indique comment utiliser le nœud Liste fichiers, ainsi que le nœud de modélisation Text Mining, pour générer la sortie du modèle de concepts. Pour plus d'informations sur l'utilisation du nœud Liste fichiers, reportez-vous à le chapitre 2.

Figure 3-10
Exemple de flux : nœud Liste fichiers avec nœud de modélisation Text Mining

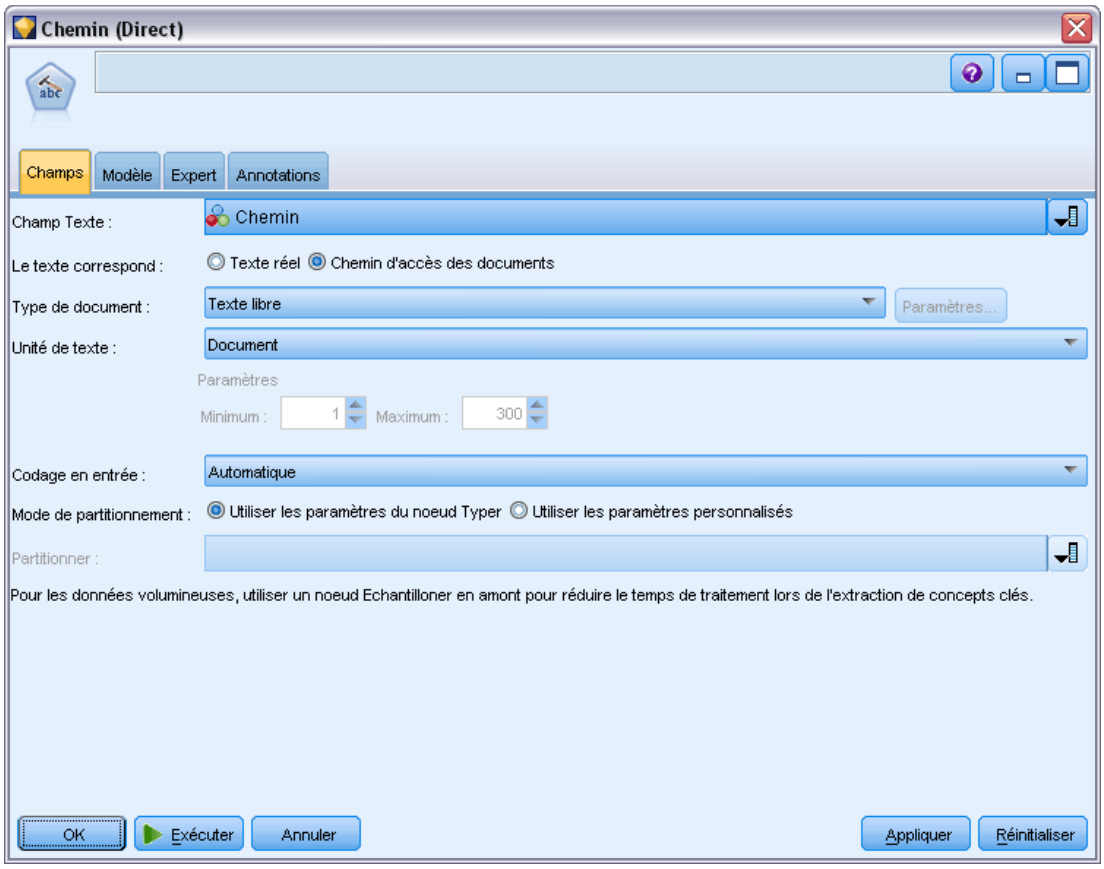
- **Noeud Liste fichiers (onglet Paramètres).** Nous avons tout d'abord ajouté ce nœud au flux pour indiquer l'emplacement de stockage des documents texte. Nous avons sélectionné le répertoire contenant tous les documents sur lesquels nous souhaitons effectuer le processus de Text Mining.

Figure 3-11
Boîte de dialogue du noeud Liste fichiers : onglet Paramètres



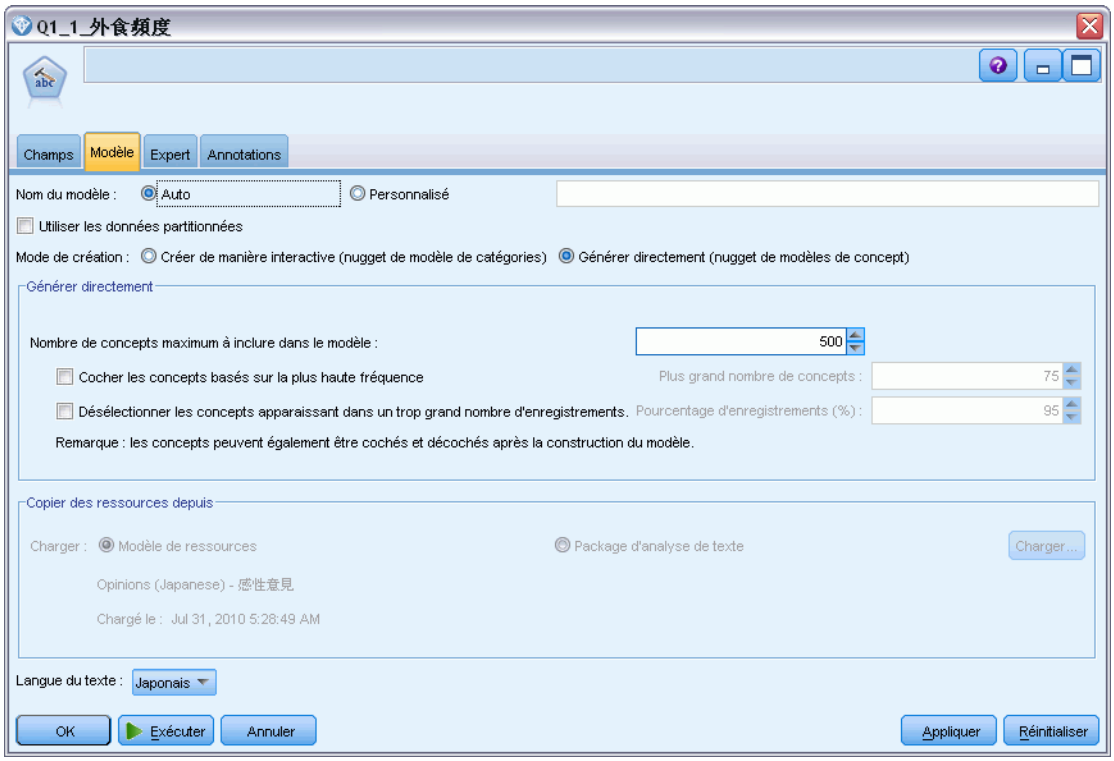
- **Noeud de Text Mining (onglet Champs).** Nous avons ensuite ajouté et connecté un noeud de Text Mining au noeud Liste fichiers. Dans ce noeud, nous avons défini le format d'entrée, le modèle de ressources et le format de sortie. Nous avons choisi le nom de champ créé à partir du noeud Liste fichiers et sélectionné l'option Le texte correspond au chemin d'accès des documents, ainsi que d'autres paramètres. Pour plus d'informations, reportez-vous à la section Utilisation du nœud Text Mining dans un flux sur p. 50.

Figure 3-12
Boîte de dialogue du noeud de Text Mining : onglet Champs



- **Nœud de Text Mining (onglet Modèle)**. Dans l'onglet Modèle, nous avons ensuite sélectionné le mode de création pour générer un nugget de modèle de concepts directement à partir de ce nœud. Vous pouvez sélectionner un autre modèle de ressources. Cependant, pour cet exemple, nous avons conservé les ressources de base Basic Resources.

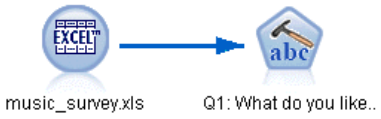
Figure 3-13
Boîte de dialogue du nœud de modélisation Text Mining : onglet Modèle



Exemple 2 : Nœuds Fichier Excel et Text Mining pour créer un modèle de catégories de façon interactive

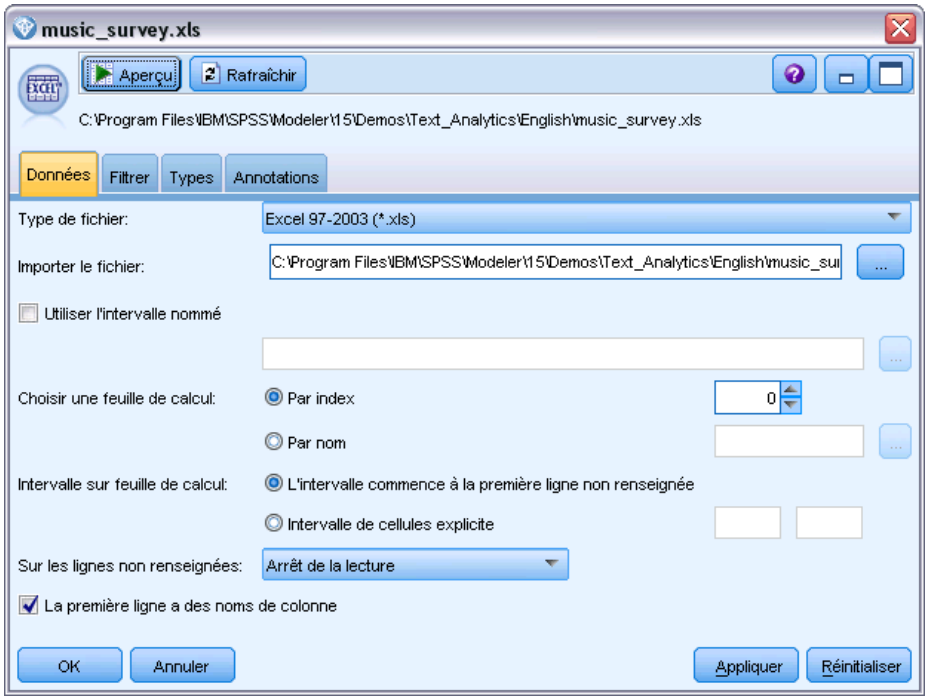
Cet exemple indique comment le nœud de Text Mining peut également lancer une session interactive. Pour plus d'informations sur la session interactive, reportez-vous au le chapitre 8.

Figure 3-14
Exemple de flux : nœud source Excel avec nœud de Text Mining (créer de manière interactive)



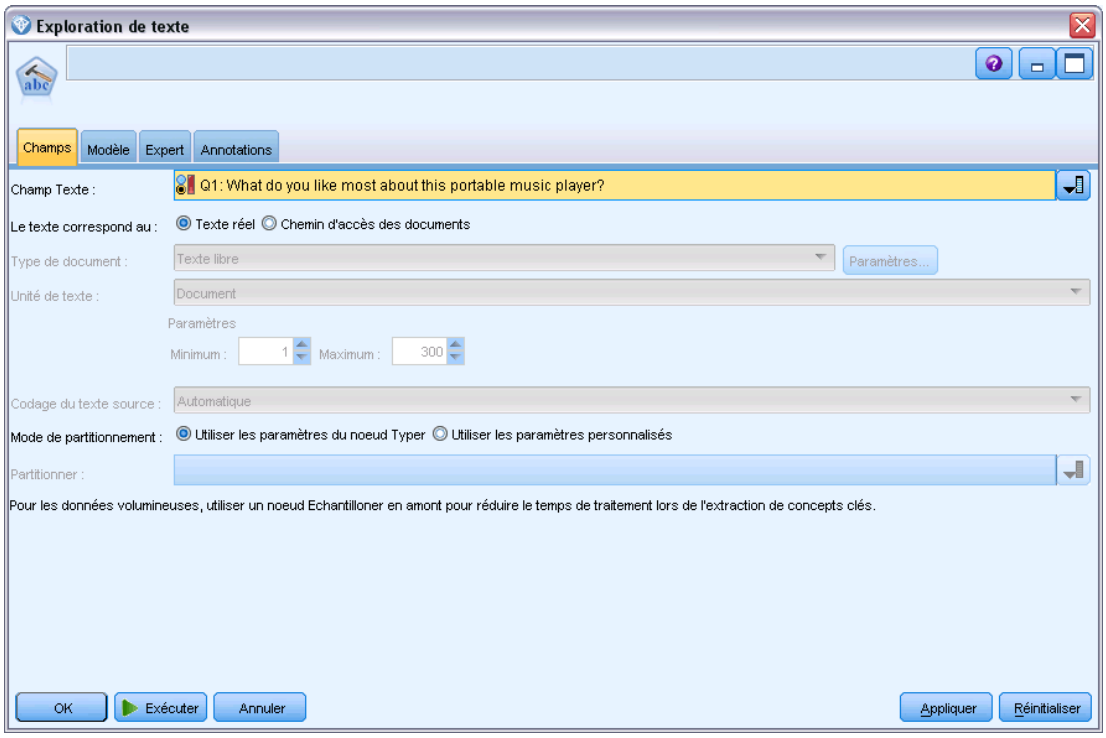
- **Nœud source Excel (onglet Données).** Nous avons tout d'abord ajouté ce nœud au flux pour indiquer l'emplacement de stockage du texte.

Figure 3-15
Boîte de dialogue de nœud source Excel : onglet Données



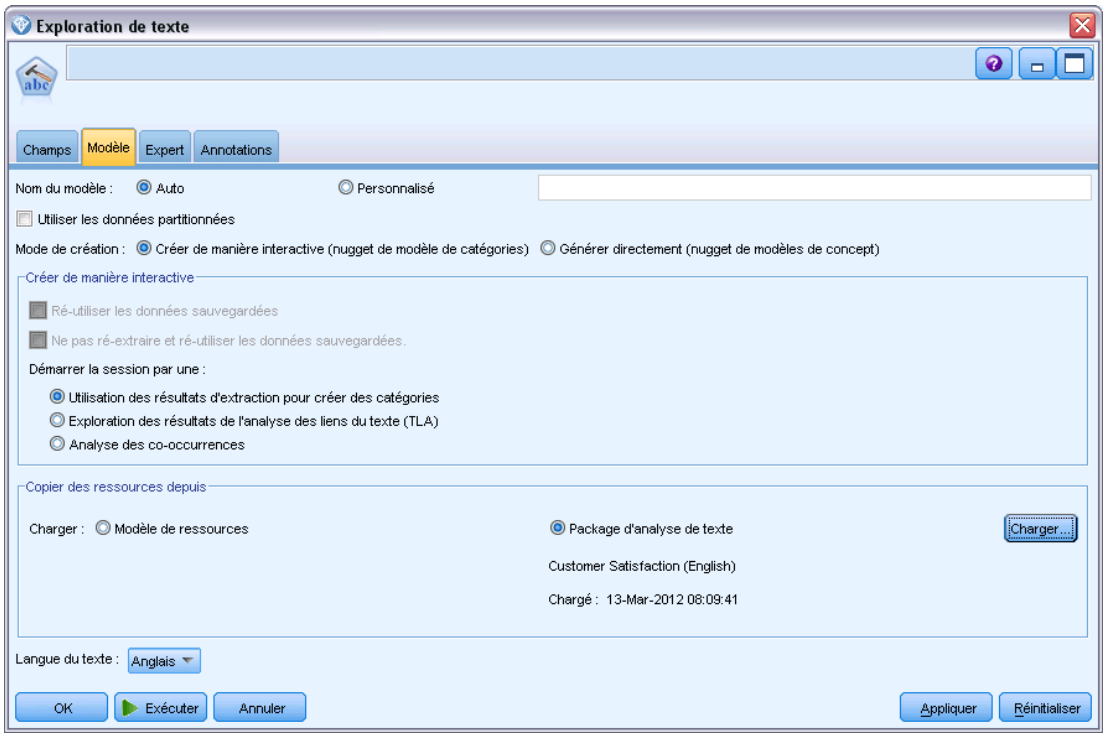
- **Nœud de Text Mining (onglet Champs).** Nous avons ensuite ajouté et connecté un nœud de Text Mining. Dans le premier onglet, nous avons défini le format d'entrée de notre choix. Nous avons sélectionné un nom de champ dans le nœud source et sélectionné l'option Le champ texte correspond au Texte réel puisque les données proviennent directement du nœud source Excel.

Figure 3-16
Boîte de dialogue du nœud de modélisation Text Mining : onglet Champs



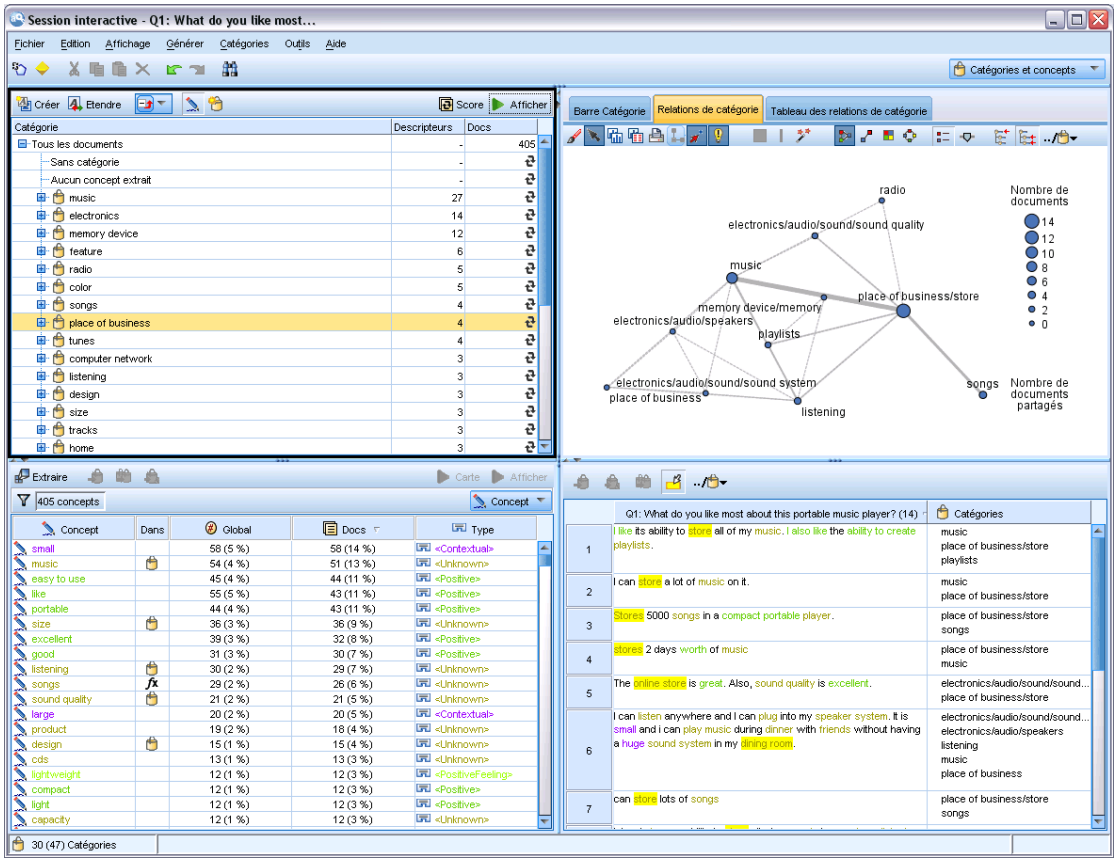
- **Nœud de Text Mining (onglet Modèle).** Dans l'onglet Modèle, nous avons ensuite décidé de créer un nugget de modèle de catégories de manière interactive et d'utiliser les résultats d'extraction pour créer automatiquement des catégories. Dans cet exemple, nous avons chargé une copie des ressources et un ensemble de catégories à partir d'un package d'analyse de texte.

Figure 3-17
Boîte de dialogue du nœud de modélisation Text Mining : onglet Modèle



- **Session interactive**. Nous avons ensuite exécuté le flux et l'interface de la session interactive s'est ouverte. A l'issue d'une extraction, nous avons commencé à explorer nos données et à améliorer nos catégories.

Figure 3-18
Session interactive



Nugget Text Mining : Modèle de concepts

Un nugget de modèles de concepts Text Mining est créé lorsque vous parvenez à exécuter un nœud de modèle Text Mining pour lequel vous avez sélectionné l'option Générer directement le modèle dans l'onglet Modèle. Un nugget de modèles de concepts Text Mining est utilisé pour la recherche en temps réel de concepts-clés dans d'autres données textuelles, telles que les notes d'un centre d'appel.

Le nugget de modèle de concepts lui-même comprend une liste de concepts qui ont été affectés à des types. Vous pouvez sélectionner n'importe lequel des concepts de ce modèle pour le scoring en fonction d'autres données. Si vous exécutez un flux contenant un nugget de modèle Text Mining, de nouveaux champs sont ajoutés aux données en fonction du mode de création sélectionné dans l'onglet Modèle du nœud de modélisation Text Mining avant la création du modèle. Pour plus d'informations, reportez-vous à la section Modèle de concepts : onglet Modèle sur p. 58.

Si le nugget de modèle a été généré à l'aide de documents traduits, le scoring sera effectuée dans la langue de traduction. De la même manière, si le nugget de modèle a été généré avec la langue Anglais, vous pouvez indiquer une langue de traduction dans le nugget de modèle, puisque les documents seront ensuite traduits en anglais.

Les nuggets de modèles Text Mining se trouvent dans la palette de nuggets de modèles (dans l'onglet Modèles situé dans la partie supérieure droite de la fenêtre IBM® SPSS® Modeler) lorsque ceux-ci sont générés.

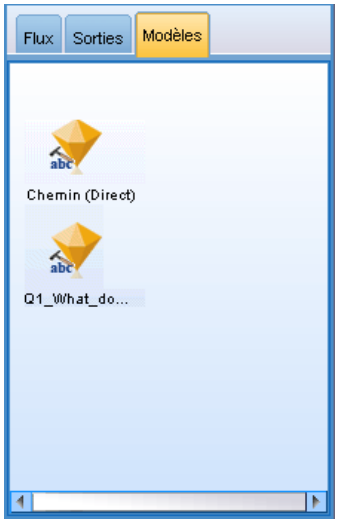
Affichage des résultats

Pour obtenir des informations sur le nugget de modèle, cliquez avec le bouton droit de la souris sur le noeud de la palette de nuggets de modèles, puis sélectionnez Parcourir dans le menu contextuel (ou Editer pour les noeuds du flux).

Ajout de modèles aux flux

Pour ajouter le nugget de modèle au flux, cliquez sur l'icône correspondante dans la palette de nuggets de modèles, puis dans l'espace de travail de flux à l'endroit où vous souhaitez placer le noeud. Vous pouvez également cliquer sur l'icône avec le bouton droit de la souris et sélectionner Ajouter au flux dans le menu contextuel. Il vous suffit alors de connecter votre flux au noeud pour pouvoir transmettre des données et générer des prévisions.

Figure 3-19
Palette de nuggets de modèles contenant un nugget de modèle Text Mining

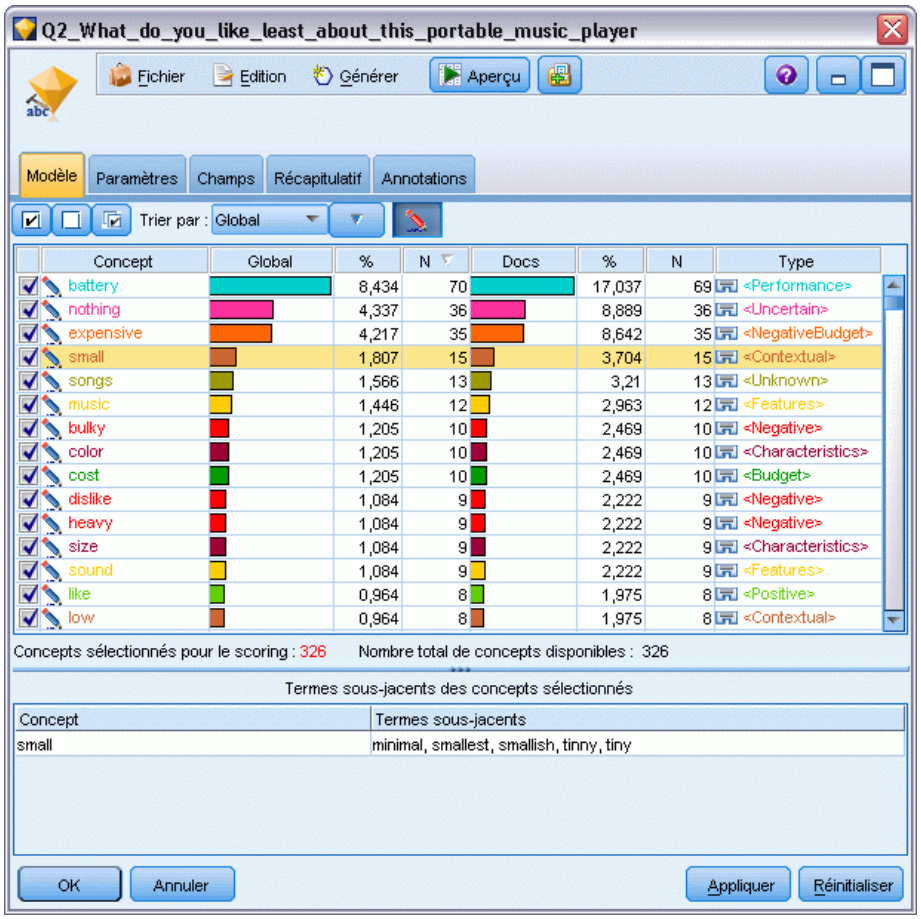


Modèle de concepts : onglet Modèle

Dans les modèles de concepts, l'ensemble des concepts ayant fait l'objet d'une extraction figurent dans l'onglet Modèle. Les concepts sont présentés sous forme de tableau, avec une ligne par concept. Cet onglet permet de sélectionner les concepts qui seront utilisés pour le scoring.

Remarque : si vous avez généré un nugget de modèle de catégories à la place, cet onglet présentera des informations différentes. Pour plus d'informations, reportez-vous à la section Nugget de modèle de catégories : onglet Modèle sur p. 73.

Figure 3-20
Boîte de dialogue de nugget de modèle de concepts : onglet Modèle



Tous les concepts sont sélectionnés pour l'affectation des documents dans les catégories par défaut, comme l'indiquent les cases à cocher de la colonne située à l'extrême gauche. Si la case est cochée, le concept est utilisé pour l'affectation des documents dans les catégories. Si elle n'est pas cochée, il est exclu de l'affectation des documents dans les catégories. Vous pouvez cocher plusieurs lignes à la fois en les sélectionnant toutes et en cliquant sur l'une des cases de la sélection.

Pour en savoir plus sur chaque concept, vous pouvez consulter les informations supplémentaires fournies dans chacune des colonnes suivantes :

Concept. Il s'agit de l'expression ou du mot principal extrait. Dans certains cas, ce concept représente le nom du concept, ainsi que d'autres termes sous-jacents associés à ce concept. Pour connaître les termes sous-jacents qui font partie d'un concept, affichez le panneau des termes sous-jacents dans cet onglet et sélectionnez le concept pour consulter les termes correspondants situés au bas de la boîte de dialogue. Pour plus d'informations, reportez-vous à la section Termes sous-jacents dans les modèles de concepts sur p. 62.

Global. Global (fréquence) fait ici référence au nombre d'occurrences d'un concept (et de tous ses termes sous-jacents) dans l'ensemble des documents/enregistrements (fréquence).

- **Diagramme en bâton.** Fréquence globale de ce concept dans les données textuelles, présentée sous la forme d'un diagramme en bâton. La barre prend la couleur du type auquel le concept est affecté pour distinguer les types de manière visuelle.
- **%.** Fréquence globale de ce concept dans les données textuelles, présentée sous la forme d'un pourcentage.
- **N.** Nombre effectif d'occurrences de ce concept dans les données textuelles.

Docs. Fait ici référence aux effectifs des documents ou des enregistrements dans lesquels le concept (et tous ses termes sous-jacents) apparaît.

- **Diagramme en bâton.** Effectifs des documents de ce concept présentés sous la forme d'un diagramme en bâton. La barre prend la couleur du type auquel le concept est affecté pour distinguer les types de manière visuelle.
- **%.** Effectifs des documents de ce concept, présentés sous la forme d'un pourcentage.
- **N.** Nombre effectif de documents ou d'enregistrements contenant ce concept.

Type. Type auquel le concept est affecté. Pour chaque concept, les colonnes Global et Docs arborent une couleur pour représenter le type auquel le concept est affecté. Un **type** correspond à un regroupement sémantique de concepts. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

Utilisation des concepts

Cliquez avec le bouton droit de la souris sur une cellule du tableau pour afficher un menu contextuel proposant les options suivantes :

- **Tout sélectionner.** Toutes les lignes du tableau sont sélectionnées.
- **Copier.** Le ou les concepts sélectionnés sont copiés dans le Presse-papiers.
- **Copier avec les champs** Les concepts sélectionnés sont copiés dans le Presse-papiers, de même que l'en-tête de colonne.
- **Activer les concepts choisis.** Coche les cases de toutes les lignes de tableau sélectionnées incluant ainsi ces concepts pour le scoring.
- **Désélectionner les concepts choisis.** Désactivez les cases de toutes les lignes de tableau sélectionnées.
- **Tout activer.** Coche toutes les cases du tableau. Par conséquent, tous les concepts sont utilisés dans les résultats finaux.
- **Tout désactiver.** Désactive toutes les cases du tableau. Les concepts désactivés ne sont pas utilisés dans les résultats finaux.
- **Inclure les concepts.** Affiche la boîte de dialogue Inclure les concepts. Pour plus d'informations, reportez-vous à la section Options pour inclure des concepts pour l'affectation des documents dans les catégories sur p. 60.

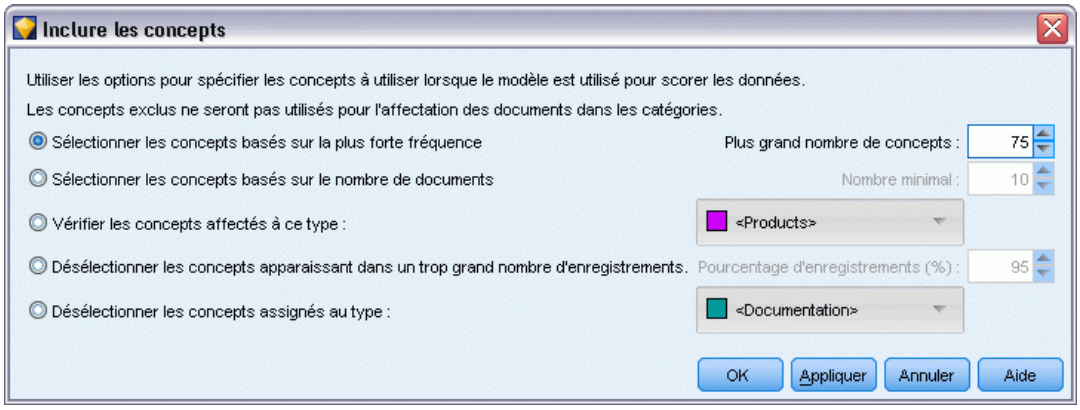
Options pour inclure des concepts pour l'affectation des documents dans les catégories

Pour activer ou désactiver les concepts qui seront utilisés pour le scoring, cliquez sur le bouton **Inclure les concepts** de la barre d'outils.

Figure 3-21
Bouton Inclure les concepts de la barre d'outils

En cliquant sur ce bouton de la barre d'outils, la boîte de dialogue Inclure les concepts s'ouvre et permet de sélectionner des concepts en fonction de règles. Tous les concepts cochés dans l'onglet Modèle seront inclus dans l'affectation des documents dans les catégories. Appliquer une règle dans cette boîte de dialogue secondaire afin de modifier les concepts qui seront utilisés pour le scoring.

Figure 3-22
Boîte de dialogue Inclure les concepts.



Vous avez le choix entre les options suivantes :

Activer les concepts en fonction de la fréquence la plus élevée. Plus grand nombre de concepts. Il s'agit du nombre de concepts qui seront cochés, en partant de celui dont la fréquence globale est la plus élevée. Le terme Fréquence fait ici référence au nombre d'occurrences d'un concept (et de tous ses termes sous-jacents) dans l'ensemble des documents/enregistrements. Il est parfois supérieur aux effectifs des enregistrements car un concept peut figurer plusieurs fois dans un même enregistrement.

Activer les concepts basés sur les effectifs des documents. Nombre minimal. Il s'agit du nombre minimal de documents nécessaires pour l'activation des concepts. Le terme effectifs des documents fait ici référence au nombre de documents/d'enregistrements dans lesquels le concept (et tous ses termes sous-jacents) apparaît.

Activer les concepts affectés au type. Sélectionnez un type dans la liste déroulante pour activer tous les concepts affectés à ce type. Les concepts sont automatiquement affectés aux types durant le processus d'extraction. Un **type** correspond à un regroupement sémantique de concepts. Les types comprennent des éléments tels que des concepts de niveau supérieur, des qualificatifs et des mots positifs et négatifs, des qualificatifs contextuels, des prénoms, des lieux, des organisations, etc. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

Désactiver les concepts apparaissant dans trop d'enregistrements. Pourcentage d'enregistrements. Désactiver les concepts apparaissant dans un pourcentage supérieur à celui indiqué pour le nombre d'enregistrements. Cette option permet d'exclure les concepts qui figurent fréquemment dans le texte ou les enregistrements, mais qui ne présentent pas d'intérêt pour l'analyse.

Désactiver les concepts affectés au type. Désactiver les concepts correspondant au type sélectionné dans la liste déroulante.

Termes sous-jacents dans les modèles de concepts

Vous pouvez consulter les termes sous-jacents définis pour les concepts que vous avez sélectionnés dans le tableau. Vous pouvez cliquer sur le bouton-bascule des termes sous-jacents dans la barre d'outils pour afficher le tableau des termes sous-jacents dans un panneau distinct, situé en bas de la boîte de dialogue.

Ces termes sous-jacents incluent les synonymes définis dans les ressources linguistiques (que ceux-ci aient été rencontrés ou non dans le texte) ainsi que toutes les formes extraites au pluriel/singulier trouvées dans le texte qui sont utilisées pour générer le nugget de modèles, les termes permutés, les termes provenant du regroupement flou, etc.

Figure 3-23
Bouton de la barre d'outils Afficher les termes sous-jacents

Remarque : Vous ne pouvez pas modifier la liste des termes sous-jacents. Cette liste est générée par l'intermédiaire de substitutions, de définitions de synonyme (dans le dictionnaire de substitutions), de regroupement flou, etc. Tous ces éléments sont définis dans les ressources linguistiques. Pour modifier la manière dont les termes sont regroupés sous un concept et comment ceux-ci sont manipulés, vous devez effectuer les modifications directement dans les ressources (dans l'Editeur de ressources dans la session interactive ou dans l'Editeur de modèle puis recharger dans le nœud) et réexécuter le flux pour obtenir un nouveau nugget de modèle comprenant les résultats mis à jour.

Cliquez avec le bouton droit de la souris sur une cellule comprenant un terme sous-jacent ou un concept pour afficher un menu contextuel proposant les options suivantes :

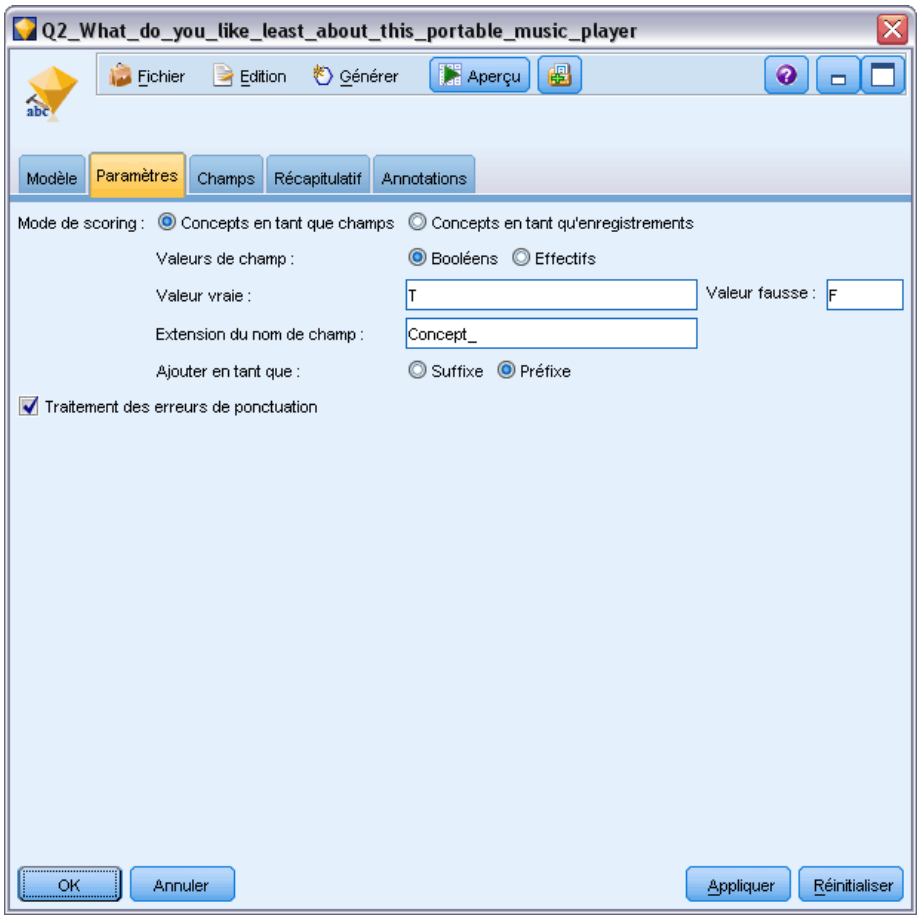
- **Copier.** La cellule sélectionnée est copiée dans le Presse-papiers.
- **Copier avec les champs.** La cellule sélectionnée est copiée dans le Presse-papiers, de même que les en-têtes de colonne.
- **Tout sélectionner.** Toutes les cellules du tableau sont sélectionnées.

Modèle de concepts : Onglet Paramètres

L'onglet Paramètres sert à définir la valeur du champ de texte des nouvelles données d'entrée, si nécessaire. Il permet également d'indiquer le modèle de données des résultats (mode de l'affectation des documents dans les catégories).

Remarque : Cet onglet n'apparaît que si le nugget de modèle figure sur l'espace de travail. Il n'existe pas si vous accédez à cette boîte de dialogue directement à partir de la palette Modèles.

Figure 3-24
Boîte de dialogue Nugget de modèle de concepts Text Mining : onglet Paramètres



Mode de scoring : Concepts en tant qu'enregistrements

Grâce à ce mode de scoring, un nouvel enregistrement est créé pour chaque paire de concept/document. Généralement, la sortie comporte plus d'enregistrements que n'en comportait l'entrée.

Outre les champs d'entrée, les nouveaux champs suivants sont ajoutés aux données :

Table 3-1
Champs de sortie de l'option Concepts en tant qu'enregistrements

Champ	Description
Concept	Contient le nom de concept extrait qui figure dans le champ des données textuelles.
Type	Indique le type du concept sous la forme d'un nom de type complet, comme *Location* ou *Person*. Un **type** correspond à un regroupement sémantique de concepts. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.
Count	Affiche le nombre d'occurrences de ce concept (et de ses termes sous-jacents) dans le corps du texte (enregistrement/document).

Lorsque vous sélectionnez cette option, toutes les autres options à l'exception de Traitement des erreurs de ponctuation sont désactivées.

Mode de scoring : Concepts en tant que champs

Dans les modèles de concepts, pour chaque enregistrement d'entrée, un enregistrement est créé pour chaque concept trouvé dans un document donné. Ainsi, les enregistrements de sortie sont aussi nombreux que les enregistrements d'entrée. Désormais, chaque enregistrement (ligne) comporte toutefois un nouveau champ (colonne) pour chaque concept sélectionné (coché) dans l'onglet Modèle. La valeur du champ de chaque concept dépend de la valeur de champ que vous sélectionnez dans cet onglet (Booléens ou Effectifs).

Valeurs de champ. Sélectionnez si vous souhaitez que le nouveau champ de chaque concept contienne un effectif ou une valeur booléenne.

- **Booléens.** Cette option permet d'obtenir des booléens aboutissant à deux valeurs distinctes dans les résultats, comme *Oui*/*Non*, *Vrai*/*Faux*, *V*/*F*, ou *1* et *2*. Les types de stockage sont définis automatiquement pour correspondre aux valeurs choisies. Par exemple, si vous entrez des valeurs numériques pour les booléens, elles seront automatiquement traitées comme valeurs entières. Les booléens disposent des types de stockage suivants : chaîne, entier, nombre réel ou date/heure. Entrez une valeur booléenne pour Vrai et pour Faux.
- **Effectifs.** Valeur utilisée pour obtenir le nombre d'occurrences du concept dans un enregistrement donné.

Extension nom de champ. Spécifiez l'extension du nom de champ. Les noms de champ générés reprennent le nom de concept et l'extension.

- **Ajouter en tant que.** Spécifiez à quel emplacement du nom de champ l'extension doit être ajoutée. Choisissez Préfixe pour ajouter l'extension en début de chaîne. Choisissez Suffixe pour ajouter l'extension en fin de chaîne.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

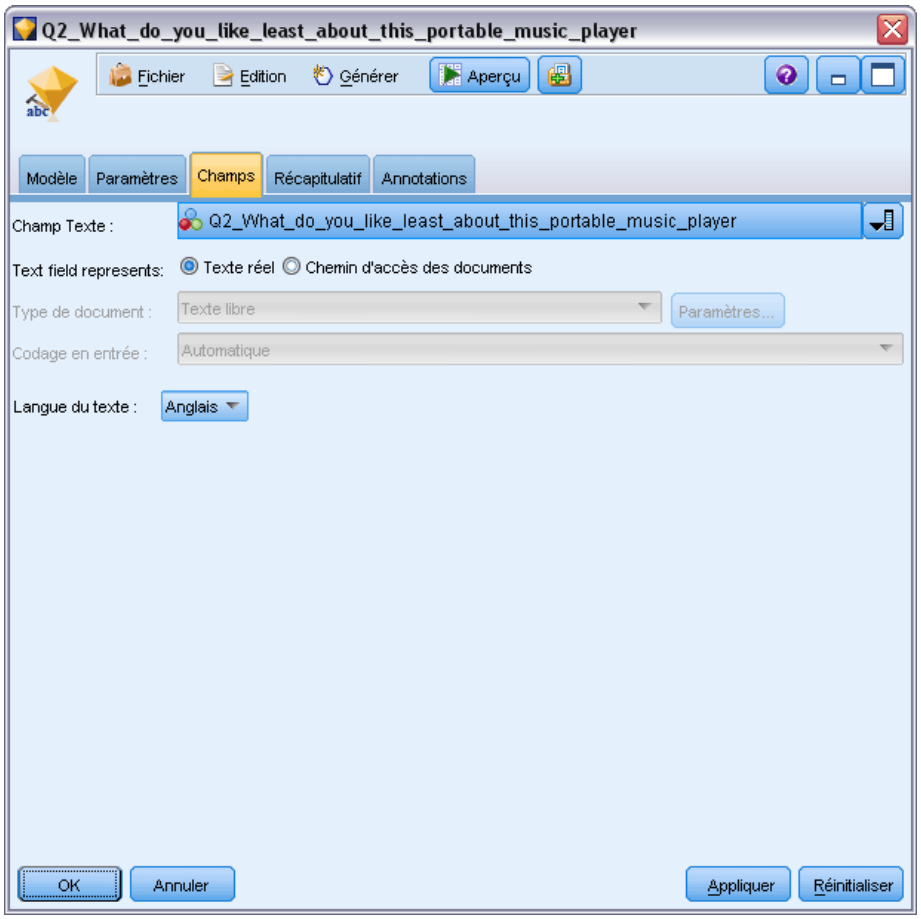
Remarque : L'option Traitement des erreurs de ponctuation ne s'applique pas au texte en japonais.

Modèle de concepts : onglet Champs

L'onglet Champs sert à définir la valeur du champ de texte des nouvelles données d'entrée, si nécessaire.

Remarque : Cet onglet n'apparaît que si le nugget de modèle figure dans le flux. Il n'apparaît pas lorsque vous accédez à cette sortie directement à partir de la palette Modèles.

Figure 3-25
Boîte de dialogue Nugget de modèle de concepts Text Mining : onglet Champs



Champ Texte. Sélectionnez le champ contenant le texte à explorer, le chemin d'accès au document ou le chemin d'accès au répertoire contenant les documents. Ce champ dépend de la source de données.

Le texte correspond au. Indique ce que contient le champ de texte spécifié dans le paramètre précédent. Les différents choix sont :

- **Texte réel.** Sélectionnez cette option si le champ contient le texte exact à partir duquel les concepts doivent être extraits.
- **Chemin d'accès des documents.** Sélectionnez cette option si le champ contient un ou plusieurs noms de chemin d'accès pour les emplacements des documents texte.

Type de document. Cette option n'est disponible que si vous indiquez que le champ texte correspond au Chemin d'accès des documents. Le type de document indique la structure du texte. Sélectionnez l'un des types suivants :

- **Texte libre.** S'utilise pour la plupart des documents ou des sources textuelles. L'ensemble de texte entier est analysé pour l'extraction. Contrairement aux autres options, il n'existe pas de paramètres supplémentaires pour cette option.

- **Texte structuré.** Utilisez ce format pour les bibliographies, les brevets et les fichiers contenant des structures standard qui peuvent être identifiées et analysées. Ce type de document est utilisé pour éviter tout ou une partie du processus d'extraction. Il permet de définir des séparateurs de termes, d'affecter des types et d'imposer une valeur de fréquence minimale. Si vous sélectionnez cette option, vous devez cliquer sur le bouton Paramètres et entrer les séparateurs de texte dans la zone Formatage de texte structuré de la boîte de dialogue Paramètres du document. Pour plus d'informations, reportez-vous à la section Paramètres de document de l'onglet Champs sur p. 35.
- **Texte XML.** Permet de signaler les balises XML qui contiennent le texte à extraire. Toutes les autres balises sont ignorées. Si vous sélectionnez cette option, vous devez cliquer sur le bouton Paramètres et indiquer de manière explicite les éléments XML contenant le texte à lire au cours du processus d'extraction dans la zone Formatage de texte XML de la boîte de dialogue Paramètres du document. Pour plus d'informations, reportez-vous à la section Paramètres de document de l'onglet Champs sur p. 35.

Encoding du texte source. Cette option n'est disponible que si vous indiquez que Le texte correspond au chemin d'accès des documents. Elle indique l'encoding de texte par défaut. Dans toutes les langues, à l'exception du japonais, l'encoding spécifié ou reconnu est converti en un encoding ISO-8859-1. Ainsi, même si vous indiquez un autre encoding, le moteur d'extraction le remplace par l'encoding ISO-8859-1 avant de traiter le texte. Les caractères qui ne figurent pas dans la définition de l'encoding ISO-8859-1 sont convertis en espaces. Pour le texte en japonais, vous pouvez choisir l'une des options de codage suivantes : SHIFT_JIS, EUC_JP, UTF-8, ou ISO-2022-JP.

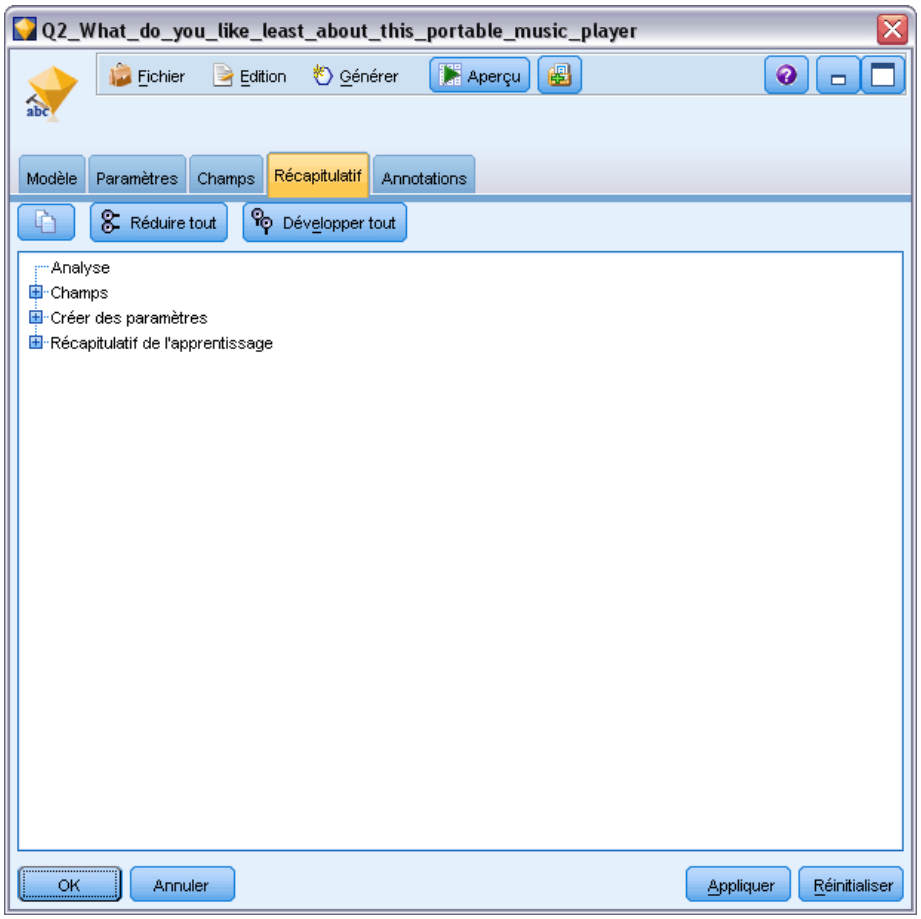
Langue du texte. Identifie la langue du texte en cours d'exploration. Il s'agit de la langue principale détectée pendant l'extraction. Si vous souhaitez acquérir la licence d'une langue prise en charge à laquelle vous n'avez pas accès actuellement, contactez votre représentant commercial.

Modèle de concepts : onglet Récapitulatif

L'onglet Récapitulatif contient des informations sur le modèle lui-même (dossier *Analyse*), sur les champs utilisés dans le modèle (dossier *Champs*), sur les paramètres utilisés pour la construction du modèle (dossier *Créer des paramètres*), ainsi que sur l'apprentissage du modèle (dossier *Récapitulatif de l'apprentissage*).

Lorsque vous accédez pour la première fois à un nœud de modélisation, l'arborescence des dossiers de l'onglet Récapitulatif est réduite. Pour afficher les résultats qui vous intéressent, utilisez la commande de développement à gauche du dossier ou cliquez sur le bouton Développer tout pour afficher tous les résultats. Pour masquer les résultats après les avoir consultés, utilisez la commande de développement pour réduire le dossier voulu ou cliquez sur le bouton Réduire tout pour réduire tous les dossiers.

Figure 3-26
Boîte de dialogue de nugget de modèle Text Mining : Onglet Récapitulatif



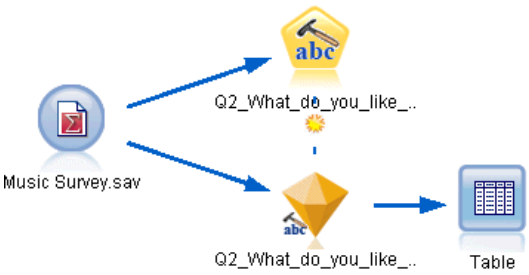
Utilisation des nuggets de modèle de concepts dans un flux

Lors de l'utilisation d'un nœud de modélisation Text Mining, vous pouvez générer soit un nugget de modèle de concepts soit un nugget de modèle de catégories (dans la session interactive). L'exemple suivant indique comment utiliser un modèle de concepts dans un flux simple.

Exemple : nœud Fichier statistiques avec le nugget de modèle de concepts

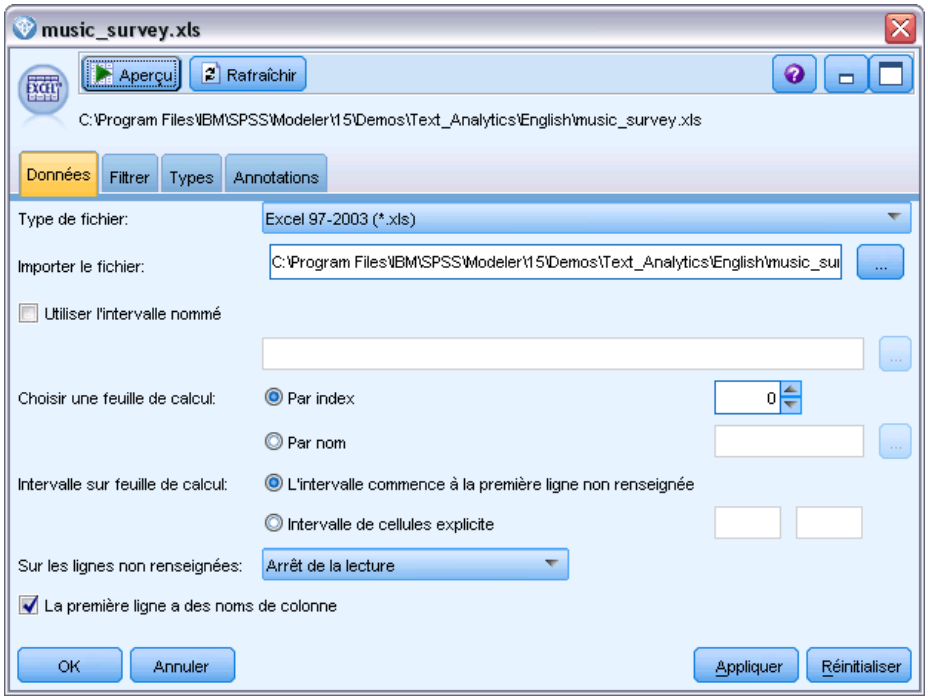
L'exemple suivant indique comment utiliser le nugget de modèle de concepts Text Mining.

Figure 3-27
Exemple de flux : nœud Fichier statistiques avec un nugget de modèle de concepts Text Mining



- **Nœud Fichier statistiques (onglet Paramètres).** Nous avons tout d'abord ajouté ce nœud au flux pour indiquer l'emplacement de stockage des documents texte.

Figure 3-28
Boîte de dialogue du nœud Fichier statistiques : onglet Données

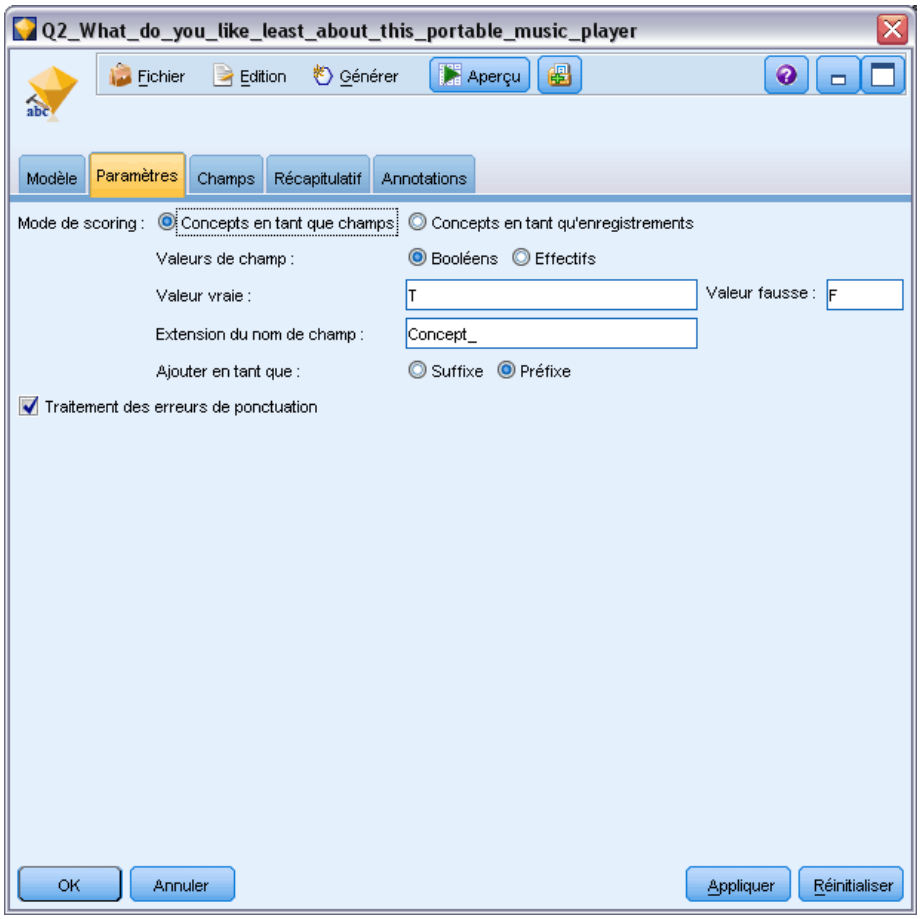


- **nugget de modèle de concepts Text Mining (onglet Modèle).** Nous avons ensuite ajouté et connecté un nugget de modèle de concepts au nœud Fichier Statistiques. Nous avons sélectionné les concepts que nous souhaitions utiliser pour scorer nos données.

Figure 3-29
Boîte de dialogue de nugget de modèle Text Mining : onglet Modèle

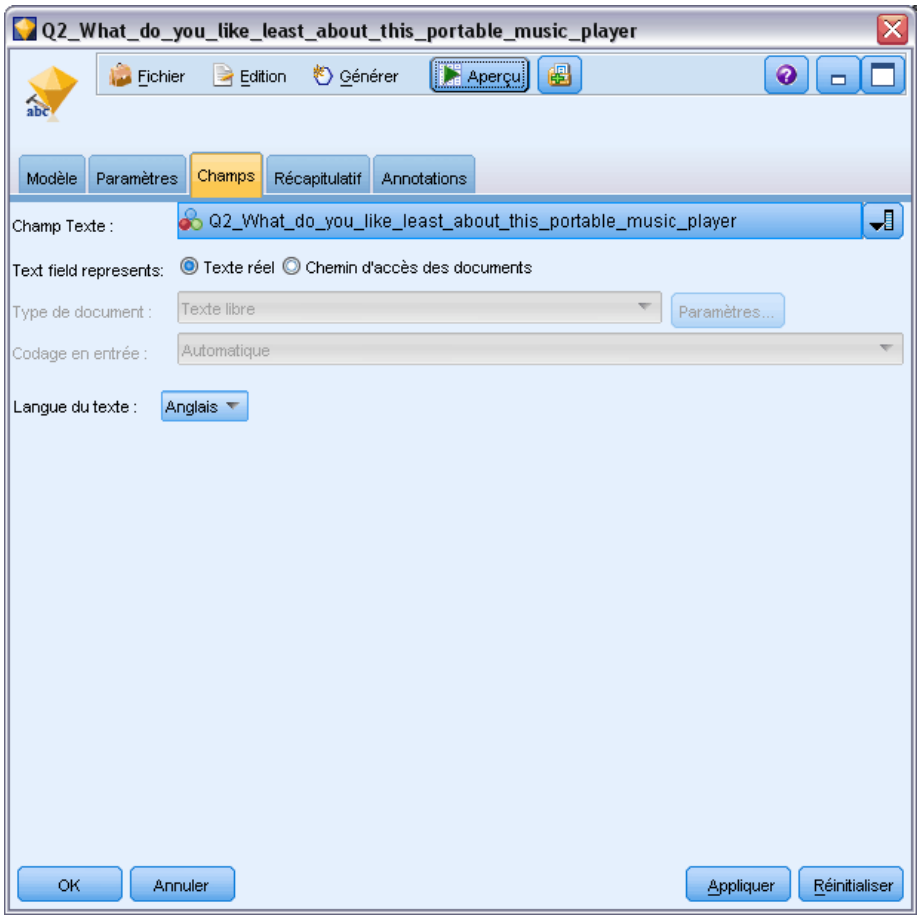
- **nugget de modèle de concepts Text Mining (onglet Paramètres).** Nous avons ensuite défini le format de sortie et sélectionné *Concepts en tant que champs*. Un nouveau champ sera créé dans la sortie pour chaque concept sélectionné dans l'onglet Modèle. Chaque nom de champ sera composé du nom du concept et du préfixe « *Concept_* »

Figure 3-30
Boîte de dialogue Nugget de modèle de concepts Text Mining : onglet Paramètres



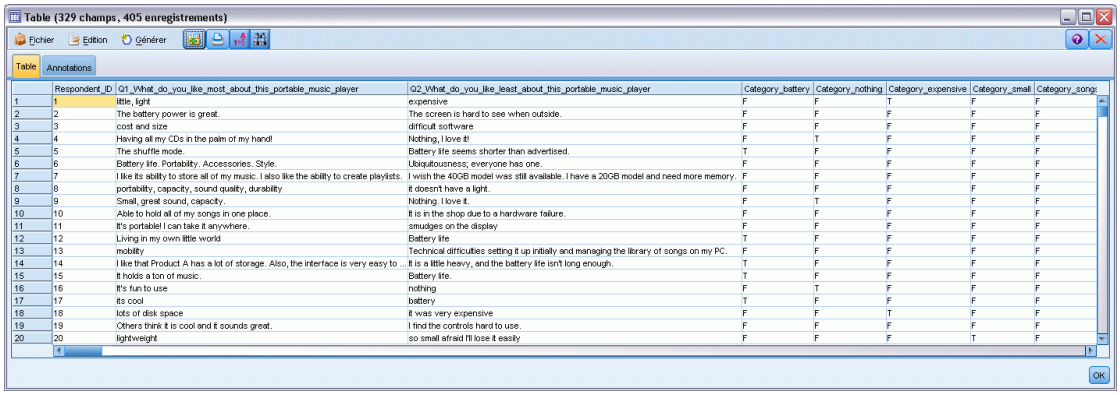
- **nugget de modèle de concepts Text Mining (onglet Champs).** Nous avons ensuite sélectionné le champ de texte, Q2_What_do_you_like_least_about_this_portable_music_player, qui est le nom de champ provenant du nœud Fichier Statistiques. Nous avons également sélectionné l'option Le champ texte correspond au texte réel.

Figure 3-31
Boîte de dialogue Nugget de modèle de concepts Text Mining : onglet Champs



▶ **Nœud Table.** Nous avons ensuite associé un nœud de table pour afficher les résultats et exécuté le flux. La sortie Table s'affiche dans l'écran.

Figure 3-32
Nous avons fait défiler les sorties de table pour afficher les booléens de concepts



	Respondent_ID	Q1_What_do_you_like_most_about_this_portable_music_player	Q2_What_do_you_like_least_about_this_portable_music_player	Category_battery	Category_nothing	Category_expensive	Category_small	Category_songs
1	1	little, light	expensive	F	F	T	F	F
2	2	The battery power is great.	The screen is hard to see when outside.	F	F	F	F	F
3	3	cost and size	difficult software	F	F	F	F	F
4	4	Having all my CDs in the palm of my hand!	Nothing, I love it!	F	T	F	F	F
5	5	The shuffle mode.	Battery life seems shorter than advertised.	T	F	F	F	F
6	6	Battery life. Portability. Accessories. Style.	Ubiquitousness; everyone has one.	F	F	F	F	F
7	7	I like its ability to store all of my music. I also like the ability to create playlists.	I wish the 40GB model was still available. I have a 20GB model and need more memory.	F	F	F	F	F
8	8	portability, capacity, sound quality, durability	it doesn't have a light.	F	F	F	F	F
9	9	Small, great sound, capacity.	Nothing. I love it.	F	T	F	F	F
10	10	Able to hold all of my songs in one place.	It is in the shop due to a hardware failure.	F	F	F	F	F
11	11	It's portable! I can take it anywhere.	smudges on the display	F	F	F	F	F
12	12	Living in my own little world	Battery life	T	F	F	F	F
13	13	mobility	Technical difficulties setting it up initially and managing the library of songs on my PC.	F	F	F	F	F
14	14	I like that Product A has a lot of storage. Also, the interface is very easy to ...	It is a little heavy, and the battery life isn't long enough.	T	F	F	F	F
15	15	It holds a ton of music.	Battery life.	T	F	F	F	F
16	16	It's fun to use	nothing	F	T	F	F	F
17	17	its cool	battery	T	F	F	F	F
18	18	lots of disk space	it was very expensive	F	F	T	F	F
19	19	Others think it is cool and it sounds great.	I find the controls hard to use.	F	F	F	F	F
20	20	lightweight	so small afraid I'll lose it easily	F	F	F	T	F

Nugget Text Mining : Modèle de catégories

Un nugget de modèle de catégories Text Mining est créé lorsque vous générez un modèle de catégories dans la session interactive. Ce nugget de modélisation contient un ensemble de catégories dont la définition se compose de concepts, de types, de patrons TLA et/ou de règles de catégorie. Le nugget permet de regrouper en catégories des réponses à des enquêtes, des entrées de blog, d'autres fils de nouvelles ou d'autres données textuelles.

Si vous lancez une session interactive dans le nœud de modélisation, vous pouvez explorer les résultats de l'extraction, adapter les ressources, mettre au point vos catégories avant de générer des modèles de catégories. Si vous exécutez un flux contenant un nugget de modèle Text Mining, de nouveaux champs sont ajoutés aux données en fonction du mode de création sélectionné dans l'onglet Modèle du nœud de modélisation Text Mining avant la création du modèle. Pour plus d'informations, reportez-vous à la section Nugget de modèle de catégories : onglet Modèle sur p. 73.

Si le nugget de modèle a été généré à l'aide de documents traduits, le scoring sera effectuée dans la langue de traduction. De la même manière, si le nugget de modèle a été généré avec la langue Anglais, vous pouvez indiquer une langue de traduction dans le nugget de modèle, puisque les documents seront ensuite traduits en anglais.

Les nuggets de modèles Text Mining se trouvent dans la palette de nuggets de modèles (dans l'onglet Modèles situé dans la partie supérieure droite de la fenêtre IBM® SPSS® Modeler) lorsque ceux-ci sont générés.

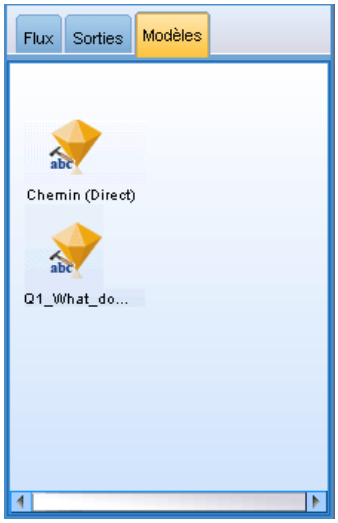
Affichage des résultats

Pour obtenir des informations sur le nugget de modèle, cliquez avec le bouton droit de la souris sur le noeud de la palette de nuggets de modèles, puis sélectionnez Parcourir dans le menu contextuel (ou Editer pour les noeuds du flux).

Ajout de modèles aux flux

Pour ajouter le nugget de modèle au flux, cliquez sur l'icône correspondante dans la palette de nuggets de modèles, puis dans l'espace de travail de flux à l'endroit où vous souhaitez placer le noeud. Vous pouvez également cliquer sur l'icône avec le bouton droit de la souris et sélectionner Ajouter au flux dans le menu contextuel. Il vous suffit alors de connecter votre flux au noeud pour pouvoir transmettre des données et générer des prévisions.

Figure 3-33
Palette de nuggets de modèles contenant un nugget de modèle Text Mining



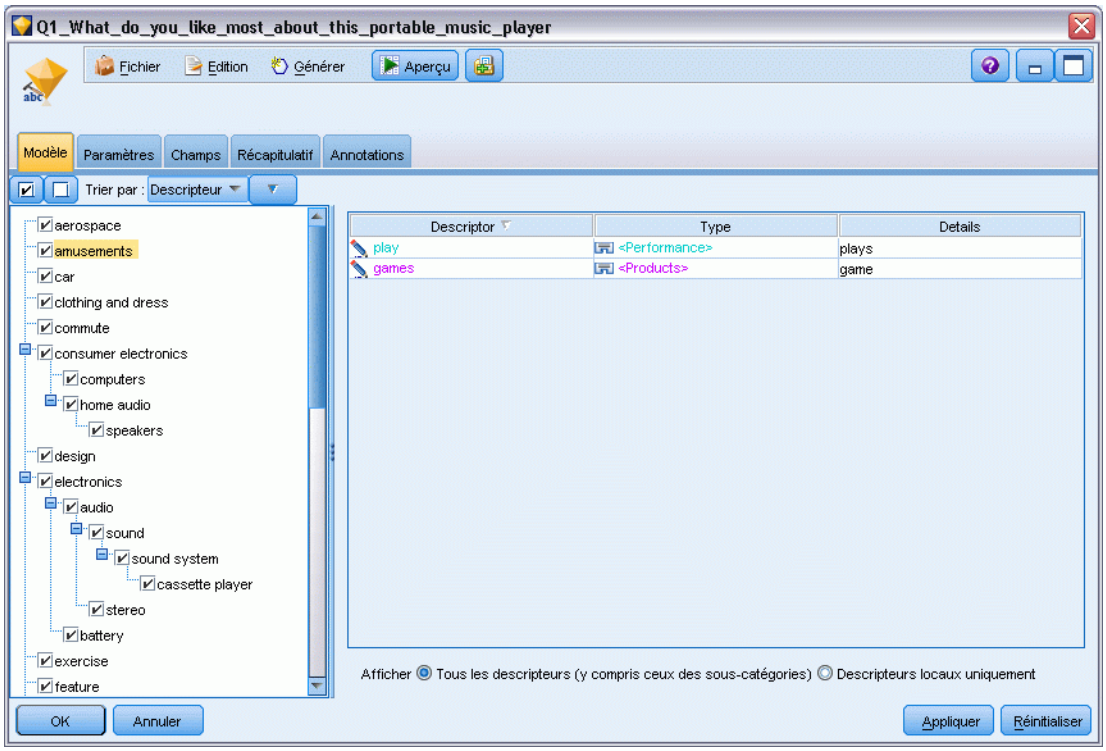
Nugget de modèle de catégories : onglet Modèle

Dans le cas des modèles de catégories, l'onglet Modèle affiche, à gauche, la liste des catégories du modèle de catégories et, à droite, les descripteurs d'une catégorie sélectionnée. Chaque catégorie comprend un certain nombre de descripteurs. Pour chaque catégorie sélectionnée, les descripteurs associés apparaissent dans la table. Ces descripteurs peuvent comporter des concepts, des règles de catégorie, des types et des patrons TLA. Le type de chaque descripteur, ainsi que des exemples de ce que chaque descripteur représente, y figurent également.

Dans cet onglet, l'objectif est de sélectionner les catégories que vous souhaitez utiliser pour l'affectation des documents dans les catégories. Dans un modèle de catégories, l'affectation des documents dans les catégories des documents et des enregistrements s'effectue par catégorie. Si un document ou un enregistrement contient au moins un descripteur dans son texte ou des termes sous-jacents, ce document ou cet enregistrement est alors affecté à la catégorie auquel le descripteur appartient. Ces termes sous-jacents incluent les synonymes définis dans les ressources linguistiques (que ceux-ci aient été rencontrés ou non dans le texte) ainsi que tous les termes extraits au pluriel/singulier trouvés dans le texte qui sont utilisées pour générer le nugget de modèles, les termes permutés, les termes provenant du regroupement flou, etc.

Remarque : Si vous avez généré un nugget de modèle de concepts à la place, cet onglet contiendra des résultats différents. Pour plus d'informations, reportez-vous à la section Modèle de concepts : onglet Modèle sur p. 58.

Figure 3-34
Boîte de dialogue de nugget de modèle de catégories : onglet Modèle



Arborescence des catégories

Pour en savoir plus sur chaque catégorie, sélectionnez la catégorie de votre choix et passez en revue les informations correspondant aux descripteurs de cette catégorie. Pour chaque descripteur, vous pouvez consulter les informations suivantes :

- Nom du **Descripteur**. Ce champ contient une icône représentant le type de descripteur et indiquant le nom de ce dernier.

Concepts	Patrons TLA
Types	Règles de catégorie

- **Type.** Ce champ contient le nom du type du descripteur. Les types sont des regroupements de concepts similaires (regroupements sémantiques), tels que les noms d'organisation, les produits ou les opinions positives. Les règles ne sont affectées à aucun type.
- **Détails.** Ce champ contient une liste du contenu du descripteur. En fonction du nombre de correspondances, la liste complète de chaque descripteur risque de ne pas s'afficher entièrement en raison de la taille limitée de la boîte de dialogue.

Sélection et copie de catégories

Toutes les catégories de niveau supérieur sont sélectionnées par défaut pour l'affectation des documents dans les catégories, comme l'indiquent les cases à cocher du panneau de gauche. Si la case est cochée, la catégorie est utilisée pour l'affectation des documents dans les catégories. Si elle n'est pas cochée, la catégorie sera exclue de l'affectation des documents dans les catégories. Vous pouvez cocher plusieurs lignes à la fois en les sélectionnant toutes et en cliquant sur l'une des cases de la sélection. Ainsi si une catégorie ou une sous-catégorie est sélectionnée mais que l'une de ses sous-catégories n'est pas sélectionnée, la case affiche un fond bleu indiquant que la sélection des enfants est seulement partielle dans la catégorie sélectionnée.

Cliquez avec le bouton droit de la souris sur une catégorie de l'arborescence pour afficher un menu contextuel proposant les options suivantes :

- **Activer les concepts choisis.** Coche les cases de toutes les lignes de tableau sélectionnées.
- **Désélectionner les concepts choisis.** Désactivez les cases de toutes les lignes de tableau sélectionnées.
- **Tout activer.** Coche toutes les cases du tableau. Par conséquent, toutes les catégories sont utilisées dans les résultats finaux. Vous pouvez également utiliser l'icône de la case correspondante sur la barre d'outils.
- **Tout désactiver.** Désactive toutes les cases du tableau. Si vous désélectionnez une catégorie, celle-ci ne sera pas utilisée dans les résultats finaux. Vous pouvez également utiliser l'icône de la case vide correspondante sur la barre d'outils.

Cliquez avec le bouton droit de la souris sur une cellule du tableau Descripteurs pour afficher un menu contextuel proposant les options suivantes :

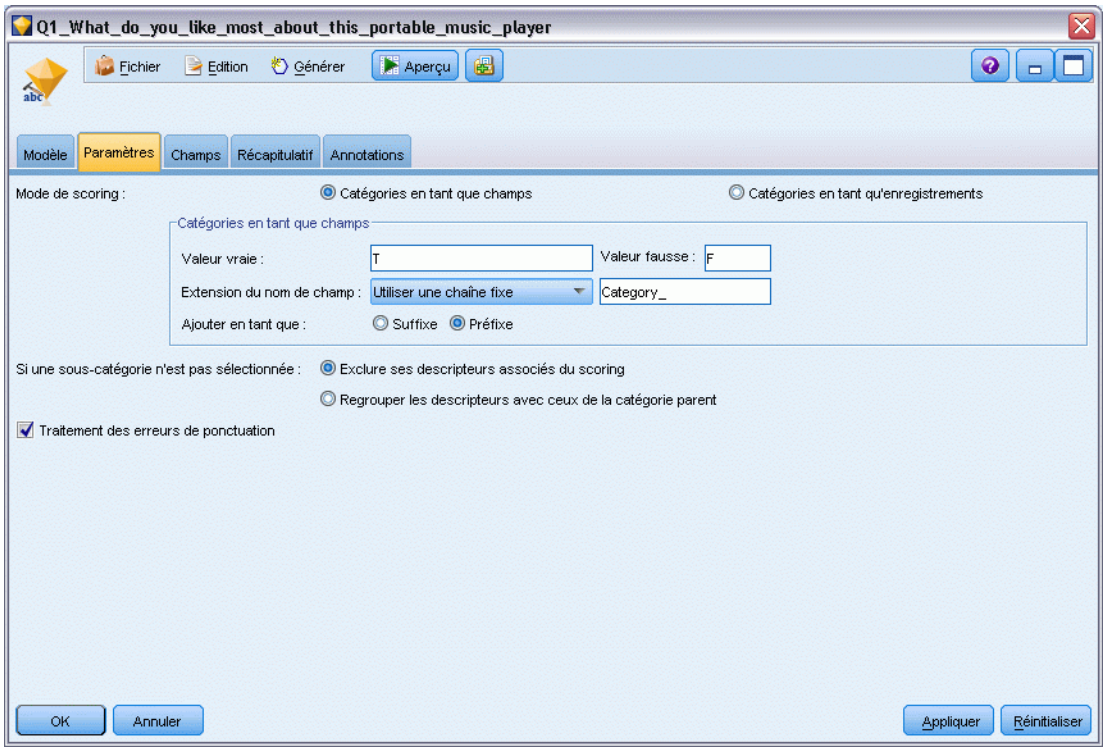
- **Copier.** Le ou les concepts sélectionnés sont copiés dans le Presse-papiers.
- **Copier avec les champs.** Le descripteur sélectionné est copié dans le Presse-papiers, de même que les en-têtes de colonne.
- **Tout sélectionner.** Toutes les lignes du tableau sont sélectionnées.

Nugget de modèle de catégories : Onglet Paramètres

L'onglet Paramètres sert à définir la valeur du champ de texte des nouvelles données d'entrée, si nécessaire. Il permet également d'indiquer le modèle de données des résultats (mode de l'affectation des documents dans les catégories).

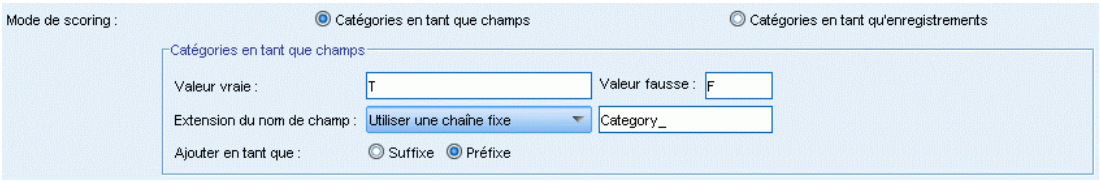
Remarque : Cet onglet n'apparaît dans la boîte de dialogue du nœud que si le nugget de modèle figure sur l'espace de travail ou dans un flux. Il n'apparaît pas lorsque vous accédez à ce nugget directement à partir de la palette Modèles.

Figure 3-35
Boîte de dialogue de nugget de modèle de catégories Text Mining : onglet Paramètres



Mode de scoring : Catégories en tant que champs

Figure 3-36
Onglet Paramètres pour « Catégories en tant que champs »



Grâce à cette option, les enregistrements de sortie sont aussi nombreux que les enregistrements d'entrée. Désormais, chaque enregistrement comporte toutefois un nouveau champ pour chaque catégorie sélectionnée (cochée) dans l'onglet Modèle. Pour chaque champ, entrez une valeur booléenne pour Vrai et pour Faux, telles que *Oui*/*Non*, *Vrai*/*Faux*, *V*/*F*, ou *1* et *2*. Les types de stockage sont définis automatiquement pour correspondre aux valeurs choisies. Par exemple, si vous entrez des valeurs numériques pour les booléens, elles seront automatiquement traitées comme valeurs entières. Les booléens disposent des types de stockage suivants : chaîne, entier, nombre réel ou date/heure.

Extension nom de champ. Vous pouvez choisir de spécifier une extension préfixe/suffixe pour le nom de champ ou vous pouvez choisir d'utiliser les codes de catégories. Les noms de champ générés reprennent le nom de catégorie et l'extension.

- **Ajouter en tant que.** Spécifiez à quel emplacement du nom de champ l'extension doit être ajoutée. Choisissez Préfixe pour ajouter l'extension en début de chaîne. Choisissez Suffixe pour ajouter l'extension en fin de chaîne.

Si une sous-catégorie n'est pas sélectionnée. Cette option permet de déterminer comment seront manipulés les descripteurs appartenant aux sous-catégories qui n'étaient pas sélectionnées pour le scoring. Il y a deux options.

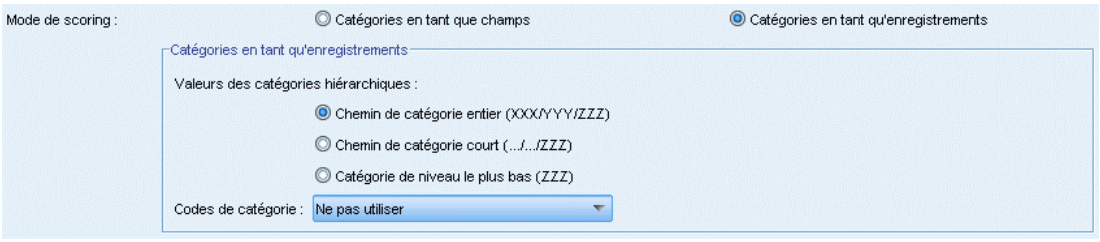
- Avec l'option **Exclure complètement ses descripteurs du scoring**, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront ignorés et ne seront pas utilisés pendant le scoring.
- Avec l'option **Agréger les descripteurs avec ceux des catégories parents**, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront utilisés comme descripteurs pour la catégorie parent (la catégorie au-dessus de cette sous-catégorie). S'il existe plusieurs niveaux de sous-catégories et que celles-ci ne sont pas sélectionnées, les descripteurs seront reportés sous la première catégorie parent disponible.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Remarque : L'option Traitement des erreurs de ponctuation ne s'applique pas au texte en japonais.

Mode de scoring : Catégories en tant qu'enregistrements

Figure 3-37
Onglet Paramètres pour « Catégories en tant qu'enregistrements »



Grâce à cette option, un nouvel enregistrement est créé pour chaque paire de category, document. Généralement, la sortie comporte plus d'enregistrements que n'en comportait l'entrée. Outre les champs d'entrée, de nouveaux champs sont également ajoutés aux données en fonction du type de modèle dont il s'agit.

Table 3-2
Champs de sortie de l'option Catégories en tant qu'enregistrements

Nouveau champ de sortie	Description
Category	Indique le nom de la catégorie à laquelle le document texte a été affecté. Si la catégorie est une sous-catégorie d'une autre catégorie, alors le chemin d'accès complet au nom de catégorie est contrôlé par la valeur choisie dans cette boîte de dialogue.

Valeurs des catégories hiérarchiques. Cette option contrôle le mode d'affichage des noms de sous-catégories dans les résultats.

- **Chemin d'accès complet aux catégories.** Cette option va générer le nom de la catégorie et le chemin d'accès complet aux catégories parents le cas échéant en utilisant des barres obliques pour séparer les noms de catégories des noms de sous-catégories.
- **Chemin d'accès court aux catégories.** Cette option va générer seulement le nom de la catégorie mais utilise des point de suspension pour afficher le nombre de catégories parents pour la catégorie en question.
- **Catégorie du niveau le plus bas.** Cette option va générer seulement le nom de la catégorie sans afficher le chemin d'accès complet ou les catégories parents.

Si une sous-catégorie n'est pas sélectionnée. Cette option permet de déterminer comment seront manipulés les descripteurs appartenant aux sous-catégories qui n'étaient pas sélectionnées pour le scoring. Il y a deux options.

- Avec l'option **Exclure complètement ses descripteurs du scoring**, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront ignorés et ne seront pas utilisés pendant le scoring.
- Avec l'option **Agréger les descripteurs avec ceux des catégories parents**, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront utilisés comme descripteurs pour la catégorie parent (la catégorie au-dessus de cette sous-catégorie). S'il existe plusieurs niveaux de sous-catégories et que celles-ci ne sont pas sélectionnées, les descripteurs seront reportés sous la première catégorie parent disponible.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Remarque : L'option Traitement des erreurs de ponctuation ne s'applique pas au texte en japonais.

Nugget de modèle de catégories : autres onglets

L'onglet Champs et l'onglet Paramètres du nugget de modèle de catégories sont identiques à ceux du nugget de modèle de concepts.

- Onglet Champs. Pour plus d'informations, reportez-vous à la section Modèle de concepts : onglet Champs sur p. 64.
- Onglet Récapitulatif. Pour plus d'informations, reportez-vous à la section Modèle de concepts : onglet Récapitulatif sur p. 66.

Utilisation des nuggets de modèle de catégories dans un flux

Le nugget de modèle de catégories Text Mining est généré à partir d'une session interactive. Vous pouvez utiliser ce nugget de modèle dans un flux.

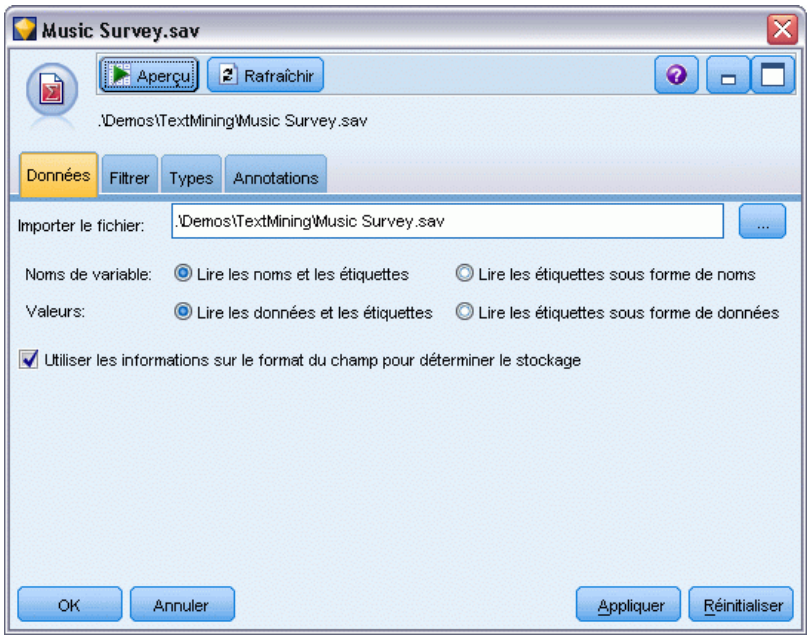
Exemple : nœud Fichier statistiques avec le nugget de modèle de catégories

L'exemple suivant indique comment utiliser le nugget de modèle Text Mining.

Figure 3-38
Exemple de flux : nœud Fichier statistiques avec un nugget de modèle de catégories Text Mining

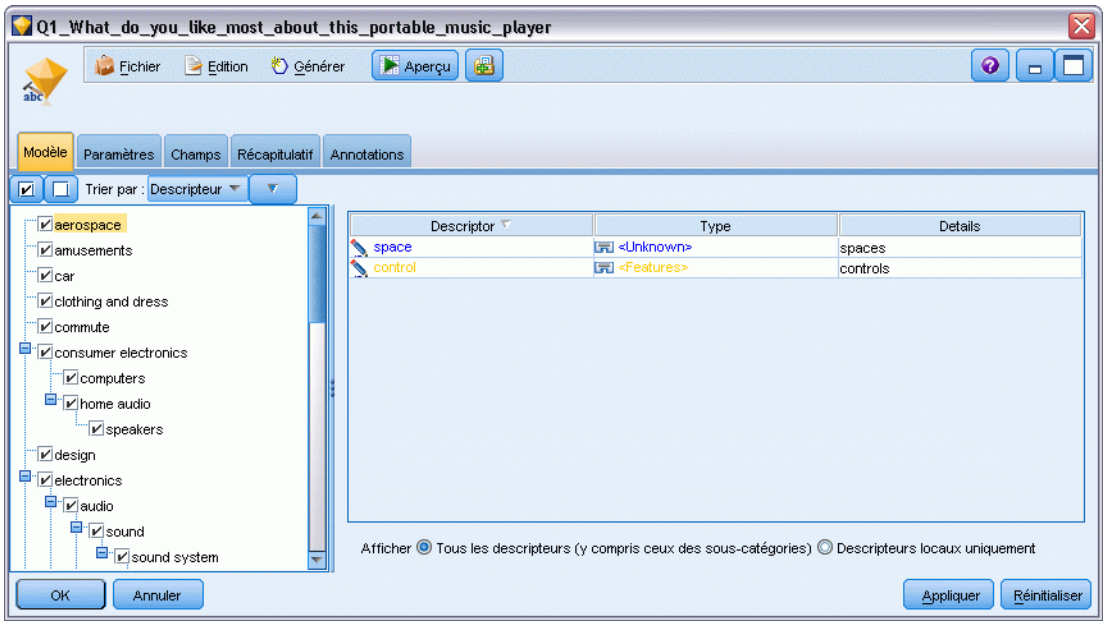
- **Nœud Fichier statistiques (onglet Paramètres).** Nous avons tout d'abord ajouté ce nœud au flux pour indiquer l'emplacement de stockage des documents texte.

Figure 3-39
Boîte de dialogue du nœud Fichier statistiques : onglet Données



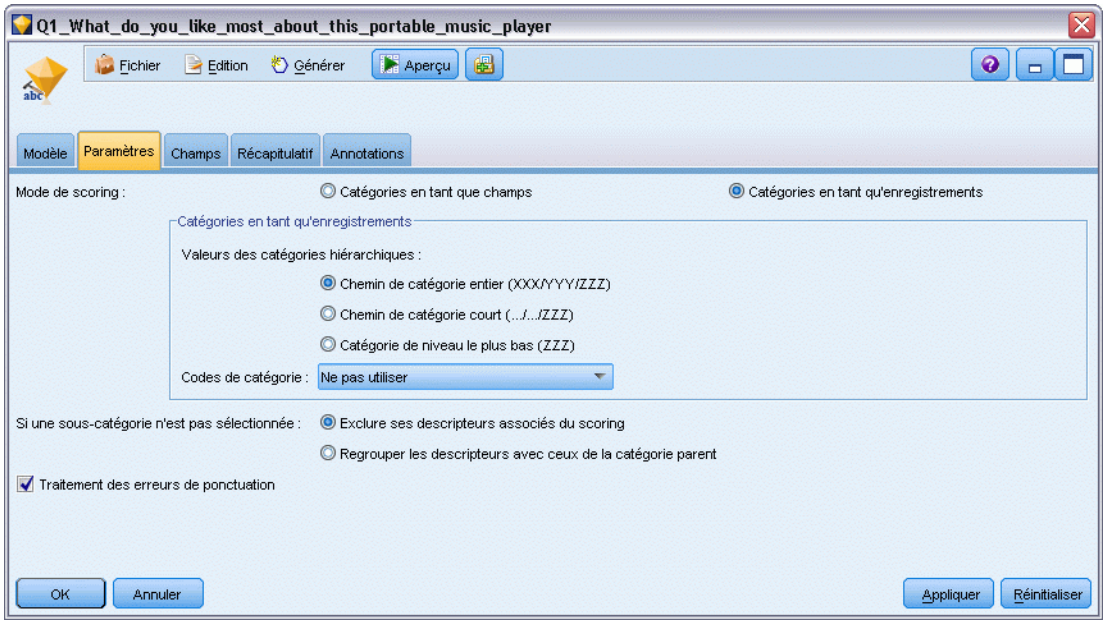
- **nugget de modèle de catégories Text Mining (onglet Modèle).** Nous avons ensuite ajouté et connecté un nugget de modèle de catégories au nœud Fichier Statistiques. Nous avons sélectionné les catégories que nous souhaitions utiliser pour scorer nos données.

Figure 3-40
Boîte de dialogue de nugget de modèle Text Mining : onglet Modèle



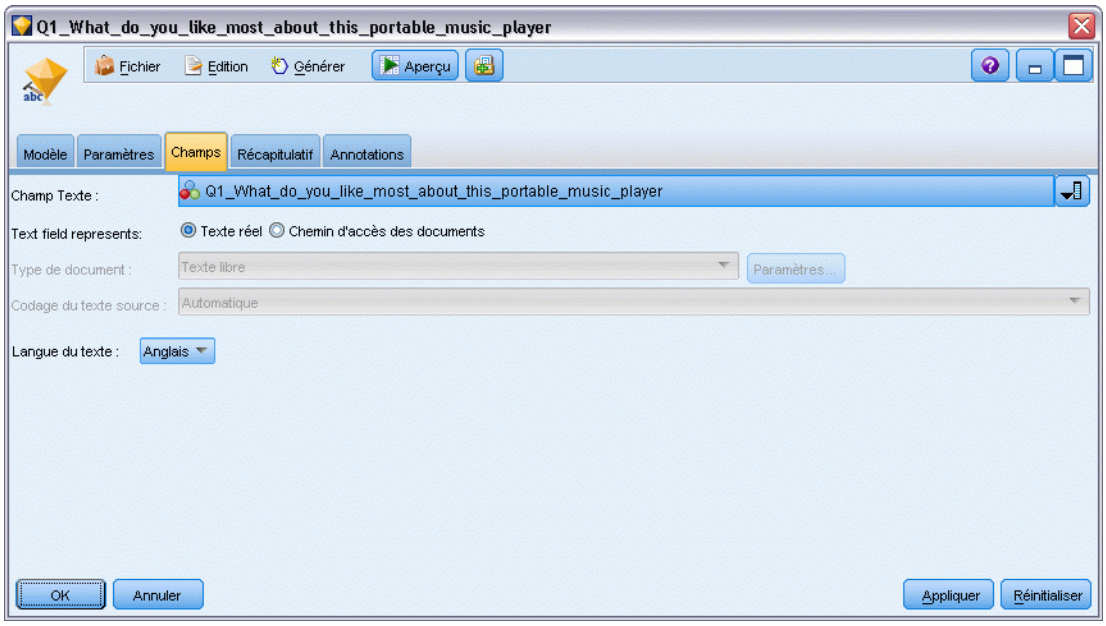
- **nugget de modèle Text Mining (onglet Paramètres).** Nous avons ensuite défini le format de sortie Catégories en tant que champs.

Figure 3-41
Boîte de dialogue de nugget de modèle de catégories : onglet Paramètres



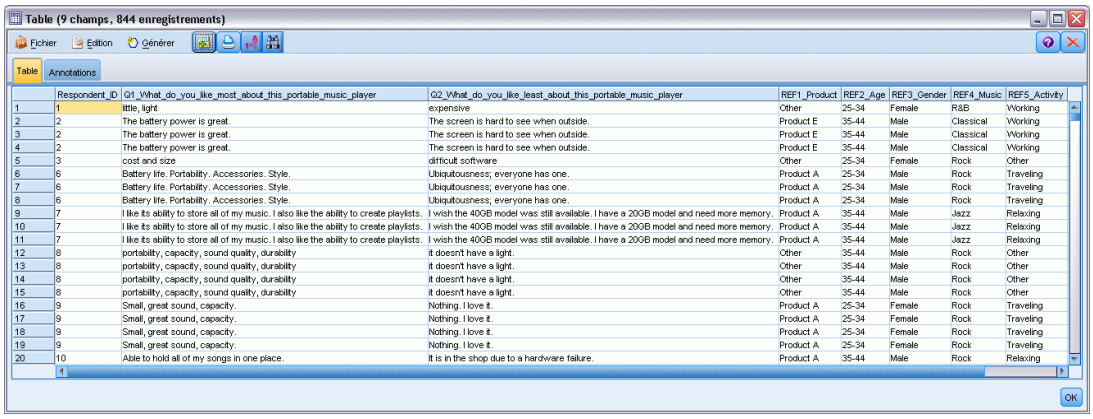
- **nugget de modèle de catégories Text Mining (onglet Champs).** Nous avons ensuite sélectionné la variable de champ de texte, correspondant au nom du champ issu du nœud Fichier statistiques, et sélectionné l'option Le texte correspond au Texte réel, ainsi que d'autres paramètres.

Figure 3-42
Boîte de dialogue de nugget de modèle Text Mining : onglet Champs



▶ **Nœud Table.** Nous avons ensuite associé un nœud de table pour afficher les résultats et exécuté le flux.

Figure 3-43
Sortie de table



Chapitre
4

Exploration des liens du texte

Nœud Analyse des liens du texte

Le nœud Analyse des liens du texte (TLA) ajoute une technologie de mise en correspondance de patrons à l'extraction de concepts du Text Mining, de façon à identifier les relations entre les concepts des données textuelles sur la base de patrons connus. Ces relations peuvent décrire l'impression d'un client vis-à-vis d'un produit, les organisations qui travaillent ensemble et même les relations entre des gènes ou des agents pharmaceutiques.

Par exemple, l'extraction du nom du produit de votre concurrent peut revêtir à vos yeux un intérêt limité. Grâce à ce nœud, vous pouvez également savoir comment les gens perçoivent le produit, du moins si ces opinions sont exprimées dans les données. Pour identifier et extraire les relations et les associations, les données textuelles sont comparées à des patrons connus.

Vous pouvez utiliser les règles de patrons TLA de certains modèles de ressources livrés avec IBM® SPSS® Modeler Text Analytics ou créer/modifier vos propres patrons. Les règles de patrons sont constituées de macros, de listes de mots et d'intervalles de mots pour former une requête booléenne, ou règle, qui est comparée à votre texte d'entrée. Lorsqu'une règle de patron TLA correspond au texte, il est possible d'extraire ce texte sous la forme d'un résultat TLA et de le restructurer sous la forme de données de sortie. Pour plus d'informations, reportez-vous à la section A propos des règles des liens du texte dans le chapitre 19 sur p. 338.

Le nœud Analyse des liens du texte propose une méthode plus directe pour identifier les résultats de patrons TLA, les extraire du texte et ajouter leurs résultats à l'ensemble de données figurant dans le flux. Mais le nœud Analyse des liens du texte ne constitue pas le seul moyen d'exécuter une analyse des liens du texte. Vous pouvez également lancer une session interactive dans le nœud de modélisation Text Mining.

Dans la session interactive, vous pouvez explorer les résultats de patrons TLA et les utiliser sous la forme de descripteurs de catégorie et/ou en apprendre davantage sur les résultats à l'aide de défilements et de graphiques. Pour plus d'informations, reportez-vous à la section Exploration de l'analyse des liens du texte dans le chapitre 12 sur p. 245. En fait, l'utilisation du nœud de Text Mining pour extraire des résultats TLA est un bon moyen d'explorer et d'affiner des modèles pour vos données, en vue d'une utilisation ultérieure directement dans le nœud Analyse des liens du texte.

La sortie peut être représentée sous la forme de 6 propriétés ou parties maximum. Les sorties des patrons japonais peuvent avoir une ou deux propriétés uniquement. Pour plus d'informations, reportez-vous à la section Sortie de nœud TLA sur p. 89.

Vous pouvez trouver ce nœud dans l'onglet SPSS Modeler Text Analytics de la palette de nœuds en bas de la fenêtre IBM® SPSS® Modeler. Pour plus d'informations, reportez-vous à la section IBM SPSS Modeler Text Analytics Noeuds dans le chapitre 1 sur p. 11.

Conditions requises. Le nœud Analyse des liens du texte accepte les données textuelles lues dans un champ à l'aide de l'un des nœuds source standard (nœud SGBD, nœud Fichier plat, etc.) ou lues dans un champ répertoriant les chemins d'accès aux documents externes générés par un nœud Liste fichiers ou un nœud Fil de nouvelles.

Puissance. Le nœud Analyse des liens du texte ne se limite pas à une simple extraction de concepts permettant de fournir des informations sur les relations *entre* les concepts, ainsi que des opinions ou des qualificatifs associés susceptibles d'apparaître dans les données.

Nœud Analyse des liens du texte : onglet Champs

Figure 4-1
Boîte de dialogue du nœud Analyse des liens du texte : onglet Champs

L'onglet Champs sert à indiquer les paramètres de champ des données dont vous allez extraire les concepts. Les paramètres pouvant être définis sont les suivants :

Champ ID. Sélectionnez le champ contenant l'identificateur des enregistrements textuels. L'identificateur doit être un entier. Le champ d'ID sert d'index aux enregistrements textuels. Utilisez un champ ID si Le texte correspond au texte à explorer. N'utilisez pas de champ d'ID si le champ texte correspond au Chemin d'accès des documents.

Champ Texte. Sélectionnez le champ contenant le texte à explorer, le chemin d'accès au document ou le chemin d'accès au répertoire contenant les documents. Ce champ dépend de la source de données.

Le texte correspond au. Indique ce que contient le champ de texte spécifié dans le paramètre précédent. Les différents choix sont :

- **Texte réel.** Sélectionnez cette option si le champ contient le texte exact à partir duquel les concepts doivent être extraits.
- **Chemin d'accès des documents.** Sélectionnez cette option si le champ contient un ou plusieurs noms de chemin d'accès pour les emplacements des documents texte.

Type de document. Cette option n'est disponible que si vous indiquez que le champ texte correspond au Chemin d'accès des documents. Le type de document indique la structure du texte. Sélectionnez l'un des types suivants :

- **Texte libre.** S'utilise pour la plupart des documents ou des sources textuelles. L'ensemble de texte entier est analysé pour l'extraction. Contrairement aux autres options, il n'existe pas de paramètres supplémentaires pour cette option.
- **Texte structuré.** Utilisez ce format pour les bibliographies, les brevets et les fichiers contenant des structures standard qui peuvent être identifiées et analysées. Ce type de document est utilisé pour éviter tout ou une partie du processus d'extraction. Il permet de définir des séparateurs de termes, d'affecter des types et d'imposer une valeur de fréquence minimale. Si vous sélectionnez cette option, vous devez cliquer sur le bouton Paramètres et entrer les séparateurs de texte dans la zone Formatage de texte structuré de la boîte de dialogue Paramètres du document. Pour plus d'informations, reportez-vous à la section Paramètres de document de l'onglet Champs dans le chapitre 3 sur p. 35.
- **Texte XML.** Permet de signaler les balises XML qui contiennent le texte à extraire. Toutes les autres balises sont ignorées. Si vous sélectionnez cette option, vous devez cliquer sur le bouton Paramètres et indiquer de manière explicite les éléments XML contenant le texte à lire au cours du processus d'extraction dans la zone Formatage de texte XML de la boîte de dialogue Paramètres du document. Pour plus d'informations, reportez-vous à la section Paramètres de document de l'onglet Champs dans le chapitre 3 sur p. 35.

Unité de texte. Cette option n'est disponible que si vous indiquez que Le texte correspond au chemin d'accès des documents et que vous sélectionnez le type de document Texte libre. Sélectionnez le mode d'extraction parmi les choix suivants :

- **Document.** Utilisez ce mode pour les documents courts et homogènes d'un point de vue sémantique (par exemple, dans le cas d'articles issus d'agences de presse).
- **Paragraphe.** Utilisez ce format pour les pages Web et les documents sans balise. Le processus d'extraction divise les documents en unités sémantiques, sur la base de certaines caractéristiques, comme des balises internes et des éléments syntaxiques. Si ce mode est sélectionné, le scoring est appliquée paragraphe par paragraphe. Par conséquent, la règle `pomme & orange` est vraie uniquement si `pomme` et `orange` se trouvent dans le même paragraphe, par exemple.

Paramètres de Paragraphe. Cette option n'est disponible que si vous avez indiqué que Le texte correspond au chemin d'accès des documents et défini l'option d'unité de texte sur Paragraphe. Indiquez les nombres maximal et minimal de caractères à utiliser dans les extractions. La taille employée est en fait arrondie de manière à inclure les caractères compris avant le point le plus proche (qu'il se situe avant ou après la limite fixée). Pour vous assurer que les associations de mots obtenues à partir du texte du groupe de documents sont représentatives, n'indiquez pas de taille d'extraction trop petite.

- **Minimum.** Indiquez le nombre minimum de caractères à utiliser dans les extractions.
- **Maximum.** Indiquez le nombre maximal de caractères à utiliser dans les extractions.

Encoding du texte source. Cette option n'est disponible que si vous indiquez que Le texte correspond au chemin d'accès des documents. Elle indique l'encoding de texte par défaut. Dans toutes les langues, à l'exception du japonais, l'encoding spécifié ou reconnu est converti en un encoding ISO-8859-1. Ainsi, même si vous indiquez un autre encoding, le moteur d'extraction le remplace par l'encoding ISO-8859-1 avant de traiter le texte. Les caractères qui ne figurent pas dans la définition de l'encoding ISO-8859-1 sont convertis en espaces. Pour le texte en japonais, vous pouvez choisir l'une des options de codage suivantes : SHIFT_JIS, EUC_JP, UTF-8, ou ISO-2022-JP.

Copiez les ressources de. Lors de l'exploration de texte, l'extraction est basée sur les paramètres de l'onglet Expert mais également sur les ressources linguistiques. Ces ressources servent de base au traitement et à l'exécution du texte pendant l'extraction afin d'obtenir des concepts, des types et des patrons TLA. Vous pouvez copier les ressources dans ce nœud à partir d'un modèle de ressources.

Un modèle de ressources est un ensemble prédéfini de bibliothèques et de ressources linguistiques et non linguistiques avancées qui ont été affinées pour un domaine ou une utilisation spécifique. Ces ressources servent de base à la gestion et le traitement des données lors de l'extraction. Cliquez sur Charger et sélectionnez le modèle à partir duquel copier vos ressources.

Les modèles sont chargés lorsque vous les sélectionnez et non lorsque le flux est exécuté. Au moment du chargement, une copie des ressources est stockée dans le nœud. Par conséquent, si vous souhaitez utiliser un modèle mis à jour, vous devez le recharger ici. Pour plus d'informations, reportez-vous à la section Copie des ressources à partir de modèles et de TAP dans le chapitre 3 sur p. 42.

Langue du texte. Identifie la langue du texte exploré. Les ressources copiées dans le noeud contrôlent les options de langue présentées. Vous pouvez sélectionner la langue correspondant aux ressources ou choisir l'option TOUTES. Nous vous recommandons vivement de spécifier la langue exacte des données textuelles ; cependant, si vous avez un doute, vous pouvez choisir l'option TOUTES. TOUTES ne s'applique pas au texte en japonais. L'option TOUTES allonge la durée d'exécution car la reconnaissance automatique de la langue est utilisée pour analyser tous les documents et enregistrements de façon à identifier d'abord la langue du texte. Avec cette option, tous les enregistrements ou documents dont la langue est prise en charge et fait l'objet d'une licence sont lus par le moteur d'extraction à l'aide des dictionnaires internes propres aux langues. Pour plus d'informations, reportez-vous à la section Identificateur de langue dans le chapitre 18 sur p. 336. Si vous souhaitez acquérir la licence d'une langue prise en charge à laquelle vous n'avez pas accès actuellement, contactez votre représentant commercial.

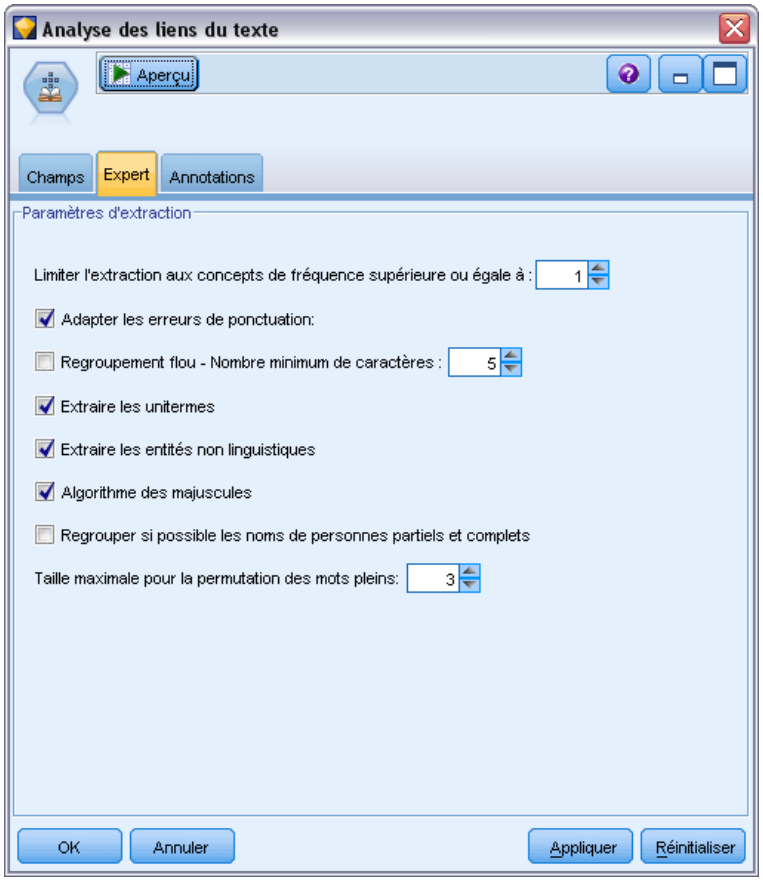
Nœud Analyse des liens du texte : Onglet Expert

Dans ce nœud, l'extraction des résultats des patrons TLA (analyse des liens du texte) est automatiquement activée. L'onglet Expert contient des paramètres supplémentaires ayant une incidence sur le mode d'extraction et de traitement du texte. Les paramètres de cette boîte de dialogue déterminent le fonctionnement de base du processus d'extraction, ainsi que quelques procédures avancées. Il existe également un certain nombre de ressources linguistiques et

d'options ayant une incidence sur les résultats de l'extraction, qui sont contrôlées par le modèle de ressources sélectionné.

Pour le texte en néerlandais, en anglais, en français, en allemand, en italien, en portugais, et en espagnol

Figure 4-2
Boîte de dialogue du nœud Analyse des liens du texte : Onglet Expert



Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Traitement des fautes de frappe. Nombre de caractères minimum requis ([n]). Cette option applique une technique de regroupement flou qui permet de regrouper les mots mal orthographiés ou d'orthographe similaire sous un seul concept. L'algorithme de regroupement flou retire temporairement toutes les voyelles (à l'exception de la première) et les consonnes doubles/triples des mots extraits pour comparer ces derniers et voir s'ils sont identiques, afin par exemple que `modélisation` et `modelisation` soient regroupés. Néanmoins, si chaque type est affecté à un type différent et que vous excluez le type `<Unknown>`, le regroupement flou n'est pas appliqué.

Vous pouvez également définir le nombre minimum de caractères *racines* requis avant d'utiliser le regroupement flou. Le nombre de caractères racine d'un terme est calculé en ajoutant l'ensemble des caractères et en soustrayant les caractères formant des suffixes flexionnels et, dans le cas des termes apparaissant sous la forme de mots composés, les déterminants et les prépositions. Par exemple, le terme `exercices` serait déterminé comme comportant 8 caractères racine dans la forme « exercice », étant donné que la lettre *s* située à la fin du mot représente une flexion (forme plurielle). De même, `jus énergétique` comporte 14 caractères racine « jus énergétique » et `conception de dessins`, 16 caractères racine « conception dessin ». Cette méthode de calcul est uniquement utilisée pour vérifier si le regroupement flou doit être appliqué, mais n'influe pas sur la mise en correspondance des mots.

Remarque : Si, d'après vous, certains mots sont ensuite groupés de manière incorrecte, vous pouvez exclure des paires de mots de cette technique en les déclarant de manière explicite dans la section Regroupement flou : Exceptions de l'onglet de ressources avancées. Pour plus d'informations, reportez-vous à la section Regroupement flou dans le chapitre 18 sur p. 328.

Extraire les unitermes. Cette option extrait les mots uniques (unitermes) dans la mesure où le mot ne fait pas déjà partie d'un mot composé et si c'est un nom ou une catégorie grammaticale non reconnue.

Extraction des entités non linguistiques. Cette option extrait les entités non linguistiques, telles que les numéros de téléphone, numéros de sécurité sociale, heures, dates, devises, chiffres, pourcentages, adresses électroniques, adresses HTTP, etc. Vous pouvez inclure ou exclure certains types d'entités non linguistiques dans la section Entités non linguistiques : configuration de l'onglet de ressources avancées. Désactivez les entités dont vous n'avez pas besoin pour éviter au moteur d'extraction un temps de traitement inutile. Pour plus d'informations, reportez-vous à la section Configuration dans le chapitre 18 sur p. 332.

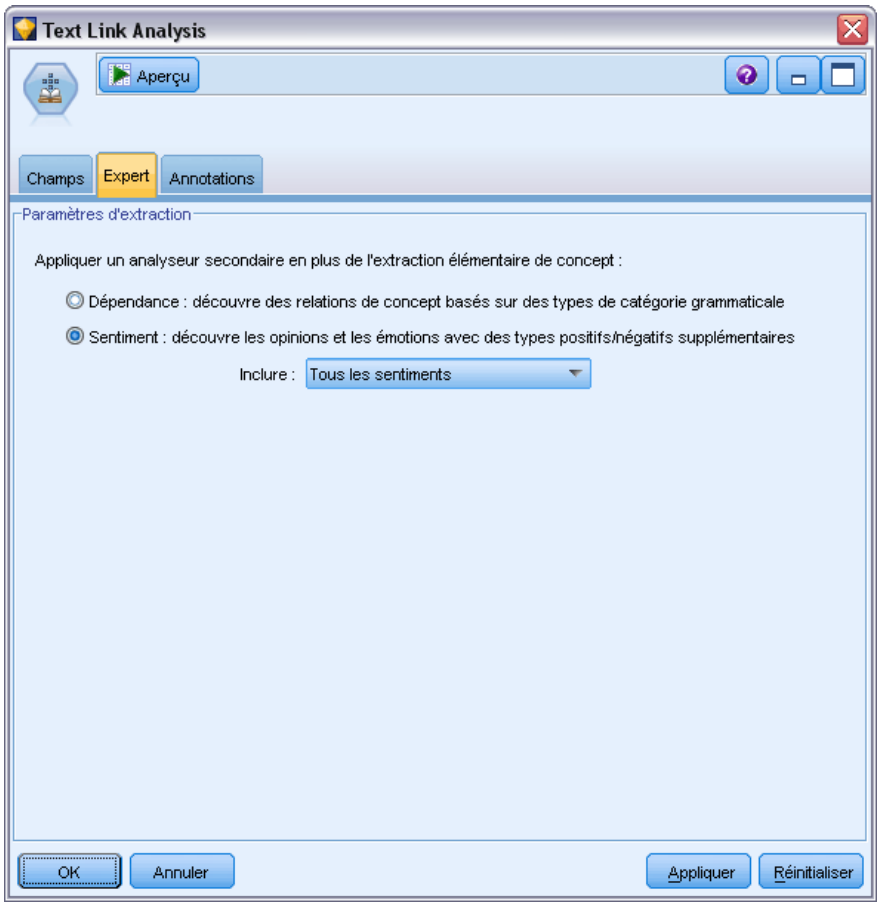
Algorithme des noms propres. Cette option extrait les termes simples et composés ne figurant pas dans les dictionnaires intégrés, si la première lettre est une majuscule. Cette option offre une bonne manière d'extraire la plupart des noms propres.

Regroupement éventuel des noms de personnes partiels et complets. Cette option regroupe les noms qui apparaissent différemment dans le texte. Cette fonction est utile car les noms sont souvent cités dans leur forme complète au début du texte puis uniquement en version abrégée. Cette option essaye de faire correspondre tout uniterme ayant le type `<Unknown>` avec le dernier mot de tout terme composé entré comme `<Person>`. Par exemple, si *martin* est trouvé et initialement saisi comme `<Unknown>`, le moteur d'extraction vérifie si un terme composé de type `<Person>` contient *martin* comme dernier mot, tel que *pierre martin*. Cette option ne s'applique pas aux prénoms, car la plupart ne sont jamais extraits comme unitermes.

Nombre de mots pleins soumis à une permutation pour le regroupement. Cette option indique le nombre maximal de mots utiles pouvant être présents lorsque s'applique la technique de permutation. Cette technique de permutation regroupe les expressions similaires qui ne diffèrent les unes des autres que par la présence de mots utiles (par exemple, de et la), quelle que soit leur flexion. Par exemple, si vous avez paramétré cette valeur sur au moins deux mots et si les deux expressions `représentants d'entreprise` et `représentants de l'entreprise` ont été extraites. Dans ce cas, les deux termes extraits sont regroupés dans la liste de concepts finale car les deux termes sont considérés comme étant les mêmes lorsque `de l'` est ignoré.

Pour les textes en japonais

Figure 4-3
Boîte de dialogue du nœud Analyse des liens du texte : Onglet Expert (texte en japonais)



Pour le texte en japonais, vous pouvez choisir l'analyse secondaire à appliquer.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Analyse secondaire. Lorsqu'une extraction est lancée, l'extraction des mots-clés de base est effectuée à l'aide de l'ensemble de types par défaut. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais dans l'annexe A sur p. 378. Mais lorsque vous sélectionnez un analyseur secondaire, vous pouvez obtenir des concepts plus nombreux et plus riches car l'extracteur inclura désormais des verbes à particules et des auxiliaires comme faisant partie du concept. Par exemple, supposons que nous avons une phrase 肩の荷が下りた, traduite par "*Ca m'a enlevé un gros poids*". Dans cet exemple, l'extraction des mots-clés de base peut extraire chaque concept séparément comme suit : 肩 (*poids*), 荷 (*gros*), 下りる (*a enlevé*), mais la relation entre ces mots n'est pas extraite. Cependant, si vous avez appliqué l'analyse de sentiment, vous pouvez extraire des concepts plus riches relatifs à un type de sentiment comme le concept =肩の荷が下りた, qui est traduit par "*avoir enlevé un gros poids*", affecté au type <良い-安心>. Dans le cas d'une analyse de sentiment, un grand nombre de types supplémentaires est

également inclus. De plus, choisir un analyseur secondaire vous permet également de générer des résultats d'analyse des liens du texte.

Remarque : Lorsqu'un analyseur secondaire est appelé, le processus d'extraction nécessite plus de temps. Pour plus d'informations, reportez-vous à la section Fonctionnement de l'extraction secondaire dans l'annexe A sur p. 371.

- **Analyse des dépendances** Choisir cette option génère des particules étendues pour les concepts d'extraction du type de base et de l'extraction des mots-clés. Vous pouvez également obtenir des résultats de patrons plus riches avec l'analyse de dépendance des liens du texte (TLA).
- **Analyse des sentiments**. Choisir cet analyseur génère l'extraction de concepts supplémentaires et, le cas échéant, l'extraction de résultats de patrons TLA. En plus des types de base, vous bénéficiez également de plus de 80 types de sentiments, notamment 嬉しい, 吉報, 幸運, 安心, 幸福, etc. Ces types permettent de découvrir des concepts et des patrons dans le texte grâce à l'expression des émotions, des sentiments et des opinions. Ce sont trois options qui dictent la cible de l'analyse des sentiments : Tous les sentiments, Sentiment représentatif uniquement et Conclusions uniquement.

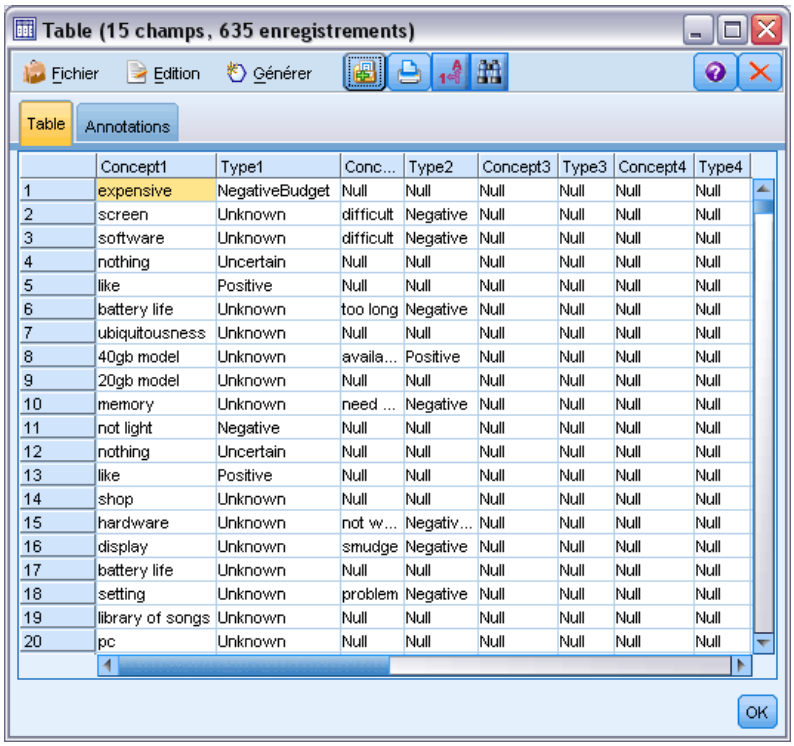
Sortie de nœud TLA

Après l'exécution du nœud Analyse des liens du texte, les données sont restructurées. Il est important de comprendre comment le Text Mining restructure les données. Pour obtenir une structure de Data mining différente, vous pouvez avoir recours aux nœuds de la palette Opérations sur les champs. Par exemple, si vous travaillez avec des données dans lesquelles chaque ligne représente un enregistrement textuel, une ligne est créée pour chaque patron découvert dans les données textuelles source. Pour chaque ligne de la sortie, il existe 15 champs :

- Six champs (Concept#, tels que Concept1, Concept2, etc. jusqu'à Concept6) représentent les concepts découverts dans la correspondance de patrons.
- Six champs (Type#, tels que Type1, Type2, etc. jusqu'à Type6) représentent le type de chaque concept.
- Nom de la règle représente le nom de la règle des liens du texte utilisée pour renvoyer le texte et générer la sortie.
- Un champ qui utilise le nom du champ ID spécifié dans le nœud et qui représente l'ID de l'enregistrement ou du document tel qu'il apparaissait dans les données d'entrée.
- Texte mis en correspondance représente la partie des données textuelles de l'enregistrement ou du document d'origine qui a été mise en correspondance avec le patron TLA.

Remarque : Les règles de patrons d'analyse des liens du texte pour le texte en japonais ne génèrent que des résultats de patrons à une ou deux propriétés.

Figure 4-4
Sortie affichée dans le nœud Table



	Concept1	Type1	Conc...	Type2	Concept3	Type3	Concept4	Type4
1	expensive	NegativeBudget	Null	Null	Null	Null	Null	Null
2	screen	Unknown	difficult	Negative	Null	Null	Null	Null
3	software	Unknown	difficult	Negative	Null	Null	Null	Null
4	nothing	Uncertain	Null	Null	Null	Null	Null	Null
5	like	Positive	Null	Null	Null	Null	Null	Null
6	battery life	Unknown	too long	Negative	Null	Null	Null	Null
7	ubiquitousness	Unknown	Null	Null	Null	Null	Null	Null
8	40gb model	Unknown	availa...	Positive	Null	Null	Null	Null
9	20gb model	Unknown	Null	Null	Null	Null	Null	Null
10	memory	Unknown	need ...	Negative	Null	Null	Null	Null
11	not light	Negative	Null	Null	Null	Null	Null	Null
12	nothing	Uncertain	Null	Null	Null	Null	Null	Null
13	like	Positive	Null	Null	Null	Null	Null	Null
14	shop	Unknown	Null	Null	Null	Null	Null	Null
15	hardware	Unknown	not w...	Negativ...	Null	Null	Null	Null
16	display	Unknown	smudge	Negative	Null	Null	Null	Null
17	battery life	Unknown	Null	Null	Null	Null	Null	Null
18	setting	Unknown	problem	Negative	Null	Null	Null	Null
19	library of songs	Unknown	Null	Null	Null	Null	Null	Null
20	pc	Unknown	Null	Null	Null	Null	Null	Null

Remarque : les flux préexistants contenant un nœud Analyse des liens du texte issu d'une version antérieure à la version 5.0 ne seront peut-être pas pleinement exécutables tant que les nœuds ne seront pas mis à jour. Certaines améliorations apportées aux versions ultérieures de IBM® SPSS® Modeler requièrent le remplacement d'anciens nœuds par leur nouvelle version, à la fois plus déployable et plus puissante.

Il est également possible d'effectuer une traduction automatique de certaines langues. Cette fonction permet d'explorer les documents figurant dans une langue que vous ne parlez pas. Si vous souhaitez utiliser la fonction de traduction, vous devez avoir accès à SDL Software as a Service (SaaS). Pour plus d'informations, reportez-vous à la section Paramètres de traduction dans le chapitre 5 sur p. 96.

Mise en cache des résultats TLA

Si vous procédez à une mise en cache, les résultats de l'analyse des liens du texte se situent dans le flux. Pour éviter de répéter l'extraction des résultats de l'analyse des liens du texte à chaque exécution du flux, sélectionnez le nœud Analyse des liens du texte, puis les options de menu suivantes Edition > Nœud >> Cache > Activer. Lors de l'exécution suivante du flux, la sortie est mise en cache dans le nœud. L'icône du nœud affiche une petite image représentant un « document » qui passe de la couleur blanche à la couleur verte lorsque le cache est rempli. Le cache est conservé pendant toute la durée de la session. Pour conserver le cache plus longtemps (après fermeture et réouverture du flux), sélectionnez le nœud, puis les options de menu suivantes :

Modifier > Nœud > Cache > Enregistrer le cache. Lors de l'ouverture suivante du flux, vous pouvez recharger le cache enregistré plutôt que d'exécuter à nouveau la traduction.

Vous pouvez également enregistrer ou activer un cache de nœud en cliquant avec le bouton droit de la souris sur le nœud et en choisissant Cache dans le menu contextuel.

Utilisation du nœud Analyse des liens du texte dans un flux

Le nœud Analyse des liens du texte sert à accéder aux données et à extraire des concepts dans un flux. Vous pouvez utiliser n'importe quel nœud source pour accéder aux données.

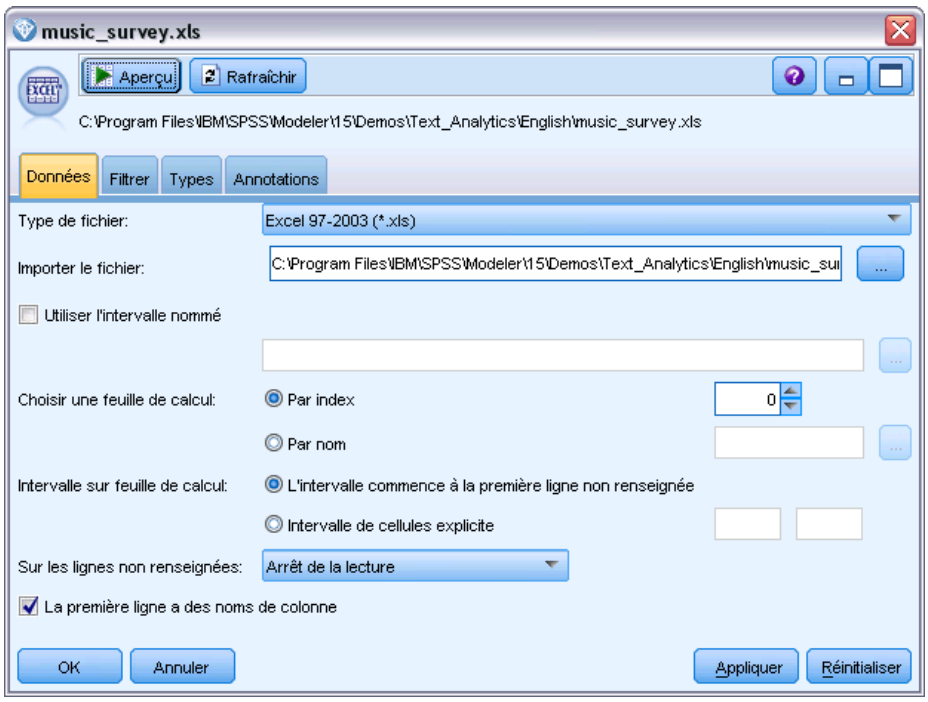
Exemple : nœud Fichier statistiques avec nœud Analyse des liens du texte

L'exemple suivant indique le mode d'utilisation du nœud Analyse des liens du texte.

Figure 4-5
Exemple : nœud Fichier statistiques avec nœud Analyse des liens du texte

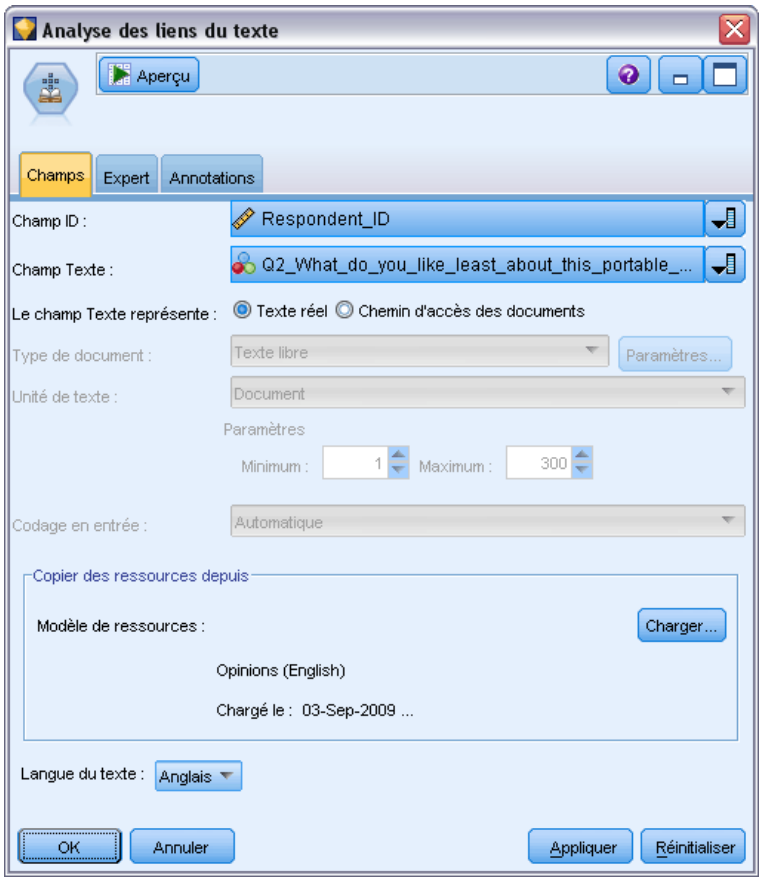
- **nœud Fichier statistiques (onglet Données).** Nous avons tout d'abord ajouté ce nœud au flux pour indiquer l'emplacement de stockage du texte.

Figure 4-6
Boîte de dialogue du nœud Fichier statistiques : onglet Données



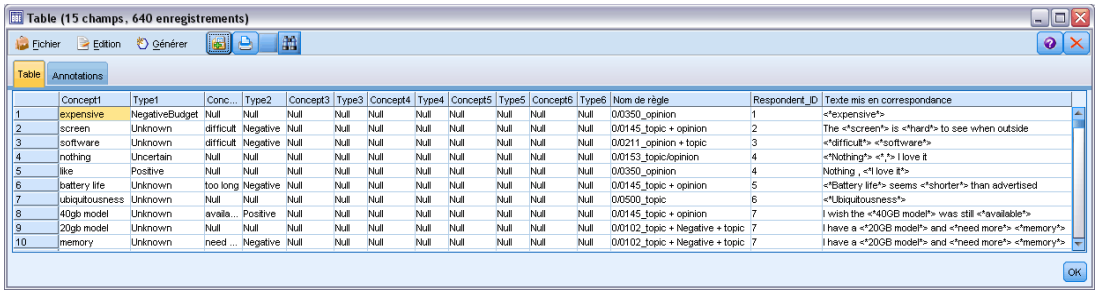
- **Nœud Analyse des liens du texte (onglet Champs).** Nous avons ensuite relié ce nœud au flux afin d'extraire les concepts en vue de l'affichage ou de la modélisation en aval. Nous avons indiqué le champ d'ID et le nom du champ de texte contenant les données, ainsi que d'autres paramètres.

Figure 4-7
Boîte de dialogue du nœud Analyse des liens du texte : onglet Champs



- **Nœud Table.** Enfin, nous avons ajouté un nœud Table pour afficher les concepts extraits des documents texte. Dans la sortie Table présentée, vous pouvez visualiser les résultats de patrons TLA trouvés dans les données après l'exécution de ce flux avec un nœud Analyse des liens du texte. Certains résultats indiquent que seul un concept/type a été mis en correspondance. D'autres résultats sont plus complexes et contiennent plusieurs types et concepts. En outre, après l'exécution des données via le nœud Analyse des liens du texte et l'extraction des concepts, plusieurs aspects des données ont changé. Les données d'origine de notre exemple contenaient 8 champs et 405 enregistrements. Après exécution du nœud Analyse des liens du texte, elles comportent 15 champs et 640 enregistrements. Il existe désormais une ligne pour chaque résultat de patron TLA trouvé. Ainsi, la ligne `ID 7` s'est transformée en trois lignes car trois résultats de patrons TLA ont été extraits. Pour fusionner les données de sortie dans vos données d'origine, vous pouvez utiliser un nœud Fusionner.

Figure 4-8
Nœud de sortie Table



	Concept1	Type1	Conc...	Type2	Concept3	Type3	Concept4	Type4	Concept5	Type5	Concept6	Type6	Nom de règle	Respondent_ID	Texte mis en correspondance
1	expensive	NegativeBudget	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	1	<*expensive*>
2	screen	Unknown	difficult	Negative	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	2	The <*screen*> is <*hard*> to see when outside
3	software	Unknown	difficult	Negative	Null	Null	Null	Null	Null	Null	Null	Null	0/0211_opinion + topic	3	<*difficult*> <*software*>
4	nothing	Uncertain	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0153_topic/opinion	4	<*Nothing*> <*,*> I love it
5	like	Positive	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	4	Nothing , <*I love it*>
6	battery life	Unknown	too long	Negative	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	5	<*Battery life*> seems <*shorter*> than advertised
7	ubiquitousness	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0500_topic	6	<*Ubiquitousness*>
8	40gb model	Unknown	availa...	Positive	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	7	I wish the <*40GB model*> was still <*available*>
9	20gb model	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>
10	memory	Unknown	need ...	Negative	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>

Chapitre
5

Traduction de texte pour l'extraction

Noeud Traduire

Vous pouvez utiliser le noeud Traduire afin de traduire en anglais un texte rédigé dans l'une des langues prises en charge, parmi lesquelles l'arabe, le chinois et le persan, pour des analyses à l'aide de IBM® SPSS® Modeler Text Analytics . Il offre ainsi la possibilité d'explorer des documents rédigés dans des langues présentant un jeu de caractères à deux octets qui, dans d'autres cas, ne seraient pas pris en charge ; en outre, il permet aux analystes d'extraire des concepts de documents rédigés en langue étrangère, et ce même s'ils ne comprennent pas la langue en question. Veuilez noter qu'il vous faut être en mesure de vous connecter à Software as a Service (SaaS) de SDL pour pouvoir utiliser le noeud Traduire.

Lorsque vous explorez du texte dans l'une de ces langues, il vous suffit d'ajouter dans votre flux un noeud Traduire avant le noeud de modélisation de Text Mining. Vous pouvez également activer la mise en cache dans le noeud Traduire pour éviter que la traduction ne soit répétée à chaque exécution du flux.

Vous pouvez trouver ce noeud dans l'onglet SPSS Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM® SPSS® Modeler. Pour plus d'informations, reportez-vous à la section IBM SPSS Modeler Text Analytics Noeuds dans le chapitre 1 sur p. 11.

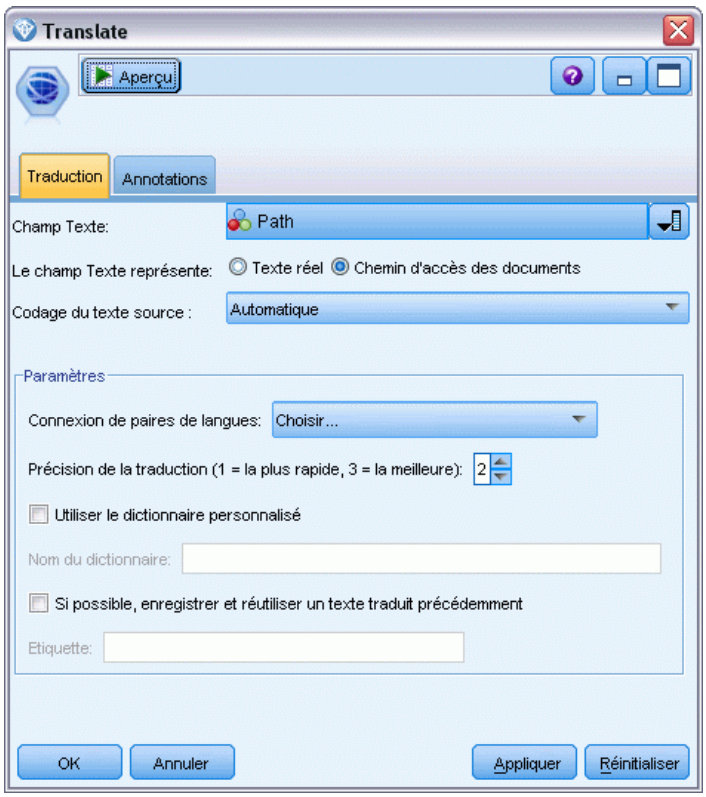
Mise en cache de la traduction. Si vous mettez la traduction en mémoire cache, le texte traduit est stocké dans le flux plutôt que dans des fichiers externes. Pour éviter de répéter la traduction à chaque exécution du flux, sélectionnez le noeud Traduire, puis les options de menu suivantes : Edition > Noeud > Cache > Activer. Lors de l'exécution suivante du flux, la traduction est mise en cache dans le noeud. L'icône du noeud affiche une petite image représentant un « document » qui passe de la couleur blanche à la couleur verte lorsque le cache est rempli. Le cache est conservé pendant toute la durée de la session. Pour conserver le cache plus longtemps (après fermeture et réouverture du flux), sélectionnez le noeud, puis les options de menu suivantes : Modifier > Noeud > Cache > Enregistrer le cache. Lors de l'ouverture suivante du flux, vous pouvez recharger le cache enregistré plutôt que d'exécuter à nouveau la traduction.

Vous pouvez également enregistrer ou activer un cache de noeud en cliquant avec le bouton droit de la souris sur le noeud et en choisissant Cache dans le menu contextuel.

Important ! Si vous essayez de récupérer des informations sur le Web avec un serveur proxy, vous devez activer ce serveur dans le fichier `net.properties` pour le serveur et le client SPSS Modeler Text Analytics . Suivez les instuctions détaillées dans ce fichier. Cela s'applique lorsque vous accédez au Web avec le noeud Fil de nouvelles ou lorsque vous récupérez une licence SDL Software as a Service (SaaS) car ces connexions utilisent Java. Pour le client, l'emplacement du fichier est *C:\Program Files\IBM\SPSS\Modeler\15\jre\lib\net.properties* par défaut. Pour le client, l'emplacement du fichier est *C:\Program Files\IBM\SPSS\Modeler\15\ext\bin\spss.TMWBServer\jre\lib\net.properties* par défaut.

Noeud Traduire : Onglet Traduction

Figure 5-1
Boîte de dialogue du noeud Traduire : onglet Champs



Champ Texte. Sélectionnez le champ contenant le texte à explorer, le chemin d'accès au document ou le chemin d'accès au répertoire contenant les documents. Ce champ dépend de la source de données. Vous pouvez spécifier tout champ de type chaîne, même ceux définis de la manière suivante : `Direction=Aucune` ou `Type=Sans type`.

Le texte correspond au. Indique ce que contient le champ de texte spécifié dans le paramètre précédent. Les différents choix sont :

- **Texte réel.** Sélectionnez cette option si le champ contient le texte exact à partir duquel les concepts doivent être extraits.
- **Chemin d'accès des documents.** Sélectionnez cette option si le champ contient un ou plusieurs chemins d'accès vers des documents externes qui contiennent le texte destiné à l'extraction. Par exemple, si un noeud Liste fichiers est utilisé pour lire une liste de documents, il convient de sélectionner cette option. Pour plus d'informations, reportez-vous à la section Noeud Liste fichiers dans le chapitre 2 sur p. 14.

Encoding du texte source. Sélectionner l'encoding du texte source. Vous pouvez commencer par sélectionner l'option Automatique mais si vous remarquez que certains fichiers ne sont pas traités correctement, nous vous recommandons de sélectionner l'encoding de la liste ci-dessous. L'option Automatique peut ne pas indentifier correctement l'encoding lorsqu'elle traite un petit texte comme de brefs enregistrements de base de données. La sortie texte de ce noeud est codée sous la forme `UTF-8`.

Paramètres. Définit les paramètres de traduction du flux.

- **Connexion des paires de langues**. Sélectionnez la paire de langues que vous souhaitez utiliser ; les paires de langues disponibles sont automatiquement affichées dans cette liste après la configuration du lien vers le service SDL dans la boîte de dialogue Paramètres de traduction. Pour plus d'informations, reportez-vous à la section Paramètres de traduction sur p. 96.
- **Précision de traduction**. Spécifiez la précision désirée en choisissant une valeur de 1 à 3 indiquant le niveau de vitesse par rapport à la précision souhaitée. Une valeur faible permet d'obtenir des résultats de traduction plus rapides mais moins précis. Une valeur élevée permet, quant à elle, d'obtenir des résultats plus précis mais après une durée de traitement plus longue. Pour optimiser le temps de traitement, nous vous recommandons de commencer par un niveau peu élevé et de l'augmenter seulement si, après avoir examiné les résultats, vous pensez avoir besoin d'une plus grande précision.
- **Utilisez le dictionnaire personnalisé**. Si vous aviez précédemment créé des dictionnaires personnalisés, détenus par SDL, vous pouvez les utiliser dans le cadre de la traduction. Pour choisir un dictionnaire personnalisé, cochez la case Utiliser un dictionnaire personnalisé et entrez le Nom du dictionnaire. Pour utiliser plusieurs dictionnaires, séparez les noms avec une virgule.
- **Si possible, enregistrer et réutiliser un texte traduit précédemment.** Indique que les résultats de traduction doivent être enregistrés et que, si le même nombre d'enregistrements/de documents existe à la prochaine exécution du flux, le contenu est supposé être le même et les résultats de traduction sont réutilisés pour écourter le temps de traitement. Si cette option est sélectionnée au moment de l'exécution et que le nombre d'enregistrements ne correspond pas au dernier contenu sauvegardé, le texte est traduit dans son intégralité, puis enregistré sous le nom d'étiquette pour la prochaine exécution. Cette option est disponible uniquement si vous avez sélectionné une langue de traduction SDL.

  Remarque : Si le texte est stocké dans le flux, vous pouvez également activer la mise en cache dans un noeud Traduire. Dans ce cas, les résultats de la traduction sont réutilisés mais tout ce qui se trouve en amont est également ignoré chaque fois que le cache est disponible.
- **Etiquette.** Si vous sélectionnez l'option Si possible, enregistrer et réutiliser un texte traduit précédemment, vous devez spécifier un nom d'étiquette pour les résultats. Cette étiquette permet d'identifier le texte préalablement traduit. Si aucune étiquette n'est spécifiée, un avertissement est ajouté aux propriétés du flux lors de l'exécution du flux et aucune réutilisation n'est possible.

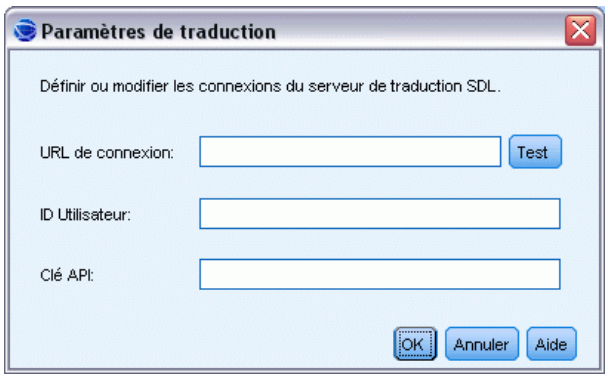
Paramètres de traduction

Dans cette boîte de dialogue, vous pouvez définir et gérer les connexions de la traduction SDL Software as a Service (SaaS) que vous pouvez réutiliser à chaque traduction. Lorsqu'une connexion a été définie ici, vous pouvez rapidement choisir une connexion de paire de langues au moment de la traduction sans avoir à saisir de nouveau tous les paramètres de connexion.

Une connexion de paire de langues identifie les langue source et langue de traduction ainsi que les détails de la connexion URL vers le serveur. Par exemple, *Chinois - Anglais* signifie que le texte source est en chinois et que la traduction effectuée est en anglais. Vous devez définir manuellement chacune des connexions auxquelles vous accéderez grâce aux services en ligne SDL.

Important ! Si vous essayez de récupérer des informations sur le Web avec un serveur proxy, vous devez activer ce serveur dans le fichier `net.properties` pour le serveur et le client IBM® SPSS® Modeler Text Analytics . Suivez les instuctions détaillées dans ce fichier. Cela s'applique lorsque vous accédez au Web avec le noeud Fil de nouvelles ou lorsque vous récupérez une licence SDL Software as a Service (SaaS) car ces connexions utilisent Java. Pour le client, l'emplacement du fichier est *C:\Program Files\IBM\SPSS\Modeler\15\jre\lib\net.properties* par défaut. Pour le client, l'emplacement du fichier est *C:\Program Files\IBM\SPSS\Modeler\15\ext\bin\spss.TMWBServer\jre\lib\net.properties* par défaut.

Figure 5-2
Boîte de dialogue Paramètres de traduction



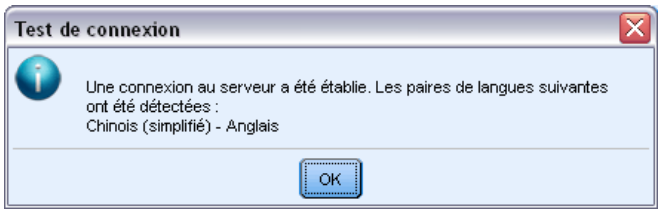
Connexion URL. Entrez l'URL de la connexion SDL Software as a Service.

ID utilisateur. Entrez l'ID unique qui vous a été fourni par SDL.

Clé API. Entrez la clé qui vous a été fournie par SDL.

Test. Cliquez sur Tester pour vérifier que la connexion est configurée correctement et pour afficher les paires de langues trouvées sur cette connexion.

Figure 5-3
Message de connexion réussie



Utilisation du noeud Traduire

Pour extraire des concepts à partir de langues de traduction prises en charge, telles que Arabe, Chinois ou Persan, ajoutez simplement un noeud Traduire avant un noeud de Text Mining dans votre flux.

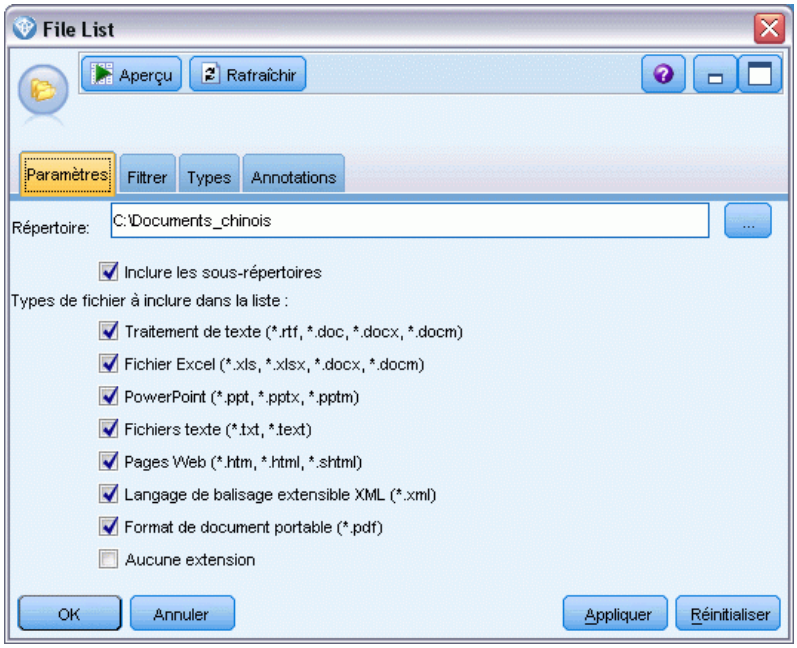
Exemple : Traduction du texte de documents externes

Si le texte à traduire fait partie d'un ou de plusieurs fichiers externes, un noeud Liste fichiers peut être utilisé pour lire une liste de noms. Dans ce cas, le noeud Traduire est ajouté entre le noeud Liste fichiers et les noeuds Text Mining suivants ; la sortie correspond alors à l'emplacement dans lequel réside le texte traduit.

Figure 5-4
Exemple de flux : Noeud Liste fichiers et noeud Traduire

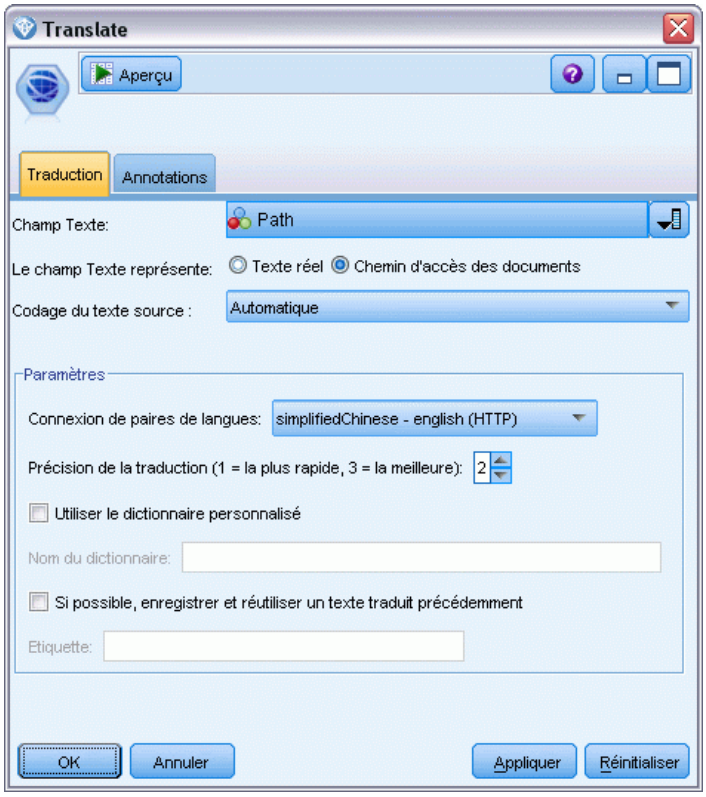
- **Noeud Liste fichiers (onglet Paramètres).** Dans le noeud Liste fichiers, nous avons sélectionné les fichiers source.

Figure 5-5
Boîte de dialogue du noeud Liste fichiers : onglet Paramètres



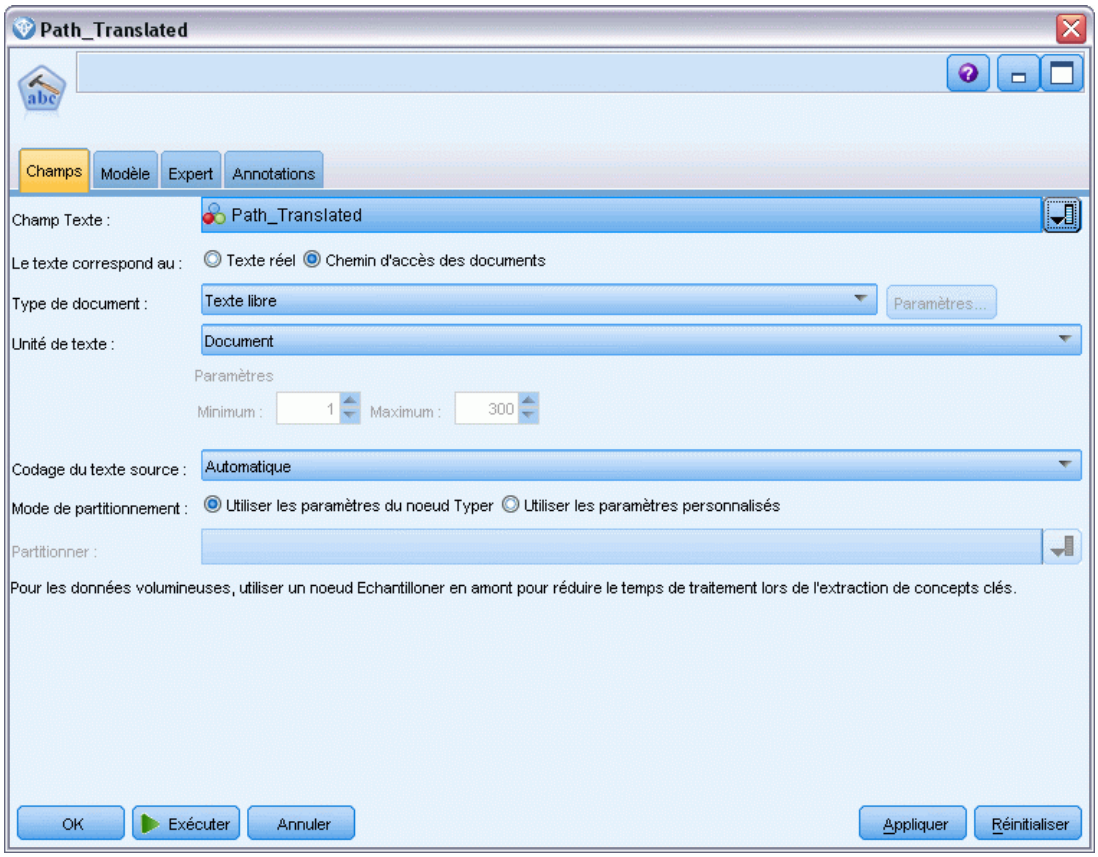
- **Noeud Traduire (onglet Traduction).** Nous avons ensuite ajouté et connecté un noeud Traduire. Dans le noeud, nous avons sélectionné le champ généré par le noeud Liste fichiers (nommé par défaut *Path*), qui définit l'emplacement d'origine des fichiers. Dans le même onglet, nous sélectionnons une connexion de paire de langues prédéfinie.

Figure 5-6
Boîte de dialogue du noeud Traduire : Onglet Traduction



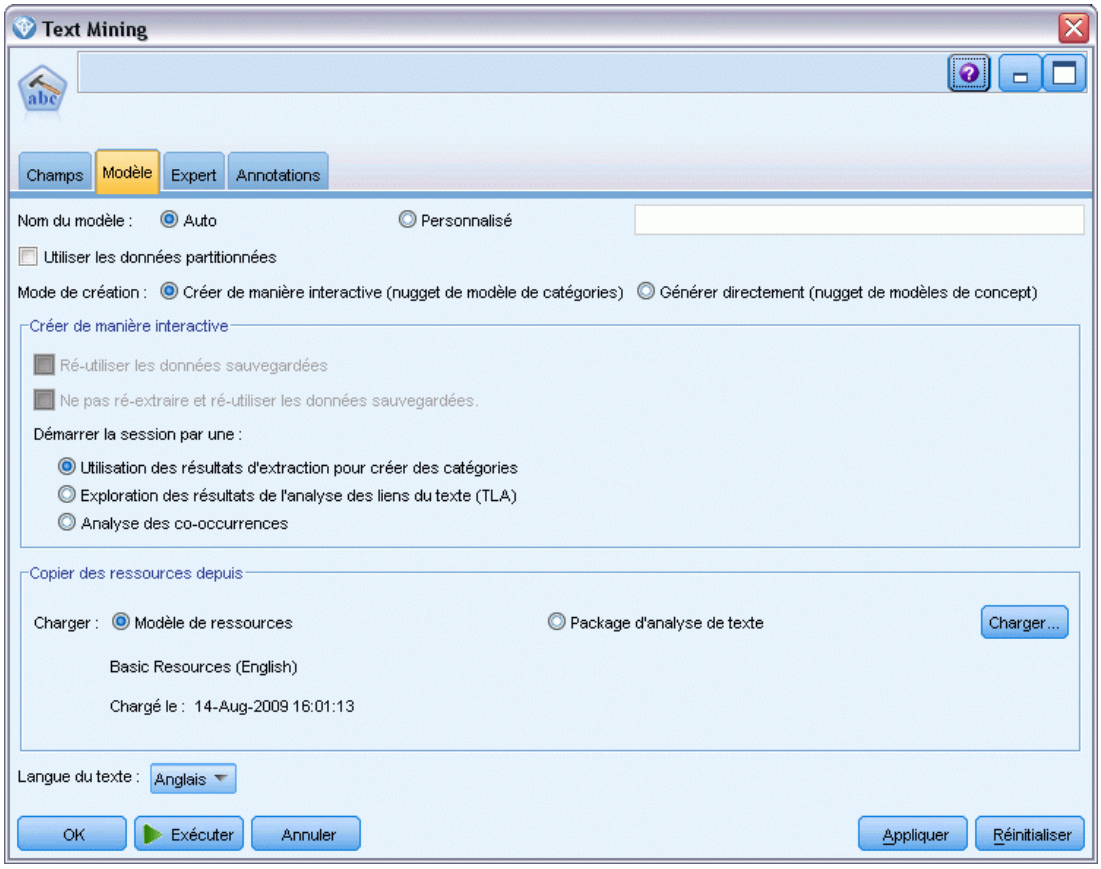
- **Noeud de Text Mining (onglet Champs).** Dans les noeuds Text Mining suivants, nous avons sélectionné le nouveau champ généré par le noeud Traduire, nommé d'après le champ de texte issu du noeud Liste fichiers et suivi de *_Translated*, qui définit l'emplacement des fichiers traduits.

Figure 5-7
Boîte de dialogue du noeud de modélisation Text Mining : onglet Champs



- **Noeud de Text Mining (onglet Modèle).** Dans l'onglet Modèle, nous avons sélectionné Anglais comme langue.

Figure 5-8
Boîte de dialogue du noeud de Text Mining : onglet Modèle



Chapitre
6

Navigation dans le texte source externe

Noeud Afficheur de fichiers

Lorsque vous explorez une collection de documents, vous pouvez spécifier le nom de chemin complet du fichier directement dans vos noeuds de modélisation Text Mining et Traduire. Cependant, lors de la sortie d'un noeud Table, vous ne pouvez afficher que le nom de chemin complet d'un document au lieu du texte qu'il contient. Le noeud Afficheur de fichiers peut être utilisé comme un analogue du noeud Table et vous permet d'accéder au texte réel de chaque document sans avoir à les fusionner en un fichier unique.

Le nœud Afficheur de fichiers peut vous permettre de mieux comprendre les résultats issus de l'extraction de texte ; en effet, il vous fournit un accès au texte source, ou au texte non traduit, dont ont été extraits les concepts, et qui serait autrement inaccessible dans le flux. Ce noeud est ajouté au flux après un noeud Liste fichiers afin d'obtenir une liste des liens vers l'ensemble des fichiers.

Ce noeud donne comme résultat une fenêtre affichant tous les documents qui ont été lus et utilisés pour extraire les concepts. L'une des icônes de la barre d'outils de cette fenêtre vous permet de lancer le rapport dans un navigateur externe répertoriant le nom des documents sous forme de liens hypertexte. Cliquez sur un lien pour ouvrir le document correspondant. Pour plus d'informations, reportez-vous à la section Utilisation du noeud Afficheur de fichiers sur p. 103.

Vous pouvez trouver ce noeud dans l'onglet IBM® SPSS® Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM® SPSS® Modeler. Pour plus d'informations, reportez-vous à la section IBM SPSS Modeler Text Analytics Noeuds dans le chapitre 1 sur p. 11.

Remarque : Lorsque vous utilisez le mode client-serveur et que les nœuds Afficheur de fichiers sont intégrés au flux, les ensembles de documents doivent être stockés dans un répertoire de serveur Web sur le serveur. Le nœud de sortie Text Mining génère la liste des documents stockés dans le répertoire du serveur Web ; les paramètres de sécurité du serveur Web gèrent donc les droits d'accès à ces documents.

Paramètres du noeud Afficheur de fichiers

La boîte de dialogue ci-dessous permet de définir les paramètres du noeud Afficheur de fichiers.

Figure 6-1
Boîte de dialogue du nœud Afficheur de fichiers : onglet Paramètres

Champ de document. Sélectionnez le champ de vos données qui contient le nom et le chemin d'accès complets des documents à afficher.

Titre de la page HTML générée. Créez un titre qui doit apparaître en haut de la page contenant la liste des documents.

Utilisation du noeud Afficheur de fichiers

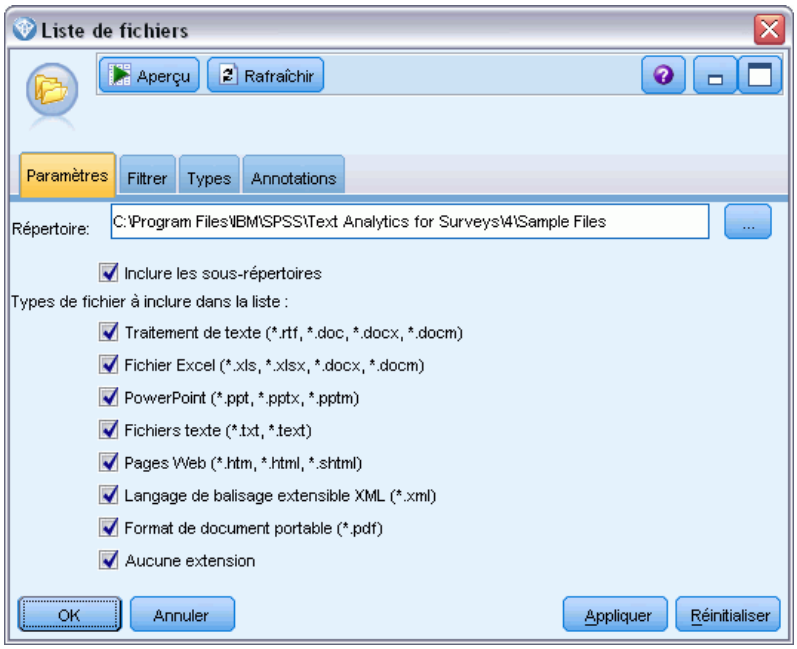
L'exemple suivant indique le mode d'utilisation du noeud Afficheur de fichiers.

Exemple : Noeud Liste fichiers et noeud Afficheur de fichiers

Figure 6-2
Flux illustrant l'utilisation d'un noeud Afficheur de fichiers

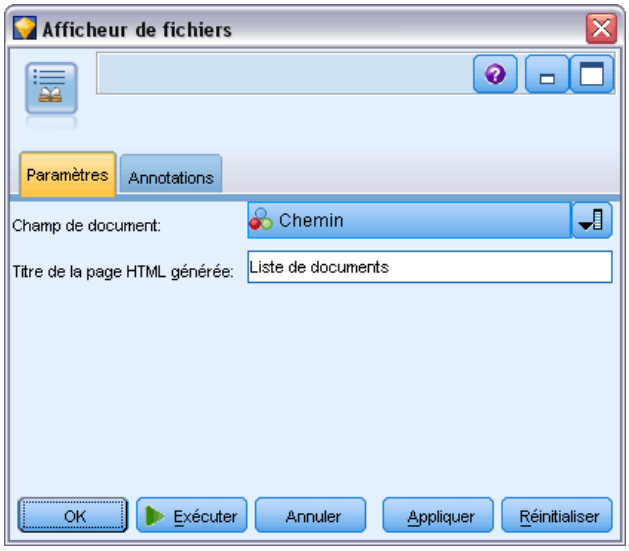
- **Noeud Liste fichiers (onglet Paramètres).** Nous avons tout d'abord ajouté ce nœud afin d'indiquer l'emplacement des documents.

Figure 6-3
Boîte de dialogue du noeud Liste fichiers : onglet Paramètres



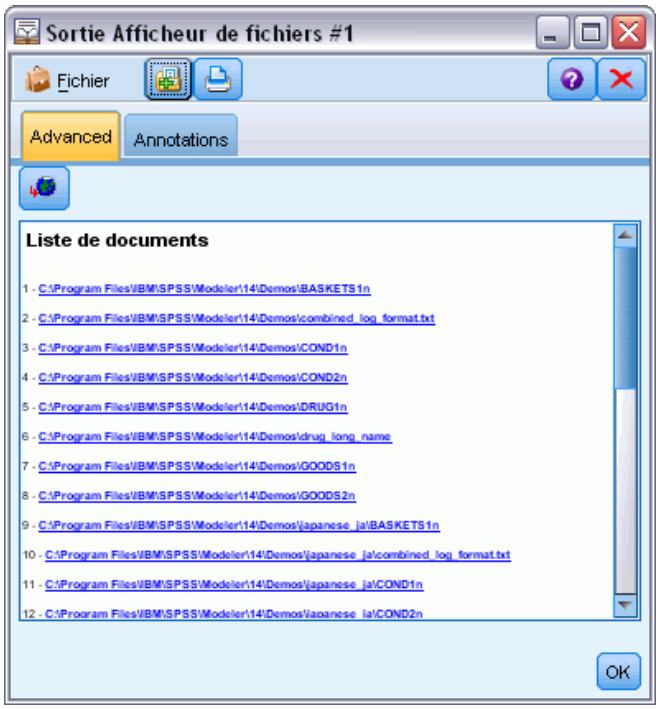
- **Nœud Afficheur de fichiers (onglet Paramètres).** Nous avons ensuite relié le nœud Afficheur de fichiers pour produire une liste HTML des documents.

Figure 6-4
Boîte de dialogue du nœud Afficheur de fichiers : onglet Paramètres



- **Boîte de dialogue Sortie Afficheur de fichiers.** Nous avons exécuté le flux qui sort la liste de documents dans une nouvelle fenêtre.

Figure 6-5
Sortie Afficheur de fichiers



- Pour voir les documents, nous avons cliqué sur le bouton de la barre d'outils représentant un globe avec une flèche rouge. Une liste de liens hypertexte de document est alors apparue dans notre navigateur.

Chapitre
7

Propriétés des noeuds pour la génération de scripts

IBM® SPSS® Modeler dispose d'un langage de génération de scripts qui vous permet d'exécuter des flux à partir de la ligne de commande. Ici, vous pouvez en savoir plus sur les propriétés des noeuds spécifiques à chacun des noeuds fournis avec IBM® SPSS® Modeler Text Analytics . Pour plus d'informations sur l'ensemble standard des noeuds fournis avec SPSS Modeler, reportez-vous au Guide de génération des scripts et d'automatisation.

Nœud Liste fichiers : filelistnode

Vous pouvez utiliser les propriétés du tableau ci-dessous pour la génération de scripts. Le noeud lui-même est appelé filelistnode.

Table 7-1
propriétés de génération de scripts du noeud Liste fichiers

Propriétés de génération de scripts	Type de données
path	*chaîne*
recurse	*booléen*
word_processing	*booléen*
excel_file	*booléen*
powerpoint_file	*booléen*
text_file	*booléen*
web_page	*booléen*
xml_file	*booléen*
pdf_file	*booléen*
no_extension	*booléen*

Remarque : De plus, le paramètre « Créer une liste » n'est plus disponible et les scripts contenant cette option seront automatiquement convertis dans une sortie « Fichiers ».

Nœud Fil de nouvelles : webfeednode

Vous pouvez utiliser les propriétés du tableau ci-dessous pour la génération de scripts. Le noeud lui-même est appelé webfeednode.

Table 7-2
Propriétés de génération de scripts du noeud Fil de nouvelles

Propriétés de génération de scripts	Type de données	Description de la propriété
urls	*chaîne1* *chaîne2* *...chaînen*	Chaque URL est spécifiée dans la structure de la liste. Liste d'URL séparée par "\n"
recent_entries	*booléen*	
limit_entries	*entier*	Nombre des entrées les plus récentes à lire (par URL).
use_previous	*booléen*	Pour enregistrer et réutiliser le cache des fils de nouvelles.
use_previous_label	*chaîne*	Nom du cache des fils de nouvelles enregistré.
start_record	*chaîne*	Balise de début non RSS.
url*n*.title	*chaîne*	Pour chaque URL de la liste, vous devez également définir une propriété ici. La première propriété sera url1.title, où le chiffre correspond à sa position dans la liste des URL. Ceci est la balise de début contenant le titre du contenu.
url*n*.short_description	*chaîne*	Comme pour url*n*.title.
url*n*.description	*chaîne*	Comme pour url*n*.title.
url*n*.authors	*chaîne*	Comme pour url*n*.title.
url*n*.contributors	*chaîne*	Comme pour url*n*.title.
url*n*.published_date	*chaîne*	Comme pour url*n*.title.
url*n*.modified_date	*chaîne*	Comme pour url*n*.title.
html_alg	None HTMLCleaner	Méthode de filtrage du contenu.
discard_lines	*booléen*	Ignorer les lignes courtes. Utilisé avec min_words
min_words	*entier*	Nombre minimal de mots.
discard_words	*booléen*	Ignorer les lignes courtes. Utilisé avec min_avg_len
min_avg_len	*entier*	
discard_scw	*booléen*	Ignorer les lignes contenant de nombreux mots à caractère unique. Utilisé avec max_scw
max_scw	*entier*	Proportion maximum en pourcentage de mots à caractère unique dans une ligne
discard_tags	*booléen*	Ignorer les lignes contenant des balises spécifiques.
tags	*chaîne*	Les caractères spéciaux doivent être ignorés avec une barre oblique inversée \.
discard_spec_words	*booléen*	Ignorer les lignes contenant des chaînes spécifiques
words	*chaîne*	Les caractères spéciaux doivent être ignorés avec une barre oblique inversée \.

Nœud de Text Mining : TextMiningWorkbench

Vous pouvez utiliser les paramètres suivants pour définir ou mettre à jour un noeud via la génération de scripts. Le noeud lui-même est appelé TextMiningWorkbench.

Important ! Il est impossible d'indiquer un autre modèle de ressources via la génération de scripts. Si vous pensez avoir besoin d'un modèle, vous devez le sélectionner dans la boîte de dialogue Noeud.

Table 7-3
propriétés de génération de scripts de noeuds de modélisation Text Mining

Propriétés de génération de scripts	Type de données	Description de la propriété
texte	*champ*	
method	ReadText ReadPath	
docType	*entier*	Avec des valeurs possibles (0,1,2), où 0 = Texte libre, 1 = Texte structuré et 2 = XML
encoding	Automatique "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Les valeurs contenant des caractères spéciaux (du type "UTF-8") doivent être mises entre guillemets afin de ne pas être confondues avec un opérateur mathématique.
unity	*entier*	Avec des valeurs possibles (0,1), où 0 = Paragraphe, 1 = Document
para_min	*entier*	
para_max	*entier*	
mtag	*chaîne*	Contient tous les paramètres mtag (de la boîte de dialogue Paramètres pour les fichiers XML).
mclef	*chaîne*	Contient tous les paramètres mclef (de la boîte de dialogue Paramètres pour les fichiers texte structuré).
partition	*champ*	
custom_field	*booléen*	Indique si un champ de partitionnement sera spécifié ou non.
use_model_name	*booléen*	
model_name	*chaîne*	
use_partitioned_data	*booléen*	Si un champ de partition est défini, seules les données d'apprentissage sont utilisées pour la création du modèle.
model_output_type	Interactive Model	Interactive génère un modèle de catégories. Model génère un modèle de concepts.
use_interactive_info	*booléen*	Pour la création interactive dans une session uniquement.
reuse_extraction_results	*booléen*	Pour la création interactive dans une session uniquement.
interactive_view	Categories TLA Clusters	Pour la création interactive dans une session uniquement.
extract_top	*entier*	Ce paramètre est utilisé lorsque model_type = Concept
use_check_top	*booléen*	

Propriétés de génération de scripts	Type de données	Description de la propriété
check_top	*entier*	
use_uncheck_top	*booléen*	
uncheck_top	*entier*	
language	de en es fr it ja nl pt	
frequency_limit	*entier*	Abandonné dans la version 14.0.
concept_count_limit	*entier*	Limiter l'extraction aux concepts ayant une fréquence globale supérieure à cette valeur. Ne s'applique pas pour le texte en japonais
fix_punctuation	*booléen*	Ne s'applique pas pour le texte en japonais
fix_spelling	*booléen*	Ne s'applique pas pour le texte en japonais
spelling_limit	*entier*	Ne s'applique pas pour le texte en japonais
extract_uniterm	*booléen*	Ne s'applique pas pour le texte en japonais
extract_nonlinguistic	*booléen*	Ne s'applique pas pour le texte en japonais
upper_case	*booléen*	Ne s'applique pas pour le texte en japonais
group_names	*booléen*	Ne s'applique pas pour le texte en japonais
permutation	*entier*	Nombre maximum de mots pleins soumis à une permutation pour le regroupement (la valeur par défaut est 3). Ne s'applique pas pour le texte en japonais.
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	Pour l'extraction de texte en japonais uniquement. *Remarque* : disponible dans IBM® SPSS® Modeler Premium. 0 = Extraction secondaire de sentiments 1 = Extraction des dépendances 2 = Aucun analyseur secondaire défini.
jp_algorithm_sense_mode	0 1 2	Pour l'extraction de texte en japonais uniquement. *Remarque* : Disponible dans SPSS Modeler Premium. 0 = Uniquement les conclusions 2 = Sentiment(s) représentatif(s) uniquement 3 = Tous les sentiments.

Nugget de modèle Text Mining : TMWBModelApplier

Vous pouvez utiliser les propriétés du tableau ci-dessous pour la génération de scripts. Le nugget lui-même est appelé TMWBModelApplier.

Table 7-4
Propriétés de nugget de modèle Text Mining

Propriétés de génération de scripts	Type de données	Description de la propriété
scoring_mode	Fields Records	
field_values	Flags Counts	Cette option n'est pas disponible dans les nuggets de modèles de catégories. Pour Flags, définissez sur TRUE ou FALSE
true_value	*chaîne*	Avec les Flags, définissez la valeur true.
false_value	*chaîne*	Avec les Flags, définissez la valeur false.
extension_concept	*chaîne*	Spécifiez l'extension du nom de champ. Les noms de champ générés reprennent le nom de concept et l'extension. Spécifiez où placer cette extension à l'aide de la valeur add_as.
extension_category	*chaîne*	Extension nom de champ. Vous pouvez choisir de spécifier une extension préfixe/suffixe pour le nom de champ ou vous pouvez choisir d'utiliser les codes de catégories. Les noms de champ générés reprennent le nom de catégorie et l'extension. Spécifiez où placer cette extension à l'aide de la valeur add_as.
add_as	Suffix Prefix	
fix_punctuation	*booléen*	
excluded_subcategories_descriptors	RollUpToParent Ignore	Pour les modèles de catégories uniquement. Si une sous-catégorie n'est pas sélectionnée. Cette option permet de déterminer comment seront manipulés les descripteurs appartenant aux sous-catégories qui n'étaient pas sélectionnées pour le scoring. Il y a deux options. ■ Ignore. Avec l'option Exclure complètement ses descripteurs du scoring, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront ignorés et ne seront pas utilisés pendant le scoring. ■ RollUpToParent. Avec l'option Agréger les descripteurs avec ceux des catégories parents, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront utilisés comme descripteurs pour la catégorie parent (la catégorie au-dessus de cette sous-catégorie). S'il existe plusieurs niveaux de sous-catégories et que celles-ci ne sont pas sélectionnées, les descripteurs seront reportés sous la première catégorie parent disponible.
check_model	*booléen*	Abandonné dans la version 14
text	*champ*	
method	ReadText ReadPath	

Propriétés de génération de scripts	Type de données	Description de la propriété
docType	*entier*	Avec les valeurs possibles (0,1,2), où 0 = Full Text, 1 = Structured Text et 2 = XML
encoding	Automatique "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Les valeurs contenant des caractères spéciaux (du type "UTF-8") doivent être mises entre guillemets afin de ne pas être confondues avec un opérateur mathématique.
language	de en es fr it ja nl pt	

Nœud Analyse des liens du texte : textlinkanalysis

Vous pouvez utiliser les paramètres du tableau suivant pour définir ou mettre à jour un noeud via la génération de scripts. Le noeud lui-même est appelé textlinkanalysis.

Important ! il est impossible de spécifier un modèle de ressources via la génération de scripts. Pour sélectionner un modèle, utilisez la boîte de dialogue du noeud.

Table 7-5
Propriétés du noeud Analyse des liens du texte (TLA)

Propriétés de génération de scripts	Type de données	Description de la propriété
id_field	*champ*	
texte	*champ*	
method	ReadText ReadPath	
docType	*entier*	Avec des valeurs possibles (0,1,2), où 0 = Texte libre, 1 = Texte structuré et 2 = XML
encoding	Automatique "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Les valeurs contenant des caractères spéciaux (du type "UTF-8") doivent être mises entre guillemets afin de ne pas être confondues avec un opérateur mathématique.
unity	*entier*	Avec des valeurs possibles (0,1), où 0 = Paragraphe, 1 = Document
para_min	*entier*	
para_max	*entier*	

Propriétés de génération de scripts	Type de données	Description de la propriété
mtag	*chaîne*	Contient tous les paramètres mtag (de la boîte de dialogue Paramètres pour les fichiers XML).
mclef	*chaîne*	Contient tous les paramètres mclef (de la boîte de dialogue Paramètres pour les fichiers texte structuré).
language	de en es fr it ja nl pt	
concept_count_limit	*entier*	Limiter l'extraction aux concepts ayant une fréquence globale supérieure à cette valeur. Ne s'applique pas pour le texte en japonais
fix_punctuation	*booléen*	Ne s'applique pas pour le texte en japonais
fix_spelling	*booléen*	Ne s'applique pas pour le texte en japonais
spelling_limit	*entier*	Ne s'applique pas pour le texte en japonais
extract_uniterm	*booléen*	Ne s'applique pas pour le texte en japonais
extract_nonlinguistic	*booléen*	Ne s'applique pas pour le texte en japonais
upper_case	*booléen*	Ne s'applique pas pour le texte en japonais
group_names	*booléen*	Ne s'applique pas pour le texte en japonais
permutation	*entier*	Nombre maximum de mots pleins soumis à une permutation pour le regroupement (la valeur par défaut est 3). Ne s'applique pas pour le texte en japonais.
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	Pour l'extraction de texte en japonais uniquement. *Remarque* : disponible dans IBM® SPSS® Modeler Premium. 0 = Extraction secondaire de sentiments 1 = Extraction des dépendances 2 = Aucun analyseur secondaire défini.
jp_algorithm_sense_mode	0 1 2	Pour l'extraction de texte en japonais uniquement. *Remarque* : Disponible dans SPSS Modeler Premium. 0 = Uniquement les conclusions 2 = Sentiment(s) représentatif(s) uniquement 3 = Tous les sentiments.

Noeud Traduire : translatenode

Vous pouvez utiliser les propriétés du tableau ci-dessous pour la génération de scripts. Le noeud lui-même est appelé translatenode.

Table 7-6
Propriétés du noeud Traduire

Propriétés de génération de scripts	Type de données	Description de la propriété
text	*champ*	
method	ReadText ReadPath	
encoding	Automatic "Big5", "Big5-HKSCS" , "UTF-8", "UTF-16", "US-ASCII", "Latin1", "CP850", "CP874", "CP1250", "CP1251", "CP1252", "CP1253", "CP1254","CP1255", "CP1256","CP1257", "CP1258","GB18030", "GB2312","GBK", "eucJP","JIS7", "SHIFT_JIS", "eucKR","TSCII", "ucs2","KOI8-R", "KOI8-U","ISO8859-1", "ISO8859-2","ISO8859-3", "ISO8859-4","ISO8859-5", "ISO8859-6","ISO8859-7", "ISO8859-8", "ISO8859-8-i", "ISO8859-9","ISO8859-10", "ISO8859-13","ISO8859-14", "ISO8859-15","IBM 850", "IBM 866","Apple Roman", "TIS-620"	Les valeurs contenant des caractères spéciaux (du type "UTF-8") doivent être mises entre guillemets afin de ne pas être confondues avec un opérateur mathématique.
lw_server_type	LOC WAN HTTP	
lw_hostname	*chaîne*	
lw_port	*entier*	
url	*chaîne*	url du serveur de traduction
apiKey	*chaîne*	
user_id	*chaîne*	
lpid	*entier*	N'est pas utilisé si *language_from* ou *language_from_id* est défini.
translate_from	Arabic, Chinese, Traditional Chinese, Czech, Danish,Dutch, English,French, German, Greek, Hindi, Hungarian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Somali, Swedish	

Propriétés de génération de scripts	Type de données	Description de la propriété
translate_from_id	ara, chi, cht, cze, dan, dut, eng, fra, ger, gre, hin, hun, ita, jpn, kor, per, pol, por, rum, rus, som, spa, swe	
translate_to	English	
translate_to_id	eng	
translation_accuracy	*entier*	Indique le niveau de précision désiré pour le processus de traduction : choisir une valeur de 1 à 3
use_previous_translation	*booléen*	Indique que les résultats de traduction ont déjà été obtenus lors d'une exécution précédente et peuvent être réutilisés.
translation_label	*chaîne*	Entrez une étiquette pour identifier les résultats de traduction en vue de leur réutilisation.

Partie II: Session interactive

Chapitre
8

Mode Session interactive

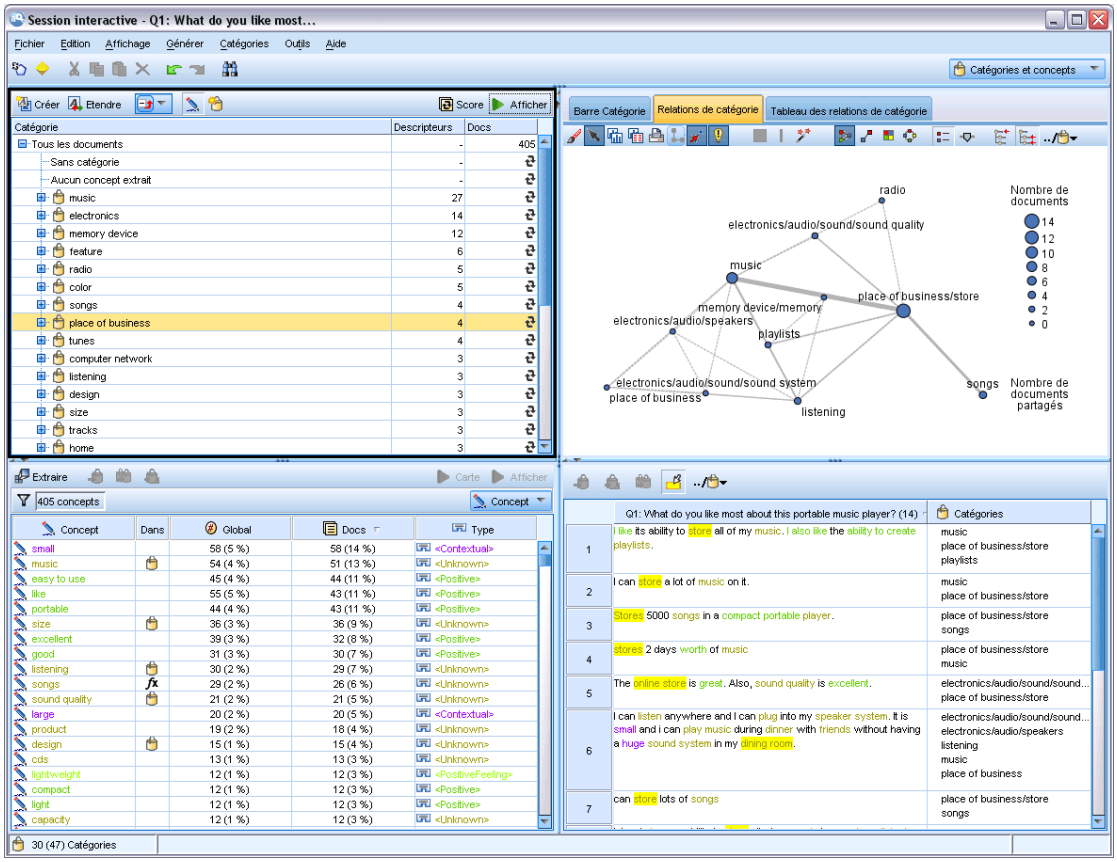
À partir d'un nœud de modélisation de text mining, vous pouvez choisir de lancer une session interactive au cours de l'exécution du flux. Dans cette session, vous pouvez extraire des concepts clés de vos données textuelles, créer des catégories et explorer des clusters et des patrons d'analyse des liens du texte, et vous pouvez également générer des modèles de catégorie. Ce chapitre présente l'interface de la session (niveau supérieur), ainsi que les principaux éléments que vous serez amenés à utiliser, notamment :

- **Résultats d'extraction.** Une fois l'extraction effectuée, les résultats extraits correspondent aux principaux mots et groupes de mots identifiés et extraits des données textuelles, également appelés *concepts*. Ces concepts sont regroupés en *types*. En utilisant ces concepts et types, vous pouvez explorer vos données et créer des catégories. Vous pouvez les gérer dans la vue **Catégories et concepts**.
- **Catégories.** Vous pouvez utiliser des descripteurs (résultats d'extraction, patrons et règles, par exemple) en tant que définition pour créer manuellement ou automatiquement un ensemble de catégories auxquelles des documents et des enregistrements sont affectés selon qu'ils contiennent ou non une partie de la définition de catégorie. Vous pouvez les gérer dans la vue **Catégories et concepts**.
- **Clusters.** Les *Clusters* représentent un regroupement de concepts. Des liens indiquant l'existence d'une relation entre ces concepts ont été établis. Ces concepts sont regroupés à l'aide d'un algorithme complexe qui s'appuie notamment sur la fréquence à laquelle deux concepts apparaissent ensemble par rapport à la fréquence à laquelle ils apparaissent séparément. Vous pouvez les gérer dans la vue **Clusters**. Vous pouvez également ajouter les concepts qui constituent un cluster à des catégories.
- **Patrons Analyse des liens du texte.** Si vous avez des règles de patrons d'analyse des liens du texte dans les ressources linguistiques ou si vous utilisez un modèle de ressources qui possède déjà certaines règles de TLA, vous pouvez alors extraire des patrons des données textuelles. Ces patrons peuvent vous permettre de découvrir des relations intéressantes entre les concepts figurant dans vos données. Vous pouvez également utiliser des patrons comme descripteurs dans vos catégories. Vous pouvez les gérer dans la vue **Analyse des liens du texte**. Pour les textes en japonais, vous devez sélectionner une analyse secondaire et activer l'extraction TLA. *Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.
- **Ressources linguistiques.** Le processus d'extraction s'appuie sur un ensemble de paramètres et de définitions linguistiques pour gérer la façon dont le texte est extrait et géré. Vous pouvez gérer ces paramètres et définitions sous la forme de modèles et de bibliothèques dans la vue **Editeur de ressources**.

Vue Catégories et concepts

L'interface de l'application comporte plusieurs vues. La vue Catégories et concepts correspond à la fenêtre dans laquelle vous pouvez créer et explorer des catégories, mais également explorer et modifier les résultats d'extraction. Le terme **Catégories** désignent un groupe d'idées et de patrons étroitement liés auxquels des documents et des enregistrements sont affectés lors d'un processus de scoring. Alors que les **concepts** se rapportent au niveau le plus basique des résultats d'extraction disponibles à utiliser en tant que blocs de construction pour vos catégories, appelés les descripteurs.

Figure 8-1
Vue Catégories et concepts



La vue Catégories et concepts comporte quatre panneaux ; vous pouvez masquer ou afficher chacun d'entre eux en sélectionnant son nom dans le menu Affichage. Pour plus d'informations, reportez-vous à la section Catégorisation des données textuelles dans le chapitre 10 sur p. 162.

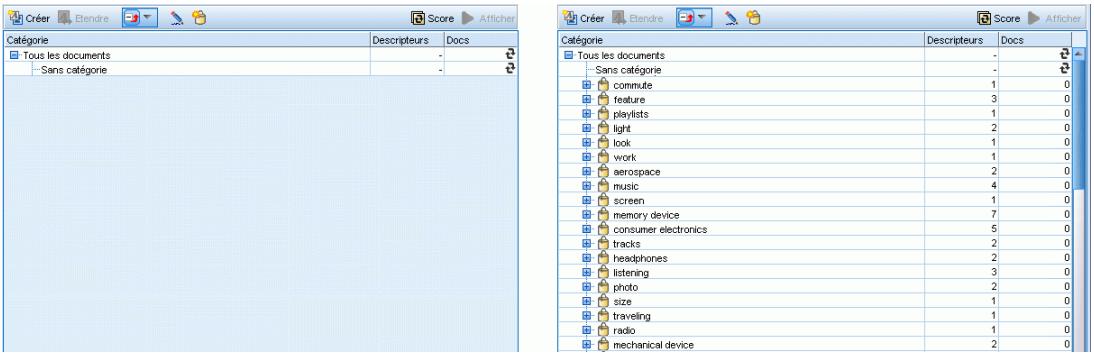
Panneau Catégories

Cette zone, située dans l'angle supérieur gauche, représente un tableau dans lequel vous pouvez gérer les catégories que vous créez. Une fois les concepts et types extraits de vos données textuelles, vous pouvez commencer à créer des catégories en appliquant des techniques (réseaux sémantiques et inclusion de concepts, par exemple) ou en procédant de façon manuelle. Lorsque

vous double-cliquez sur le nom d'une catégorie, la boîte de dialogue Définitions de catégorie s'ouvre. Tous les descripteurs qui composent la définition de cette catégorie (concepts, types, règles) y figurent. Pour plus d'informations, reportez-vous à la section Catégorisation des données textuelles dans le chapitre 10 sur p. 162. Toutes les techniques automatiques ne sont pas disponibles pour toutes les langues.

Lorsque vous sélectionnez une ligne dans ce panneau, vous pouvez afficher les informations concernant les descripteurs ou les documents/enregistrements correspondants dans les panneaux Données et Visualisation.

Figure 8-2
Vue Catégories et concepts : panneau Catégories avec et sans catégories



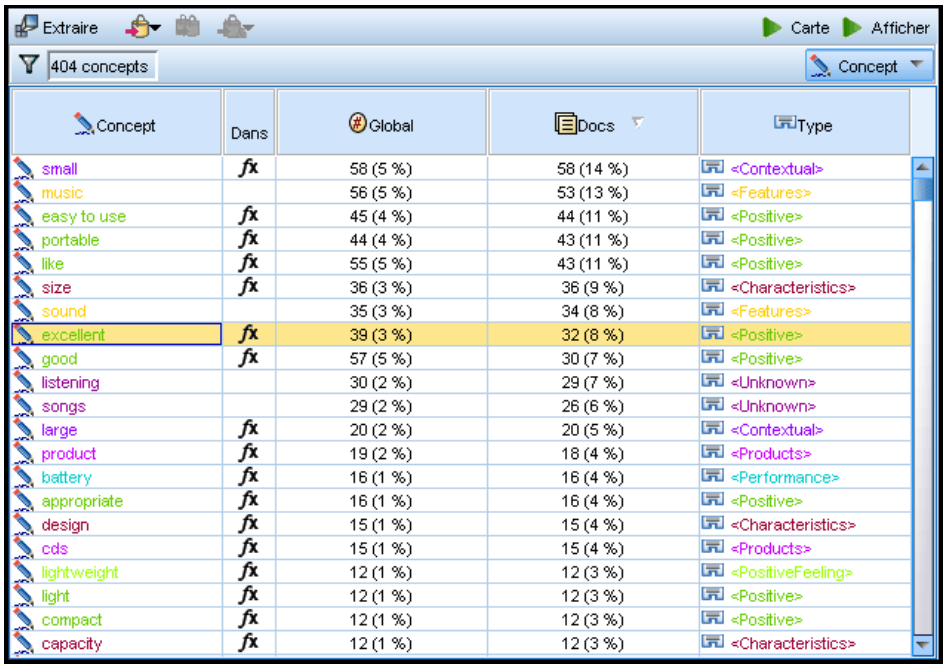
Panneau Résultats d'extraction

Cette zone, située dans l'angle inférieur gauche, présente les résultats de l'extraction. Lorsque vous exécutez une extraction, le moteur du programme d'extraction parcourt les données textuelles, identifie les concepts pertinents et affecte un type à chaque concept. Les **concepts** correspondent à des mots ou groupes de mots extraits à partir des données textuelles. Les **types** représentent des regroupements sémantiques de concepts stockés sous la forme de déclarations de types. Une fois l'extraction terminée, les concepts et les types apparaissent avec un codage couleur dans le panneau Résultats d'extraction. Pour plus d'informations, reportez-vous à la section Résultats d'extraction : concepts et types dans le chapitre 9 sur p. 139.

Vous pouvez voir l'ensemble des termes sous-jacents pour un concept en passant la souris sur le nom du concept. En procédant ainsi, une info-bulle apparaît indiquant le nom du concept et plusieurs lignes de termes qui sont groupés sous ce concept. Ces termes sous-jacents incluent les synonymes définis dans les ressources linguistiques (que ceux-ci aient été rencontrés ou non dans le texte) ainsi que tous les termes extraits au pluriel/singulier, les termes permutés, les termes provenant du regroupement flou, etc. Vous pouvez également copier ces termes ou voir l'ensemble complet des termes sous-jacents en cliquant avec le bouton droit sur le nom du concept et en choisissant l'option du menu contextuel.

Le Text Mining est un processus itératif au cours duquel les résultats de l'extraction sont passés en revue en fonction du contexte des données textuelles. Ils sont ensuite affinés afin de générer de nouveaux résultats avant d'être réévalués. Vous pouvez affiner les résultats de l'extraction en modifiant les ressources linguistiques. Vous pouvez procéder en partie à ce réglage à partir du panneau Résultats d'extraction ou du panneau de données directement, ou bien directement dans la vue Editeur de ressources. Pour plus d'informations, reportez-vous à la section La vue Editeur de ressources sur p. 129.

Figure 8-3
Vue Catégories et concepts : Panneau Résultats d'extraction après une extraction

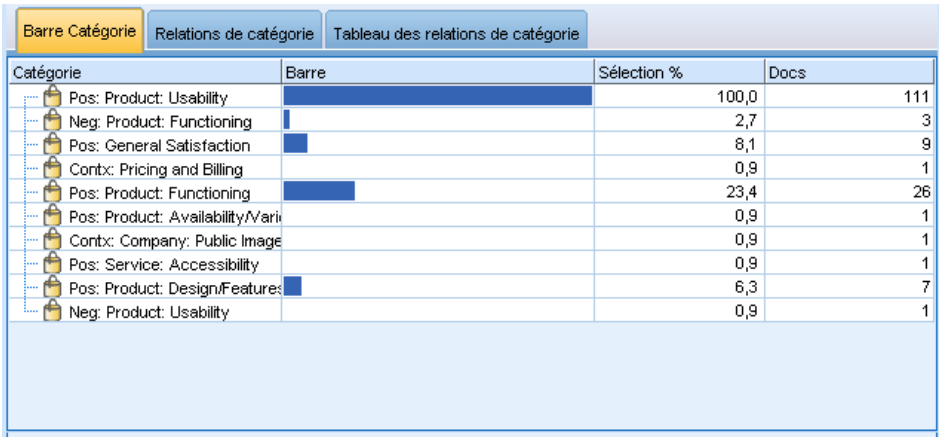


Concept	Dans	Global	Docs	Type
small	fx	58 (5 %)	58 (14 %)	<Contextual>
music		56 (5 %)	53 (13 %)	<Features>
easy to use	fx	45 (4 %)	44 (11 %)	<Positive>
portable	fx	44 (4 %)	43 (11 %)	<Positive>
like	fx	55 (5 %)	43 (11 %)	<Positive>
size	fx	36 (3 %)	36 (9 %)	<Characteristics>
sound		35 (3 %)	34 (8 %)	<Features>
excellent	fx	39 (3 %)	32 (8 %)	<Positive>
good	fx	57 (5 %)	30 (7 %)	<Positive>
listening		30 (2 %)	29 (7 %)	<Unknown>
songs		29 (2 %)	26 (6 %)	<Unknown>
large	fx	20 (2 %)	20 (5 %)	<Contextual>
product	fx	19 (2 %)	18 (4 %)	<Products>
battery	fx	16 (1 %)	16 (4 %)	<Performance>
appropriate	fx	16 (1 %)	16 (4 %)	<Positive>
design	fx	15 (1 %)	15 (4 %)	<Characteristics>
cds	fx	15 (1 %)	15 (4 %)	<Products>
lightweight	fx	12 (1 %)	12 (3 %)	<PositiveFeeling>
light	fx	12 (1 %)	12 (3 %)	<Positive>
compact	fx	12 (1 %)	12 (3 %)	<Positive>
capacity	fx	12 (1 %)	12 (3 %)	<Characteristics>

Panneau Visualisation

Cette zone, située dans l'angle supérieur droit, présente, selon plusieurs perspectives, les éléments communs apparaissant dans la catégorisation des documents/enregistrements. Chaque graphique ou diagramme fournit des informations similaires, mais les présente d'une façon différente ou avec un niveau de détail différent. Vous pouvez vous appuyer sur ces graphiques et diagrammes pour analyser les résultats de la catégorisation, et affiner les catégories ou générer des rapports. Par exemple, un graphique peut révéler des catégories trop similaires (lorsqu'elles ont en commun plus de 75 % de leurs enregistrements, par exemple) ou trop différentes. Le contenu d'un graphique ou d'un diagramme dépend des éléments sélectionnés dans les autres panneaux. Pour plus d'informations, reportez-vous à la section Graphiques et diagrammes de catégorie dans le chapitre 13 sur p. 255.

Figure 8-4
Vue Catégories et concepts : panneau Visualisation



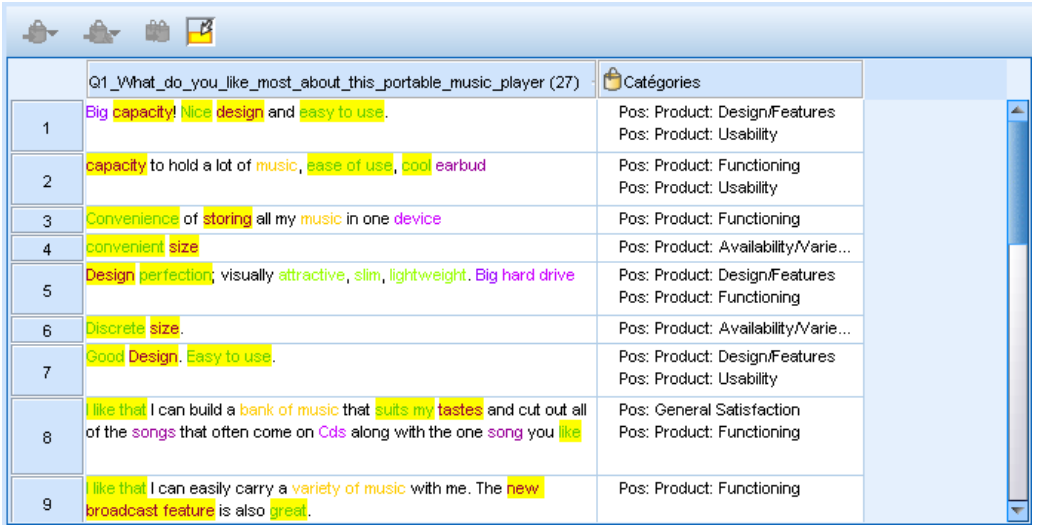
Catégorie	Barre	Sélection %	Docs
Pos: Product: Usability		100,0	111
Neg: Product: Functioning		2,7	3
Pos: General Satisfaction		8,1	9
Contx: Pricing and Billing		0,9	1
Pos: Product: Functioning		23,4	26
Pos: Product: Availability/Vari		0,9	1
Contx: Company: Public Image		0,9	1
Pos: Service: Accessibility		0,9	1
Pos: Product: Design/Features		6,3	7
Neg: Product: Usability		0,9	1

Panneau Données

Le panneau Données est situé dans l'angle inférieur droit. Ce panneau présente un tableau contenant les documents ou les enregistrements correspondant à une sélection dans une autre zone de la vue. En fonction des éléments sélectionnés, seul le texte correspondant apparaît dans le panneau Données. Une fois votre sélection effectuée, cliquez sur un bouton Afficher pour remplir le panneau de données à l'aide du texte correspondant.

Si un autre panneau contient une sélection, les documents ou enregistrements correspondants représentent les concepts mis en surbrillance en couleur pour vous permettre de les repérer plus facilement dans le texte. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher une info-bulle contenant le nom du concept sous lequel il a été extrait et le type auquel il a été affecté. Pour plus d'informations, reportez-vous à la section Le panneau Données dans le chapitre 10 sur p. 174.

Figure 8-5
Vue Catégories et concepts : panneau Données



	Q1_What_do_you_like_most_about_this_portable_music_player (27)	Catégories
1	Big capacity! Nice design and easy to use.	Pos: Product: Design/Features Pos: Product: Usability
2	capacity to hold a lot of music, ease of use, cool earbud	Pos: Product: Functioning Pos: Product: Usability
3	Convenience of storing all my music in one device	Pos: Product: Functioning
4	convenient size	Pos: Product: Availability/Varie...
5	Design perfection; visually attractive, slim, lightweight. Big hard drive	Pos: Product: Design/Features Pos: Product: Functioning
6	Discrete size.	Pos: Product: Availability/Varie...
7	Good Design. Easy to use.	Pos: Product: Design/Features Pos: Product: Usability
8	I like that I can build a bank of music that suits my tastes and cut out all of the songs that often come on Cds along with the one song you like	Pos: General Satisfaction Pos: Product: Functioning
9	I like that I can easily carry a variety of music with me. The new broadcast feature is also great.	Pos: Product: Functioning

Recherche dans la vue Catégories et concepts

Il peut s'avérer nécessaire de localiser rapidement des informations dans une section particulière. Avec la barre d'outils Rechercher, vous pouvez entrer la chaîne à rechercher et définir un autre critère de recherche comme la sensibilité à la casse ou la direction de la recherche. Puis vous pouvez choisir le panneau dans lequel effectuer la recherche.

Figure 8-6
Barre d'outils Rechercher dans la vue Catégories et concepts

Pour utiliser la fonction de recherche

- Dans la vue Catégories et concepts, choisissez Modifier > Rechercher dans les menus. La barre d'outils Rechercher apparaît au-dessus du panneau Catégories et des panneaux Visualisation.
- Saisissez la chaîne de mots que vous recherchez dans la zone de texte. Vous pouvez contrôler la casse, les correspondances partielles et le sens de la recherche à l'aide des boutons de la barre d'outils.
- Dans la barre d'outils, cliquez sur le nom du panneau dans lequel effectuer la recherche. Si une correspondance est détectée, le texte est mis en surbrillance dans la fenêtre.
- Pour rechercher la correspondance suivante, cliquez de nouveau sur le nom du panneau.

La vue Clusters

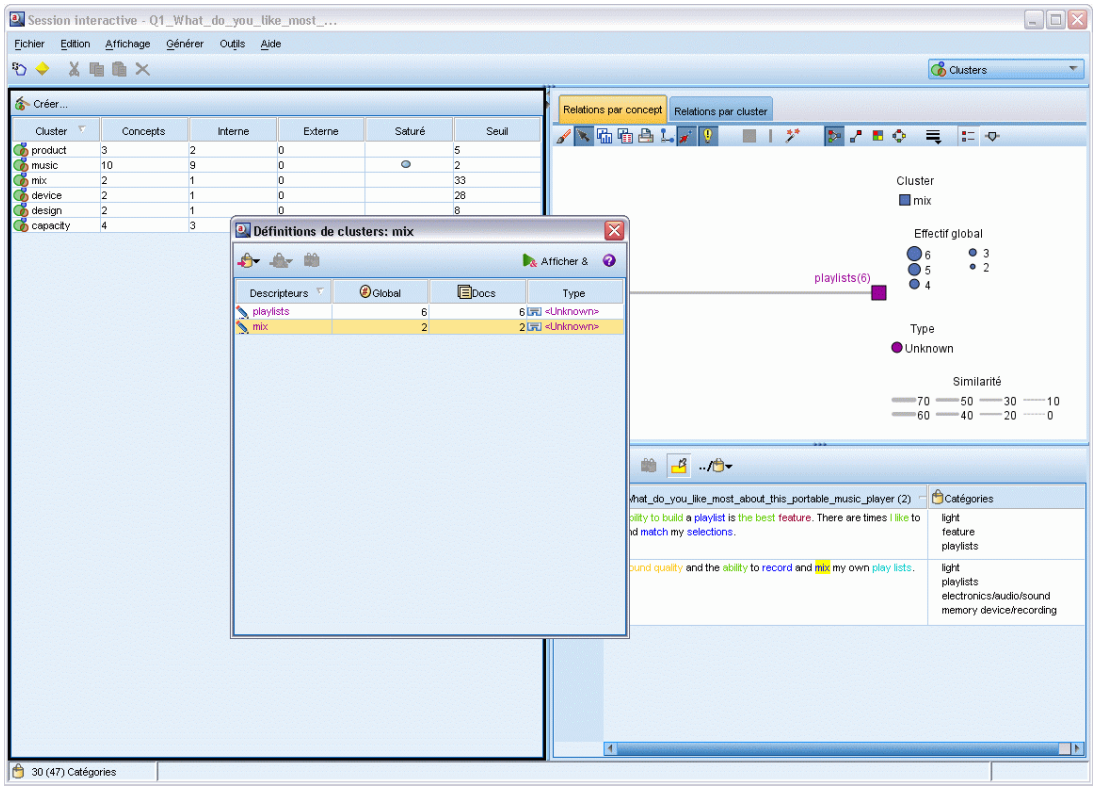
Dans la vue Clusters, vous pouvez créer et explorer les résultats de cluster trouvés dans vos données textuelles. Les **clusters** constituent des regroupements de concepts générés par des algorithmes de classification non supervisée qui reposent sur la fréquence d'occurrence des concepts et sur la fréquence à laquelle ils apparaissent ensemble. Les clusters ont pour objectif de regrouper les concepts apparaissant ensemble, alors que les catégories ont pour objectif de regrouper les documents ou les enregistrements en fonction des correspondances existant entre le texte et les descripteurs (concept, règles, patrons) pour chaque catégorie.

Plus les concepts figurant dans un cluster apparaissent ensemble, moins ils apparaissent avec d'autres concepts et plus le cluster permet d'identifier des relations intéressantes entre les concepts. Deux concepts font l'objet d'une cooccurrence lorsqu'ils apparaissent tous les deux (ou que l'un de leurs synonymes ou termes apparaît) dans le même document ou enregistrement. Pour plus d'informations, reportez-vous à la section Analyse des clusters dans le chapitre 11 sur p. 237.

Vous pouvez créer des clusters et les explorer dans un ensemble de diagrammes et de graphiques susceptibles de révéler les relations existant entre des concepts, découverte qui prendrait autrement beaucoup trop de temps. Alors que vous ne pouvez pas ajouter de clusters entiers à des catégories, vous pouvez ajouter les concepts qui figurent dans un cluster à une catégorie à l'aide de la boîte de dialogue Définitions du cluster. Pour plus d'informations, reportez-vous à la section Définitions du cluster dans le chapitre 11 sur p. 243.

Vous pouvez modifier les paramètres de classification non supervisée de manière à orienter les résultats. Pour plus d'informations, reportez-vous à la section Création de clusters dans le chapitre 11 sur p. 238.

Figure 8-7
vue Clusters



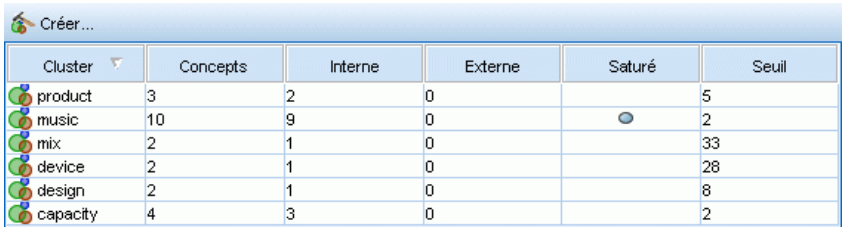
La vue Clusters est organisée en trois panneaux ; vous pouvez masquer ou afficher chacun d'entre eux en sélectionnant son nom dans le menu Affichage. En règle générale, seuls les panneaux Clusters et Visualisation sont affichés.

Panneau Clusters

Ce panneau, situé à gauche, présente les clusters trouvés dans les données textuelles. Pour générer des résultats de classification, cliquez sur le bouton Créer. Les clusters sont générés par un algorithme de classification non supervisée qui tente d'identifier les concepts qui apparaissent fréquemment ensemble.

A chaque nouvelle extraction, les résultats de cluster sont effacés et vous devez générer une nouvelle fois les clusters pour obtenir les tous derniers résultats. Lorsque vous générez les clusters, vous pouvez modifier certains paramètres (nombre maximal de clusters à créer, nombre maximal de concepts pouvant exister ou nombre maximal de liens avec des concepts extérieurs, par exemple). Pour plus d'informations, reportez-vous à la section Exploration des Clusters dans le chapitre 11 sur p. 242.

Figure 8-8
Vue Clusters : panneau Clusters



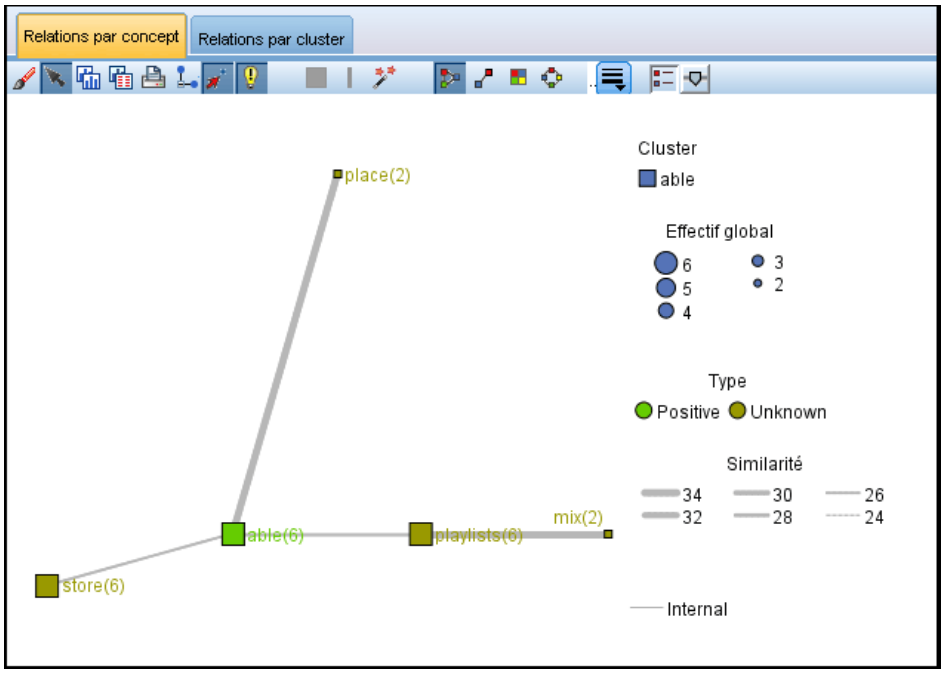
Panneau Visualisation

Ce panneau, qui se trouve dans le coin supérieur droit, offre deux perspectives concernant la classification : un graphique Relations par concept et un graphique Relations par cluster. Si ce panneau est masqué, vous pouvez l'afficher à partir du menu Affichage (Affichage > Visualisation). En fonction des éléments sélectionnés dans le panneau Clusters, vous pouvez afficher les interactions correspondantes entre les clusters ou à l'intérieur des clusters. Les résultats sont présentés sous plusieurs formes :

- **Relations par concept.** Graphique Relations représentant l'ensemble des concepts des clusters sélectionnés, ainsi que les concepts liés en dehors du cluster.
- **Relations par cluster.** Graphique Relations représentant les liens existant entre les clusters sélectionnés et d'autres clusters, ainsi que les liens entre ces autres clusters.

Remarque : Pour pouvoir afficher un graphique Relations par cluster, vous devez avoir créé des clusters présentant des liens externes. Les liens externes sont des liens entre des paires de concepts dans des clusters distincts (un concept au sein d'un cluster et un concept en dehors, dans un autre cluster). Pour plus d'informations, reportez-vous à la section Graphiques Cluster dans le chapitre 13 sur p. 259.

Figure 8-9
Vue Clusters : panneau Visualisation



Panneau Données

Le panneau Données, situé dans l'angle inférieur droit, est masqué par défaut. Vous ne pouvez pas afficher le volet Données depuis le volet Clusters étant donné que ces clusters couvrent plusieurs documents/enregistrements. ce qui rend les résultats sans intérêt. Vous pouvez toutefois consulter les données correspondant à une sélection dans la boîte de dialogue Définitions du cluster. En fonction des éléments sélectionnés dans cette boîte de dialogue, seul le texte correspondant figure dans le panneau Données. Une fois les éléments sélectionnés, cliquez sur le bouton Afficher & pour remplir le panneau de données avec les documents ou les enregistrements contenant l'ensemble des concepts.

Les documents ou enregistrements correspondants indiquent les concepts en surbrillance colorée pour vous aider à les identifier facilement dans le texte. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher le concept sous lequel il a été extrait et le type auquel il a été affecté. Le panneau Données peut comporter plusieurs colonnes ; la colonne correspondant au champ de texte est toujours affichée. Elle porte le nom du champ de texte utilisé lors de l'extraction ou le nom d'un document lorsque les données textuelles figurent dans plusieurs fichiers différents. D'autres colonnes sont disponibles. Pour plus d'informations, reportez-vous à la section Le panneau Données dans le chapitre 10 sur p. 174.

La vue Analyse des liens du texte

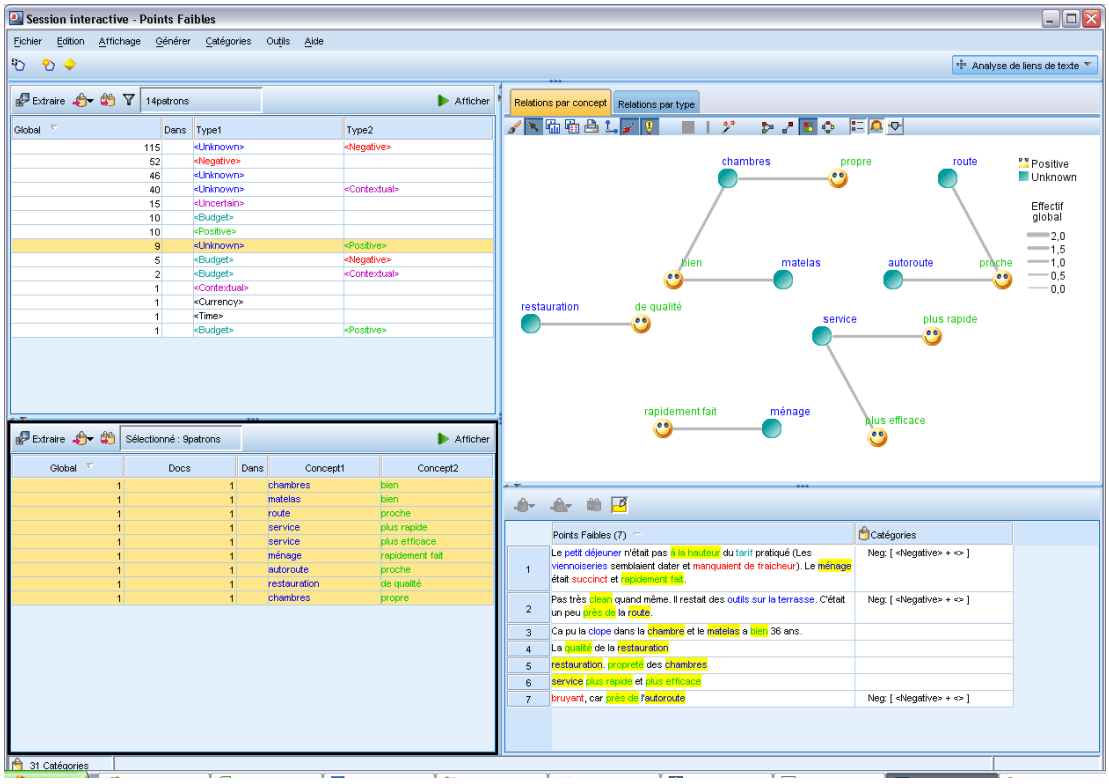
Dans la vue Analyse des liens du texte, vous pouvez créer et explorer les patrons d'analyse des liens du texte trouvés dans vos données textuelles. L'analyse des liens du texte est une technologie de mise en correspondance de patrons qui vous permet de définir des règles de TLA et de les comparer aux concepts extraits et aux relations trouvées dans le texte.

Les patrons s'avèrent particulièrement utiles lorsque vous tentez de découvrir des relations entre des concepts ou des opinions sur un sujet donné. En voici quelques exemples : extraction d'opinions sur des produits à partir de données d'enquête, extraction de relations génomiques à partir de rapports de recherche médicale ou extraction de relations entre des personnes ou des lieux à partir de renseignements.

Une fois les patrons TLA extraits, vous pouvez les explorer dans le panneau Données ou Visualisation et même les ajouter à des catégories dans la vue Catégories et concepts. Pour pouvoir extraire les résultats de l'analyse des liens du texte, des règles d'analyse des liens du texte doivent être définies dans le modèle de ressources ou dans les bibliothèques que vous utilisez. Pour plus d'informations, reportez-vous à la section A propos des règles des liens du texte dans le chapitre 19 sur p. 338.

Si vous extrayez des résultats de patrons d'analyse des liens du texte, les résultats s'affichent dans cette vue. Sinon, vous devez cliquer sur le bouton Extraire et sélectionner l'option pour permettre l'extraction de ces patrons.

Figure 8-10
Vue Analyse des liens du texte

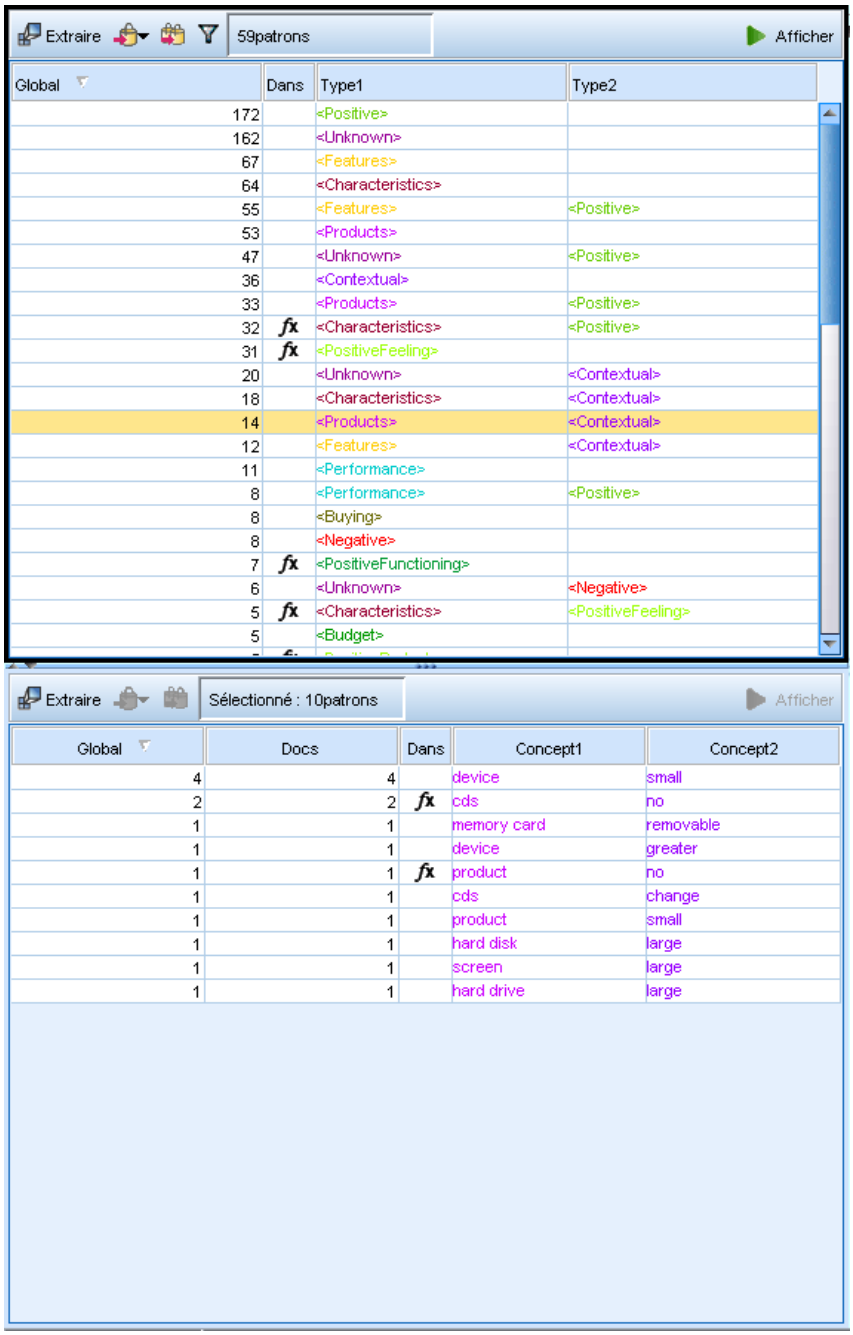


La vue Analyse des liens du texte est organisée en quatre panneaux ; vous pouvez masquer ou afficher chacun d'entre eux en sélectionnant son nom dans le menu Affichage. Pour plus d'informations, reportez-vous à la section Exploration de l'analyse des liens du texte dans le chapitre 12 sur p. 245.

Panneaux Patrons de type et Patrons de concept

Dans ces deux panneaux interconnectés situés à gauche, vous pouvez explorer et sélectionner des résultats de patrons TLA. Les patrons comportent jusqu'à six types ou six concepts. Veuillez noter que pour du texte en japonais, les patrons sont des séries d'un ou de deux types ou concepts maximum. La règle de patrons d'analyse des liens du texte définie dans les ressources linguistiques détermine la complexité des résultats de patrons. Pour plus d'informations, reportez-vous à la section A propos des règles des liens du texte dans le chapitre 19 sur p. 338.*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Figure 8-11
Vue Analyse des liens du texte : Panneaux Patrons de type et Patrons de concept



Les résultats de patrons sont d'abord regroupés au niveau du type, puis ils sont divisés en patrons de concept. C'est la raison pour laquelle il existe deux panneaux de résultats différents : Patrons de type (en haut à gauche) et Patrons de concept (en bas à gauche).

- **Patrons de type.**Le panneau Patrons de type présente les patrons extraits comportant au moins deux types associés correspondant à une règle de patrons TLA. Les patrons de type se présentent sous la forme `<Organization>` + `<Location>` + `<Positive>`, ce qui permet d'obtenir un commentaire positif concernant une organisation située dans une location particulière.
- **Patrons de concept.** Le panneau Patrons de concept présente les patrons extraits au niveau du concept pour tous les patrons de type actuellement sélectionnés dans le panneau Patrons de type situé au-dessus. Les patrons de concept suivent une structure de type `hôtel + paris + merveilleux`.

Comme avec les résultats d'extraction dans la vue Catégories et concepts, vous pouvez vérifier les résultats ici. Si vous souhaitez affiner les types et concepts qui constituent ces patrons, procédez aux modifications dans le panneau Résultats d'extraction de la vue Catégories et concepts ou directement dans l'éditeur de ressources, puis exécutez une nouvelle extraction des patrons.

Panneau Visualisation

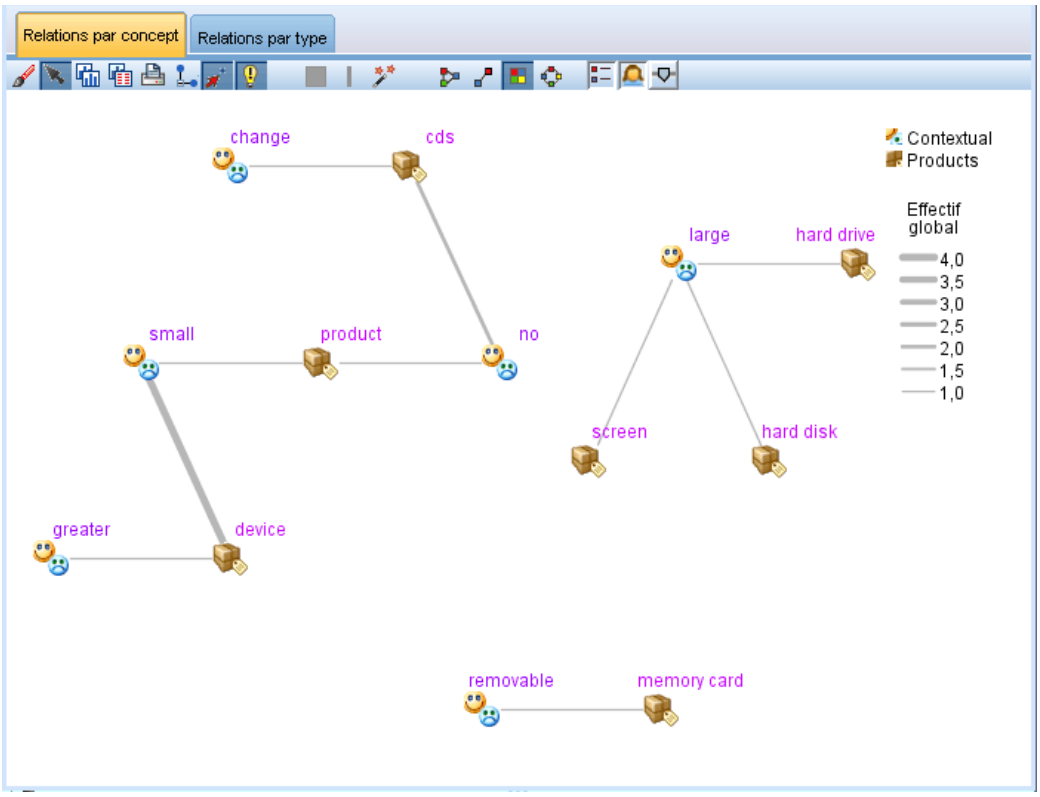
Ce panneau, situé dans l'angle supérieur droit de la vue Analyse des liens du texte, représente un graphique Relations des patrons sélectionnés sous forme de patrons de type ou de patrons de concept. Si ce panneau est masqué, vous pouvez l'afficher à partir du menu Affichage (Affichage > Visualisation). En fonction des éléments sélectionnés dans d'autres panneaux, vous pouvez afficher les interactions correspondantes entre les documents/enregistrements et les patrons.

Les résultats sont présentés sous plusieurs formes :

- **Graphique de concept.** Ce graphique présente tous les concepts figurant dans les patrons sélectionnés. Dans un graphique de concept, l'épaisseur des lignes et la taille des nœuds (lorsque les icônes de type ne sont pas représentées) indiquent le nombre d'occurrences globales figurant dans la table sélectionnée.
- **Graphique de type.** Ce graphique présente tous les types figurant dans les patrons sélectionnés. Dans un graphique de type, l'épaisseur des lignes et la taille des nœuds (lorsque les icônes de type ne sont pas représentées) indiquent le nombre d'occurrences globales figurant dans la table sélectionnée. Les nœuds sont représentés soit par une couleur de type, soit par une icône.

Pour plus d'informations, reportez-vous à la section Graphiques Analyse des liens du texte dans le chapitre 13 sur p. 261.

Figure 8-12
Analyse des liens du texte : panneau Visualisation



Panneau Données

Le panneau Données est situé dans l'angle inférieur droit. Ce panneau présente un tableau contenant les documents ou les enregistrements correspondant à une sélection dans une autre zone de la vue. En fonction des éléments sélectionnés, seul le texte correspondant apparaît dans le panneau Données. Une fois votre sélection effectuée, cliquez sur un bouton Afficher pour remplir le panneau de données à l'aide du texte correspondant.

Si un autre panneau contient une sélection, les documents ou enregistrements correspondants représentent les concepts mis en surbrillance en couleur pour vous permettre de les repérer plus facilement dans le texte. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher une info-bulle contenant le nom du concept sous lequel il a été extrait et le type auquel il a été affecté. Pour plus d'informations, reportez-vous à la section Le panneau Données dans le chapitre 10 sur p. 174.

La vue Editeur de ressources

IBM® SPSS® Modeler Text Analytics utilise un moteur d'extraction fiable pour saisir avec rapidité et précision les principaux concepts des données textuelles. Ce moteur s'appuie essentiellement sur les ressources linguistiques pour déterminer la quantité de données textuelles non structurées à analyser et à interpréter.

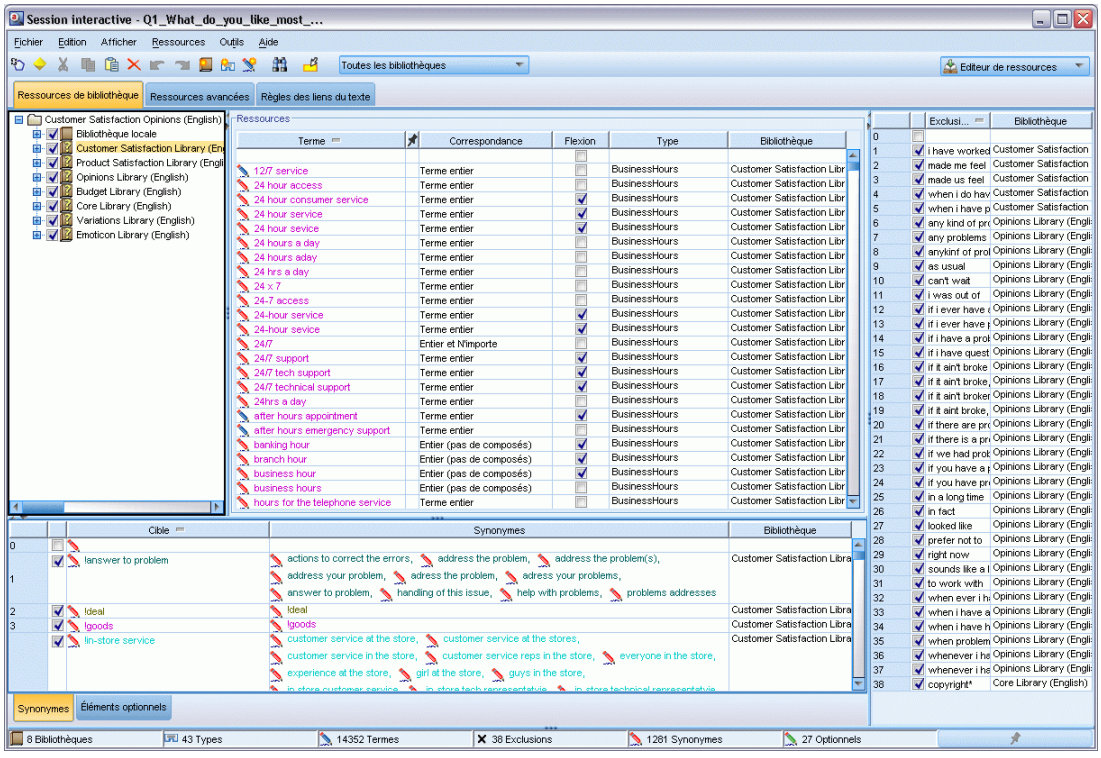
Dans la vue de l'Editeur de ressources, vous pouvez afficher et affiner les ressources linguistiques utilisées pour extraire les concepts, les regrouper en types, découvrir des patrons dans les données textuelles, et bien plus encore. SPSS Modeler Text Analytics propose plusieurs modèles de ressources préconfigurés. Pour certaines langues, vous pouvez également utiliser les ressources dans un package d'analyse de texte. Pour plus d'informations, reportez-vous à la section Utilisation des packages d'analyse de texte dans le chapitre 10 sur p. 224.

Comme ces ressources risquent de ne pas toujours être parfaitement adaptées au contexte de vos données, vous pouvez créer, modifier et gérer vos propres ressources pour un contexte ou un domaine particulier dans l'Editeur de ressources. Pour plus d'informations, reportez-vous à la section Utilisation des bibliothèques dans le chapitre 16 sur p. 290.

Pour simplifier le processus d'adaptation de vos ressources linguistiques, vous pouvez effectuer les tâches courantes relatives aux dictionnaires directement à partir de la vue Catégories et concepts à l'aide des menus contextuels des panneaux Résultats d'extraction et Données. Pour plus d'informations, reportez-vous à la section Affinage des résultats de l'extraction dans le chapitre 9 sur p. 154.

Remarque : L'interface des ressources correspondantes pour le texte en japonais est légèrement différente. l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium. Pour plus d'informations, reportez-vous à la section Modification des ressources pour du texte en japonais dans l'annexe A sur p. 374.

Figure 8-13
Vue Editeur de ressources



Les opérations que vous effectuez dans la vue Editeur de ressources concernent la gestion et l'adaptation des ressources linguistiques. Ces ressources sont stockées sous la forme de modèles et de bibliothèques. La vue de l'Editeur de ressources comporte quatre parties : le panneau Arborescence des bibliothèques, le panneau Déclaration de types, le panneau Dictionnaire de substitutions et le panneau Dictionnaire d'exclusions.

Remarque : Pour plus d'informations, reportez-vous à la section Interface de l'éditeur dans le chapitre 15 sur p. 275.

Définition des options

Vous pouvez définir les options générales de IBM® SPSS® Modeler Text Analytics dans la boîte de dialogue Options. Cette boîte de dialogue comporte les onglets suivants :

- **Session.** Cet onglet contient les options générales et les séparateurs.
- **Affichage.** Cet onglet contient les options relatives aux couleurs utilisées dans l'interface.
- **Sons.** Cet onglet contient les options relatives aux signaux sonores.

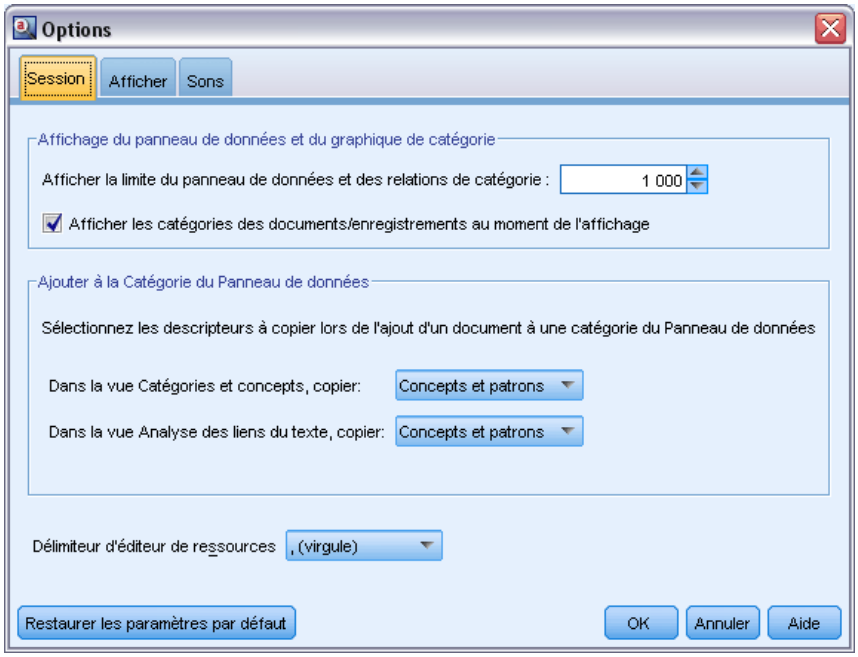
Pour éditer les options

- A partir des menus, sélectionnez Outils> Options. La boîte de dialogue Options apparaît.
- Sélectionnez l'onglet contenant les informations à modifier.
- Modifiez les options.
- Cliquez sur OK pour enregistrer les modifications.

Options : onglet Session

Cet onglet permet de définir certains paramètres de base.

Figure 8-14
Boîte de dialogue Options : onglet Session



Affichage du panneau de données et du graphique de catégorie. Ces options affectent la manière dont les données sont présentées dans le panneau de données et dans le panneau Visualisation dans la vue Catégories et concepts.

- **Limite d'affichage du panneau de données et des relations de catégorie.** Cette option permet de définir le nombre maximal de documents à afficher ou à utiliser pour remplir les panneaux de données ou les graphiques et les diagrammes de la vue Catégories et concepts.
- **Afficher les catégories pour les documents/enregistrement lors de l'affichage.** Si cette option est sélectionnée, les documents ou les enregistrements font l'objet d'un scoring à chaque fois que vous cliquez sur Afficher de sorte que toutes les catégories auxquelles ils appartiennent peuvent être affichées dans la colonne Catégorie du panneau de données, ainsi que dans les graphiques de catégorie. Dans certains cas, notamment avec des ensembles de données importants, vous pouvez désactiver cette option pour que les données et les graphiques s'affichent plus rapidement.

Ajouter à la Catégorie du Panneau de données. Ces options affectent les éléments ajoutés aux catégories lors de l'ajout de documents et d'enregistrements du Panneau de données.

- **Dans la vue Catégories et concepts, copier.** Ajouter un document ou un enregistrement du Panneau de données dans cette vue copiera soit uniquement les concepts soit à la fois les concepts et les patrons.
- **Dans la vue Analyse des liens du texte, copier.** Ajouter un document ou un enregistrement du Panneau de données dans cette vue copiera soit uniquement les patrons soit à la fois les concepts et les patrons.

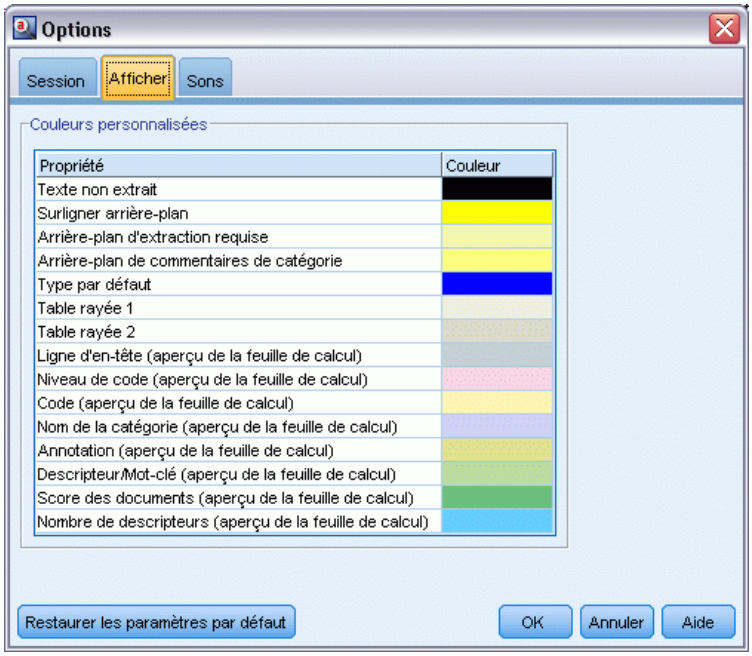
Séparateur de l'éditeur de ressources. Sélectionnez le caractère à utiliser en tant que séparateur lors de la saisie d'éléments, tels que des concepts, des synonymes et des éléments optionnels dans la vue Editeur de ressources.

Options : Onglet Affichage

Dans cet onglet, vous pouvez éditer les options qui affectent la présentation globale de l'application et les couleurs utilisées pour différencier les éléments.

Remarque : Pour donner au modèle une apparence classique ou l'apparence d'une des versions précédentes, ouvrez la boîte de dialogue Options utilisateur dans le menu Outils de la fenêtre principale IBM® SPSS® Modeler.

Figure 8-15
Boîte de dialogue Options : onglet Couleurs



Couleurs personnalisées. Editez les couleurs des éléments apparaissant à l'écran. Vous pouvez modifier la couleur de chacun des éléments du tableau. Pour indiquer une couleur personnalisée, cliquez sur la zone Couleur à droite de l'élément à modifier et choisissez une couleur dans la liste déroulante des couleurs.

- **Texte non extrait.** Données textuelles, qui n'ont pas encore été extraites, visibles dans le panneau Données.
- **Surligner l'arrière-plan.** Couleur d'arrière-plan de sélection de texte utilisée lors de la sélection d'éléments dans les panneaux ou de texte dans le panneau Données.
- **Arrière-plan d'extraction nécessaire.** Couleur d'arrière-plan des panneaux Résultats d'extraction, Patrons et Clusters indiquant que des modifications ont été apportées aux bibliothèques et qu'une extraction est nécessaire.
- **Arrière-plan des commentaires de catégorie.** Couleur d'arrière-plan de catégorie apparaissant à l'issue d'une opération.
- **Type par défaut.** Couleur par défaut des types et des concepts apparaissant dans le panneau de données et le panneau Résultats d'extraction. Cette couleur sera appliquée à tous les types personnalisés que vous créerez dans l'éditeur de ressources. Vous pouvez remplacer cette

couleur par défaut pour les déclarations de types personnalisées en éditant leurs propriétés dans l'Editeur de ressources. Pour plus d'informations, reportez-vous à la section Création de types dans le chapitre 17 sur p. 306.

- **Table rayée 1.** Première des deux couleurs utilisées en alternance dans le tableau de la boîte de dialogue Editer les concepts forcés afin de différencier chaque ensemble de lignes.
- **Table rayée 2.** Seconde des deux couleurs utilisées en alternance dans le tableau de la boîte de dialogue Editer les concepts forcés afin de différencier chaque ensemble de lignes.

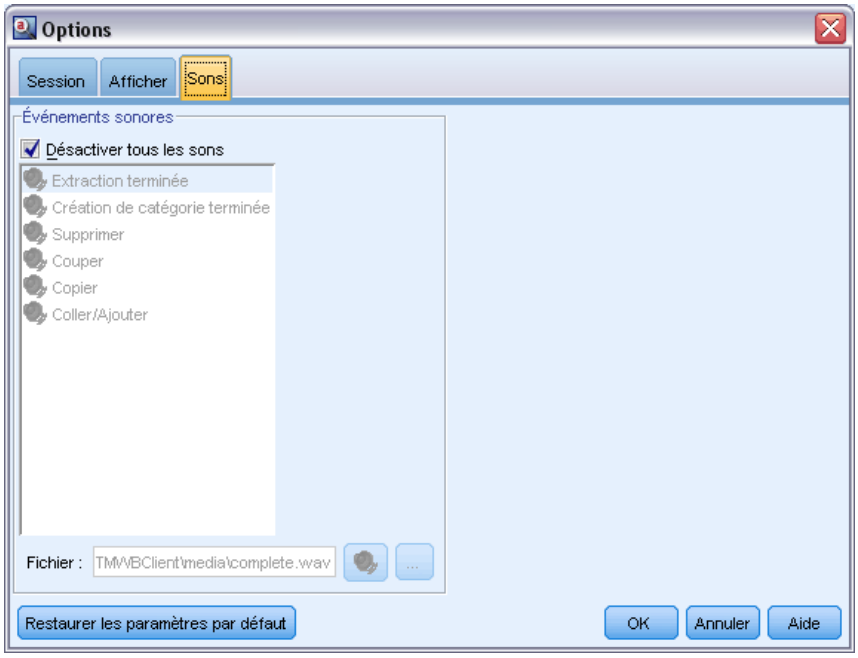
Remarque : Si vous cliquez sur le bouton Rétablir les valeurs par défaut, toutes les options de cette boîte de dialogue reprennent les valeurs qui leur étaient affectées lors de la première installation de ce produit.

Options : onglet Sons

Cet onglet permet d'éditer les options concernant les sons. Dans Sons, vous pouvez indiquer un son à utiliser pour vous avertir lorsqu'un événement se produit. Un grand nombre de sons sont disponibles. Utilisez le bouton ... pour sélectionner un son. Les fichiers *.wav* utilisés pour créer des sons pour IBM® SPSS® Modeler Text Analytics sont stockés dans le sous-répertoire */media* du répertoire d'installation. Si vous ne souhaitez pas que les sons soient lus, sélectionnez Désactiver tous les sons. Les sons sont désactivés par défaut.

Remarque : Si vous cliquez sur le bouton Rétablir les valeurs par défaut, toutes les options de cette boîte de dialogue reprennent les valeurs qui leur étaient affectées lors de la première installation de ce produit.

Figure 8-16
Boîte de dialogue Options : onglet Sons



Microsoft Internet Explorer Paramètres de l'aide

Paramètres Microsoft Internet Explorer

La plupart des fonctions d'Aide de cette application utilisent des techniques basées sur Microsoft Internet Explorer. Certaines versions d'Internet Explorer (y compris celle fournie avec Microsoft Windows XP, Service Pack 2) bloquent par défaut ce qu'elles considèrent comme du « contenu actif » dans les fenêtres Internet Explorer de votre ordinateur local. Ce paramètre par défaut risque de bloquer du contenu dans les fonctions d'Aide. Pour visualiser tout le contenu de l'Aide, modifiez le comportement par défaut d'Internet Explorer.

- Dans les menus Internet Explorer, choisissez :
  Outils > Options Internet...
- Cliquez sur l'onglet Avancé.
- Accédez à la section Sécurité.
- Sélectionnez (activez) Autoriser le contenu actif à s'exécuter dans les fichiers de la zone Ordinateur local.

Génération de nuggets de modèle et de nœuds de modélisation

Lors d'une session interactive, vous pouvez utiliser le travail que vous avez effectué pour générer les éléments suivants :

- **Un nœud de modélisation Text Mining.** Un nœud de modélisation généré à partir d'une session interactive est un nœud de Text Mining dont les paramètres et les options reflètent ceux enregistrés dans la session interactive ouverte. Cela peut être utile lorsque vous ne disposez plus du nœud de Text Mining d'origine ou que vous souhaitez créer une version. Pour plus d'informations, reportez-vous à la section Text Mining pour les concepts et les catégories dans le chapitre 3 sur p. 30.
- **Un nugget de modèle de catégories.** Un nugget de modèle créé à partir d'une session interactive est un nugget de modèle de catégories. Vous devez disposer d'au moins une catégorie dans la vue Catégories et concepts pour pouvoir générer un modèle de catégories. Pour plus d'informations, reportez-vous à la section Nugget Text Mining : Modèle de catégories dans le chapitre 3 sur p. 72.

Pour générer un nœud de modélisation Text Mining

- Dans les menus, sélectionnez Générer > Générer le nœud de modélisation. Un nœud de modélisation Text Mining est ajouté à l'espace de travail en utilisant l'ensemble des paramètres actuellement définis dans la session interactive. Le nom du nœud est indiqué après le champ de texte.

Pour générer un nugget de modèle de catégories

- Dans les menus, sélectionnez Générer > Générer le modèle. Un nugget de modèle est généré directement dans la palette Modèle avec le nom par défaut.

Mise à jour des nœuds de modélisation et enregistrement

Lors d'une session interactive, nous vous conseillons de mettre à jour le nœud de modélisation de temps en temps afin d'enregistrer vos modifications. Vous devez également mettre à jour le nœud de modélisation chaque fois que vous avez terminé de travailler dans la session interactive et que vous souhaitez enregistrer votre travail. Lorsque vous mettez à jour le nœud de modélisation, le contenu de la session interactive est enregistré dans le nœud de Text Mining à l'origine de la session interactive. Cela n'a pas pour effet de fermer la fenêtre de sortie.

Important ! Cette mise à jour n'enregistrera pas votre flux. Pour enregistrer votre flux, rendez-vous dans la fenêtre principale de IBM® SPSS® Modeler après la mise à jour du nœud de modélisation.

Pour mettre à jour un nœud de modélisation

▶ Dans les menus, sélectionnez Fichier > Mettre à jour le nœud de modélisation. Le nœud de modélisation est mis à jour avec les paramètres d'extraction et de création, ainsi qu'avec les options et les catégories que vous avez créées.

Fermeture et fin de sessions

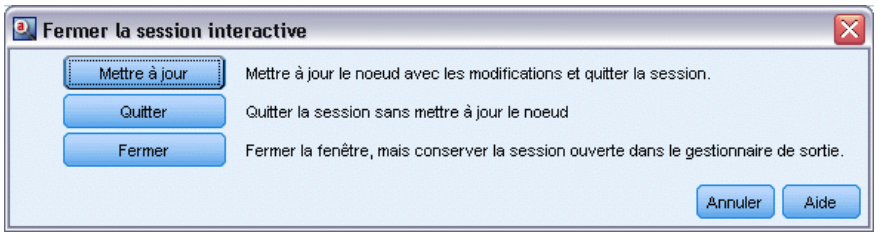
Lorsque vous avez terminé de travailler dans votre session, vous pouvez la quitter de trois manières différentes :

- **Enregistrer.** Cette option permet d'enregistrer au préalable votre travail dans le nœud de modélisation d'origine pour les sessions futures, et de publier les bibliothèques pour les réutiliser dans d'autres sessions. Pour plus d'informations, reportez-vous à la section Partage de bibliothèques dans le chapitre 16 sur p. 298. Une fois l'enregistrement terminé, la fenêtre de la session est fermée, et la session est supprimée du gestionnaire de sortie de la fenêtre IBM® SPSS® Modeler.
- **Quitter.** Cette option supprime tout travail non enregistré, ferme la fenêtre de la session et supprime la session du gestionnaire de sortie de la fenêtre SPSS Modeler. Pour libérer de la mémoire, nous vous conseillons d'enregistrer tout travail important et de quitter la session.
- **Fermer.** Cette option n'enregistre pas ni ne supprime votre travail. Elle ferme la fenêtre de la session, mais la session reste active. Vous pouvez rouvrir la fenêtre de la session en sélectionnant cette session dans le gestionnaire de sortie de la fenêtre SPSS Modeler.

Pour fermer une session interactive

▶ Dans les menus, sélectionnez Fichier > Fermer.

Figure 8-17
Boîte de dialogue Fermer la session interactive



Accessibilité via le clavier

L'interface de la session interactive propose des raccourcis clavier afin de rendre les fonctionnalités du produit plus accessibles. Vous pouvez appuyer simultanément sur la touche Alt + la touche appropriée pour activer les menus de la fenêtre (par exemple, Alt+F pour accéder au menu Fichier) ou sur la touche Tab pour passer d'une commande à l'autre dans une boîte de dialogue. Dans cette section, nous étudierons les raccourcis clavier, qui proposent une autre manière de naviguer. Il existe d'autres raccourcis clavier pour l'interface IBM® SPSS® Modeler.

Table 8-1
Raccourcis clavier génériques

Touche de raccourci	**Fonction**
Ctrl+1	Afficher le premier onglet d'un panneau à onglets.
Ctrl+2	Afficher le second onglet d'un panneau à onglets.
Ctrl+A	Sélectionner tous les éléments du panneau actif.
Ctrl+C	Copier le texte sélectionné dans le Presse-papiers.
Ctrl+E	Lancer une extraction dans les vues Catégories et concepts et Analyse des liens du texte.
Ctrl+F	Afficher la barre d'outils Rechercher dans l'l'Editeur de ressources/Editeur de modèle, si elle ne l'est pas déjà, et la rendra active.
Ctrl+I	Dans la vue Catégories et concepts, lancer la boîte de dialogue Définitions de catégorie pour la catégorie sélectionnée. Dans la vue Clusters, lancer la boîte de dialogue Définitions du cluster pour le cluster sélectionné.
Ctrl+R	Ouvrir la boîte de dialogue Ajouter des termes dans l'Editeur de ressources/Editeur de modèle.
Ctrl+T	Ouvrir la boîte de dialogue Propriétés de type afin de créer un type dans l'Editeur de ressources/Editeur de modèle.
Ctrl+V	Coller le contenu du Presse-papiers.
Ctrl+X	Couper les éléments sélectionnés dans l'Editeur de ressources/Editeur de modèle.
Ctrl+Y	Rétablir la dernière action effectuée dans la vue.
Ctrl+Z	Annuler la dernière action effectuée dans la vue.
F1	Afficher l'Aide ou, dans une boîte de dialogue, afficher l'aide contextuelle d'un élément.
F2	Activer ou désactiver le mode Edition dans les cellules du tableau.
F6	Activer les différents panneaux principaux dans la vue active.
F8	Activer les barres de fractionnement du panneau afin de le redimensionner.
F10	Développer le menu Fichier principal.
flèche haut, flèche bas	Redimensionner le panneau verticalement lorsque la barre de fractionnement est sélectionnée.
flèche gauche, flèche droite	Redimensionner le panneau horizontalement lorsque la barre de fractionnement est sélectionnée.
Origine, Fin	Redimensionner les panneaux à une taille minimale ou maximale lorsque la barre de fractionnement est sélectionnée.
Tabulation	Passer à l'élément suivant dans la fenêtre, le panneau ou la boîte de dialogue.
Maj+F10	Afficher le menu contextuel d'un élément.
Maj+Tab	Passer à l'élément précédent dans la fenêtre ou la boîte de dialogue.
Maj+flèche	Sélectionner les caractères dans le champ Editer en mode Edition (F2).

Touche de raccourci	Fonction
Ctrl+Tab	Activer la zone principale suivante dans la fenêtre.
Maj+Ctrl+Tab	Activer la zone principale précédente dans la fenêtre.

Raccourcis pour les boîtes de dialogue

Plusieurs touches de raccourci et de lecteur d'écran peuvent être utiles lorsque vous utilisez les boîtes de dialogue. Lorsque vous ouvrez une boîte de dialogue, vous pouvez appuyer sur la touche de tabulation pour activer la première commande et lancer le lecteur d'écran. La liste exhaustive des raccourcis clavier et de lecteur d'écran spéciaux est fournie dans le tableau suivant.

Table 8-2
Raccourcis pour boîtes de dialogue

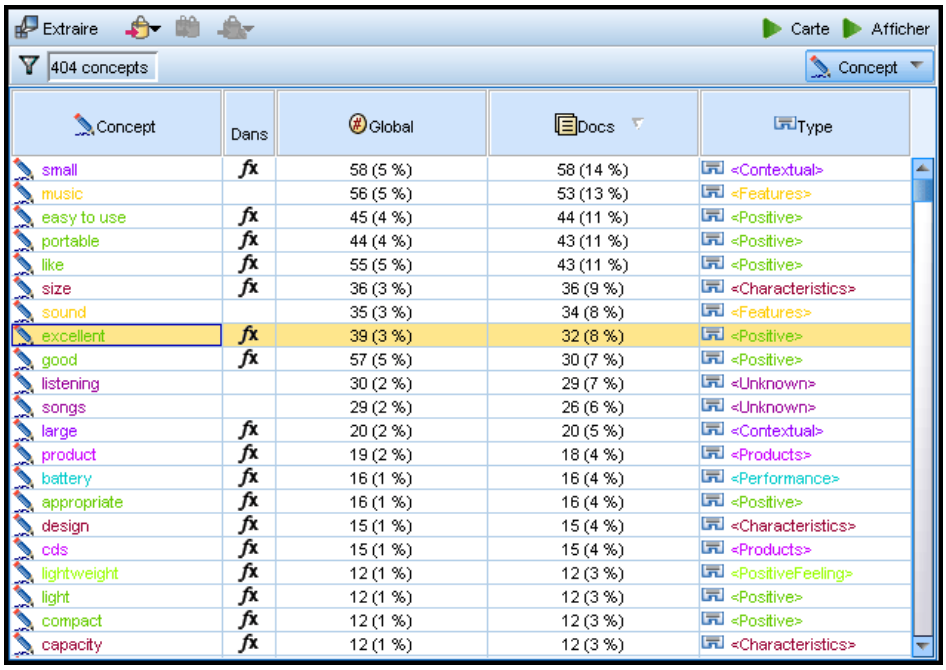
Touche de raccourci	Fonction
Tabulation	Passer à l'élément suivant dans la fenêtre ou la boîte de dialogue.
Ctrl+Tab	Passer d'une zone de texte à l'élément suivant.
Maj+Tab	Passer à l'élément précédent dans la fenêtre ou la boîte de dialogue.
Maj+Ctrl+Tab	Passer d'une zone de texte à l'élément précédent.
barre d'espacement	Sélectionner la commande ou le bouton actif.
Echap	Annuler les modifications et fermer la boîte de dialogue.
Entrée	Valider les modifications et fermer la boîte de dialogue (équivalent au bouton OK). Si vous êtes dans une zone de texte, vous devez tout d'abord appuyer sur Ctrl+Tab pour la quitter.

Chapitre
9

Extraction de concepts et de types

Chaque fois que vous exécutez un flux qui lance la session interactive, une extraction des données textuelles de ce flux est automatiquement effectuée. Le résultat final de cette extraction correspond à un ensemble de concepts, de types, voire de patrons lorsque les ressources linguistiques contiennent des patrons d'analyse de liens du texte (TLA). Vous pouvez afficher et utiliser les concepts et les types dans le panneau Résultats d'extraction. Pour plus d'informations, reportez-vous à la section Fonctionnement de l'extraction dans le chapitre 1 sur p. 7.

Figure 9-1
Panneau Résultats d'extraction après une extraction



Concept	Dans	Global	Docs	Type
small	fx	58 (5 %)	58 (14 %)	<Contextual>
music		56 (5 %)	53 (13 %)	<Features>
easy to use	fx	45 (4 %)	44 (11 %)	<Positive>
portable	fx	44 (4 %)	43 (11 %)	<Positive>
like	fx	55 (5 %)	43 (11 %)	<Positive>
size	fx	36 (3 %)	36 (9 %)	<Characteristics>
sound		35 (3 %)	34 (8 %)	<Features>
excellent	fx	39 (3 %)	32 (8 %)	<Positive>
good	fx	57 (5 %)	30 (7 %)	<Positive>
listening		30 (2 %)	29 (7 %)	<Unknown>
songs		29 (2 %)	26 (6 %)	<Unknown>
large	fx	20 (2 %)	20 (5 %)	<Contextual>
product	fx	19 (2 %)	18 (4 %)	<Products>
battery	fx	16 (1 %)	16 (4 %)	<Performance>
appropriate	fx	16 (1 %)	16 (4 %)	<Positive>
design	fx	15 (1 %)	15 (4 %)	<Characteristics>
cds	fx	15 (1 %)	15 (4 %)	<Products>
lightweight	fx	12 (1 %)	12 (3 %)	<PositiveFeeling>
light	fx	12 (1 %)	12 (3 %)	<Positive>
compact	fx	12 (1 %)	12 (3 %)	<Positive>
capacity	fx	12 (1 %)	12 (3 %)	<Characteristics>

Si vous souhaitez affiner vos résultats d'extraction, vous pouvez modifier les ressources linguistiques et procéder à une réextraction. Pour plus d'informations, reportez-vous à la section Affinage des résultats de l'extraction sur p. 154. Le processus d'extraction s'appuie sur les ressources et sur les paramètres éventuels de la boîte de dialogue Extraire pour déterminer la manière d'extraire et d'organiser les résultats. Vous pouvez utiliser les résultats d'extraction pour définir la plupart, voire l'ensemble, de vos définitions de catégorie.

Résultats d'extraction : concepts et types

Le processus d'extraction analyse les données textuelles, puis identifie les concepts pertinents, les extrait et les affecte à des types. Une fois l'extraction terminée, les résultats apparaissent dans le panneau Résultats d'extraction, dans l'angle inférieur gauche de la vue Catégories et concepts. La

première fois que vous lancez une session, c'est le modèle de ressources linguistiques que vous avez sélectionné dans le noeud qui sert à extraire et à organiser ces concepts et ces types.

Les concepts, les types et les patrons TLA extraits sont désignés collectivement par le terme **résultats d'extraction** et servent de descripteurs ou blocs de construction pour vos catégories. Vous pouvez également utiliser des concepts, des types et des patrons dans vos règles de catégorie. De plus, les techniques automatiques s'appuient sur des concepts et des types pour créer des catégories.

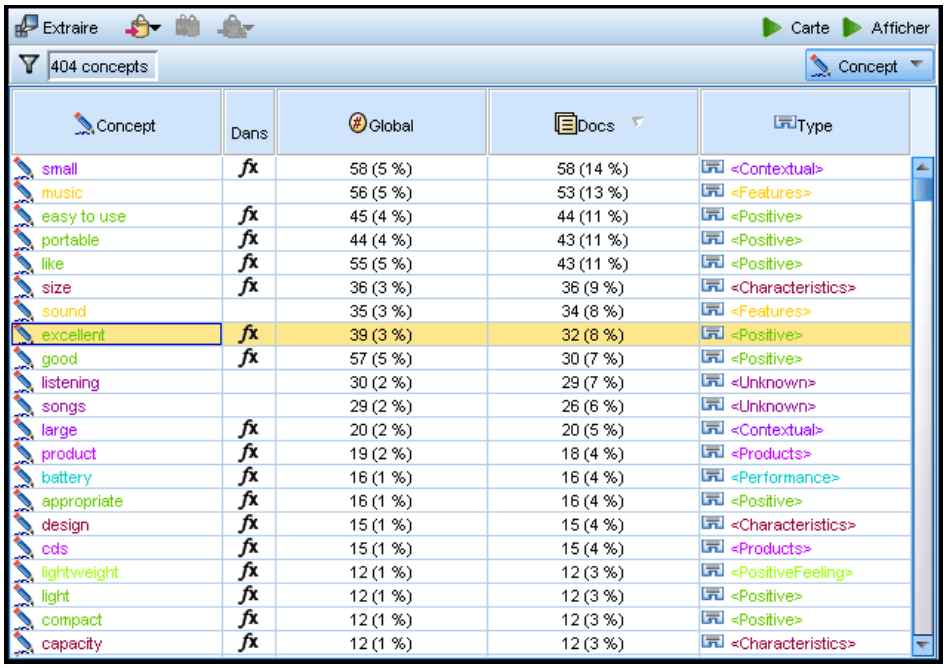
L'analyse de Text mining est un processus itératif au cours duquel les résultats de l'extraction sont passés en revue en fonction du contexte des données textuelles, puis affinés afin de générer de nouveaux résultats avant d'être réévalués. Après l'extraction, passez en revue les résultats et apportez les modifications que vous estimez nécessaires en modifiant les ressources linguistiques. Vous pouvez directement affiner une partie des ressources dans le panneau Résultats d'extraction, , le panneau Données ou dans les boîtes de dialogue Définitions de catégorie ou Définitions du cluster. Pour plus d'informations, reportez-vous à la section Affinage des résultats de l'extraction sur p. 154. Vous pouvez également travailler directement dans la vue de l'Editeur de ressources. Pour plus d'informations, reportez-vous à la section La vue Editeur de ressources dans le chapitre 8 sur p. 129.

Après cette optimisation, vous pouvez procéder à une nouvelle extraction pour voir les nouveaux résultats. En affinant vos résultats d'extraction dès le départ, vous êtes assuré d'obtenir à chaque nouvelle extraction des résultats identiques et parfaitement adaptés au contexte des données dans vos définitions de catégorie. De cette manière, l'attribution des documents/enregistrements à vos définitions de catégorie est plus précise et plus à même d'être répétée.

Concepts

Lors du processus d'extraction, les données textuelles sont analysées pour découvrir des mots isolés intéressants ou pertinents (par exemple, `élection` ou `paix`) et des groupes de mots (par exemple, `élection présidentielle`, `élection du président` ou `traités de paix`). Ces mots et groupes de mots sont appelés des *termes*. En utilisant les ressources linguistiques, les termes pertinents sont extraits et les termes similaires sont regroupés sous un terme principal appelé **concept**.

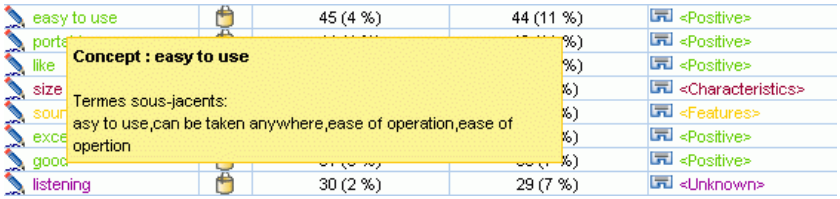
Figure 9-2
Panneau Résultats d'extraction après une extraction



Concept	Dans	Global	Docs	Type
small	fx	58 (5 %)	58 (14 %)	<Contextual>
music		56 (5 %)	53 (13 %)	<Features>
easy to use	fx	45 (4 %)	44 (11 %)	<Positive>
portable	fx	44 (4 %)	43 (11 %)	<Positive>
like	fx	55 (5 %)	43 (11 %)	<Positive>
size	fx	36 (3 %)	36 (9 %)	<Characteristics>
sound		35 (3 %)	34 (8 %)	<Features>
excellent	fx	39 (3 %)	32 (8 %)	<Positive>
good	fx	57 (5 %)	30 (7 %)	<Positive>
listening		30 (2 %)	29 (7 %)	<Unknown>
songs		29 (2 %)	26 (6 %)	<Unknown>
large	fx	20 (2 %)	20 (5 %)	<Contextual>
product	fx	19 (2 %)	18 (4 %)	<Products>
battery	fx	16 (1 %)	16 (4 %)	<Performance>
appropriate	fx	16 (1 %)	16 (4 %)	<Positive>
design	fx	15 (1 %)	15 (4 %)	<Characteristics>
cds	fx	15 (1 %)	15 (4 %)	<Products>
lightweight	fx	12 (1 %)	12 (3 %)	<PositiveFeeling>
light	fx	12 (1 %)	12 (3 %)	<Positive>
compact	fx	12 (1 %)	12 (3 %)	<Positive>
capacity	fx	12 (1 %)	12 (3 %)	<Characteristics>

Vous pouvez voir l'ensemble des termes sous-jacents pour un concept en passant la souris sur le nom du concept. En procédant ainsi, une info-bulle apparaît indiquant le nom du concept et plusieurs lignes de termes qui sont groupés sous ce concept. Ces termes sous-jacents incluent les synonymes définis dans les ressources linguistiques (que ceux-ci aient été rencontrés ou non dans le texte) ainsi que tous les termes extraits au pluriel/singulier, les termes permutés, les termes provenant du regroupement flou, etc. Vous pouvez également copier ces termes ou voir l'ensemble complet des termes sous-jacents en cliquant avec le bouton droit sur le nom du concept et en choisissant l'option du menu contextuel.

Figure 9-3
Panneau Résultats d'extraction après une extraction



Par défaut, les concepts sont affichés en minuscules et triés dans l'ordre décroissant en fonction des effectifs des documents (colonne Docs.). Quand les concepts sont extraits, un type leur est affecté pour regrouper les concepts similaires. Ils apparaissent sous différents codes de couleurs en fonction de ce type. Le choix des couleurs s'effectue sous les propriétés du type, dans l'Editeur de ressources. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

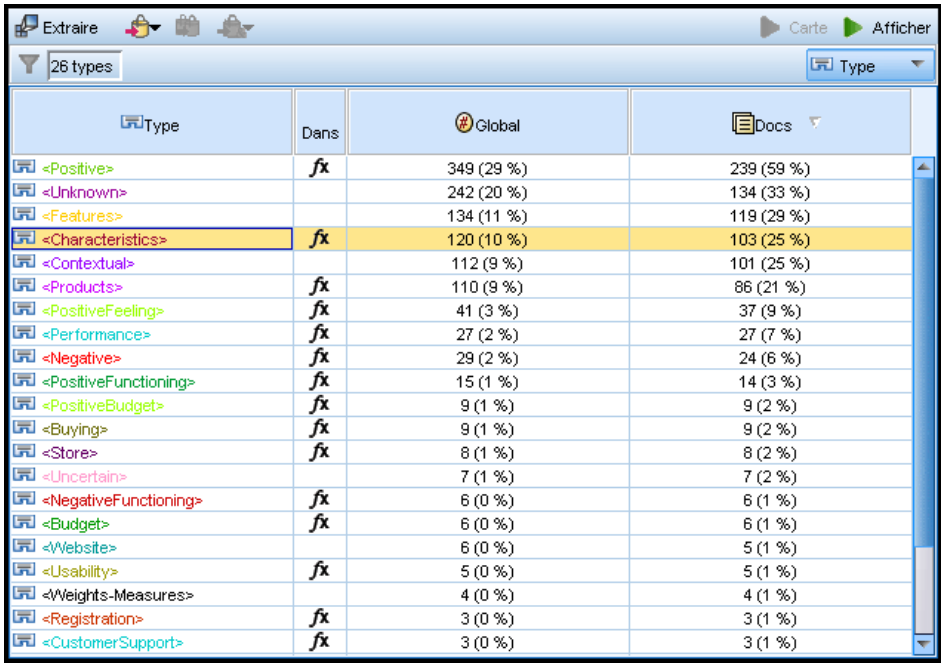
Lorsqu'un concept, un type ou un patron est utilisé dans une définition de catégorie, une icône apparaît dans la colonne de tri Dans.

Types

Les **types** correspondent à des regroupements sémantiques de concepts. Quand les concepts sont extraits, un type leur est affecté pour regrouper les concepts similaires. Plusieurs types intégrés sont fournis avec IBM® SPSS® Modeler Text Analytics , dont `<Location>`, `<Organization>`, `<Person>`,`<Positive>`, `<Negative>` etc. Par exemple, le type `<Location>` regroupe des mots clés géographiques et des lieux. Ce type est affecté à des concepts tels que `chicago`, `paris` et `tokyo`. Dans la plupart des langues, les concepts qui ne se trouvent dans aucune déclaration de type mais sont extraits du texte sont automatiquement dotés du type `<Unknown>` ; cependant, pour le texte en japonais ils sont automatiquement dotés du type <名詞>*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium. Pour plus d'informations, reportez-vous à la section Types intégrés dans le chapitre 17 sur p. 306.

Quand vous sélectionnez la vue Type, les types extraits apparaissent par défaut dans l'ordre décroissant, par fréquence globale. Vous pouvez également remarquer que les types font l'objet d'un codage couleur pour être plus faciles à distinguer. Les couleurs font partie des propriétés du type. Pour plus d'informations, reportez-vous à la section Création de types dans le chapitre 17 sur p. 306. Vous pouvez aussi créer vos propres types.

Figure 9-4
Vue Type : Panneau Résultats d'extraction



Type	Dans	Global	Docs
<Positive>	fx	349 (29 %)	239 (59 %)
<Unknown>		242 (20 %)	134 (33 %)
<Features>		134 (11 %)	119 (29 %)
<Characteristics>	fx	120 (10 %)	103 (25 %)
<Contextual>		112 (9 %)	101 (25 %)
<Products>	fx	110 (9 %)	86 (21 %)
<PositiveFeeling>	fx	41 (3 %)	37 (9 %)
<Performance>	fx	27 (2 %)	27 (7 %)
<Negative>	fx	29 (2 %)	24 (6 %)
<PositiveFunctioning>	fx	15 (1 %)	14 (3 %)
<PositiveBudget>	fx	9 (1 %)	9 (2 %)
<Buying>	fx	9 (1 %)	9 (2 %)
<Store>	fx	8 (1 %)	8 (2 %)
<Uncertain>		7 (1 %)	7 (2 %)
<NegativeFunctioning>	fx	6 (0 %)	6 (1 %)
<Budget>	fx	6 (0 %)	6 (1 %)
<Website>		6 (0 %)	5 (1 %)
<Usability>	fx	5 (0 %)	5 (1 %)
<Weights-Measures>		4 (0 %)	4 (1 %)
<Registration>	fx	3 (0 %)	3 (1 %)
<CustomerSupport>	fx	3 (0 %)	3 (1 %)

Patrons

Vous pouvez également extraire des patrons de vos données textuelles. Toutefois, vous devez disposer d'une bibliothèque qui contient des règles de patrons TLA (analyse des liens du texte) dans l'Editeur de ressources. Vous devez choisir d'extraire ces patrons dans le paramètre de noeud SPSS Modeler Text Analytics ou dans la boîte de dialogue Extraire à l'aide de l'option Extraction

avec analyse des liens du texte. Pour plus d'informations, reportez-vous à la section Exploration de l'analyse des liens du texte dans le chapitre 12 sur p. 245.

Extraction de données

Chaque fois qu'une extraction est nécessaire, le panneau Résultats d'extraction devient jaune et le message Appuyer sur le bouton Extraire pour extraire les concepts apparaît sous la barre d'outils de ce panneau.

Vous aurez besoin de procéder à une extraction si vous n'avez pas encore de résultat d'extraction, si vous avez modifié les ressources linguistiques et devez mettre à jour les résultats d'extraction, ou si vous avez réouvert une session dans laquelle vous n'aviez pas enregistré les résultats de l'extraction (Outils > Options).

Remarque : Si vous modifiez le nœud source de votre flux après la mise en cache des résultats d'extraction à l'aide de l'option Utiliser le travail d'une session..., vous devrez exécuter une nouvelle extraction une fois la session interactive démarrée, pour obtenir les résultats d'extraction mis à jour.

Lorsque vous effectuez une extraction, un indicateur de progression apparaît pour vous fournir des informations sur l'état de l'extraction. Pendant ce temps, le moteur du programme d'extraction analyse toutes les données textuelles et identifie les termes et patrons pertinents puis les extrait et les attribue à un type. Ensuite, le moteur tente de regrouper les synonymes sous un terme principal, appelé un concept. Une fois le processus terminé, les concepts, les types et les patrons obtenus apparaissent dans le panneau Résultats d'extraction.

Le résultat du processus d'extraction correspond à un ensemble de concepts, de types et de patrons d'analyse des liens du texte (TLA) si la fonction a été activée. Vous pouvez afficher et utiliser ces concepts et ces types dans le panneau Résultats d'extraction de la vue Catégories et concepts. Si vous avez extrait des patrons TLA, vous pouvez les visualiser dans la vue Analyse des liens du texte.

Remarque : Il existe un rapport entre la taille de votre ensemble de données et la durée nécessaire à l'exécution du processus d'extraction. Vous pouvez toujours envisager d'insérer un nœud Echantillon en amont ou d'optimiser la configuration de votre ordinateur.

Pour extraire des données

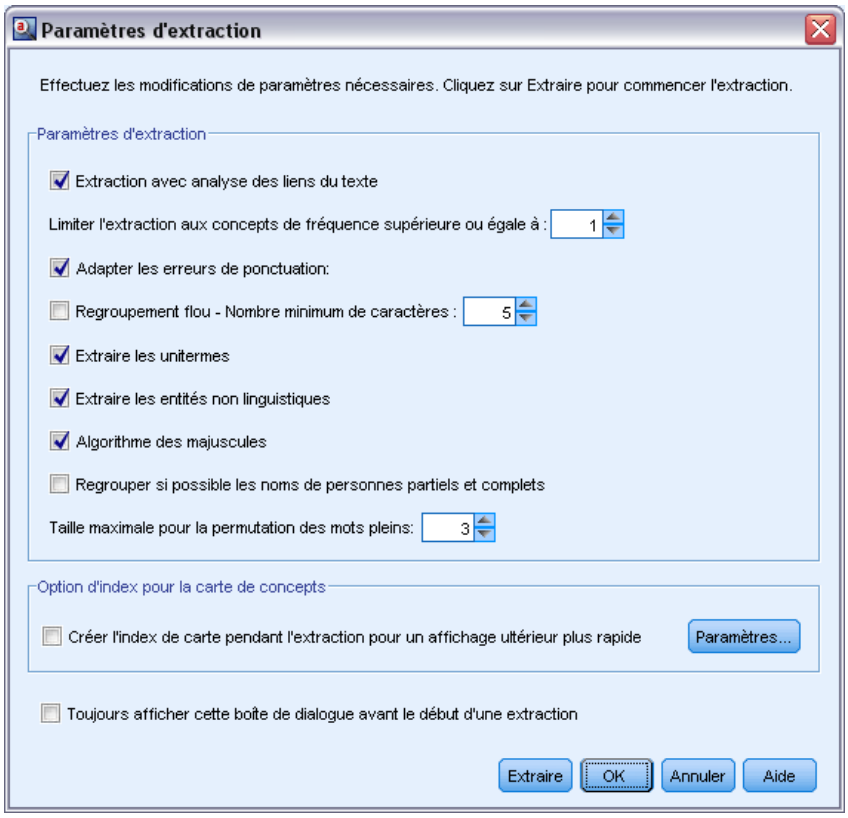
- A partir des menus, sélectionnez Outils> Extraire. Vous pouvez également cliquer sur le bouton Extraire de la barre d'outils.
- Si vous choisissez de toujours afficher la boîte de dialogue Paramètres d'extraction, elle apparait pour vous permettre d'apporter des modifications. Vous trouverez plus loin dans cette rubrique des descriptions de chaque paramètre.
- Cliquez sur Extraire pour lancer le processus d'extraction. Dès le début de l'extraction, une boîte de dialogue indique la progression du processus. Après l'extraction, les résultats apparaissent dans le panneau Résultats d'extraction. Par défaut, les concepts sont affichés en minuscules et triés dans l'ordre décroissant en fonction des effectifs des documents (colonne Docs.).

Vous pouvez examiner les résultats à l'aide des options de la barre d'outils pour trier les résultats différemment, les filtrer, ou encore afficher une autre vue (concepts ou types). Vous pouvez aussi redéfinir vos résultats d'extraction en travaillant avec les ressources linguistiques. Pour plus d'informations, reportez-vous à la section Affinage des résultats de l'extraction sur p. 154.

Pour le texte en néerlandais, en anglais, en français, en allemand, en italien, en portugais, et en espagnol

La boîte de dialogue Paramètres d'extraction contient des options d'extraction élémentaires.

Figure 9-5
Boîte de dialogue Paramètres d'extraction



Extraction avec analyse des liens du texte. Indique que vous souhaitez extraire les patrons TLA de vos données textuelles. L'option suppose également que vous disposez de règles de patrons TLA dans l'une de vos bibliothèques de l'éditeur de ressources. Cette option risque d'augmenter considérablement la durée de l'extraction. Pour plus d'informations, reportez-vous à la section Exploration de l'analyse des liens du texte dans le chapitre 12 sur p. 245.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Traitement des fautes de frappe. Nombre de caractères minimum requis ([n]). Cette option applique une technique de regroupement flou qui permet de regrouper les mots mal orthographiés ou d'orthographe similaire sous un seul concept. L'algorithme de regroupement flou retire temporairement toutes les voyelles (à l'exception de la première) et les consonnes doubles/triples des mots extraits pour comparer ces derniers et voir s'ils sont identiques, afin par exemple que `modélisation` et `modelisation` soient regroupés. Néanmoins, si chaque type est affecté à un type différent et que vous excluez le type `<Unknown>`, le regroupement flou n'est pas appliqué.

Vous pouvez également définir le nombre minimum de caractères *racines* requis avant d'utiliser le regroupement flou. Le nombre de caractères racine d'un terme est calculé en ajoutant l'ensemble des caractères et en soustrayant les caractères formant des suffixes flexionnels et, dans le cas des termes apparaissant sous la forme de mots composés, les déterminants et les prépositions. Par exemple, le terme `exercices` serait déterminé comme comportant 8 caractères racine dans la forme « exercice », étant donné que la lettre *s* située à la fin du mot représente une flexion (forme plurielle). De même, `jus énergétique` comporte 14 caractères racine « jus énergétique » et `conception de dessins`, 16 caractères racine « conception dessin ». Cette méthode de calcul est uniquement utilisée pour vérifier si le regroupement flou doit être appliqué, mais n'influe pas sur la mise en correspondance des mots.

Remarque : Si, d'après vous, certains mots sont ensuite groupés de manière incorrecte, vous pouvez exclure des paires de mots de cette technique en les déclarant de manière explicite dans la section Regroupement flou : Exceptions de l'onglet de ressources avancées. Pour plus d'informations, reportez-vous à la section Regroupement flou dans le chapitre 18 sur p. 328.

Extraire les unitermes. Cette option extrait les mots uniques (unitermes) dans la mesure où le mot ne fait pas déjà partie d'un mot composé et si c'est un nom ou une catégorie grammaticale non reconnue.

Extraction des entités non linguistiques. Cette option extrait les entités non linguistiques, telles que les numéros de téléphone, numéros de sécurité sociale, heures, dates, devises, chiffres, pourcentages, adresses électroniques, adresses HTTP, etc. Vous pouvez inclure ou exclure certains types d'entités non linguistiques dans la section Entités non linguistiques : configuration de l'onglet de ressources avancées. Désactivez les entités dont vous n'avez pas besoin pour éviter au moteur d'extraction un temps de traitement inutile. Pour plus d'informations, reportez-vous à la section Configuration dans le chapitre 18 sur p. 332.

Algorithme des noms propres. Cette option extrait les termes simples et composés ne figurant pas dans les dictionnaires intégrés, si la première lettre est une majuscule. Cette option offre une bonne manière d'extraire la plupart des noms propres.

Regroupement éventuel des noms de personnes partiels et complets. Cette option regroupe les noms qui apparaissent différemment dans le texte. Cette fonction est utile car les noms sont souvent cités dans leur forme complète au début du texte puis uniquement en version abrégée. Cette option essaye de faire correspondre tout uniterme ayant le type `<Unknown>` avec le dernier mot de tout terme composé entré comme `<Person>`. Par exemple, si *martin* est trouvé et initialement saisi comme `<Unknown>`, le moteur d'extraction vérifie si un terme composé de type `<Person>` contient *martin* comme dernier mot, tel que *pierre martin*. Cette option ne s'applique pas aux prénoms, car la plupart ne sont jamais extraits comme unitermes.

Nombre de mots pleins soumis à une permutation pour le regroupement. Cette option indique le nombre maximal de mots utiles pouvant être présents lorsque s'applique la technique de permutation. Cette technique de permutation regroupe les expressions similaires qui ne diffèrent les unes des autres que par la présence de mots utiles (par exemple, de et la), quelle que soit leur flexion. Par exemple, si vous avez paramétré cette valeur sur au moins deux mots et si les deux expressions `représentants d'entreprise` et `représentants de l'entreprise` ont été extraites. Dans ce cas, les deux termes extraits sont regroupés dans la liste de concepts finale car les deux termes sont considérés comme étant les mêmes lorsque `de l'` est ignoré.

Option de l'index pour la carte de concept Indique que vous souhaitez créer l'index de la carte au moment de l'extraction afin que les cartes de concept puissent être rapidement tracées plus tard. Pour modifier les paramètres de l'index, cliquez sur Paramètres. Pour plus d'informations, reportez-vous à la section Création d'index de cartes de concepts sur p. 153.

Toujours afficher cette boîte de dialogue avant le début d'une extraction. Vous permet de spécifier si vous souhaitez que la boîte de dialogue Paramètres d'extraction apparaisse à chaque fois que vous effectuez une extraction, si vous ne voulez pas qu'elle s'affiche sauf si vous l'ouvrez via le menu Outils ou si vous voulez être interrogé à chaque fois que vous effectuez une extraction si vous souhaitez modifier les paramètres d'extraction.

Pour les textes en japonais

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

La boîte de dialogue Paramètres d'extraction contient des options d'extraction de base pour le texte en japonais. Par défaut, les paramètres sélectionnés dans la boîte de dialogue sont les mêmes que ceux sélectionnés dans l'onglet Expert du nœud de modélisation Text Mining. Pour pouvoir utiliser du texte en japonais, vous devez utiliser le texte comme entrée et choisir un modèle de langue japonaise ou un package d'analyse de texte dans l'onglet Modèle du nœud Text Mining. Pour plus d'informations, reportez-vous à la section Copie des ressources à partir de modèles et de TAP dans le chapitre 3 sur p. 42.

Figure 9-6
Boîte de dialogue Paramètres d'extraction pour le texte en japonais

Remarque : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Analyse secondaire. Lorsqu'une extraction est lancée, l'extraction des mots-clés de base est effectuée à l'aide de l'ensemble de types par défaut. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais dans l'annexe A sur p. 378. Mais lorsque vous sélectionnez un analyseur secondaire, vous pouvez obtenir des concepts plus nombreux et plus riches car l'extracteur inclura désormais des verbes à particules et des auxiliaires comme faisant partie du concept. Par exemple, supposons que nous avons une phrase 肩の荷が下りた, traduite par "*Ca m'a enlevé un gros poids*". Dans cet exemple, l'extraction des mots-clés de base peut extraire chaque concept séparément comme suit : 肩 (*poids*), 荷 (*gros*), 下りる (*a enlevé*), mais la relation entre ces mots n'est pas extraite. Cependant, si vous avez appliqué l'analyse de sentiment, vous pouvez extraire des concepts plus riches relatifs à un type de sentiment comme le concept =肩の荷が下りた, qui est traduit par "*avoir enlevé un gros poids*", affecté au type <良い-安心>. Dans le cas d'une analyse de sentiment, un grand nombre de types supplémentaires est également inclus. De plus, choisir un analyseur secondaire vous permet également de générer des résultats d'analyse des liens du texte.

Remarque : Lorsqu'un analyseur secondaire est appelé, le processus d'extraction nécessite plus de temps. Pour plus d'informations, reportez-vous à la section Fonctionnement de l'extraction secondaire dans l'annexe A sur p. 371.

- **Analyse des dépendances** Choisir cette option génère des particules étendues pour les concepts d'extraction du type de base et de l'extraction des mots-clés. Vous pouvez également obtenir des résultats de patrons plus riches avec l'analyse de dépendance des liens du texte (TLA).
- **Analyse des sentiments**. Choisir cet analyseur génère l'extraction de concepts supplémentaires et, le cas échéant, l'extraction de résultats de patrons TLA. En plus des types de base, vous bénéficiez également de plus de 80 types de sentiments, notamment 嬉しい, 吉報, 幸運, 安心, 幸福, etc. Ces types permettent de découvrir des concepts et des patrons dans le texte

grâce à l'expression des émotions, des sentiments et des opinions. Ce sont trois options qui dictent la cible de l'analyse des sentiments : Tous les sentiments, Sentiment représentatif uniquement et Conclusions uniquement.

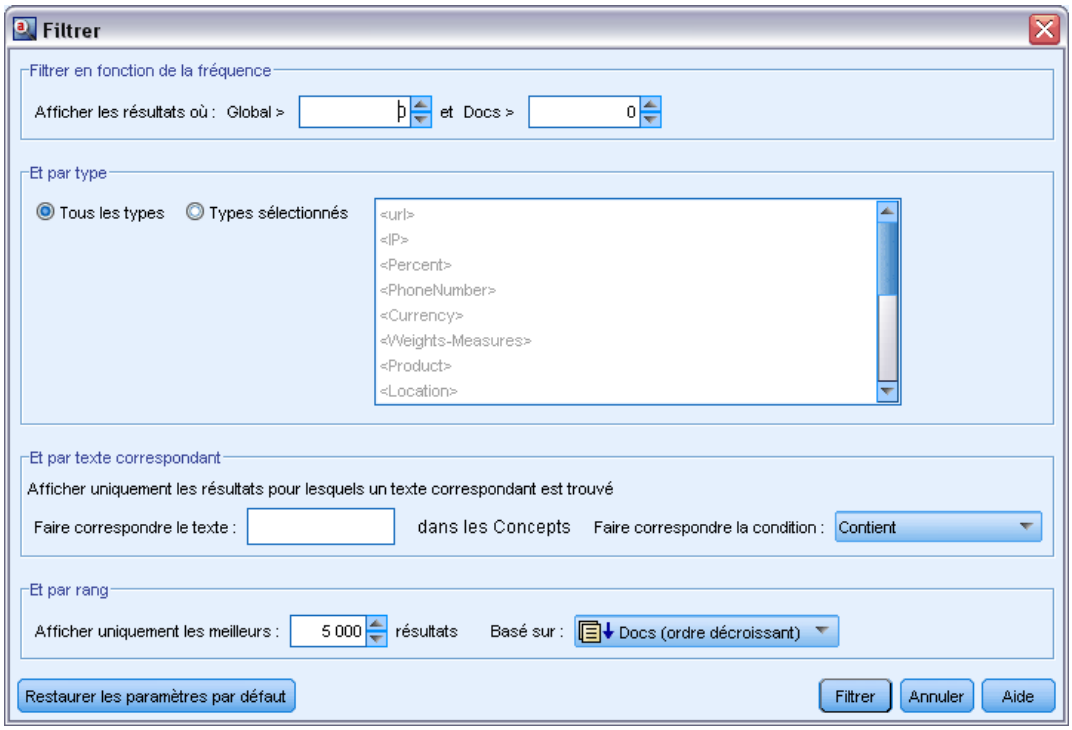
- **Pas d'analyseur secondaire**. Cette option désactive tous les analyseurs secondaires. Cette option ne peut pas être sélectionnée si l'option Extraction avec analyse des liens du texte a été sélectionnée car un deuxième analyseur est nécessaire pour obtenir les résultats TLA.

Extraction avec analyse des liens du texte. Indique que vous souhaitez extraire les patrons TLA de vos données textuelles. L'option suppose également que vous disposez de règles de patrons TLA dans l'une de vos bibliothèques de l'éditeur de ressources. Cette option risque d'augmenter considérablement la durée de l'extraction. De plus, un analyseur secondaire doit être sélectionné afin d'extraire les résultats de patrons TLA. Pour plus d'informations, reportez-vous à la section Exploration de l'analyse des liens du texte dans le chapitre 12 sur p. 245.

Filtrage des résultats d'extraction

Lorsque vous travaillez sur des ensembles de données très volumineux, le processus d'extraction peut renvoyer des millions de résultats. Pour de nombreux utilisateurs, cette quantité peut compliquer l'examen des résultats. Par conséquent, afin de mettre en évidence les résultats les plus intéressants, vous pouvez les filtrer à l'aide de la boîte de dialogue Filtrer dans le panneau Résultats d'extraction.

Figure 9-7
Boîte de dialogue Filtrer (du panneau Résultats d'extraction)



N'oubliez pas que tous les paramètres de cette boîte de dialogue Filtrer sont utilisés pour filtrer les résultats d'extraction disponibles pour les catégories.

Filtrer en fonction de la fréquence. Vous pouvez appliquer un filtre afin de n'afficher que les résultats présentant une certaine valeur de fréquence globale ou de documents.

- La **fréquence globale** est le nombre total d'apparitions d'un concept dans l'ensemble de documents ou d'enregistrements. Cette valeur apparaît dans la colonne Global.
- La **fréquence de documents** est le nombre total de documents ou d'enregistrements dans lesquels un concept apparaît. Cette valeur est visible dans la colonne Docs.

Par exemple, si le concept `otan` est apparu 800 fois dans 500 enregistrements, nous en déduisons qu'il présente une fréquence globale de 800 et une fréquence de document de 500.

Et par type. Vous pouvez appliquer un filtre qui n'affiche que les résultats appartenant à certains types. Vous pouvez choisir tous les types ou uniquement des types spécifiques.

Et par texte correspondant. Vous pouvez également appliquer un filtre n'affichant que les résultats correspondant à la règle que vous définissez ici. Entrez l'ensemble de caractères devant être renvoyés dans le champ Texte correspondant et sélectionnez la condition dans laquelle il faut appliquer la correspondance.

Table 9-1
Conditions de correspondance de texte

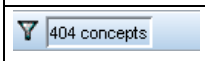
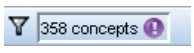
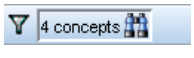
Condition	Description
Contient	Le texte est mis en correspondance si la chaîne apparaît n'importe où. (Option par défaut)
Commence par	Le texte est seulement mis en correspondance si le concept ou le type commence par le texte entré.
Se termine par	Le texte est seulement mis en correspondance si le concept ou le type se termine par le texte entré.
Correspon-dance exacte	Toute la chaîne doit concorder avec le nom du concept ou du type.

Et par rang. Vous pouvez également appliquer un filtre pour n'afficher qu'un certain nombre de concepts en fonction de leur fréquence globale (Global) ou de leur fréquence de documents (Docs), dans l'ordre croissant ou décroissant.

Résultats affichés dans le panneau Résultats d'extraction

Voici des exemples illustrant la manière dont les résultats filtrés peuvent s'afficher dans la barre d'outils du panneau Résultats d'extraction :

Table 9-2
Exemples d'informations sur les filtres

Informations sur les filtres	Description
	La barre d'outils indique le nombre de résultats. Comme aucun filtre de correspondance de texte n'est défini et que le maximum n'est pas atteint, aucune icône supplémentaire n'apparaît.
	La barre d'outils montre que les résultats ont été limités au maximum défini dans le filtre, à savoir 300 dans le cas présent. La présence d'une icône violette indique que le nombre maximal de concepts est atteint. Placez le curseur sur l'icône pour obtenir plus d'informations.
	La barre d'outils montre que les résultats ont été limités à l'aide d'un filtre de correspondance de texte. Cela est indiqué par l'icône d'une loupe.

Pour filtrer les résultats

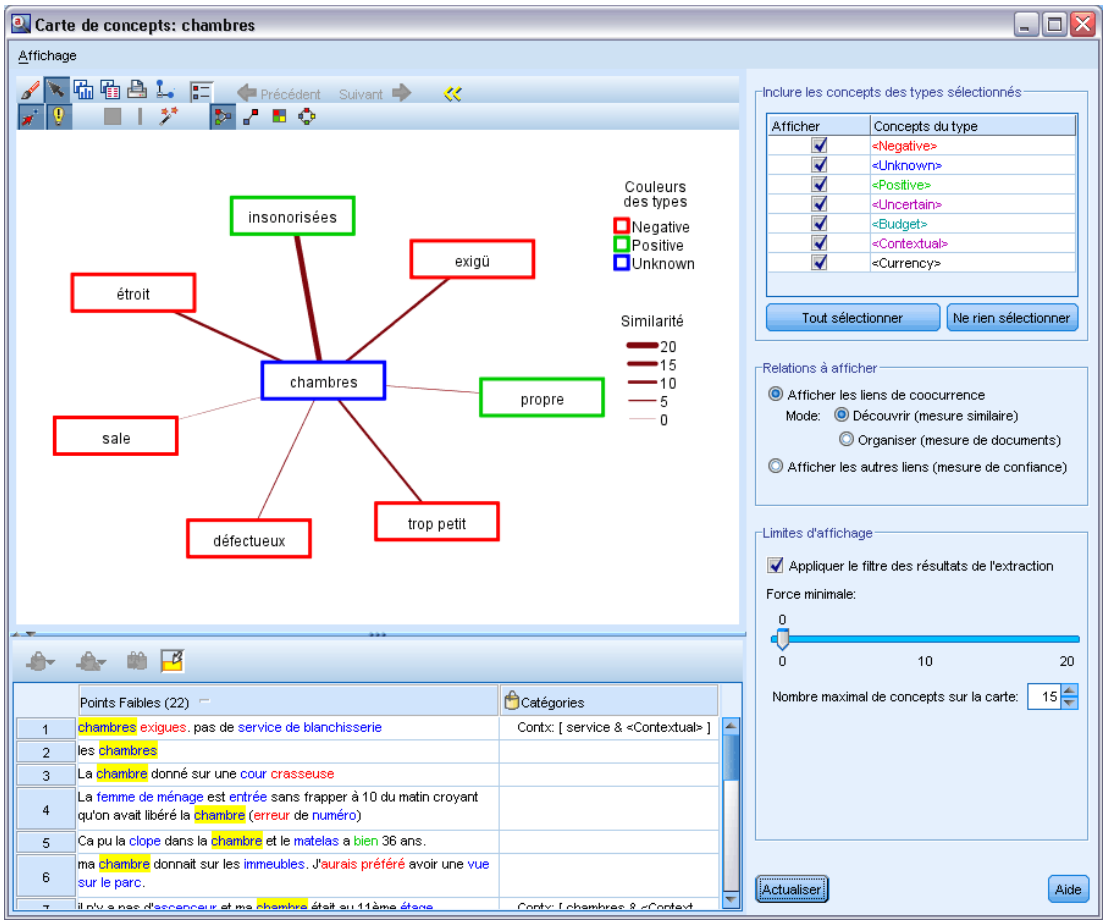
- A partir des menus, sélectionnez Outils> Filtrer. La boîte de dialogue Filtrer apparaît.
- Sélectionnez et affinez les filtres à utiliser.
- Cliquez sur OK pour appliquer les filtres et visualiser les nouveaux résultats dans le panneau Résultats d'extraction.

Exploration des cartes de concept

Vous pouvez créer une carte de concepts pour explorer les interrelations entre les concepts. En sélectionnant un concept unique et en cliquant sur Carte, une fenêtre de carte de concept s'ouvre et vous permet d'explorer l'ensemble des concepts associés au concept sélectionné. Vous pouvez filtrer les concepts à afficher en modifiant les paramètres tels que les types à inclure, les types de relations à rechercher, etc.

Important ! Avant de pouvoir créer une carte, il convient de générer un index. Cette opération peut durer plusieurs minutes. Cependant, une fois que vous avez généré l'index, vous n'avez pas à le générer de nouveau jusqu'à ce que vous procédiez à une nouvelle extraction. Si vous souhaitez que l'index soit généré automatiquement à chaque extraction, sélectionnez cette option dans les paramètres d'extraction. Pour plus d'informations, reportez-vous à la section Extraction de données sur p. 143.

Figure 9-8
Une carte de concepts pour le concept sélectionné



Pour visualiser une carte de concepts

- Dans le panneau Résultats d'extraction, sélectionnez un concept unique.
- Dans la barre d'outils de ce panneau, cliquez sur le bouton Carte. Si l'index de la carte a déjà été généré, la carte de concept s'ouvre dans une boîte de dialogue distincte. Si l'index de la carte n'a pas été généré ou qu'il est obsolète, l'index doit être reconstruit. Ce processus peut durer plusieurs minutes.
- Cliquez sur la carte pour l'explorer. Si vous double-cliquez sur un concept lié, la carte se redessine automatiquement et vous montre les concepts liés pour le concept sur lequel vous venez de double-cliquer.
- La barre d'outils supérieure propose quelques outils de base pour la carte tels que le retour à une carte précédente, des liens de filtrage en fonction de la puissance des relations ainsi que l'ouverture de la boîte de dialogue du filtre pour contrôler les types de concepts qui apparaissent ainsi que les types de relations à représenter. Une seconde ligne dans la barre d'outils contient les

outils d'édition de graphiques. Pour plus d'informations, reportez-vous à la section Utilisation des palettes et des barres d'outils de graphiques dans le chapitre 13 sur p. 264.

- Si les types de liens trouvés ne vous donnent pas satisfaction, consultez de nouveau les paramètres de cette carte qui se trouvent à droite de la carte.

Paramètres de carte : Inclure les concepts des types sélectionnés

Seuls les concepts appartenant aux types sélectionnés dans le tableau apparaissent sur la carte. Pour masquer les concepts d'un certain type, désélectionnez ce type dans le tableau.

Paramètres de carte : Relations à afficher

Afficher les liens de cooccurrence. Si vous souhaitez afficher des liens de cooccurrence, choisissez ce mode. Ce mode influe sur la façon dont la puissance des liens est calculée.

- *Découvrir (mesure de similarité)*. Avec cette mesure, la puissance du lien est calculée à l'aide d'un calcul plus complexe qui prend en compte la fréquence à laquelle deux concepts apparaissent séparément et celle à laquelle ils apparaissent ensemble. Une valeur de puissance élevée signifie que deux concepts ont tendance à apparaître plus souvent ensemble que séparément. Avec cette formule, toutes les valeurs de virgule flottante sont converties en nombres entiers.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Où C_I est le nombre de documents ou d'enregistrements dans lequel apparaît le concept `I`.

C_J est le nombre de documents ou d'enregistrements dans lequel apparaît le concept `J`.

C_{IJ} est le nombre de documents ou d'enregistrements dans lequel la paire de concepts `I` et `J` est co-occurrente dans l'ensemble de documents.

- *Organiser (mesure de document)*. La puissance de ces liens avec cette mesure est déterminée par le nombre brut de cooccurrences. En général, plus deux concepts sont fréquents, plus ils ont de chances d'apparaître ensemble. Une valeur de puissance élevée signifie que deux concepts apparaissent souvent ensemble.

Afficher les autres liens (mesure de confiance). Vous pouvez choisir d'autres liens à afficher ; ils peuvent être sémantiques, de dérivation (morphologiques) ou d'inclusion (syntaxiques) et sont associés au nombre d'étapes supprimées d'un concept par rapport au concept auquel il est associé. Ces liens peuvent vous aider à affiner les ressources, particulièrement en cas de synonymes ou d'ambiguïtés.Pour les descriptions courtes de chacune de ces techniques de regroupement, consultez Paramètres linguistiques avancés sur p. 181

Remarque : N'oubliez pas que si ces liens n'ont pas été sélectionnés lors de la création de l'index ou si aucune relation n'a été trouvée, alors aucun lien ne sera affiché. Pour plus d'informations, reportez-vous à la section Création d'index de cartes de concepts sur p. 153.

Paramètres de carte : Limites d'affichage

Appliquer le filtre des résultats d'extraction. Si vous ne souhaitez pas utiliser tous les concepts, vous pouvez utiliser le filtre du panneau des résultats d'extraction pour limiter l'affichage. Sélectionnez ensuite cette option et IBM® SPSS® Modeler Text Analytics recherchera les concepts associés à l'aide de l'ensemble de filtres. Pour plus d'informations, reportez-vous à la section Filtrage des résultats d'extraction sur p. 148.

Force minimale. Définissez ici la force minimale des liens. Tous les concepts associés ayant une force de relation inférieure à cette limite n'apparaîtront pas sur la carte.

Nombre de concepts maximal sur la carte. Définissez le nombre maximal de relations à afficher sur la carte.

Création d'index de cartes de concepts

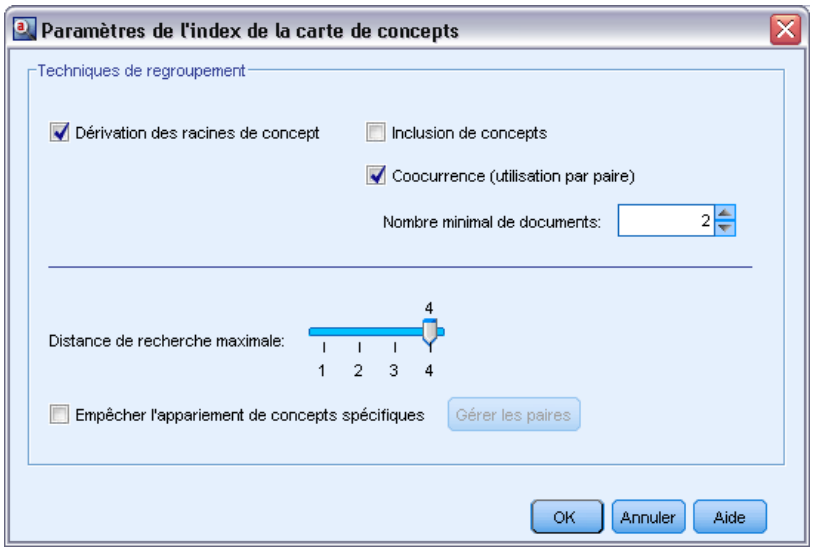
Avant de pouvoir créer une carte, il convient de générer un index de relations de concept. Chaque fois que vous créez une carte de concept, IBM® SPSS® Modeler Text Analytics fait référence à cet index. Vous pouvez choisir les relations à indexer en sélectionnant les techniques dans cette boîte de dialogue

Techniques de regroupement. Choisissez une ou plusieurs techniques. Pour des descriptions brèves de chacune de ces techniques, voyez A propos des Techniques linguistiques sur p. 186 Toutes les techniques ne sont pas disponibles dans toutes les langues.

Éviter l'appariement de concepts spécifiques. Cochez cette case pour arrêter le processus de regroupement ou d'appariement de deux concepts dans les résultats. Pour créer ou gérer des paires de concepts, cliquez sur Gérer les paires. Pour plus d'informations, reportez-vous à la section Gestion des paires d'exceptions de liens dans le chapitre 10 sur p. 185.

La création de l'index peut durer plusieurs minutes. Cependant, une fois l'index généré, il est inutile de le générer de nouveau jusqu'à la prochaine extraction ou si vous souhaitez modifier des paramètres pour inclure plus de relations. Si vous souhaitez générer un index à chaque extraction, vous pouvez sélectionner une option dans les paramètres d'extraction. Pour plus d'informations, reportez-vous à la section Extraction de données sur p. 143.

Figure 9-9
Paramètres de l'index



Affinage des résultats de l'extraction

L'extraction est un processus itératif selon lequel vous pouvez extraire des résultats, les examiner, les modifier, puis procéder à une nouvelle extraction pour les mettre à jour. Etant donné que la précision et la continuité sont essentielles à la réussite du Text Mining et de la catégorisation, le fait d'affiner les résultats de l'extraction dès le départ garantit qu'à chaque nouvelle extraction, vous obtenez exactement les mêmes résultats dans vos définitions de catégorie. De cette manière, l'attribution des documents et des enregistrements à vos catégories est plus précise et plus répétable.

Les résultats de l'extraction font office de blocs de construction pour les catégories. Lorsque vous créez des catégories à l'aide de ces résultats d'extraction, les documents et les enregistrements sont automatiquement attribués aux catégories s'ils contiennent du texte qui correspond à un ou plusieurs descripteurs de catégorie. Bien que vous puissiez commencer la catégorisation avant d'affiner les ressources linguistiques, il est utile d'examiner les résultats de l'extraction au moins une fois au préalable.

Durant l'examen des résultats, vous pouvez trouver des éléments pour lesquels vous souhaitez que le moteur du programme d'extraction se comporte de manière différente. Prenons les exemples suivants :

- **Synonymes non reconnus.** Supposons que vous trouviez plusieurs concepts qui, selon vous, sont synonymes, par exemple `intelligent`, `astucieux`, `brillant` et `ingénieux`, et qu'ils apparaissent tous en tant que concepts individuels dans les résultats d'extraction. Vous pouvez créer une définition de synonyme dans laquelle les concepts `astucieux`, `brillant` et `ingénieux` sont regroupés sous le concept cible `intelligent`. Cette action regroupe ainsi tous les concepts avec `intelligent` et la fréquence globale est alors supérieure. Pour plus d'informations, reportez-vous à la section Ajout de synonymes sur p. 156.

- **Modification du type des concepts.** Supposons que les concepts de vos résultats d'extraction apparaissent sous un type et vous souhaitez qu'ils soient affectés à un autre type. Dans un autre exemple, imaginez que vous trouviez 15 concepts relatifs aux légumes dans vos résultats d'extraction et que vous souhaitiez tous les ajouter à un nouveau type appelé `<Légume>`. Dans la plupart des langues, les concepts qui ne se trouvent dans aucune déclaration de type mais sont extraits du texte sont automatiquement dotés du type `<Unknown>` ; cependant, pour le texte en japonais ils sont automatiquement dotés du type <名詞>*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium. Vous pouvez ajouter des concepts aux types. Pour plus d'informations, reportez-vous à la section Ajout de concepts à des types sur p. 158.
- **Concepts non pertinents.** Supposons qu'un concept extrait présente une fréquence très élevée, ce qui signifie qu'il apparaît dans de nombreux enregistrements ou documents. Toutefois, vous considérez que ce concept n'est pas pertinent pour l'analyse. Vous pouvez l'exclure de l'extraction. Pour plus d'informations, reportez-vous à la section Exclusion de concepts de l'extraction sur p. 160.
- **Correspondances incorrectes.** Supposons que lors de l'examen des enregistrements ou des documents qui contiennent un certain concept, vous découvriez que deux mots ont été regroupés de façon incorrecte, par exemple `faculté` et `facilité`. Cette correspondance peut être due à un algorithme interne, connu sous le nom de regroupement flou, qui ignore temporairement les consonnes et voyelles doubles ou triples, de façon à regrouper les fautes d'orthographe courantes. Vous pouvez ajouter ces mots à une liste de paires de mots qui ne doivent pas être regroupés. Pour plus d'informations, reportez-vous à la section Regroupement flou dans le chapitre 18 sur p. 328. Le regroupement flou n'est pas disponible pour le texte en japonais.
- **Concepts non extraits.** Supposons que vous remarquiez, au moment de l'examen du texte du document ou de l'enregistrement, que quelques mots ou expressions n'ont pas été extraits alors que vous vous attendiez à ce qu'ils le soient. Souvent, ces mots sont des verbes ou des adjectifs qui ne présentent pas d'intérêt pour vous. Cependant, il arrive que vous souhaitiez utiliser un mot ou une expression, qui n'a pas été extrait, comme partie intégrante d'une définition de catégorie. Pour extraire le concept, vous pouvez imposer un terme dans une déclaration de types. Pour plus d'informations, reportez-vous à la section Extraction de mots imposée sur p. 161.

Il est possible d'exécuter un grand nombre de ces modifications directement à partir du panneau Résultats d'extraction, du panneau Données, de la boîte de dialogue Définitions de catégorie ou de la boîte de dialogue Définitions du cluster en sélectionnant un ou plusieurs éléments et en cliquant avec le bouton droit de la souris pour accéder aux menus contextuels.

Une fois les modifications apportées, la couleur d'arrière-plan du panneau change pour indiquer que vous devez procéder à une nouvelle extraction pour visualiser ces modifications. Pour plus d'informations, reportez-vous à la section Extraction de données sur p. 143. Si vous travaillez avec des ensembles de données volumineux, il peut s'avérer plus efficace de procéder à une nouvelle extraction après plusieurs modifications plutôt qu'après chaque modification.

Remarque : vous pouvez visualiser l'ensemble complet des ressources linguistiques modifiables utilisées pour produire les résultats de l'extraction dans la vue de l'Editeur de ressources (Affichage > Editeur de ressources). Dans cette vue, les ressources apparaissent sous la forme de bibliothèques et de dictionnaires. Vous pouvez personnaliser les concepts et les types directement au sein des

bibliothèques et des dictionnaires. Pour plus d'informations, reportez-vous à la section Utilisation des bibliothèques dans le chapitre 16 sur p. 290.

Ajout de synonymes

Les **synonymes** sont des mots ayant le même sens. Les synonymes sont souvent utilisés pour regrouper des termes et leurs abréviations, ou pour réunir les mots fréquemment mal orthographiés sous la version correcte du mot. Grâce à l'utilisation de synonymes, la fréquence du concept cible est plus élevée, ce qui facilite largement la recherche d'informations similaires qui se présentent sous différentes formes dans vos données textuelles.

Les modèles et bibliothèques de ressources linguistiques fournis avec le produit contiennent de nombreux synonymes prédéfinis. Néanmoins, si vous découvrez des synonymes non reconnus, vous pouvez les définir de façon à ce qu'ils soient reconnus lors de la prochaine extraction.

La première étape consiste à décider du concept cible, ou concept principal. Le **concept cible** est l'expression ou le mot sous lequel vous souhaitez regrouper tous les termes synonymes dans les résultats finaux. Au cours de l'extraction, les synonymes sont regroupés sous ce concept cible. La deuxième étape consiste à identifier tous les synonymes de ce concept. Le concept cible vient remplacer tous les synonymes dans l'extraction finale. Pour être un synonyme, un terme doit être extrait. En revanche, il n'est pas nécessaire que le concept cible soit extrait pour que la substitution se produise. Par exemple, si vous souhaitez que le terme `astucieux` soit remplacé par `intelligent`, alors `astucieux` est le synonyme et `intelligent` est le concept cible.

Si vous créez une définition de synonyme, un nouveau concept cible est ajouté au dictionnaire. Vous devez ensuite ajouter des synonymes à ce concept cible. Lorsque vous créez ou éditez des synonymes, ces modifications sont enregistrées dans les dictionnaires de synonymes de l'Editeur de ressources. Pour visualiser la totalité du contenu de ces dictionnaires de synonymes ou pour apporter un nombre important de modifications, travaillez plutôt directement dans l'Editeur de ressources. Pour plus d'informations, reportez-vous à la section Dictionnaires des substitutions/synonymes dans le chapitre 17 sur p. 315.

Les nouveaux synonymes sont automatiquement stockés dans la première bibliothèque répertoriée dans l'arborescence de bibliothèques de la vue de l'Editeur de ressources—par défaut, il s'agit de la *bibliothèque locale*.

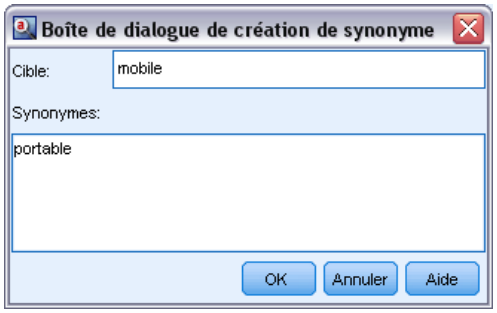
Remarque : si vous recherchez une définition de synonyme et que vous ne la trouvez ni via les menus contextuels ni directement dans l'Editeur de ressources, une correspondance a peut-être été obtenue à partir d'une technique interne de regroupement flou. Pour plus d'informations, reportez-vous à la section Regroupement flou dans le chapitre 18 sur p. 328.

Pour créer un synonyme

- Dans le panneau Résultats d'extraction, dans le panneau Données, la boîte de dialogue Définitions de catégorie ou la boîte de dialogue Définitions du cluster, sélectionnez les concepts pour lesquels vous souhaitez créer un nouveau synonyme.

- Dans les menus, choisissez Edition > Ajouter au synonyme > Nouveau. La boîte de dialogue Créer un synonyme apparaît.

Figure 9-10
Boîte de dialogue Créer un synonyme



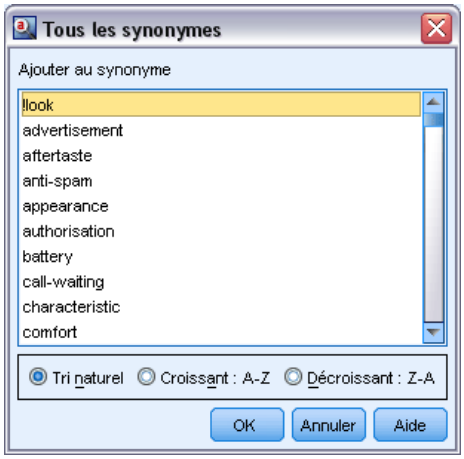
- Entrez un concept cible dans la zone de texte Cible. Il s'agit du concept sous lequel tous les synonymes seront regroupés.
- Si vous souhaitez ajouter davantage de synonymes, entrez-les dans la zone de liste Synonymes. Utilisez le séparateur global pour séparer chaque terme synonyme. Pour plus d'informations, reportez-vous à la section Options : onglet Session dans le chapitre 8 sur p. 131.
- Si vous travaillez avec du texte en japonais, choisissez un type pour ces synonymes en sélectionnant le nom du type dans le champ Synonymes du type. Cependant, la cible prend le type attribué pendant l'extraction. Mais si la cible n'a pas été extraite en tant que concept, alors le type répertorié dans cette colonne est affecté à la cible dans les résultats de l'extraction.

  Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.
- Cliquez sur OK pour appliquer les modifications. La boîte de dialogue se ferme et la couleur d'arrière-plan du panneau Résultats d'extraction change, indiquant que vous devez procéder à une nouvelle extraction pour visualiser les modifications. Si vous prévoyez plusieurs modifications, apportez-les avant de procéder à une nouvelle extraction.

Pour ajouter des éléments à un synonyme

- Dans le panneau Résultats d'extraction, dans le panneau Données, la boîte de dialogue Définitions de catégorie ou la boîte de dialogue Définitions du cluster, sélectionnez les concepts à ajouter à une définition de synonyme existante.
- Dans les menus, choisissez Edition > Ajouter au synonyme > . Le menu affiche un ensemble de synonymes, plaçant en début de liste les synonymes créés le plus récemment. Sélectionnez le nom du synonyme auquel vous souhaitez ajouter les concepts sélectionnés. Si vous trouvez le synonyme recherché, sélectionnez-le ; les concepts sélectionnés sont ajoutés à cette définition de synonyme. Si vous ne le trouvez pas, sélectionnez Plus pour afficher la boîte de dialogue Tous les synonymes.

Figure 9-11
Boîte de dialogue Tous les synonymes



- Dans la boîte de dialogue Tous les synonymes, vous pouvez trier la liste selon l'ordre de tri naturel (ordre de création), ou selon l'ordre croissant ou décroissant. Sélectionnez le nom du synonyme auquel vous souhaitez ajouter les concepts sélectionnés et cliquez sur OK. La boîte de dialogue se ferme et les concepts sont ajoutés à la définition de synonyme.

Ajout de concepts à des types

Durant chaque processus d'extraction, les concepts extraits sont affectés à des types de façon à regrouper les termes qui présentent un élément commun. IBM® SPSS® Modeler Text Analytics est livré avec de nombreux types intégrés. Pour plus d'informations, reportez-vous à la section Types intégrés dans le chapitre 17 sur p. 306.Dans la plupart des langues, les concepts qui ne se trouvent dans aucune déclaration de type mais sont extraits du texte sont automatiquement dotés du type `<Unknown>` ; cependant, pour le texte en japonais ils sont automatiquement dotés du type <名詞>*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Lors de l'analyse de vos résultats, vous pouvez découvrir que certains concepts apparaissent dans un type que vous voulez attribuer à un autre ou qu'un groupe de mots appartient en fait à un nouveau type propre. Dans ces cas de figure, vous pouvez réattribuer les concepts à un autre type ou encore créer un type. Vous ne pouvez pas créer de nouveaux types pour du texte en japonais.

Par exemple, supposons que vous travaillez avec des données d'enquête liées aux automobiles et que vous souhaitez procéder à une catégorisation en vous centrant sur différentes parties des véhicules. Vous pouvez créer un type appelé `<Tableau de bord>` pour regrouper tous les concepts liés aux indicateurs et boutons situés sur le tableau de bord des véhicules. Ensuite, vous pouvez attribuer des concepts tels que `jauge de carburant`, `chauffage`, `radio` et `compteur kilométrique` à ce nouveau type.

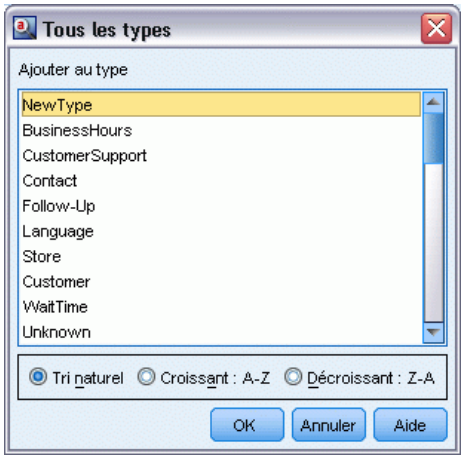
Autre exemple : supposons que vous travaillez avec des données d'enquête liées aux universités et que l'extraction a défini `Jean Moulin` (l'université) en tant que type `<Person>` plutôt qu'en tant que type `<Organization>`. Dans ce cas, vous pouvez ajouter ce concept au type `<Organization>`.

Lorsque vous créez un type ou ajoutez des concepts à une liste de termes d'un type, ces modifications sont enregistrées dans des déclarations de types dans les bibliothèques de ressources linguistiques de l'Editeur de ressources. Pour visualiser le contenu de ces bibliothèques ou pour apporter un nombre important de modifications, travaillez plutôt directement dans l'Editeur de ressources. Pour plus d'informations, reportez-vous à la section Ajout de termes dans le chapitre 17 sur p. 308.

Pour ajouter un concept à un type

- Dans le panneau Résultats d'extraction, dans le panneau Données, la boîte de dialogue Définitions de catégorie ou la boîte de dialogue Définitions du cluster, sélectionnez les concepts à ajouter à un type existant.
- Cliquez avec le bouton droit de la souris pour ouvrir le menu contextuel.
- Dans les menus, choisissez Edition > Ajouter au type >. Le menu affiche un ensemble de types, plaçant en début de liste les types créés le plus récemment. Sélectionnez le nom du type auquel vous souhaitez ajouter les concepts sélectionnés. Si vous trouvez le nom du type recherché, sélectionnez-le ; les concepts sélectionnés sont ajoutés à ce type. Si vous ne le trouvez pas, sélectionnez Plus pour afficher la boîte de dialogue Tous les types.

Figure 9-12
Boîte de dialogue Tous les types



- Dans la boîte de dialogue Tous les types, vous pouvez trier la liste selon l'ordre de tri naturel (ordre de création) ou selon l'ordre croissant ou décroissant. Sélectionnez le nom du type auquel vous souhaitez ajouter les concepts sélectionnés et cliquez sur OK. La boîte de dialogue se ferme et les concepts sont ajoutés au type en tant que termes.

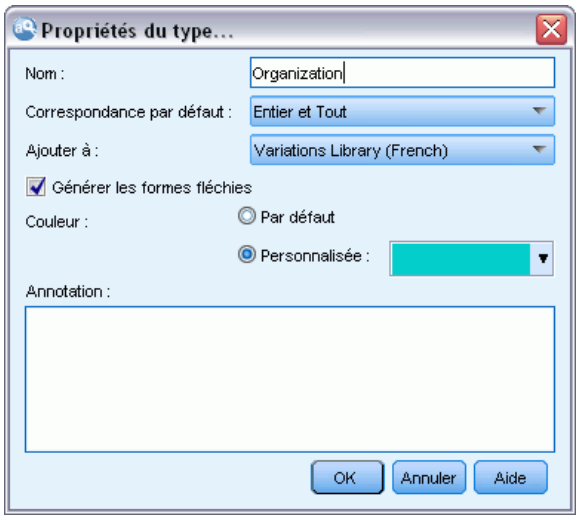
Remarque : Avec du texte en japonais, il existe des cas où la modification du type d'un terme ne modifiera pas le type auquel il sera finalement affecté dans la liste d'extraction finale. Cela est dû au fait que les dictionnaires internes ont priorité pendant l'extraction pour les termes de base.

Remarque : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Pour créer un type

- Dans le panneau Résultats d'extraction, dans le panneau Données, la boîte de dialogue Définitions de catégorie ou la boîte de dialogue Définitions du cluster, sélectionnez les concepts pour lesquels vous souhaitez créer un nouveau type.
- Dans les menus, choisissez Edition > Ajouter au type > Nouveau. La boîte de dialogue Propriétés de type apparaît.

Figure 9-13
Boîte de dialogue Propriétés de type



- Dans la zone de texte Nom, entrez le nom de ce nouveau type et apportez éventuellement des modifications aux autres champs. Pour plus d'informations, reportez-vous à la section Création de types dans le chapitre 17 sur p. 306.
- Cliquez sur OK pour appliquer les modifications. La boîte de dialogue se ferme et la couleur d'arrière-plan du panneau Résultats d'extraction change, indiquant que vous devez procéder à une nouvelle extraction pour visualiser les modifications. Si vous prévoyez plusieurs modifications, apportez-les avant de procéder à une nouvelle extraction.

Exclusion de concepts de l'extraction

Lors de l'examen des résultats, vous pouvez éventuellement découvrir des concepts dont vous ne souhaitiez pas l'extraction ou l'utilisation par une technique de création de catégorie automatisée. Dans certains cas, ces concepts présentent une fréquence très élevée et ne sont pas du tout pertinents pour votre analyse. Vous pouvez dès lors marquer un concept à exclure de l'extraction finale. En règle générale, les concepts que vous entrez dans cette liste sont des mots ou des expressions de liaison utilisés pour la continuité du texte, mais qui ne lui apportent rien d'important et risquent d'encombrer les résultats de l'extraction. En ajoutant ces concepts au dictionnaire d'exclusions, vous êtes assuré qu'ils ne seront jamais extraits.

Le processus d'exclusion implique que toutes les variantes du concept exclu disparaissent des résultats lors de la prochaine extraction. Si ce concept apparaît déjà en tant que descripteur dans une catégorie, il restera dans la catégorie avec un effectif nul après la nouvelle extraction.

Lorsque vous excluez, ces modifications sont enregistrées dans un dictionnaire d'exclusions dans l'Editeur de ressources. Pour visualiser toutes les définitions d'exclusion et les éditer directement, travaillez plutôt directement dans l'Editeur de ressources. Pour plus d'informations, reportez-vous à la section Dictionnaires d'exclusions dans le chapitre 17 sur p. 321.

Remarque : Avec du texte en japonais, il existe des cas où l'exclusion d'un terme ou d'un type ne l'exclura pas. Cela est dû au fait que les dictionnaires internes ont priorité pendant l'extraction pour les termes de base des ressources en japonais.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Pour exclure des concepts

- Dans le panneau Résultats d'extraction, dans le panneau Données, la boîte de dialogue Définitions de catégorie ou la boîte de dialogue Définitions du cluster, sélectionnez les concepts à exclure de l'extraction.
- Cliquez avec le bouton droit de la souris pour ouvrir le menu contextuel.
- Sélectionnez Exclure de l'extraction. Le concept est ajouté au dictionnaire d'exclusions dans l'Editeur de ressources et la couleur d'arrière-plan du panneau Résultats d'extraction change, indiquant que vous devez procéder à une nouvelle extraction pour visualiser les modifications. Si vous prévoyez plusieurs modifications, apportez-les avant de procéder à une nouvelle extraction.

  Remarque. Les mots exclus sont automatiquement stockés dans la première bibliothèque répertoriée dans l'arborescence de bibliothèques de l'Editeur de ressources—par défaut, il s'agit de la *bibliothèque locale*.

Extraction de mots imposée

Lors de l'examen des données textuelles dans le panneau Données après l'extraction, vous pouvez découvrir que certains mots ou expressions n'ont pas été extraits. Souvent, ces mots sont des verbes ou des adjectifs qui ne présentent pas d'intérêt pour vous. Cependant, il arrive que vous souhaitiez utiliser un mot ou une expression, qui n'a pas été extrait, comme partie intégrante d'une définition de catégorie.

Si vous souhaitez que ces mots et expressions soient extraits, vous pouvez imposer un terme dans une bibliothèque de types. Pour plus d'informations, reportez-vous à la section Ajout des termes forcés dans le chapitre 17 sur p. 312.

Important ! Le marquage d'un terme dans un dictionnaire comme étant imposé n'est pas infaillible. En effet, même si vous avez explicitement ajouté un terme à un dictionnaire, il arrive qu'il n'apparaisse pas dans le panneau Résultats d'extraction après la nouvelle extraction ou qu'il apparaisse bien, mais pas exactement tel que vous l'avez déclaré. Bien que cet événement soit rare, il peut avoir lieu lorsqu'un mot ou une expression a déjà été extrait dans le cadre d'une expression plus longue. Pour éviter cela, appliquez l'option de mise en correspondance Entier (pas de composés) à ce terme dans la déclaration de types. Pour plus d'informations, reportez-vous à la section Ajout de termes dans le chapitre 17 sur p. 308.

Chapitre
10
Catégorisation des données textuelles

Dans la vue Catégories et concept, vous pouvez créer des **catégories** qui représentent, essentiellement, des rubriques ou des concepts de niveau supérieur qui captureront les principales idées, connaissances et attitudes exprimées dans le texte.

En ce qui concerne la version de IBM® SPSS® Modeler Text Analytics 14, les catégories peuvent également avoir une structure hiérarchique, ce qui signifie qu'elles peuvent contenir des sous-catégories et que ces sous-catégories peuvent elles-même contenir des sous-catégories, et ainsi de suite. Vous pouvez importer des structures de catégories prédéfinies, nommées auparavant plans de codage, avec des catégories hiérarchiques ou bien créer ces catégories hiérarchiques à l'intérieur du produit.

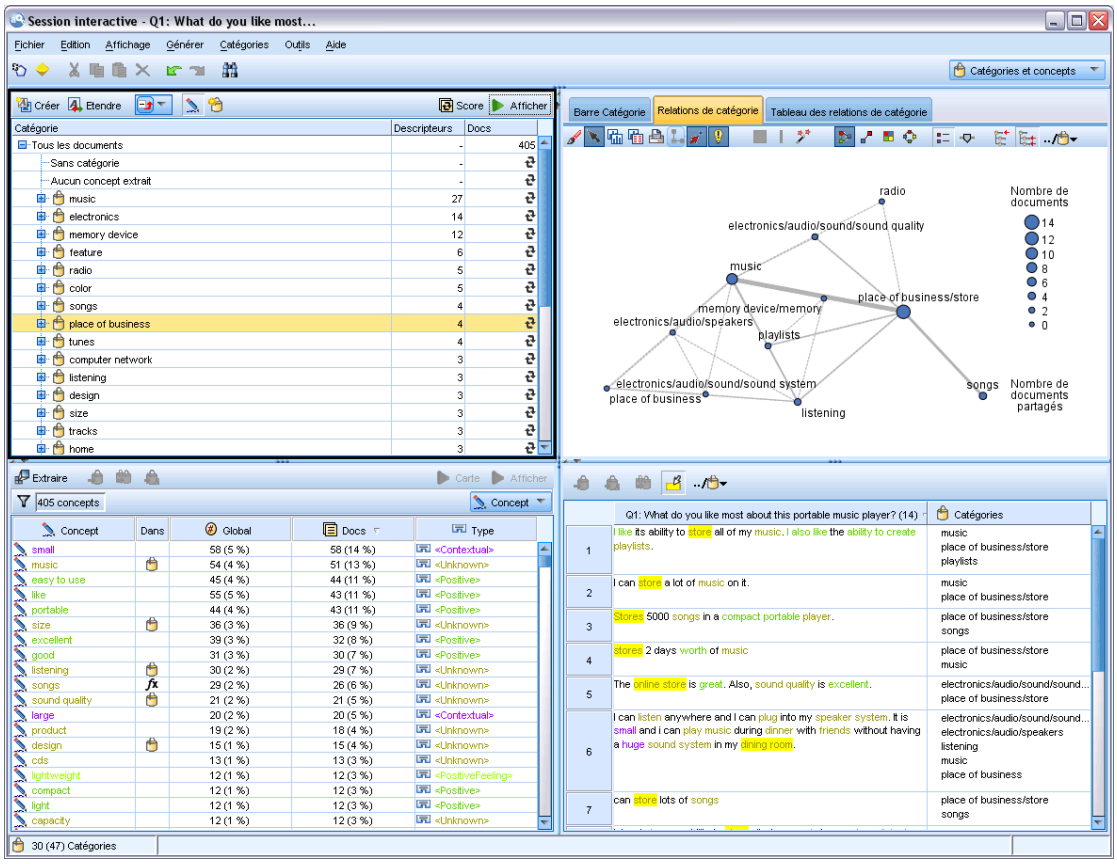
En effet, les catégories hiérarchiques vous permettent de créer une structure en arborescence contenant une ou plusieurs sous-catégories afin d'y regrouper plus clairement les éléments par groupe de concepts ou de rubriques. Prenons un exemple simple avec des activités de loisir ; pour répondre à une question telle que *Quelle activité aimeriez-vous faire si vous aviez davantage de temps ?* vous pourriez avoir des catégories principales comme *sport*, *art et artisanat*, *pêche*, etc. ; à un niveau inférieur, sous *sport*, vous pourriez avoir des sous-catégories pour voir s'il s'agit de *jeux de ballon*, *sports nautiques*, etc.

Les catégories sont constituées d'un ensemble de descripteurs, tels que des *concepts*, des *types*, des *patrons* et des *règles de catégorie*. Ensemble, ces descripteurs permettent d'identifier si un document ou un enregistrement appartient ou non à une catégorie. Le texte d'un document ou d'un enregistrement peut être analysé afin de déterminer s'il correspond à un descripteur. Si une correspondance est détectée, le document/l'enregistrement est attribué à cette catégorie. Ce processus est appelé **catégorisation**.

Vous pouvez créer des catégories, les utiliser et les explorer à l'aide des données affichées dans les quatre panneaux de la vue Catégories et concepts, que vous pouvez masquer ou afficher en sélectionnant leur nom dans le menu Affichage.

- **Panneau Catégories.** Dans ce panneau, créez et gérez vos catégories. Pour plus d'informations, reportez-vous à la section Le panneau Catégories sur p. 164.
- **Panneau Résultats d'extraction.** Dans ce panneau, explorez et utilisez les concepts et les types extraits. Pour plus d'informations, reportez-vous à la section Résultats d'extraction : concepts et types dans le chapitre 9 sur p. 139.
- **Panneau Visualisation.** Dans ce panneau, explorez visuellement vos catégories et analysez la manière dont elles interagissent. Pour plus d'informations, reportez-vous à la section Graphiques et diagrammes de catégorie dans le chapitre 13 sur p. 255.
- **Panneau Données.** Dans ce panneau, explorez et passez en revue le texte contenu dans les documents et les enregistrements qui correspondent aux sélections effectuées. Pour plus d'informations, reportez-vous à la section Le panneau Données sur p. 174.

Figure 10-1
Vue Catégories et concepts



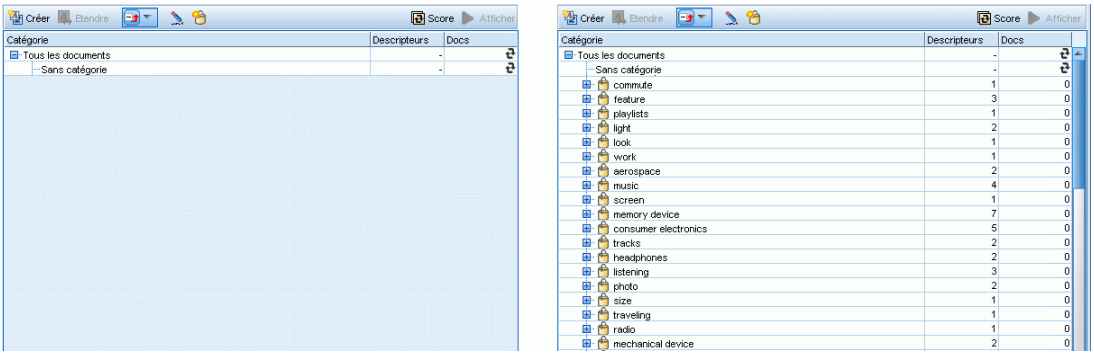
Vous pouvez commencer avec un ensemble de catégories provenant d'un package d'analyse de texte (TAP), ou importer depuis un fichier de catégories prédéfinies, vous pouvez également avoir besoin de créer votre propre ensemble. Les catégories peuvent être créées automatiquement à l'aide des techniques fiables et automatisées qui utilisent les résultats d'extraction (concepts, types et patrons) pour générer des catégories et leurs descripteurs. Les catégories peuvent aussi être créées manuellement en utilisant des informations supplémentaires que vous pouvez avoir au sujet des données. Toutefois, vous pouvez uniquement créer des catégories manuellement ou les affiner via la session interactive. Pour plus d'informations, reportez-vous à la section Nœud de Text Mining : onglet Modèle dans le chapitre 3 sur p. 37. Vous pouvez également créer des définitions de catégorie manuellement en faisant glisser les résultats de l'extraction dans les catégories. Vous pouvez enrichir ces catégories ou une catégorie vide par l'ajout de règles de catégorie à une catégorie, par l'utilisation de vos propres catégories prédéfinies ou par une combinaison.

Chaque technique convient à un type de données et à certaines situations. Cependant, il est souvent judicieux de combiner plusieurs méthodes dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. En cours de catégorisation, vous pouvez également envisager d'apporter d'autres modifications aux ressources linguistiques.

Le panneau Catégories

Le panneau Catégories est la zone dans laquelle vous pouvez créer et gérer vos catégories. Ce panneau est situé en haut à gauche de la vue Catégories et concepts. Après avoir extrait les concepts et les types de vos données textuelles, vous pouvez créer des catégories automatiquement à l'aide de technique telles que l'inclusion de concept, la co-occurrence, etc. ou vous pouvez les créer manuellement. Pour plus d'informations, reportez-vous à la section Création de catégories sur p. 177.

Figure 10-2
panneau Catégories avec et sans catégories



Chaque fois qu'une catégorie est créée ou mise à jour, les documents ou enregistrements peuvent être évalués en cliquant sur le bouton Score pour déterminer si le texte correspond à un descripteur dans une catégorie donnée. Si une correspondance est détectée, le document oul'enregistrement est attribué à cette catégorie. Le résultat final est que la plupart, sinon l'intégralité, des documents ou des enregistrements est affectée à des catégories en fonction des descripteurs des catégories.

Table d'arborescence des catégories

La table d'arborescence de ce panneau présente l'ensemble de catégories, des sous-catégories et des descripteurs. L'arborescence possède également plusieurs colonnes présentant des informations pour chaque élément de l'arborescence. Les colonnes pouvant être affichées sont les suivantes :

- **Code.** Répertorie la valeur de code de chaque catégorie. Cette colonne est sélectionnée par défaut. Vous pouvez afficher cette colonne avec les menus : Affichage > Panneau Catégories.
- **Catégorie.** Contient l'arborescence des catégories avec le nom de la catégorie et des sous-catégories. De plus si l'on clique sur l'icône de la barre d'outils des descripteurs, l'ensemble des descripteurs sera aussi affiché.
- **Descripteurs.** Donne le nombre de descripteurs qui la définissent. Ce nombre ne comprend pas le nombre de descripteurs dans les sous-catégories. Aucun nombre n'est donné lorsqu'un nom de descripteur apparaît dans la colonne Catégories. Vous pouvez afficher ou masquer les descripteurs eux-mêmes dans l'arborescence à l'aide des menus : Affichage > Panneau Catégories > Tous les descripteurs.
- **Docs.** Après le scoring, cette colonne affiche le nombre de documents ou d'enregistrements qui sont catégorisés dans une catégorie et toutes ses sous-catégories. Ainsi, si 5 enregistrements correspondent à votre catégorie de niveau supérieur en fonction de ses descripteurs et 7 autres

enregistrements correspondent à une sous-catégorie en fonction de ses descripteurs, le nombre de documents total pour la catégorie de niveau supérieur est la somme des deux : dans ce cas, ce nombre est de 12. Mais, si le même enregistrement correspondait à la catégorie de niveau supérieure et à sa sous-catégorie, alors ce nombre serait 11.

S'il n'existe aucune catégorie, le tableau contient tout de même deux lignes. La première ligne, Tous les documents, correspond au nombre total de documents ou d'enregistrements. Une seconde ligne, Sans catégorie, représente le nombre de documents/d'enregistrements devant être catégorisés.

Pour chaque catégorie présente dans le panneau, une icône représentant un petit seau jaune précède le nom de la catégorie. Si vous double-cliquez sur une catégorie ou choisissez Affichage > Définitions des catégories dans les menus, la boîte de dialogue Définitions des catégories s'ouvre et présente tous les éléments, appelés **descripteurs**, qui la définissent, comme les concepts, les types, les patrons et les règles de catégorie. Pour plus d'informations, reportez-vous à la section A propos des catégories sur p. 172. Par défaut, la table d'arborescence des catégories n'affiche pas les descripteurs dans les catégories. Si vous souhaitez voir les descripteurs directement dans l'arborescence plutôt que dans la boîte de dialogue Définitions de catégories, cliquez sur le bouton bascule à l'aide de l'icône du stylo dans la barre d'outils. Lorsque le bouton bascule est sélectionné, vous pouvez développer votre arborescence pour afficher également les descripteurs.

Scoring des catégories

La colonne Documents. dans la table d'arborescence des catégories affiche le nombre de documents ou d'enregistrements qui sont catégorisés dans cette catégorie spécifique. Si les nombres sont périmés ou ne sont pas calculés, une icône apparaît dans cette colonne. Vous pouvez cliquer sur Scorer dans la barre d'outils du panneau pour recalculer le nombre de documents. Gardez à l'esprit que le processus de scoring peut prendre un certain temps lorsque vous utilisez des ensembles de données volumineux.

Sélection des catégories dans l'arborescence

Lorsque vous effectuez des sélections dans l'arborescence, vous ne pouvez sélectionner que des catégories Frère, c'est-à-dire, si vous sélectionnez des catégories de niveau supérieur, vous ne pouvez pas également sélectionner une sous-catégorie. Ou si vous sélectionnez 2 sous-catégories d'une catégorie donnée, vous ne pouvez pas sélectionner en même temps une sous-catégorie d'une autre catégorie. La sélection d'une catégorie discontinue provoquera la perte de votre sélection précédente.

Affichage dans les panneaux Données et Visualisation

Lorsque vous sélectionnez une ligne dans le tableau, vous pouvez cliquer sur le bouton Afficher pour que les panneaux Visualisation et Données soient actualisés avec les informations correspondant à votre sélection. Si un panneau n'est pas visible, le fait de cliquer sur Afficher l'ouvre.

Réglage de vos catégories

La catégorisation peut ne pas générer des résultats parfaits pour vos données lors de votre première tentative, et il se peut que vous souhaitiez supprimer des catégories ou les combiner avec d'autres catégories. Vous pouvez également vous apercevoir, en consultant les résultats de l'extraction,

que certaines catégories n'ayant pas été créées vous seraient utiles. Dans ce cas, vous pouvez apporter des modifications manuelles aux résultats afin de les adapter à votre contexte. Pour plus d'informations, reportez-vous à la section Edition et réglage des catégories sur p. 231.

Stratégies et méthodes de création de catégories

Si vous n'avez pas encore effectué d'extraction ou que vos résultats d'extraction ne sont pas à jour, l'utilisation de l'une des techniques de création ou d'extension de catégorie vous invitera automatiquement à effectuer une extraction. Après avoir appliqué une technique, les concepts et les types qui ont été regroupés dans une catégorie restent disponibles et pourront être classés par le biais d'autres techniques. Cela signifie que vous pouvez voir un concept dans plusieurs catégories sauf si vous choisissez de ne pas les réutiliser.

Afin de vous aider à créer les catégories les plus pertinentes, veuillez examiner les points suivants :

- **Méthodes de création de catégories**
- **Stratégies de création de catégories**
- **Conseils pour la création de catégories**

Méthodes de création de catégories

Puisque chaque ensemble de données est unique, le nombre de méthodes de création de catégories et l'ordre dans lequel vous les appliquez peut varier. De plus, dans la mesure où vos objectifs d'exploration de texte peuvent être différents d'un ensemble de données à un autre, vous pouvez tester les différentes méthodes afin de déterminer celle qui génère les meilleurs résultats pour les données textuelles en question. Aucune des techniques automatiques ne pourra parfaitement catégoriser vos données; c'est pourquoi nous recommandons de trouver et d'appliquer une ou plusieurs techniques automatiques qui correspondent aux besoins de vos données.

Outre l'utilisation de packages d'analyse de texte (TAP,*.*tap*) avec des ensembles de catégories prédéfinis, vous pouvez également catégoriser vos réponses à l'aide de toute combinaison des méthodes suivantes :

- **Techniques de création automatique.** Plusieurs options de catégorie basées sur la linguistique et la fréquence sont disponibles pour créer automatiquement des catégories. Pour plus d'informations, reportez-vous à la section Création de catégories sur p. 177.
- **Techniques d'extension automatiques.** Plusieurs techniques linguistiques sont disponibles pour étendre les catégories existantes en ajoutant et en améliorant les descripteurs afin qu'ils capturent davantage d'enregistrements. Pour plus d'informations, reportez-vous à la section Extension de catégories sur p. 194.
- **Techniques manuelles.** Il existe plusieurs méthodes manuelles, telles que le glisser-déposer. Pour plus d'informations, reportez-vous à la section Création de catégories manuellement sur p. 199.

Stratégies de création de catégories

La liste de stratégies suivante n'est en aucun cas exhaustive mais elle peut vous donner une idée de l'approche de la création de vos catégories.

- Lorsque vous définissez le noeud de Text Mining, sélectionnez un ensemble de catégories dans un package d'analyse de texte (TAP) afin de débuter votre analyse avec des catégories prédéfinies. Il est possible que ces catégories suffisent à l'analyse de votre texte depuis le début. Cependant, si vous souhaitez ajouter d'autres catégories, vous pouvez modifier les paramètres de création de catégories (Catégories > Configurer les paramètres). Ouvrez la boîte de dialogue Paramètres avancés : Linguistique et choisissez l'option d'entrée Catégorie Résultats d'extraction inutilisés et créez les catégories supplémentaires.
- Lorsque vous définissez un noeud, sélectionnez un ensemble de catégories dans un TAP à partir de la vue Catégories et concepts de la session interactive. Ensuite, faites glisser et déposer les concepts ou patrons inutilisés dans les catégories qui vous semblent appropriées. Ensuite, étendez les catégories existantes que vous venez de modifier (Catégories > Etendre les catégories) pour obtenir davantage de descripteurs associés aux descripteurs de catégorie existants.
- Créez des catégories automatiquement à l'aide des paramètres linguistiques avancés (Catégories > Créer des catégories). Ensuite, affinez manuellement les catégories en supprimant des descripteurs, des catégories ou en fusionnant des catégories similaires jusqu'à ce que vous soyez satisfait des catégories résultantes. De plus, si vous créez des catégories **sans** utiliser l'option Généraliser avec des caractères génériques lorsque cela est possible, vous pouvez également essayer de simplifier automatiquement les catégories à l'aide de la fonctionnalité Etendre les catégories, en activant l'option Généraliser.
- Importez un fichier de catégories prédéfinies avec des noms de catégories très descriptifs et/ou des annotations. De plus, si à l'origine vous importez **sans** choisir l'option d'importer ou de générer des descripteurs à partir des noms de catégorie, vous pouvez ensuite utiliser la boîte de dialogue Etendre les catégories et choisir l'option Etendre les catégories vides avec des descripteurs générés à partir du nom de la catégorie. Puis, étendez ces catégories une deuxième fois, mais utilisez alors les techniques de regroupement.
- Créez manuellement un premier ensemble de catégories en triant les concepts ou patrons de concept par fréquence puis en faisant glisser et déposer les plus intéressants dans le panneau Catégories. Une fois que vous avez cet ensemble initial de catégories, utilisez la fonction Etendre (Catégories > Etendre les catégories) pour développer et affiner toutes les catégories sélectionnées afin qu'elles comprennent les autres descripteurs associés et correspondent ainsi à plus d'enregistrements.

Après avoir appliqué ces techniques, nous vous recommandons de passer en revue les catégories résultantes et d'utiliser des techniques manuelles pour effectuer des changements mineurs, supprimer toute mauvaise réaffectation ou ajouter les enregistrements ou mots ayant été oubliés. En outre, puisque l'utilisation de techniques différentes peut être à l'origine de catégories redondantes, vous avez la possibilité de fusionner ou de supprimer des catégories si nécessaire.
Pour plus d'informations, reportez-vous à la section Edition et réglage des catégories sur p. 231.

Conseils pour la création de catégories

Afin de vous aider à créer de meilleures catégories, vous pouvez examiner certains conseils qui peuvent vous aider à prendre des décisions à propos de votre approche.

Conseils concernant le rapport Catégorie-Document

Il est rare que les catégories auxquelles les documents et les enregistrements sont affectés s'excluent mutuellement dans l'analyse de texte qualitative, et ce, pour au moins deux raisons :

- En premier lieu, il est généralement admis que plus un document ou un enregistrement texte est long, plus les idées et opinions qu'il exprime sont diverses. Ainsi, un document ou un enregistrement a plus de chances de se voir attribuer plusieurs catégories.
- En second lieu, il existe souvent plusieurs manières de regrouper ou d'interpréter des documents ou des enregistrements texte qui ne sont pas logiquement distincts. Dans le cas d'une enquête comportant une question ouverte sur les opinions politiques de la personne interrogée, nous pourrions créer des catégories, telles que *libéral* et *conservateur*, ou *Républicain* et *Démocrate*, ainsi que des catégories plus spécifiques, telles que *libéral pour l'aspect social*, *conservateur pour l'aspect fiscal* et ainsi de suite. Il n'est pas nécessaire que ces catégories s'excluent mutuellement, ni qu'elles soient exhaustives.

Conseils sur le nombre de catégories à créer

La création de catégories doit découler directement des données : lorsque vous voyez un élément intéressant concernant vos données, vous pouvez créer une catégorie qui représente ces informations. En général, le nombre de catégories que vous pouvez créer n'est pas limité. En revanche, si vous créez trop de catégories, vous aurez des difficultés à les gérer. Deux principes sont applicables :

- **Fréquence des catégories.** Pour qu'une catégorie soit utile, elle doit contenir une quantité minimale de documents ou d'enregistrements. Un ou deux documents peuvent contenir des éléments intéressants. Cependant, si cela ne concerne qu'un ou deux documents sur 1 000, les informations qu'ils contiennent, peuvent ne pas être pas assez fréquentes par rapport à la quantité de documents pour présenter une utilité pratique.
- **Complexité.** Plus vous créez de catégories, plus vous avez d'informations à examiner et à résumer après analyse. Toutefois, un trop grand nombre de catégories risque de créer une trop grande complexité sans apporter de détails utiles.

Malheureusement, aucune règle ne détermine à partir de quel nombre les catégories sont jugées trop nombreuses, ni ne définit le nombre minimal d'enregistrements par catégorie. C'est à vous d'effectuer ces déterminations, en fonction des exigences de votre situation.

Nous pouvons toutefois vous fournir nos conseils et vous indiquer où commencer. Bien que le nombre de catégories ne doive pas être trop élevé, il est préférable d'en avoir trop que pas assez lors des premières étapes de l'analyse. Il est plus simple de regrouper des catégories relativement similaires que de diviser des observations en catégories nouvelles. La stratégie consistant à travailler d'un plus grand nombre à un moins grand nombre de catégories est donc généralement la meilleure. Etant donné la nature itérative du Text Mining et la facilité avec laquelle il peut être exécuté avec ce logiciel, il est acceptable de créer plus de catégories au départ.

Choix des meilleurs descripteurs

Les informations suivantes vous fourniront quelques conseils afin de choisir ou de créer les meilleurs descripteurs (concepts, types, patrons TLA et règles de catégorie) pour vos catégories. Les descripteurs sont les blocs de construction des catégories. Lorsqu'une partie ou l'ensemble du texte d'un document ou d'un enregistrement correspond à un descripteur, le document ou l'enregistrement est mis en correspondance avec la catégorie.

Un descripteur n'est mis en correspondance avec des documents ou des enregistrements que s'il contient ou correspond à un concept ou à un patron extrait. Par conséquent, utilisez les concepts, les types, les patrons et les règles de catégorie de la manière indiquée dans les paragraphes suivants.

Les concepts représentant un ensemble de termes sous-jacents (en plus de se représenter eux-mêmes) qui peuvent comprendre un ensemble vaste de termes allant des formes singulier/pluriel, aux synonymes, ou aux variations orthographiques, seul le concept lui-même doit être utilisé en tant que descripteur ou comme partie d'un descripteur. Pour en savoir plus sur les termes sous-jacents de chaque concept donné, cliquez sur le nom du concept dans le panneau Résultats d'extractionde la vue Catégories et concepts. Lorsque vous placez la souris sur le nom d'un concept, une info-bulle apparaît et affiche tous les termes sous-jacents trouvés dans votre texte lors de la dernière extraction. Les concepts ne possèdent pas tous des termes sous-jacents. Par exemple, si `voiture` et `véhicule` sont des synonymes mais que `voiture` est extrait comme concept et `véhicule` comme terme sous-jacent, alors vous devez utiliser `voiture` comme descripteur, puisqu'il permettra de mettre en correspondance le document ou les enregistrements qui contiennent également le terme `véhicule`.

Les concepts et les types utilisés comme descripteurs

Vous devez utiliser un concept comme descripteur lorsque vous souhaitez trouver tous les documents ou enregistrements contenant ce concept (ou ses termes sous-jacents). Dans ce cas, l'utilisation d'une règle de catégorie plus complexe n'est pas nécessaire puisque le nom du concept exact est suffisant. Souvenez-vous que lorsque vous utilisez des ressources qui extraient des opinions, il se peut que les concepts changent lors de l'extraction des patrons TLA, afin de capturer le sens exact de la phrase (reportez-vous à l'exemple de la section qui suit concernant les patrons TLA).

Par exemple, une réponse à une enquête indiquant les fruits préférés de chaque personne telles que "*Les pommes et l'ananas sont les meilleurs fruits*" peut engendrer l'extraction de `pommes` et `ananas`. En ajoutant le concept `pommes` en tant que descripteur à votre catégorie, toutes les réponses contenant le concept `pommes` (ou ses termes sous-jacents) sont mises en correspondance avec cette catégorie.

Toutefois, si vous voulez juste connaître les réponses qui mentionnent le terme *pommes* de quelque manière que ce soit, vous pouvez écrire une règle de catégorie, comme par exemple `* pommes *`, pour capturer toutes les réponses contenant des concepts tels que `pommes`, `jus de pommes`, ou `tarte aux pommes`.

Vous pouvez également capturer tous les documents ou enregistrements qui contiennent des concepts de même type en utilisant un type directement en tant que descripteur comme par exemple `<Fruit>`. Remarque : vous ne pouvez pas utiliser la fonction * avec les types.

Pour plus d'informations, reportez-vous à la section Résultats d'extraction : concepts et types dans le chapitre 9 sur p. 139.

Patrons d'analyse des liens du texte (TLA) utilisés comme descripteurs

Utilisez un résultat de patron TLA comme descripteur si vous souhaitez capturer des idées plus nuancées ou plus fines. Lorsque le texte est analysé lors de l'extraction TLA, le texte (document ou enregistrement) est traité phrase par phrase ou proposition par proposition, plutôt que dans son ensemble. En prenant en considération l'ensemble des parties d'une phrase, l'analyse des liens du texte peut identifier les opinions, les relations entre deux éléments, ou une négation, par exemple et comprendre ainsi le sens exact de la phrase. Vous pouvez utiliser des patrons de concepts ou des patrons de type en tant que descripteurs. Pour plus d'informations, reportez-vous à la section Patrons de type et Patrons de concept dans le chapitre 12 sur p. 247.

Par exemple, si l'on considère le texte "*la pièce n'était pas propre*", les concepts suivants peuvent en être extraits : `pièce` et `propre`. Toutefois, si l'extraction TLA avait été activée dans les paramètres de l'extraction, elle aurait pu détecter que `propre` est utilisé de manière négative et correspond en fait au concept `pas propre`, qui est aussi synonyme du concept `sale`. Vous pouvez voir dans cet exemple que l'utilisation du concept `propre` seul, en tant que descripteur sera mis en correspondance avec ce texte, mais capturera aussi d'autres documents ou enregistrements qui mentionnent la propreté. Par conséquent, il peut être plus judicieux d'utiliser le patron de concept TLA `sale` comme concept de sortie, car il sera mis en correspondance avec ce texte et sera plus approprié en tant que descripteur.

Règles métier de catégories utilisées comme descripteurs

Les règles de catégorie sont des instructions qui classifient automatiquement les documents ou les enregistrements en une catégorie basée sur une expression logique à l'aide de concepts, de types et de patrons extraits ou d'opérateurs booléens. Par exemple, vous pouvez créer une expression qui signifie *inclure tous les enregistrements contenant le concept extrait* `ambassade` *mais ne pas inclure* `argentine`*dans cette catégorie*.

Vous pouvez créer et utiliser des règles de catégorie en tant que descripteurs dans vos catégories pour exprimer différentes idées à l'aide de `&`, `|`, et `!()` Booléens. Pour obtenir des informations sur la syntaxe de ces règles et la manière de les rédiger et de les modifier, consultez Utilisation des règles de catégorie sur p. 201

- Utilisez une règle de catégorie avec l'opérateur booléen `&` (AND) pour trouver des documents ou des enregistrements dans lesquels se trouvent 2 concepts ou plus. Les concepts connectés par des opérateurs `&` ne doivent pas obligatoirement se trouver dans la même phrase, ils peuvent apparaître n'importe où dans le même document ou enregistrement pour être mis en correspondance avec la catégorie. Par exemple, si vous créez la règle de catégorie suivante : `nourriture & bon marché` comme descripteur, elle correspondra à un enregistrement comportant le texte "*la nourriture était très chère, mais les chambres étaient bon marché*" bien que le terme `nourriture` ne soit pas le nom auquel s'applique l'adjectif `bon marché`, et ce car le texte contient à la fois les termes `nourriture` et `bon marché`.
- Utilisez une règle de catégorie comportant l'opérateur booléen `!()` (NOT) comme descripteur pour trouver des documents ou des enregistrements dans lesquels se trouvent certains termes mais pas d'autres Ceci peut permettre de ne pas regrouper des informations qui semblent être reliées en fonction des mots mais pas en fonction du contexte. Par exemple, si vous créez la règle de catégorie suivante : `<Société> & !(ibm)` comme descripteur, elle correspondra au texte suivant : *SPSS Inc. est une société fondée en 1967* et ne correspondra pas au texte suivant : *la société de logiciels a été achetée par IBM.*.

- Utilisez une règle de catégorie avec l'opérateur booléen (OR) pour trouver des documents ou des enregistrements contenant un concept ou type parmi plusieurs. Par exemple, si vous créez la règle de catégorie suivante : `(personnel|équipe|employés|collaborateurs) & mauvais` comme descripteur, elle correspondra à tout document ou enregistrement dans lequel l'un de ces noms de trouvent en même temps que le concept de `mauvais`.
- Utilisez les types dans les règles de catégorie afin de les rendre plus génériques et plus déployables. Par exemple, si vous travaillez sur des données hôtelières, il se peut que vous soyez intéressé par l'opinion des clients sur le personnel de l'hôtel. Des termes associés peuvent inclure des mots comme réceptionniste, serveur, serveuse, réception, etc... Dans ce cas, vous pouvez créer un nouveau type nommé `<PersonnelHôtel>` et lui ajouter tous les termes précédents. Alors qu'il est possible de créer une règle de catégorie pour chaque type d'employés tels que `[* serveuse* & aimable]`, `[* réception* & amical]`, `[* réceptionniste * & accueillant]`, vous pouvez également créer une règle de catégorie unique, plus générale à l'aide du type `<PersonnelHôtel>` pour capturer toutes les réponses contenant des opinions positives sur le personnel de l'hôtel sous la forme de `[<PersonnelHôtel> & <Positif>]`.

Remarque : Vous pouvez utiliser à la fois + et `&` dans les règles de catégorie lorsque des patrons TLA sont inclus dans ces règles. Pour plus d'informations, reportez-vous à la section Utilisation des patrons TLA dans les règles de catégorie sur p. 203.

Exemple des différences de mise en correspondance lors de l'utilisation des concepts, des patrons TLA ou des règles de catégories comme descripteurs.

L'exemple suivant montre la manière dont l'utilisation des concepts, des règles de catégorie ou des patrons TLA en tant que descripteurs affecte la classification des documents ou des enregistrements en catégories. Supposons que vous ayez les cinq enregistrements suivants :

- A : *"personnel du restaurant fantastique, nourriture excellente et chambres propres et confortables."*
- B : *"le personnel du restaurant était horrible, mais les chambres étaient propres."*
- C : *"Chambres confortables et propres."*
- D : *"Ma chambre n'était pas très propre."*
- E : *"Propre."*

Dans la mesure où les enregistrements contiennent le mot *propre* et que vous souhaitez capturer cette information, vous pouvez créer un des descripteurs indiqué dans le tableau ci-dessous. En fonction de la nature des informations que vous essayez de capturer, vous pouvez utiliser un type de descripteur et voir les résultats produits.

Table 10-1
Mise en correspondance des exemples d'enregistrements avec les descripteurs

Descripteur	A	B	C	D	E	Explication
`propre`	*mise en correspondance*	*mise en correspondance*	*mise en correspondance*	*mise en correspondance*	*mise en correspondance*	Le descripteur est un concept extrait. Chaque enregistrement contient le concept `propre`, même l'enregistrement D puisque sans TLA, le concept "*pas propre*" ne signifie pas automatiquement `sale` dans les règles TLA.
`propre + .`	-	-	-	-	*mise en correspondance*	Le descripteur est un patron TLA qui représente `propre` lui-même. Ne correspond qu'à l'enregistrement dans lequel `propre` a été extrait sans concept associé lors de l'extraction TLA.
`[propre]`	*mise en correspondance*	*mise en correspondance*	*mise en correspondance*	-	*mise en correspondance*	Le descripteur est une règle de catégorie qui cherche une règle TLA contenant `propre` seul ou associé à autre chose. Correspond à tous les enregistrements dans lesquels une sortie TLA contenant `propre` a été trouvée, que `propre` soit lié à un autre concept, comme par exemple `chambre`, ou non et indépendamment de sa position dans la phrase.

A propos des catégories

Les **catégories** font référence à un groupe de concepts, d'opinions ou d'attitudes étroitement associés. Pour être utile, une catégorie doit également être décrite par une étiquette ou une phrase courte évoquant l'essentiel de sa signification.

Par exemple, si vous analysez des réponses de consommateurs à une enquête sur une nouvelle lessive, vous pouvez créer une catégorie étiquetée *odeur* qui contient toutes les réponses décrivant l'odeur de ce produit. Mais cette catégorie ne fera pas la différence entre ceux qui ont aimé l'odeur et ceux qui ne l'ont pas aimée. Dans la mesure où IBM® SPSS® Modeler Text Analytics permet d'extraire des opinions si vous utilisez les ressources appropriées, vous pouvez alors créer deux autres catégories qui identifieront les personnes interrogées qui *ont aimé l'odeur* et celles qui *n'ont pas aimé l'odeur*.

Vous pouvez créer et travailler avec vos catégories dans le panneau Catégories situé dans le panneau supérieur gauche de la fenêtre de la vue Catégories et concepts. Chaque catégorie est définie par un ou plusieurs descripteurs. Les **descripteurs** sont des concepts, des types, des patrons et des règles de catégorie, utilisés pour définir une catégorie.

Si vous souhaitez voir les descripteurs qui forment une catégorie donnée, vous pouvez cliquer sur l'icône du crayon dans la barre d'outils du panneau Catégories puis développer l'arborescence pour afficher les descripteurs. Vous pouvez également sélectionner la catégorie et ouvrir la boîte de dialogue Définitions des catégories (Vue > Définitions des catégories).

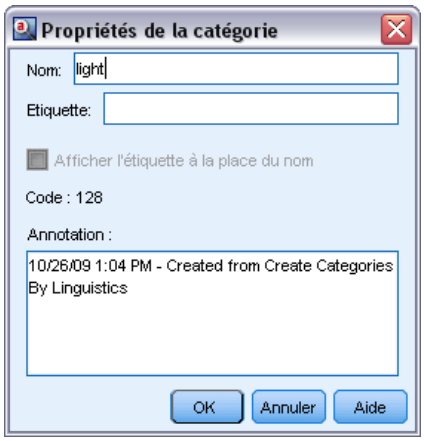
Lorsque vous créez automatiquement des catégories à l'aide des techniques de création, telles que l'inclusion de concept , les différentes techniques utiliseront les concepts et les types comme descripteurs pour créer vos catégories. Si vous extrayez des patrons TLA, ous pouvez ajouter ces

patrons ou une partie d'entre eux en tant que descripteurs de catégorie. Pour plus d'informations, reportez-vous à la section Exploration de l'analyse des liens du texte dans le chapitre 12 sur p. 245. Si vous créez des clusters, vous pouvez ajouter les concepts dans un cluster aux nouvelles catégories ou à des catégories existantes. Enfin, vous pouvez créer manuellement des règles de catégorie à utiliser en tant que descripteurs dans vos catégories. Pour plus d'informations, reportez-vous à la section Utilisation des règles de catégorie sur p. 201.

Propriétés de la catégorie

En plus des descripteurs, les catégories ont également des propriétés que vous pouvez éditer afin de renommer les catégories, d'ajouter une étiquette ou une annotation.

Figure 10-3
Boîte de dialogue Propriétés de catégorie



Les propriétés suivantes sont disponibles :

- **Nom.** Ce nom apparaît dans l'arborescence par défaut. Lorsqu'une catégorie est créée à l'aide d'une technique automatique, un nom lui est attribué automatiquement.
- **Etiquette.** L'utilisation d'étiquettes permet de créer des descriptions de catégorie plus significatives en vue de les utiliser dans d'autres produits, ou dans d'autres tableaux ou graphiques. Si vous sélectionnez l'option permettant d'afficher l'étiquette, celle-ci est alors utilisée dans l'interface pour désigner la catégorie.
- **Code.** Le numéro de code correspond à la valeur de code de cette catégorie. .
- **Annotations.** Vous pouvez ajouter une description courte pour chaque catégorie dans ce champ. Lorsqu'une catégorie est générée par la boîte de dialogue Créer des catégories, une note est ajoutée automatiquement à cette annotation. Vous pouvez aussi ajouter un échantillon de texte à une annotation directement dans le panneau Données en sélectionnant le texte et en sélectionnant Catégories > Ajouter à l'annotation dans le menu.

Le panneau Données

Lorsque vous créez des catégories, vous pouvez parfois souhaiter examiner certaines des données textuelles utilisées. Par exemple, si vous créez une catégorie dans laquelle 640 documents sont catégorisés, vous pouvez souhaiter consulter une partie ou l'intégralité de ces documents afin de découvrir le texte qui était en réalité rédigé. Vous pouvez consulter les enregistrements ou les documents dans le panneau Données, situé dans l'angle inférieur droit. S'il n'apparaît pas par défaut, sélectionnez Affichage> Panneaux > Données dans les menus.

Le panneau de données affiche une ligne par document ou enregistrement correspondant à la sélection dans le panneau Catégories, dans le panneau Résultats d'extraction ou dans la boîte de dialogue Définitions de catégorie jusqu'à une certaine limite d'affichage. Par défaut, le nombre de document oud'enregistrements affichés dans le panneau de données est limité pour vous permettre de consulter vos données plus rapidement. Cependant, vous pouvez modifier cette limite dans la boîte de dialogue Options. Pour plus d'informations, reportez-vous à la section Options : onglet Session dans le chapitre 8 sur p. 131.

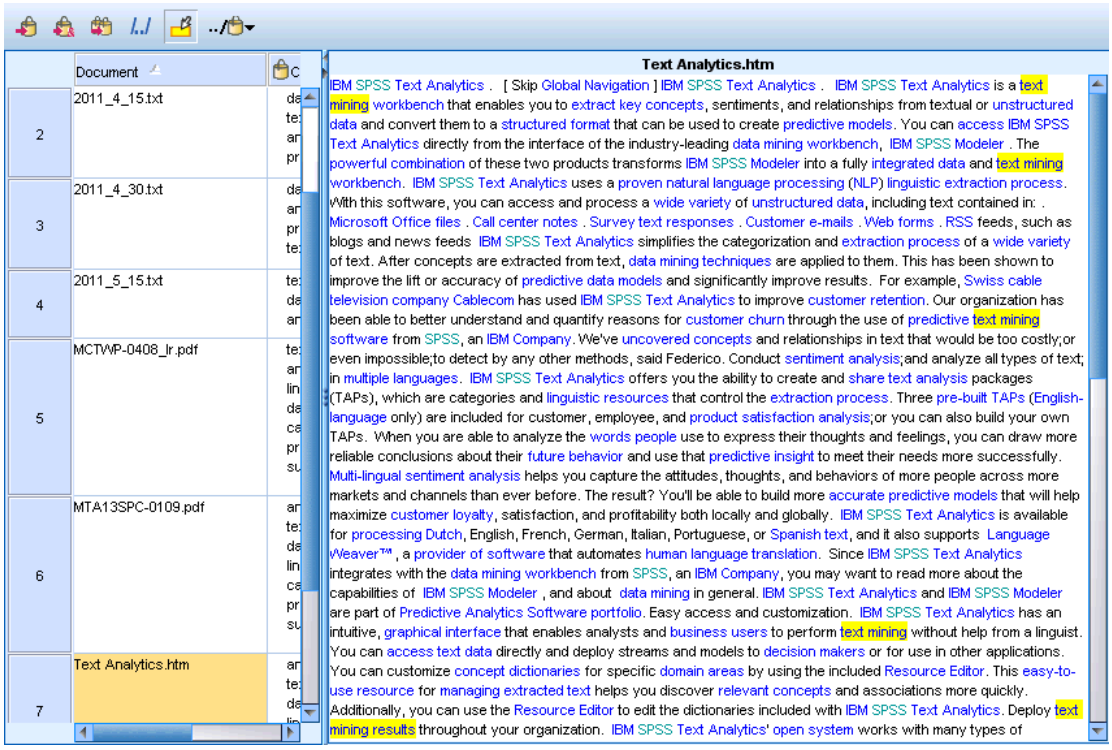
Affichage et actualisation du panneau Données

L'affichage du panneau Données n'est pas automatiquement actualisé car en présence d'ensembles de données volumineux, l'opération prendrait trop de temps. Par conséquent, chaque fois que vous effectuez une sélection dans un autre panneau de cette vue ou dans la boîte de dialogue Définitions de catégorie, cliquez sur Afficher pour actualiser le contenu du panneau Données.

Documents texte ou enregistrements

Si vos données textuelles sont sous la forme d'enregistrements et que le texte est relativement bref, le champ de texte du panneau Données affiche les informations dans leur intégralité. Cependant, si vous utilisez des enregistrements et de grands ensembles de données, la colonne du champ de texte affiche une petite partie du texte et ouvre un panneau Aperçu du texte à droite qui permet de consulter une plus grande partie du texte de l'enregistrement sélectionné dans la table, voire son intégralité. Si vos données textuelles se présentent sous la forme de documents, le panneau Données affiche le nom de fichier du document. Lorsque vous sélectionnez un document, le panneau Aperçu du texte s'ouvre et affiche le texte du document sélectionné.

Figure 10-4
Panneau de données avec panneau Aperçu du texte



Couleurs et mise en surbrillance

Chaque fois que vous affichez des données, des concepts et des descripteurs trouvés dans ces documents ou enregistrements, ils apparaissent en couleur pour vous permettre de les identifier facilement dans le texte. Le code couleur correspond aux types auxquels les concepts appartiennent. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher le concept sous lequel il a été extrait et le type auquel il a été affecté. Tout texte n'ayant pas été extrait apparaît en noir. En règle générale, ces mots non extraits sont souvent des connecteurs (*et* ou *avec*), des pronoms (*me* ou *ils*), et des verbes (*être*, *avoir* ou *prendre*).

Colonnes du panneau Données

Alors que la colonne de champ du texte est toujours visible, il est possible d'afficher également d'autres colonnes. Pour afficher d'autres colonnes, cliquez sur Affichage > Panneau Données dans les menus, puis sélectionnez la colonne que vous souhaitez afficher dans le panneau de données. Les colonnes pouvant être affichées sont les suivantes :

- **"Nom du champ de texte" (#)/Documents.** Ajoute une colonne pour les données textuelles à partir desquelles des concepts et des types ont été extraits. Si vos données sont contenues dans des documents, la colonne est appelée Documents, et seul le nom de fichier du document ou son chemin d'accès complet est visible. Pour examiner le texte de ces documents, vous devez consulter le panneau Aperçu du texte. Le nombre de lignes du panneau Données est indiqué entre parenthèses après le nom de cette colonne. Il peut arriver que les documents ou

les enregistrements ne soient pas tous affichés en raison d'une limite définie dans la boîte de dialogue Options pour optimiser la vitesse de chargement. Si la limite est atteinte, le nombre sera suivi de - Max. Pour plus d'informations, reportez-vous à la section Options : onglet Session dans le chapitre 8 sur p. 131.

- **Catégories.** Répertorie chacune des catégories à laquelle appartient un enregistrement. Lorsque cette colonne est affichée, l'actualisation du panneau Données peut prendre plus de temps afin d'afficher les informations les plus récentes.
- **Rang de pertinence.** Donne un rang pour chaque enregistrement dans une seule catégorie. Ce rang montre dans quelle mesure l'enregistrement correspond à la catégorie par rapport aux autres enregistrements dans cette catégorie. Sélectionnez une catégorie dans le panneau Catégories (panneau supérieur gauche) pour voir le rang. Pour plus d'informations, reportez-vous à la section Pertinence des catégories sur p. 176.
- **Effectifs de catégories** Répertorie le nombre de catégories auxquelles appartient un enregistrement.

Pertinence des catégories

Pour vous aider à créer de meilleures catégories, vous pouvez examiner la pertinence des documents ou des enregistrements dans chaque catégorie ainsi que la pertinence de toutes les catégories auxquelles appartient un document ou unenregistrement.

Pertinence d'une catégorie pour un enregistrement

Quand un document ou un enregistrement s'affiche dans le panneau Données, toutes les catégories auxquelles il appartient sont répertoriées dans la colonne Catégories. Quand un document ou un enregistrement appartient à plusieurs catégories, les catégories dans cette colonne s'affichent dans l'ordre de la correspondance la plus pertinente à la moins pertinente. La première catégorie est considérée comme correspondant le mieux à ce document ou à cet enregistrement. Pour plus d'informations, reportez-vous à la section Le panneau Données sur p. 174.

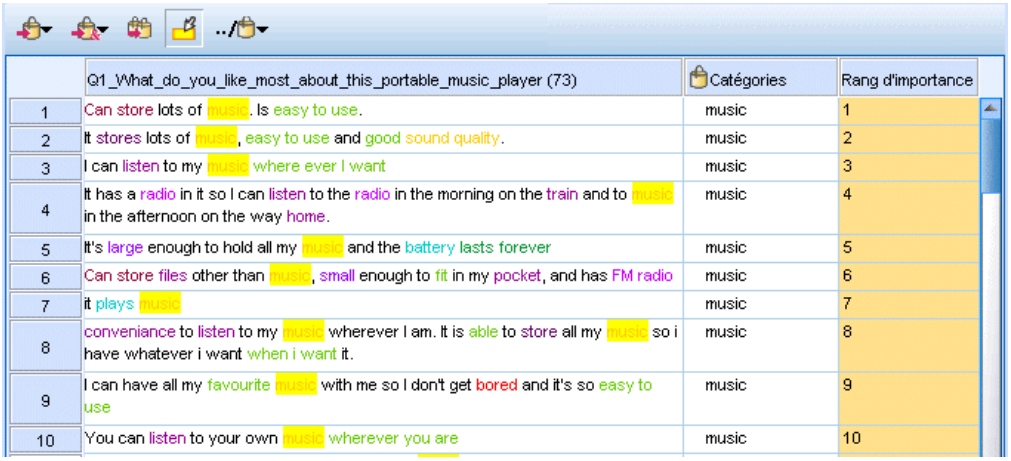
Pertinence d'un enregistrement pour une catégorie

Quand vous sélectionnez une catégorie, vous pouvez examiner la pertinence de chacun de ses enregistrements dans la colonne Rang de pertinence dans le panneau Données. Ce rang de pertinence indique à quel point le document ou l'enregistrement correspond à la catégorie sélectionnée par rapport aux autres enregistrements de cette catégorie. Pour voir le rang des enregistrements pour une seule catégorie, sélectionnez celle-ci dans le panneau Catégories (panneau supérieur gauche) et le rang du document ou de l'enregistrement s'affiche dans la colonne. Cette colonne n'est pas visible par défaut mais vous pouvez choisir de l'afficher. Pour plus d'informations, reportez-vous à la section Le panneau Données sur p. 174.

Plus le rang de l'enregistrement est bas, plus il correspond à la catégorie sélectionnée de façon à ce que 1 soit la meilleure correspondance. Si plusieurs enregistrements ont la même pertinence, chacun est affiché avec le même rang suivi d'un signe égal (=) pour montrer qu'ils ont la même pertinence. Par exemple, vous pouvez avoir les rangs suivants 1=, 1=, 3, 4, etc., ce qui signifie qu'il existe deux enregistrements considérés de manière égale comme étant les meilleures correspondances pour cette catégorie.

Astuce : Vous pouvez ajouter le texte de l'enregistrement le plus pertinent à l'annotation de catégorie pour donner une meilleure description de la catégorie. Ajoutez le texte directement à partir du panneau Données en sélectionnant le texte et en choisissant Catégories > Ajouter à l'annotation dans le menu.

Figure 10-5
Panneau Données affichant les catégories et le rang de pertinence

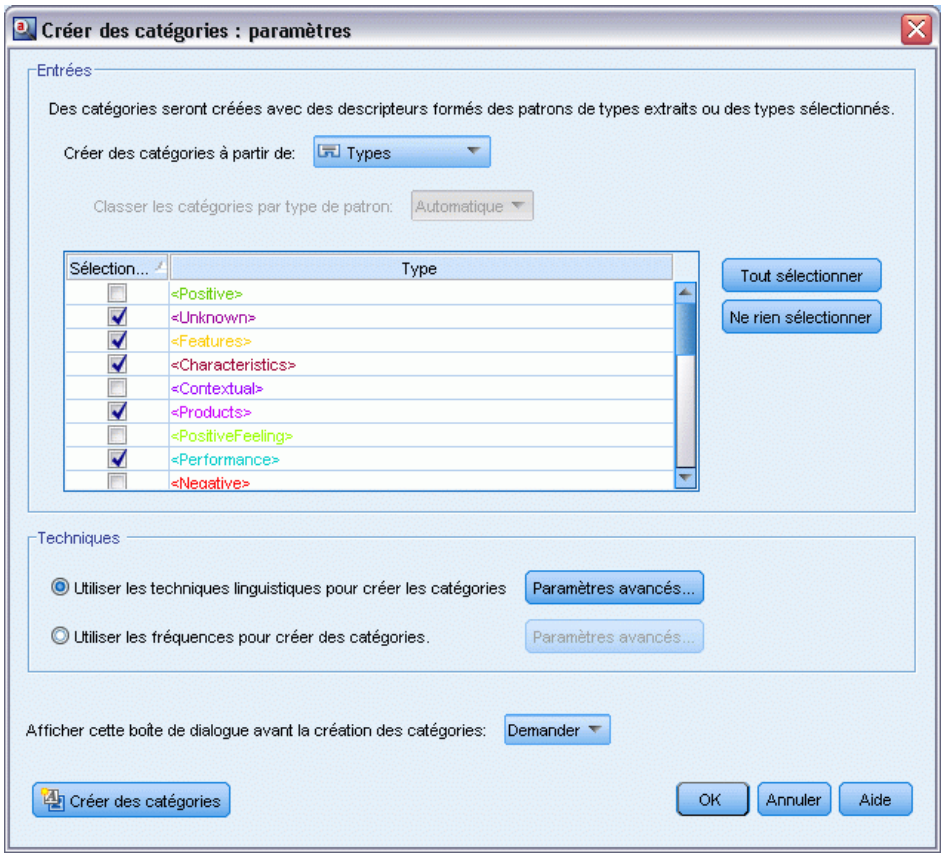


	Q1_What_do_you_like_most_about_this_portable_music_player (73)	Catégories	Rang d'importance
1	Can store lots of music. Is easy to use.	music	1
2	It stores lots of music, easy to use and good sound quality.	music	2
3	I can listen to my music where ever I want	music	3
4	It has a radio in it so I can listen to the radio in the morning on the train and to music in the afternoon on the way home.	music	4
5	It's large enough to hold all my music and the battery lasts forever	music	5
6	Can store files other than music, small enough to fit in my pocket, and has FM radio	music	6
7	it plays music	music	7
8	conveniance to listen to my music wherever I am. It is able to store all my music so i have whatever i want when i want it.	music	8
9	I can have all my favourite music with me so I don't get bored and it's so easy to use	music	9
10	You can listen to your own music wherever you are	music	10

Création de catégories

Bien que vous puissiez avoir des catégories d'un package d'analyse de texte, vous pouvez aussi créer automatiquement des catégories à l'aide de diverses techniques linguistiques et de fréquence. Dans la boîte de dialogue Créer des paramètres de catégorie, vous pouvez appliquer les techniques linguistiques et de fréquence automatiques pour créer des catégories à partir de concepts ou de patrons de concept.

Figure 10-6
Boîte de dialogue Créer des catégories



En général, les catégories peuvent être constituées de différents types de descripteurs (types, concepts, patrons TLA, règles de catégorie). Quand vous créez des catégories à l'aide des techniques de création de catégorie automatiques, les catégories créées sont nommées selon un concept ou un patron de concept (en fonction de l'entrée que vous sélectionnez) et chacune d'elle contient un ensemble de descripteurs. Ces descripteurs peuvent avoir la forme de règles de catégorie ou de concepts et comprennent tous les concepts associés découverts par les techniques.

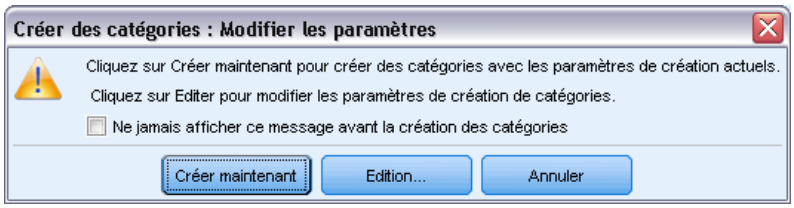
Après avoir créé les catégories, vous pouvez en apprendre beaucoup à leur sujet en les examinant dans le panneau Catégories ou en les explorant à travers les graphiques et les diagrammes. Vous pouvez utiliser des techniques manuelles pour effectuer des changements mineurs, supprimer toute mauvaise réaffectation, ou ajouter les enregistrements ou mots ayant été oubliés. Après avoir appliqué une technique, les concepts, les types et les patrons qui ont été regroupés dans une catégorie peuvent encore être classés par le biais d'autres techniques. En outre, puisque l'utilisation de techniques différentes peut également être à l'origine de catégories redondantes ou inappropriées, vous avez la possibilité de fusionner ou de supprimer des catégories.
Pour plus d'informations, reportez-vous à la section Edition et réglage des catégories sur p. 231.

Important ! Dans les versions précédentes, les règles de cooccurrence et de synonymes étaient entre crochets. Dans la version actuelle, les crochets indiquent désormais un résultat de patrons d'analyse des liens du texte. Les règles de cooccurrence et de synonymes sont maintenant entourées de parenthèses, comme dans `(enceintes acoustiques|enceintes)`.

Pour créer des catégories

▶ Dans les menus, sélectionnez Catégories > Créer des catégories. Un message apparaît, sauf si vous avez choisi de ne pas recevoir d'invite.

Figure 10-7
Invite avant la création



▶ Choisissez si vous voulez créer maintenant ou éditer d'abord les paramètres.

- Cliquez sur Créer Maintenant pour commencer à créer des catégories à l'aide des paramètres actuels. Les paramètres sélectionnés par défaut sont souvent suffisants pour commencer le processus de catégorisation. Le processus de création de catégories commence et une boîte de dialogue de progression apparaît.
- Cliquez sur Editer pour examiner et modifier les paramètres de création.

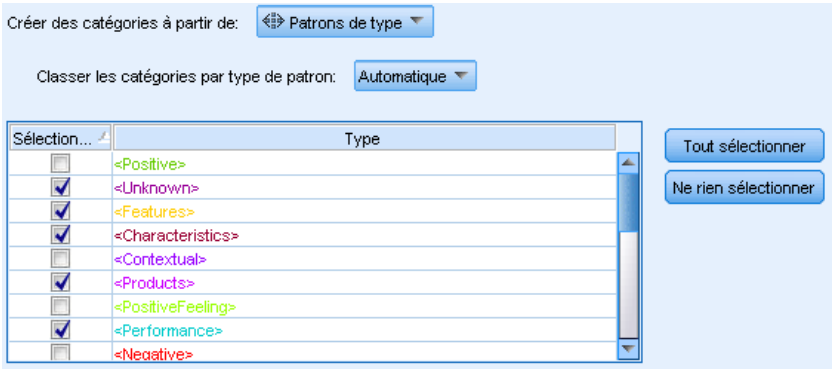
Remarque : le nombre de catégories maximum pouvant être affichées est de 10 000. Un avertissement apparaît si ce nombre est atteint ou dépassé. Si cela se produit, modifiez les options Créer ou développer des catégories afin de réduire le nombre de catégories créées.

Entrées

Les catégories sont créées à partir de descripteurs dérivés de patrons de type ou de types. Dans le tableau, vous pouvez sélectionner les types ou patrons individuels à inclure dans le processus de création de catégories.

Patrons de type. Si vous sélectionnez les patrons de type, des catégories sont créées à partir de patrons plutôt qu'à partir de types et de concepts uniquement. De cette manière, tous les enregistrements ou documents contenant un patron de concept appartenant au patron de type sélectionné sont classés en catégories. Ainsi, si vous sélectionnez le patron de type `<Budget>` et `<Positive>` dans le tableau, des catégories telles que `coût & <Positive>` ou `taux & excellent` peuvent être créées.

Figure 10-8
Boîte de dialogue Créer des catégories montrant les patrons de type disponibles



Lorsque des patrons de type sont utilisés comme entrées pour la création de catégories automatiques, les techniques identifient parfois plusieurs façons de former la structure des catégories. Techniquement, il n'existe pas une bonne façon de créer des catégories, mais une structure peut être plus adaptée à votre analyse qu'une autre. Pour aider à personnaliser la sortie dans ce cas, vous pouvez choisir un type préféré. Toutes les catégories de niveau supérieur créées proviendront d'un concept du type sélectionné ici (et pas d'un autre type). Chaque sous-catégorie contiendra un patron des liens du texte de ce type. Choisissez ce type dans le champ Structurer les catégories par type de patron : et le tableau sera mis à jour et n'affichera que les patrons applicables contenant le type sélectionné. La plupart du temps, le type `<Unknown>` sera présélectionné. Ainsi tous les patrons contenant le type `<Unknown>` (pour du texte qui n'est pas en japonais) seront sélectionnés. Pour le texte en japonais, <名詞> sera présélectionné par le programme. *Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium. Le tableau affiche les types dans l'ordre décroissant en commençant par celui contenant le plus grand nombre d'enregistrements ou de documents (effectifs des documents).

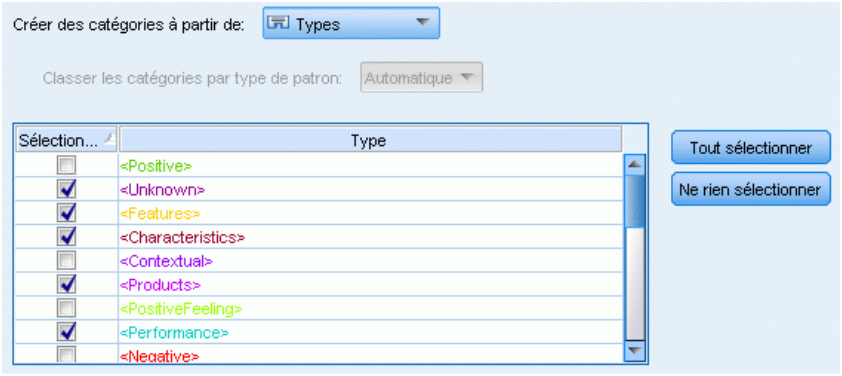
Types. Si vous sélectionnez les types, les catégories seront créées à partir des concepts appartenant aux types sélectionnés. Ainsi, si vous sélectionnez le type `<Budget>` dans le tableau, des catégories telles que `coût` ou `prix` peuvent être créées car `coût` et `prix` sont des concepts attribués au type `<Budget>`.

Par défaut, seuls les types qui capturent le plus d'enregistrements ou de documents sont sélectionnés. Cette présélection vous permet de considérer rapidement les types les plus intéressants et d'éviter de créer des catégories sans intérêt. Le tableau affiche les types dans l'ordre décroissant en commençant par celui contenant le plus grand nombre d'enregistrements ou de documents (effectifs des documents). Les types de la bibliothèque `Opinions` sont désélectionnés par défaut dans le tableau des types.

Les entrées que vous choisissez ont une influence sur les catégories obtenues. Lorsque vous choisissez d'utiliser Types comme entrée, vous pouvez plus facilement voir les concepts associés. Par exemple, si vous créez des catégories en utilisant Types comme entrée, vous pourrez obtenir une catégorie `Fruit` avec des concepts comme `ananas`, `poire`, `agrumes`, `orange` etc. Si vous choisissez Patrons de type comme entrée et que vous sélectionnez le patron `<Unknown> + <Positive>`, par exemple, alors vous pouvez obtenir une catégorie `fruit + <Positive>` avec une ou deux sortes de fruits comme `fruit + savoureux` et `pomme + bonne`. Ce deuxième résultat n'affiche que 2 patrons de concept car les autres occurrences de fruit ne sont

pas nécessairement considérés comme positives. Et bien que cela puisse être suffisant pour vos donnés textuelles actuelles, dans les enquêtes longitudinales où différents ensembles de documents sont utilisés, il est conseillé d'ajouter manuellement d'autres descripteurs comme `agrume + positif` ou d'utiliser des types. La simple utilisation de types comme entrée vous permettra de trouver tous les fruits possibles.

Figure 10-9
Boîte de dialogue Créer des catégories montrant les types disponibles



Techniques

Puisque chaque ensemble de données est unique, le nombre de méthodes et l'ordre dans lequel vous les appliquez peut varier. Dans la mesure où vos objectifs d'exploration de texte peuvent être différents d'un ensemble de données à un autre, vous pouvez tester les différentes techniques afin de déterminer celle qui génère les meilleurs résultats pour les données textuelles en question.

Vous n'avez pas besoin d'être un expert de ces paramètres pour les utiliser. Par défaut, les paramètres les plus communs et moyens sont déjà sélectionnés. C'est pourquoi vous pouvez contourner la boîte de dialogue Paramètres avancés et créer directement vos catégories. De même, si vous effectuez des modifications ici, vous n'avez pas besoin de revenir à la boîte de dialogue Paramètres à chaque fois car les derniers paramètres sont toujours conservés.

Sélectionnez les techniques linguistiques ou de fréquence et cliquez sur le bouton Paramètres avancés pour afficher les paramètres des techniques sélectionnées. Aucune des techniques automatiques ne pourra parfaitement catégoriser vos données; c'est pourquoi nous recommandons de trouver et d'appliquer une ou plusieurs techniques automatiques qui correspondent aux besoins de vos données. Vous ne pouvez pas créer de catégories en utilisant en même temps les techniques de fréquence et linguistique.

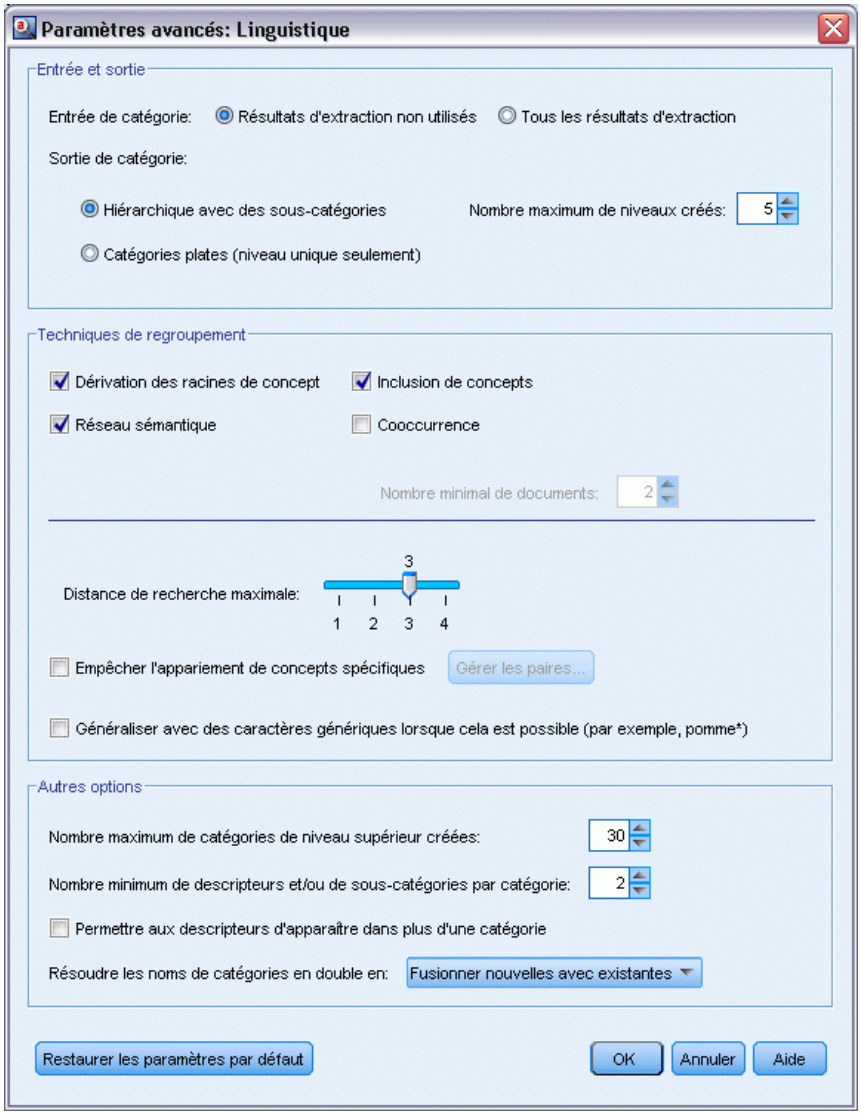
- **Techniques linguistiques avancées**. Pour plus d'informations, reportez-vous à sur p. 181.
- **Techniques de fréquence avancées**. Pour plus d'informations, reportez-vous à sur p. 192.

Paramètres linguistiques avancés

Quand vous créez des catégories, vous pouvez choisir parmi diverses techniques de création de catégorie linguistiques avancées, dont la *dérivation des racines de concept*(non disponible en japonais), l'*inclusion de concepts*, les *réseaux sémantiques* (anglais uniquement) et les *règles de cooccurrence*. Ces techniques peuvent être utilisées individuellement ou conjointement pour créer des catégories.

Notez que puisque chaque ensemble de données est unique, le nombre de méthodes et l'ordre dans lequel vous les appliquez peut varier. Dans la mesure où vos objectifs d'exploration de texte peuvent être différents d'un ensemble de données à un autre, vous pouvez tester les différentes techniques afin de déterminer celle qui génère les meilleurs résultats pour les données textuelles en question. Aucune des techniques automatiques ne pourra parfaitement catégoriser vos données; c'est pourquoi nous recommandons de trouver et d'appliquer une ou plusieurs techniques automatiques qui correspondent aux besoins de vos données.

Figure 10-10
Paramètres avancés : La boîte de dialogue Linguistique pour créer des catégories



Entrée et sortie

Entrée de catégorie. Sélectionnez à partir de quoi les catégories seront créées :

- Résultats d'extraction non utilisés. Cette option permet aux catégories d'être créées à partir des résultats d'extraction qui ne sont pas encore utilisés dans une catégorie existante. Ceci réduit la tendance des enregistrements à correspondre à plusieurs catégories et limite le nombre de catégories produites.
- Tous les résultats d'extraction. Cette option permet aux catégories d'êtres créées à l'aide de tous les résultats d'extraction. Ceci est particulièrement utile quand aucune ou peu de catégories existent déjà.

Sortie de catégorie. Sélectionnez la structure générale des catégories qui seront créées :

- Hiérarchique avec des sous-catégories. Cette option active la création de sous-catégories et de sous-sous-catégories. Vous pouvez définir la profondeur de vos catégories en choisissant le nombre de niveaux maximum (champ Nombre maximum de niveaux créés) pouvant être créés. Si vous choisissez 3, les catégories peuvent contenir des sous-catégories et ces sous-catégories peuvent également contenir des sous-catégories.
- Catégories plates (un seul niveau). Cette option n'active qu'un seul niveau de catégories à créer, ce qui signifie qu'aucune sous-catégorie ne sera créée.

Techniques de regroupement

Chaque technique disponible convient à certains types de données et de situation. Cependant, il est souvent judicieux de combiner des techniques dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. Aussi est-il possible de voir un concept figurer dans plusieurs catégories ou de rencontrer des catégories redondantes.

Dérivation des racines de concept. Cette technique crée des catégories en sélectionnant un concept et en recherchant les autres concepts y étant associés. Pour cela, elle analyse si certains composants du concept sont liés d'un point de vue morphologique ou s'ils partagent des racines. Cette technique est très utile pour l'identification de concepts de mots composés synonymes, car les concepts de chaque catégorie générée sont synonymes ou leur signification est proche. Elle utilise des données de différentes longueurs et génère un nombre inférieur de catégories compactes. Par exemple, le concept `opportunités d'avancer` serait regroupé avec les concepts `opportunité d'avancement` et `opportunité d'un avancement`. Pour plus d'informations, reportez-vous à la section Dérivation des racines de concept sur p. 187.Vous ne pouvez pas utiliser cette option pour le texte en japonais.

Réseau sémantique. Cette technique commence en identifiant les sens possibles de chaque concept à partir de son index complet de relations existant entre les mots puis crée des catégories en regroupant les concepts associés. Cette technique est plus performante lorsque les concepts sont connus dans le réseau sémantique et qu'ils ne sont pas trop ambigus. Son efficacité est cependant amoindrie lorsque le texte contient des termes spécialisés dont le réseau n'a pas connaissance. Par exemple, le concept `pomme granny smith` pourrait être regroupé avec `pomme gala` et `pomme golden` car il s'agit de soeurs de la granny smith. Pour donner un autre exemple, le concept `animal` pourrait être regroupé avec `chat` et `kangourou` car il s'agit d'hyponymes d'`animal`. Cette technique est uniquement disponible pour les textes en anglais dans cette version. Pour plus d'informations, reportez-vous à la section Réseaux sémantiques sur p. 189.

Inclusion de concept. Cette technique crée des catégories en regroupant les concepts multitermes (mots composés) selon qu'ils contiennent ou non des mots qui sont des sous-ensembles ou des super-ensembles d'un mot dans l'autre. Par exemple, le concept `siège` serait regroupé avec `siège de sécurité`, `siège couchette` et `commande de siège éjectable`. Pour plus d'informations, reportez-vous à la section Inclusion de concepts sur p. 188.

Cooccurrence. Cette technique crée des catégories à partir des cooccurrences trouvées dans le texte. Ainsi, lorsque des concepts ou des patrons de concept apparaissent souvent ensemble dans des documents et des enregistrements, la cooccurrence reflète une relation sous-jacente qui a vraisemblablement de l'intérêt dans vos définitions de catégorie. Lorsque des mots font l'objet d'une cooccurrence de manière significative, une règle de cooccurrence est créée et peut être utilisée comme descripteur de catégorie pour une nouvelle sous-catégorie. Par exemple, si de nombreux enregistrements contiennent les mots `prix` et `disponibilité`, (mais que peu d'enregistrements contiennent l'un sans l'autre) alors ces concepts peuvent être regroupés dans une règle de cooccurrence, (`prix &disponible`) et ils peuvent être affectés à une sous-catégorie de la catégorie `prix` par exemple.Pour plus d'informations, reportez-vous à la section Règles de cooccurrence sur p. 191.

- **Nombre minimum de documents.**. Pour aider à déterminer l'intérêt des coocurrences, déterminez le nombre minimum de documents ou d'enregistrements devant contenir une cooccurrence donnée pour être utilisés en tant que descripteurs dans une catégorie.

Distance de recherche maximale. Sélectionnez jusqu'où vous souhaitez que les techniques effectuent la recherche avant de créer des catégories. Plus la valeur est basse, moins vous obtenez de résultats. Cependant, ces résultats produisent moins de parasites, et sont davantage susceptibles d'être liés ou associés de façon significative. Plus la valeur est élevée, plus vous obtenez de résultats. Toutefois, ces résultats risquent d'être moins fiables ou moins pertinents. Bien que cette option soit généralement appliquée à toutes les techniques, son effet est maximal sur les coocurrences et les réseaux sémantiques.

Éviter l'appariement de concepts spécifiques. Cochez cette case pour arrêter le processus de regroupement ou d'appariement de deux concepts dans les résultats. Pour créer ou gérer des paires de concepts, cliquez sur Gérer les paires... Pour plus d'informations, reportez-vous à la section Gestion des paires d'exceptions de liens sur p. 185.

Généralisez avec des caractères génériques lorsque cela est possible. Sélectionnez cette option pour permettre au produit de créer des règles génériques dans les catégories à l'aide du caractère générique astérisque. Par exemple, au lieu de produire plusieurs descripteurs tels que `[tarte aux pommes + .]` et `[tarte aux fraises + .]`, l'utilisation de caractères génériques peut donner `[tarte * + .]`. Si vous généralisez avec des caractères génériques, vous obtenez souvent exactement le même nombre d'enregistrements ou de documents que précédemment. Toutefois, cette option a l'avantage de réduire le nombre de descripteurs de catégorie et de les simplifier. De plus, cette option augmente les possibilités de catégoriser davantage d'enregistrements ou de documents en utilisant ces catégories sur de nouvelles données textuelles (par exemple dans les enquêtes longitudinales/par vagues).

Autres options de création de catégories

En plus de sélectionner les techniques de regroupement à appliquer, vous pouvez éditer plusieurs autres options de création comme suit :

Le nombre maximum de catégories de niveau supérieur **créées.** Cette option sert à limiter le nombre de catégories pouvant être générées lorsque vous cliquez sur le bouton Créer des catégories. Dans certains cas, vous pouvez obtenir de meilleurs résultats si vous réglez cette valeur élevée puis supprimez les catégories sans intérêt.

Le nombre minimum de descripteurs et/ou de sous-catégories par catégorie. Utilisez cette option pour définir le nombre minimum de descripteurs et de sous-catégories qu'une catégorie doit contenir pour être créée. Cette option permet de limiter la création de catégories qui ne capturent pas un nombre assez important d'enregistrements ou de documents.

¨**Permettre aux descripteurs d'apparaître dans plusieurs catégories.** Lorsqu'elle est sélectionnée, cette option permet aux descripteurs d'être utilisés dans plusieurs des catégories qui seront créées ensuite. Cette option est généralement sélectionnée car les éléments entrent fréquemment ou "naturellement" dans deux catégories ou plus, et le fait de leur donner cette possibilité permet d'obtenir des catégories d'une plus grande qualité. Si vous ne sélectionnez pas cette option, vous réduisez le chevauchement d'enregistrements dans plusieurs catégories et en fonction du type de données que vous avez, ceci peut être souhaitable. Toutefois, avec la plupart des types de données, le fait de limiter les descripteurs à une seule catégorie entraîne une perte de la qualité ou de la diversité des catégories. Par exemple, imaginons que vous avez un concept `fabricant sièges-auto`. Avec cette option, ce concept peut apparaître dans une catégorie basée sur le texte `sièges-auto` et dans une autre basée sur `fabricant`. Mais si cette option n'est pas sélectionnée, bien que vous puissiez quand même obtenir les deux catégories, le concept `fabricant sièges-auto` n'apparaîtra comme descripteur que dans la catégorie à laquelle il correspond le mieux en fonction de plusieurs facteurs, notamment du nombre d'enregistrements dans lesquels `sièges-auto` et `fabricant` apparaissent respectivement.

Résoudre les noms de catégorie en double. Choisissez la manière de manipuler les nouvelles catégories ou sous-catégories dont le nom sera identique dans des catégories existantes. Vous pouvez fusionner les nouvelles catégories ou sous-catégories (et leur descripteurs) avec les catégories existantes qui portent le même nom. Vous pouvez également choisir d'ignorer la création de ces catégories si un nom en double se rencontre dans les catégories existantes.

Gestion des paires d'exceptions de liens

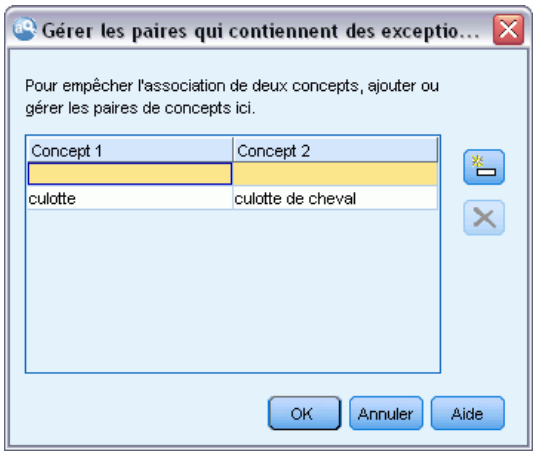
Au cours de la création de catégorie, de la classification et du regroupement de concepts, les algorithmes internes regroupent les mots par associations connues. Pour ne pas associer deux concepts par paires, ou pour ne pas les lier, vous pouvez activer cette fonctionnalité dans la boîte de dialogue Paramètres avancés de création de catégories, la boîte de dialogue Créer des clusters , et la boîte de dialogue Paramètres de l'index de la carte de concepts puis cliquer sur le bouton Gérer les paires.

Dans la boîte de dialogue Gérer les exceptions de liens, vous pouvez ajouter, modifier ou supprimer les paires de concepts. Tapez une paire par ligne. Si vous entrez les paires ici, cela évitera la génération d'association par paires lors de la création ou de l'extension des catégories, de

la classification, ou du regroupement de concepts. Saisissez les mots exactement comme vous les voulez, par exemple la version accentuée d'un mot n'est pas la même que la version non accentuée.

Par exemple, si vous voulez vous assurer que `foudre` et `coup de foudre` ne soient pas regroupés, vous pouvez ajouter la paire dans une ligne séparée du tableau :

Figure 10-11
Boîte de dialogue Gérer des paires d'exceptions de liens



A propos des Techniques linguistiques

Quand vous créez ou développez des catégories, vous pouvez choisir parmi diverses techniques de création de catégorie linguistiques avancées, dont la*dérivation des racines de concept*(non disponible en japonais), l'*inclusion de concepts*, les *réseaux sémantiques* (anglais uniquement) et les *règles de cooccurrence*. Ces techniques peuvent être utilisées individuellement ou conjointement pour créer des catégories.

Vous n'avez pas besoin d'être un expert de ces paramètres pour les utiliser. Par défaut, les paramètres les plus communs et moyens sont déjà sélectionnés. Vous pouvez contourner la boîte de dialogue Paramètres avancés et créer ou étendre directement vos catégories. De même, si vous effectuez des modifications ici, vous n'avez pas besoin de revenir à la boîte de dialogue Paramètres à chaque fois car les derniers paramètres sont toujours conservés.

Néanmoins, notez que puisque chaque ensemble de données est unique, le nombre de méthodes et l'ordre dans lequel vous les appliquez peut varier. Dans la mesure où vos objectifs d'exploration de texte peuvent être différents d'un ensemble de données à un autre, vous pouvez tester les différentes techniques afin de déterminer celle qui génère les meilleurs résultats pour les données textuelles en question. Aucune des techniques automatiques ne pourra parfaitement catégoriser vos données; c'est pourquoi nous recommandons de trouver et d'appliquer une ou plusieurs techniques automatiques qui correspondent aux besoins de vos données.

Les principales techniques linguistiques automatiques pour la création de catégories sont :

- **Dérivation des racines de concept.** Cette technique crée des catégories en sélectionnant un concept et en recherchant les autres concepts y étant associés. Pour cela, elle analyse si certains composants du concept sont liés d'un point de vue morphologique. Pour plus d'informations, reportez-vous à la section Dérivation des racines de concept sur p. 187.Vous ne pouvez pas utiliser cette option pour le texte en japonais.

- **Inclusion de concept.** Cette technique crée des catégories en sélectionnant un concept et en recherchant les autres concepts le contenant. Pour plus d'informations, reportez-vous à la section Inclusion de concepts sur p. 188.
- **Réseau sémantique.** Cette technique commence en identifiant les sens possibles de chaque concept à partir de son index complet de relations existant entre les mots puis crée des catégories en regroupant les concepts associés. Pour plus d'informations, reportez-vous à la section Réseaux sémantiques sur p. 189. Cette option n'est disponible que pour le texte en anglais.
- **Cooccurrence.** Cette technique crée des règles de cooccurrence qui peuvent être utilisées pour créer une nouvelle catégorie, étendre une catégorie ou comme entrée pour une autre technique de catégorie. Pour plus d'informations, reportez-vous à la section Règles de cooccurrence sur p. 191.

Dérivation des racines de concept

Remarque : Cette technique n'est pas disponible pour le texte en japonais.

La technique de dérivation des racines de concept crée des catégories en sélectionnant un concept et en recherchant les autres concepts y étant associés. Pour cela, elle analyse si certains composants du concept sont liés d'un point de vue morphologique. Un composant est un mot. La technique tente de regrouper les concepts en observant les terminaisons (suffixes) de chaque composant d'un concept et en recherchant les autres concepts ayant pu être créés à partir d'eux. Ainsi, lorsque des mots dérivent d'autres mots, il est vraisemblable qu'ils partagent la même signification ou qu'ils s'en rapprochent. Des règles internes, propres à chaque langue, identifient les terminaisons. Par exemple, le concept `opportunités d'avancer` serait regroupé avec les concepts `opportunité d'avancement` et `opportunité d'un avancement`.

Vous pouvez utiliser la dérivation des racines de concept sur n'importe quel type de texte. Utilisée seule, elle renvoie un nombre assez peu élevé de catégories et chacune d'entre elles contient généralement peu de concepts. Les concepts de chaque catégorie sont synonymes ou associés dans le cadre de la situation. Cet algorithme peut s'avérer utile, même si vous créez les catégories manuellement ; les synonymes qu'il repère peuvent être synonymes des concepts qui vous intéressent particulièrement.

Remarque : Vous pouvez éviter que les concepts ne se groupent entre eux en les spécifiant de manière explicite. Pour plus d'informations, reportez-vous à la section Gestion des paires d'exceptions de liens sur p. 185.

Fractionnement des termes en composants et suppression de la flexion

Lorsque vous appliquez les techniques de dérivation des racines de concept ou d'inclusion de concepts, les termes sont tout d'abord fractionnés en composants (mots), puis la flexion des composants est supprimée. Lorsque vous appliquez une technique, les concepts et leurs termes associés sont chargés et fractionnés en composants en fonction de leurs séparateurs, tels que les espaces, les traits d'union et les apostrophes. Par exemple, le terme `administrateur système` est fractionné en composants comme suit : `{administrateur, système}`.

Cependant, certaines parties du terme d'origine peuvent ne pas être utilisées. C'est ce que l'on appelle des mots vides. Voici certains des composants pouvant être ignorés en français : `un`, `une`, `et`, `par`, `pour`, `de`, `du`, `des`, `dans`, `ou`, `le`, `à` et `avec`.

Par exemple, le terme `analyse des données` est constitué de l'ensemble de composants `{données, analyse}` et tant les mots `de` et `le` sont considérés comme pouvant être ignorés. Par ailleurs, l'ordre des mots dans un ensemble de composants n'est pas significatif. Ainsi, les trois termes suivants peuvent être équivalents : `voile pour bateau de compétition`, `compétition de bateau à voile` et `bateau à voile de compétition`, dans la mesure où ils ont tous le même ensemble de composants `{voile, bateau, compétition}`. Chaque fois que deux termes sont considérés comme étant équivalents, les concepts correspondants sont fusionnés pour constituer un nouveau concept qui référence l'ensemble des termes.

De plus, puisque les composants d'un terme peuvent être déclinés, des règles propres à la langue sont appliquées en interne pour identifier les termes équivalents, quelle que soit la variation occasionnée par leur flexion (formes au pluriel, par exemple). De cette manière, les termes `niveau de fiabilité` et `niveaux de fiabilité` peuvent être identifiés comme étant équivalents, puisque la forme au singulier sans flexion est `niveau`.

Fonctionnement de la dérivation des racines de concept

Après que les termes ont été fractionnés en composants et que leur flexion a été supprimée (voir la section précédente), l'algorithme de dérivation des racines de concept analyse les terminaisons, ou suffixes, des composants pour retrouver leur radical, puis regroupe les concepts dont les radicaux sont identiques ou similaires. Les terminaisons sont identifiées à l'aide d'un ensemble de règles de dérivation linguistique propres à la langue du texte. Par exemple, une règle de dérivation concernant un texte rédigé en langue française détermine qu'un composant de concept présentant le suffixe `ique` peut être dérivé d'un concept ayant le même radical, mais se terminant par le suffixe `ie`. Grâce à cette règle (et à la suppression de la flexion), l'algorithme peut regrouper les concepts `étude épidémiologique` et `études épidémiologiques`.

Etant donné que les termes ont déjà été fractionnés en composants et que les composants pouvant être ignorés (par exemple, `d'`) ont été identifiés, l'algorithme de dérivation des racines de concept est également capable de regrouper le concept `étude d'épidémiologie` avec `étude épidémiologique`.

L'ensemble de règles de dérivation de composant a été choisi de telle sorte que la plupart des concepts regroupés par cet algorithme soient synonymes : les concepts `étude épidémiologique`, `études épidémiologiques`, `étude d'épidémiologie` sont tous des termes équivalents. Pour augmenter le degré d'exhaustivité, certaines règles de dérivation laissent l'algorithme regrouper des concepts qui sont associés dans le cadre de la situation. Par exemple, l'algorithme peut regrouper des concepts tels que `réparer les voitures` et `réparation de voitures`.

Inclusion de concepts

La technique d'inclusion de concepts crée des catégories en prenant un concept et, à l'aide d'algorithmes de série lexicale, identifie les concepts inclus dans d'autres concepts. Ainsi, lorsque les mots d'un concept constituent un sous-ensemble d'un autre concept, ces algorithmes reflètent

une relation sémantique sous-jacente. L'inclusion est une technique puissante qui peut être utilisée avec n'importe quel type de texte.

Elle fonctionne bien en conjonction avec des réseaux sémantiques, mais elle peut être utilisée séparément. L'inclusion de concepts peut aussi fournir de meilleurs résultats lorsque les documents ou les enregistrements contiennent de nombreux termes appartenant à un domaine spécifique. Ceci se confirme particulièrement lorsque vous avez préalablement affiné les dictionnaires de telle sorte que les termes spéciaux soient extraits et regroupés de façon appropriée (avec leurs synonymes).

Fonctionnement de l'inclusion de concept

Avant l'application de l'algorithme d'inclusion de concept, les termes sont fragmentés en composants et leur flexion est supprimée. Pour plus d'informations, reportez-vous à la section Dérivation des racines de concept sur p. 187. Ensuite, l'algorithme d'inclusion de concepts analyse les ensembles de composants. Pour chacun d'entre eux, l'algorithme recherche un ensemble de composants qui soit un sous-ensemble du premier.

Par exemple, si vous disposez du concept `déjeuner diététique`, qui comporte l'ensemble de composants `{déjeuner, diététique}`, et le concept `déjeuner`, qui comporte l'ensemble de composants `{déjeuner}`, l'algorithme en conclurait que `déjeuner diététique` est un type de `déjeuner` et les regrouperait.

Pour citer un exemple plus étendu, si le concept `siège` figure dans le panneau Résultats d'extraction et que vous appliquez cet algorithme, les concepts tels que `siège de sécurité`, `siège en cuir`, `siège couchette`, `commande de siège éjectable`, `siège-auto coque` et `instructions pour siège-auto` seraient également regroupés dans cette catégorie.

Puisque les termes sont déjà fractionnés en composants et que les composants pouvant être ignorés (par exemple, `en` et `d'`) ont été identifiés, l'algorithme d'inclusion de concepts reconnaît que le concept `cours d'espagnol avancé` inclut le concept `cours en espagnol`.

Remarque : Vous pouvez éviter que les concepts ne se groupent entre eux en les spécifiant de manière explicite. Pour plus d'informations, reportez-vous à la section Gestion des paires d'exceptions de liens sur p. 185.

Réseaux sémantiques

Dans cette version, la technique de réseau sémantique n'est disponible que pour les textes rédigés en anglais.

Cette technique crée des catégories à l'aide d'un réseau intégré de relations entre les mots. De ce fait, cette technique peut produire d'excellents résultats lorsque les termes sont précis et ne sont pas trop ambigus. En revanche, ne vous attendez pas à identifier de nombreux liens entre des concepts hautement techniques/spécialisés. En travaillant avec des concepts de ce type, il est probable que les techniques d'inclusion de concepts et de dérivation des racines de concept s'avèrent plus utiles.

Fonctionnement du réseau sémantique

La technique de réseau sémantique vise à exploiter les relations connues existant entre les mots pour créer des catégories de synonymes ou d'hyponymes. Un **hyponyme** est un concept considéré comme étant un type de second concept, de façon à ce qu'une relation hiérarchique, également appelée relation ISA, soit établie. Par exemple, si `animal` est un concept, `chat` et `kangourou` sont des hyponymes d'`animal`, puisque ce sont des types d'animaux.

Outre les relations de synonyme et d'hyponyme, la technique du réseau sémantique examine également les liens partie-tout existant entre les concepts du type `<Location>`. Par exemple, cette technique regroupe les concepts `normandie`, `provence` et `france` en une seule catégorie, car la Normandie et la Provence sont des parties de la France.

Les réseaux sémantiques commencent par identifier les sens possibles de chaque concept du réseau. Lorsque des concepts sont identifiés comme étant synonymes ou hyponymes, ils sont regroupés dans une seule catégorie. Par exemple, la technique crée une seule catégorie contenant ces trois concepts : `pomme à couteau`, `pomme à croquer` et `granny smith`, puisque le réseau sémantique contient les informations suivantes : 1) `pomme à croquer` est synonyme de `pomme à couteau` et 2) `granny smith` est un type de `pomme à couteau` (c'est-à-dire un hyponyme de `pomme à couteau`).

Considérés individuellement, de nombreux concepts, en particulier les termes univoques, sont ambigus. Par exemple, le concept `buffet` peut désigner un type de repas ou un meuble. Si l'ensemble des concepts inclut `repas`, `meuble` et `buffet`, l'algorithme est forcé de choisir entre regrouper `buffet` avec `repas` ou avec `meuble`. Sachez que, dans certains cas, les choix effectués par l'algorithme peuvent ne pas être appropriés dans le contexte d'un ensemble particulier d'enregistrements ou de documents.

La technique du réseau sémantique peut offrir de bien meilleures performances que l'inclusion de concepts avec certains types de données. Alors que le réseau sémantique et l'inclusion de concepts reconnaissent que `pinceau rouge` est une sorte de `pinceau`, seul le réseau sémantique reconnaît que `rouleau` est également une sorte de`pinceau`.

Les réseaux sémantiques peuvent être utilisés conjointement avec les autres techniques. Par exemple, imaginons que vous ayez sélectionné les techniques de réseau sémantique et d'inclusion, et que le réseau sémantique ait regroupé le concept `professeur` avec le concept `tuteur` (car un tuteur est un type de professeur). L'algorithme d'inclusion peut regrouper le concept `directeur d'étude` avec `tuteur`. En conséquence, les deux algorithmes collaborent afin de produire une catégorie de sortie contenant les trois concepts : `tuteur`, `directeur d'étude` et `professeur`.

Options du réseau sémantique

Il existe de nombreux paramètres supplémentaires pouvant être intéressants avec cette technique.

- Changer la distance de recherche maximale. Sélectionnez jusqu'où vous souhaitez que les techniques effectuent la recherche avant de créer des catégories. Plus la valeur est basse, moins vous obtenez de résultats. Cependant, ces résultats produisent moins de parasites, et sont davantage susceptibles d'être liés ou associés entre eux de façon significative. Plus la valeur est élevée, plus vous obtenez de résultats. Toutefois, ces résultats risquent d'être moins fiables ou moins pertinents.

Par exemple, en fonction de la distance, l'algorithme recherche de `pain au chocolat` jusqu'à `viennoiserie` (son parent), puis `petit pain` (grand-parent) et vers le haut jusqu'à `pain`.

En réduisant la distance de recherche, cette technique produit des catégories plus petites qui peuvent faciliter le travail lorsque les catégories produites sont trop grandes ou regroupent trop d'éléments.

Important ! Il est recommandé, lorsque vous utilisez cette technique, de ne pas appliquer l'option Nombre de caractères minimum requis (qui figure dans l'onglet Expert du nœud ou dans la boîte de dialogue Extraire) pour les regroupements flous, car certains mauvais regroupements peuvent avoir un impact très négatif sur les résultats.

Règles de cooccurrence

Les règles de cooccurrence vous permettent d'identifier et de regrouper les concepts étroitement liés au sein de l'ensemble de documents ou d'enregistrements. Ainsi, lorsque des concepts apparaissent souvent ensemble dans des documents et des enregistrements, la cooccurrence reflète une relation sous-jacente qui a vraisemblablement de l'intérêt dans vos définitions de catégories. Cette technique crée des règles de cooccurrence qui peuvent être utilisées pour créer une nouvelle catégorie, étendre une catégorie ou comme entrée pour une autre technique de catégorie. Deux concepts co-existent fortement s'ils apparaissent fréquemment ensemble dans un ensemble d'enregistrements et rarement séparément dans les autres enregistrements. Cette technique génère de bons résultats avec de plus grands ensembles de données contenant au moins plusieurs centaines de documents ou d'enregistrements.

Par exemple, si de nombreux enregistrements contiennent les mots `prix` et `disponibilité`, ces concepts peuvent être regroupés dans une règle de cooccurrence, `(prix & disponible)`. Dans un autre exemple, si les concepts `beurre`, `confiture`, `tartine` apparaissent plus souvent ensemble que séparément, ils sont regroupés dans une règle de cooccurrence de concepts `(beurre&confiture & tartine)`.

Important ! Dans les versions précédentes, les règles de cooccurrence et de synonymes étaient entre crochets. Dans la version actuelle, les crochets indiquent désormais un résultat de patrons d'analyse des liens du texte. Les règles de cooccurrence et de synonymes sont maintenant entourées de parenthèses, comme dans `(enceintes acoustiques|enceintes)`.

Fonctionnement des règles de cooccurrence

Cette technique analyse les documents ou les enregistrements à la recherche d'au moins deux concepts apparaissant souvent ensemble. Au moins deux concepts présentent une forte cooccurrence s'ils apparaissent souvent l'un avec l'autre dans un ensemble de documents ou d'enregistrements, et s'ils apparaissent rarement séparément dans d'autresdocuments ou enregistrements.

Une règle de catégorie se crée dès que le système identifie des concepts cooccurrents. Ces règles comportent au moins deux concepts reliés par l'opérateur booléen &. Il s'agit d'instructions logiques qui classent automatiquement un document ou un enregistrement dans une catégorie, si l'ensemble des concepts qu'elles régissent sont cooccurrents dans ces documents ou enregistrements.

Options des règles de cooccurrence

Si vous utilisez la technique des règles de cooccurrence, vous pouvez régler plusieurs paramètres ayant une influence sur les règles obtenues :

- Changer la distance de recherche maximale. Sélectionnez la distance de recherche de cooccurrences. Si vous augmentez la distance de recherche, la valeur de similarité minimum requise pour chaque cooccurrence diminue et par conséquent, de nombreuses règles de cooccurrence peuvent être produites, mais celles avec une valeur de similarité basse auront généralement peu d'importance. Lorsque vous diminuez la distance de recherche, la valeur de similarité requise minimum augmente et par conséquent, moins de règles de cooccurrences sont produites mais elles ont tendance à être plus importantes (plus fortes).
- Nombre minimum de documents.. Le nombre minimum d'enregistrements ou de documents qui doivent contenir une paire de concepts donnée pour qu'elle soit considérée comme une cooccurrence. Plus cette option a une valeur basse, plus il est facile de trouver des cooccurrences. Augmenter la valeur produit des cooccurrences moins nombreuses mais plus importantes. Par exemple, imaginons que les concepts « pomme » et « poire » se trouvent ensemble dans 2 enregistrements (et qu'aucun des deux ne se trouve dans d'autres enregistrements). Avec le Nombre minimum d'de documents. défini sur 2 (par défaut), la technique de cooccurrences créera une règle de catégorie (pomme et poire). Si la valeur passe à 3, la règle ne sera plus créée.

Remarque : avec de petits ensembles de données (< 1 000 réponses), il est possible qu'il n'y ait pas de cooccurrences avec les paramètres par défaut. Si c'est le cas, essayez d'augmenter la valeur de la distance de recherche.

Remarque : Vous pouvez éviter que les concepts ne se groupent entre eux en les spécifiant de manière explicite. Pour plus d'informations, reportez-vous à la section Gestion des paires d'exceptions de liens sur p. 185.

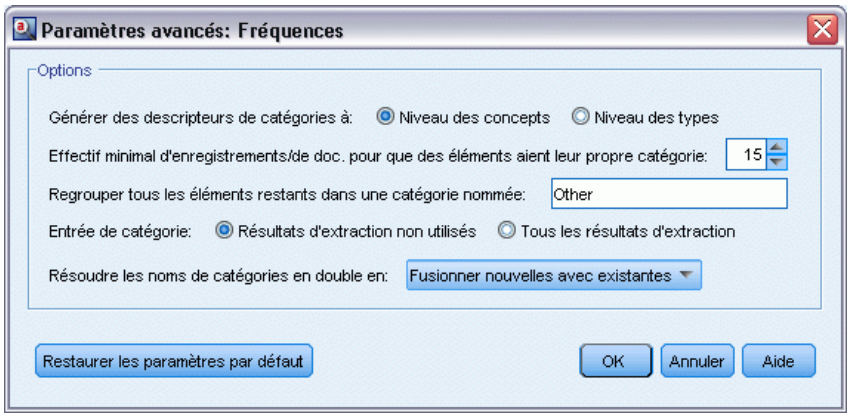
Paramètres de fréquence avancés

Vous pouvez créer des catégories en fonction d'une technique de fréquence simple et mécanique. Avec cette technique, vous pouvez créer une catégorie pour chaque élément (type, concept ou patron) situé en amont des effectifs des enregistrements ou des documents. Vous pouvez également créer une catégorie regroupant tous les termes moins fréquents. Les effectifs désignent le nombre d'enregistrements ou de documents contenant le concept extrait (et tous ses synonymes), le type ou le patron dans la question par rapport au nombre total d'occurrences d'un concept, d'un type ou d'un patron dans l'ensemble du texte.

Le regroupement d'éléments fréquents peut produire des résultats intéressants, car il peut indiquer une réponse courante ou importante. Elle est très efficace si elle a été exécutée sur les résultats d'extraction inutilisés après que d'autres techniques ont été appliquées. Une autre application consiste à exécuter cette technique immédiatement après l'extraction lorsqu'aucune autre catégorie n'existe, éditer les résultats pour supprimer les catégories sans intérêt, puis étendre ces catégories afin de leur faire correspondre toujours plus d'enregistrements ou de documents. Pour plus d'informations, reportez-vous à la section Extension de catégories sur p. 194.

A la place de cette technique, vous pouvez également trier les concepts et les patrons de concept par nombre décroissant des enregistrements ou des documents dans le panneau Résultats d'extraction puis en faisant glisser et en déposant les premiers dans le panneau Catégories pour créer les catégories correspondantes.

Figure 10-12
Paramètres avancés : boîte de dialogue des fréquences



Créer des descripteurs de catégories. Sélectionnez le type d'entrée pour les descripteurs. Pour plus d'informations, reportez-vous à la section Création de catégories sur p. 177.

- Niveau de concept. Le fait de sélectionner cette option signifie que les fréquences de concepts ou de patrons de concept seront utilisées. Les concepts sont utilisés si les types ont été sélectionnés comme entrée pour la création de catégorie et les patrons de concept sont utilisés si les patrons de type ont été sélectionnés. En général, appliquer cette technique au niveau du concept produira des résultats plus précis car les concepts et patrons de concept représentent un niveau de mesure moins élevé.
- Niveau de type. Le fait de sélectionner cette option signifie que les fréquences de types ou de patrons de type seront utilisées. Les types sont utilisés si les types ont été sélectionnés comme entrée pour la création de catégorie et les patrons de type sont utilisés si les patrons de type ont été sélectionnés. L'application de cette technique au niveau de type vous permet d'obtenir une vue rapide du type d'informations présentes données.

Effectif minimum de documents pour que les éléments aient leur propre catégorie. Cette option vous permet de créer des catégories à partir des éléments fréquents. Cette option limite les résultats aux seules catégories qui contiennent un descripteur que l'on rencontre dans au moins X enregistrements ou X documents, où X est la valeur à saisir pour cette option.

Regroupez tous les éléments restants dans une catégorie nommée. Cette option vous permet de regrouper tous les concepts ou types peu fréquents en une seule catégorie 'fourre-tout' portant le nom de votre choix. Par défaut, cette catégorie se nomme *Autre*.

Entrée de catégorie. Sélectionnez le groupe auquel appliquer ces techniques :

- Résultats d'extraction non utilisés. Cette option permet aux catégories d'être créées à partir des résultats d'extraction qui ne sont pas encore utilisés dans une catégorie existante. Ceci réduit la tendance des enregistrements à correspondre à plusieurs catégories et limite le nombre de catégories produites.
- Tous les résultats d'extraction. Cette option permet aux catégories d'êtres créées à l'aide de tous les résultats d'extraction. Ceci est particulièrement utile quand aucune ou peu de catégories existent déjà.

Résoudre les noms de catégorie en double. Choisissez la manière de manipuler les nouvelles catégories ou sous-catégories dont le nom sera identique dans des catégories existantes. Vous pouvez fusionner les nouvelles catégories ou sous-catégories (et leur descripteurs) avec les catégories existantes qui portent le même nom. Vous pouvez également choisir d'ignorer la création de ces catégories si un nom en double se rencontre dans les catégories existantes.

Extension de catégories

L'extension est un processus au cours duquel des descripteurs sont ajoutés ou améliorés automatiquement pour « agrandir » les catégories existantes. L'objectif est de produire une meilleure catégorie qui capture les enregistrements ou documents associés qui n'ont pas été attribués à cette catégorie à l'origine.

Les techniques de regroupement automatiques que vous sélectionnez tenteront d'identifier les concepts, patrons TLA et règles de catégorie associées aux descripteurs de catégorie existants. Ces nouveaux concepts, patrons et règles de catégorie sont ensuite ajoutés comme nouveaux descripteurs ou ajoutés aux descripteurs existants. Les techniques de regroupement pour l'extension incluent la *dérivation des racines de concept* (non disponible en japonais) , l'*inclusion de concepts*, les *réseaux sémantiques* (uniquement en anglais) et les *règles de cooccurrence*. La méthode Etendre les catégories vides avec des descripteurs générés à partir du nom de la catégorie génère des descripteurs à l'aide des mots dans les noms de catégories. Ainsi, plus les noms de catégories sont descriptifs, meilleurs sont les résultats.

Remarque : Les techniques de fréquence ne sont pas disponibles lors de l'extension des catégories.

L'extension est une excellente manière d'améliorer vos catégories de manière interactive. Voici quelques exemples de cas où vous pouvez étendre une catégorie :

- Après avoir glissé/déposé les patrons de concept pour créer des catégories dans le panneau Catégories
- Après avoir créé des catégories manuellement et avoir ajouté des règles de catégorie et des descripteurs simples
- Après avoir importé un fichier de catégories prédéfini dans lequel les catégories avaient des noms très descriptifs
- Après avoir affiné les catégories provenant du TAP que vous avez choisi

Vous pouvez étendre plusieurs fois une catégorie. Par exemple, si vous avez importé un fichier de catégorie prédéfinie avec des noms très descriptifs, vous pouvez effectuer l'extension avec l'option Étendre les catégories vides avec des descripteurs créés à partir du nom de la catégorie pour obtenir un premier ensemble de descripteurs puis étendre de nouveau ces catégories.

Néanmoins, dans d'autres cas, l'extension multiple peut entraîner une catégorie trop générique si les descripteurs sont de plus en plus étendus. Comme les techniques de création et d'extension de regroupement utilisent des algorithmes sous-jacents similaires, l'extension directement après la création de catégories ne produira probablement pas de résultats plus intéressants.

Astuces :

- Si vous tentez une extension et ne voulez pas utiliser les résultats, vous pouvez toujours annuler l'opération (Edition > Annuler) immédiatement après l'extension.
- L'extension peut produire au moins deux règles de catégorie dans une catégorie qui correspondent exactement au même ensemble de documents, les règles étant créées indépendamment pendant le processus. Si besoin est, vous pouvez afficher les catégories et supprimer les redondances en modifiant manuellement la description des catégories. Pour plus d'informations, reportez-vous à la section Modification des descripteurs de catégorie sur p. 233.

Pour étendre des catégories

▶ Dans le panneau Catégories, sélectionnez les catégories à étendre.

▶ Dans les menus, sélectionnez Catégories > Etendre des catégories. Un message apparaît, sauf si vous avez choisi de ne pas recevoir d'invite.

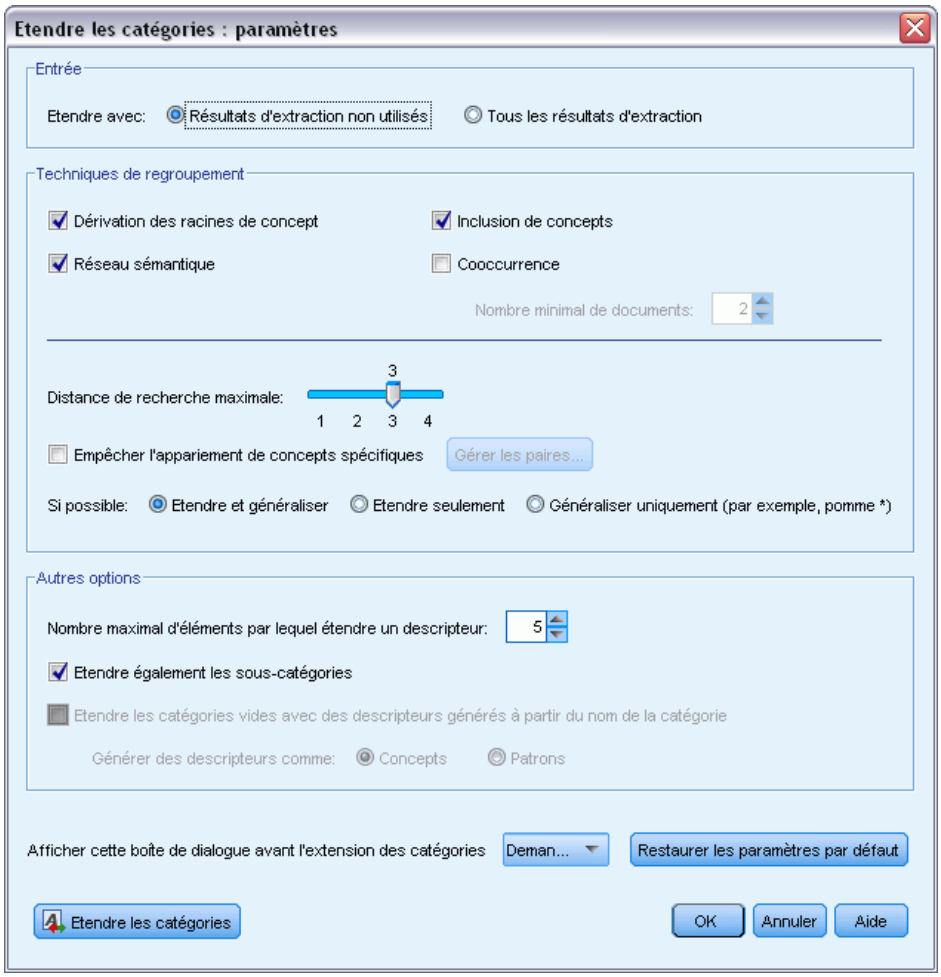
▶ Choisissez si vous voulez créer maintenant ou éditer d'abord les paramètres.

- Cliquez sur Etendre pour commencer à étendre des catégories à l'aide des paramètres actuels. Le processus commence et une boîte de dialogue de progression apparaît.
- Cliquez sur Editer pour examiner et modifier les paramètres.

Après la tentative d'extension, toutes les catégories pour lesquelles de nouveaux descripteurs ont été trouvés sont signalées par le mot Etendue dans le panneau Catégories, afin que vous puissiez les identifier rapidement. Le texte Etendue reste affiché jusqu'à ce que vous étendiez à nouveau la catégorie, que vous la modifiez d'une autre manière, ou la supprimiez via le menu contextuel.

Remarque : le nombre de catégories maximum pouvant être affichées est de 10 000. Un avertissement apparaît si ce nombre est atteint ou dépassé. Si cela se produit, modifiez les options Créer ou développer des catégories afin de réduire le nombre de catégories créées.

Figure 10-13
Boîte de dialogue Etendre les catégories



Chaque technique disponible lors de la création ou l'extension de catégories convient à certains types de données et de situations. Cependant, il est souvent judicieux de combiner plusieurs techniques dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. Dans la session interactive, les concepts et les types qui ont été regroupés dans une catégorie seront toujours disponibles la prochaine fois que vous créerez des catégories. Aussi est-il possible de voir un concept figurer dans plusieurs catégories ou de rencontrer des catégories redondantes.

Etendre avec. Sélectionnez l'entrée à utiliser pour étendre les catégories :

- Résultats d'extraction non utilisés. Cette option permet aux catégories d'être créées à partir des résultats d'extraction qui ne sont pas encore utilisés dans une catégorie existante. Ceci réduit la tendance des enregistrements à correspondre à plusieurs catégories et limite le nombre de catégories produites.
- Tous les résultats d'extraction. Cette option permet aux catégories d'êtres créées à l'aide de tous les résultats d'extraction. Ceci est particulièrement utile quand aucune ou peu de catégories existent déjà.

Techniques de regroupement

Pour les descriptions courtes de chacune de ces techniques, reportez-vous à la rubrique Paramètres linguistiques avancés sur p. 181. Ces techniques comprennent :

- Dérivation des racines de concept*(non disponible pour le japonais)*
- Réseau sémantique (*texte anglais uniquement et non utilisé si l'option Généraliser uniquement est sélectionnée.*)
- Inclusion de concepts
- Cooccurrence et sous-option Nombre minimal de documents.

Plusieurs types sont définitivement exclus de la technique de réseau sémantique car ils ne renverront pas de résultats pertinents. Ils comprennent `<Positive>`, `<Negative>`, `<IP>`, d'autres types non-linguistiques, etc.

Distance de recherche maximale. Sélectionnez jusqu'où vous souhaitez que les techniques effectuent la recherche avant de créer des catégories. Plus la valeur est basse, moins vous obtenez de résultats. Cependant, ces résultats produisent moins de parasites, et sont davantage susceptibles d'être liés ou associés de façon significative. Plus la valeur est élevée, plus vous obtenez de résultats. Toutefois, ces résultats risquent d'être moins fiables ou moins pertinents. Bien que cette option soit généralement appliquée à toutes les techniques, son effet est maximal sur les coocurrences et les réseaux sémantiques.

Éviter l'appariement de concepts spécifiques. Cochez cette case pour arrêter le processus de regroupement ou d'appariement de deux concepts dans les résultats. Pour créer ou gérer des paires de concepts, cliquez sur Gérer les paires... Pour plus d'informations, reportez-vous à la section Gestion des paires d'exceptions de liens sur p. 185.

Lorsque cela est possible : choisissez s'il faut simplement effectuer une extension, généraliser les descripteurs avec des caractères génériques ou effectuer ces deux opérations.

- **Etendre et généraliser.** Cette option permet d'étendre les catégories sélectionnées et de généraliser les descripteurs. Lorsque vous choisissez de généraliser, le produit crée des règles de catégorie génériques dans les catégories à l'aide du caractère générique astérisque. Par exemple, au lieu de produire plusieurs descripteurs tels que `[tarte aux pommes + .]` et `[tarte aux fraises + .]`, l'utilisation de caractères génériques peut donner `[tarte * + .]`. Si vous généralisez avec des caractères génériques, vous obtenez souvent exactement le même nombre d'enregistrements ou de documents que précédemment. Toutefois, cette option a l'avantage de réduire le nombre de descripteurs de catégorie et de les simplifier. De plus, cette option augmente les possibilités de catégoriser davantage d'enregistrements ou de documents en utilisant ces catégories sur de nouvelles données textuelles (par exemple dans les enquêtes longitudinales/par vagues).
- **Etendre uniquement**. Cette option étend vos catégories sans généralisation. Il peut être utile de choisir d'abord l'option Etendre uniquement pour les catégories créées manuellement puis d'étendre de nouveau ces mêmes catégories à l'aide de l'option Etendre et généraliser.
- **Généraliser uniquement.** Cette option généralisera les descripteurs sans étendre vos catégories de quelque autre façon.

  Remarque : sélectionner cette option désactive l'option Réseau sémantique car l'option Réseau sémantique est uniquement disponible lorsqu'une description doit être étendue.

Autres options d'extension de catégories

En plus de sélectionner les techniques de regroupement à appliquer, vous pouvez éditer les options suivantes :

Nombre maximum d'éléments pour étendre un descripteur. Lors de l'extension d'un descripteur avec des éléments (concepts, types et autres expressions), définissez le nombre maximum d'éléments pouvant être ajoutés à un seul descripteur. Si vous fixez cette limite à 10, un maximum de 10 éléments supplémentaires peuvent être ajoutés à un descripteur existant. Si plus de 10 éléments doivent être ajoutés, les techniques arrêtent d'ajouter de nouveaux éléments après l'ajout du dixième. Ceci peut raccourcir la liste de descripteur mais ne garantit pas que les éléments les plus intéressants aient été utilisés en premier. Vous préférez peut-être réduire la taille de l'extension sans pénaliser la qualité en utilisant l'option Généraliser avec des caractères génériques lorsque cela est possible. Cette option ne s'applique qu'aux descripteurs qui contiennent les booléens `&` (AND) ou `!` (NOT).

Etendre également les sous-catégories. Cette option étendra également toutes les sous-catégories sous les catégories sélectionnées.

Etendre les catégories vides avec des descripteurs générés à partir du nom de la catégorie. Cette méthode s'applique uniquement aux catégories vides qui ont 0 descripteurs. Si une catégorie contient déjà des descripteurs, elle ne sera pas étendue de cette manière. Cette option tente de créer automatiquement des descripteurs pour chaque catégorie en fonction des mots constituant le nom de la catégorie. Le nom de catégorie est analysé pour voir si les mots du nom correspondent à des concepts extraits. Si un concept est reconnu, il est utilisé pour trouver les patrons de concept correspondants et ceux-ci sont utilisés pour générer des descripteurs pour la catégorie. Cette option produit les meilleurs résultats lorsque les noms de catégories sont à la fois longs et descriptifs. C'est une méthode rapide pour générer des descripteurs de catégories, qui à leur tour permettent à la catégorie de capturer les enregistrements contenant ces descripteurs. Cette option est particulièrement utile quand vous importez des catégories d'ailleurs ou quand vous créez des catégories manuellement avec de longs noms descriptifs.

Générer des descripteurs en tant que. Cette option s'applique uniquement si l'option précédente est sélectionnée.

- **Concepts.** Choisissez cette option pour produire les résultats des descripteurs sous la forme de concepts, qu'ils aient été extraits ou non du texte source.
- **Patrons.** Choisissez cette option pour produire les résultats des descripteurs sous la forme de patrons, que les résultats des patrons ou tout autre patron aient été extraits ou non.

Création de catégories manuellement

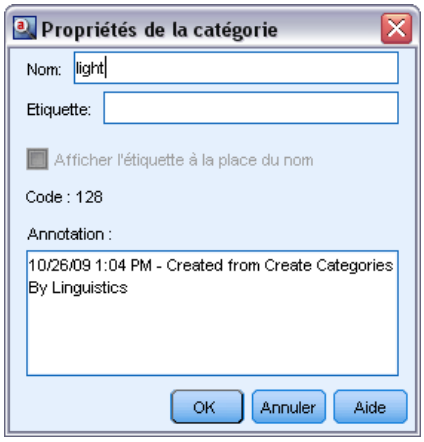
En plus de créer des catégories à l'aide des techniques de création de catégories automatiques que sont l'éditeur de règles, vous pouvez également créer des catégories manuellement. Les méthodes manuelles suivantes existent :

- Création d'une catégorie vide dans laquelle vous ajouterez des éléments un par un. Pour plus d'informations, reportez-vous à la section Création de catégories ou attribution d'un nouveau nom aux catégories sur p. 199.
- Déplacement de termes, de types et de patrons dans le panneau des catégories. Pour plus d'informations, reportez-vous à la section Création de catégories par la méthode Glisser-déposer sur p. 200.

Création de catégories ou attribution d'un nouveau nom aux catégories

Vous pouvez créer des catégories vides en vue d'y ajouter des concepts et des types. Vous pouvez également renommer vos catégories.

Figure 10-14
Boîte de dialogue Propriétés de catégorie



Pour créer une catégorie vide

- Accédez au panneau Catégories.
- Dans les menus, choisissez Catégories > Créer une catégorie vide. La boîte de dialogue apparaît.
- Entrez un nom dans le champ Nom pour cette catégorie.

▶ Cliquez sur OK pour valider ce nom et fermer la boîte de dialogue. La boîte de dialogue se ferme et un nouveau nom de catégorie apparaît dans le panneau.

Vous pouvez maintenant commencer à ajouter des éléments à cette catégorie. Pour plus d'informations, reportez-vous à la section Ajout de descripteurs aux catégories sur p. 232.

Pour renommer une catégorie

▶ Sélectionnez une catégorie, puis choisissez Catégories > Renommer la catégorie. La boîte de dialogue apparaît.

▶ Entrez un nouveau nom dans le champ Nom pour cette catégorie.

▶ Cliquez sur OK pour valider ce nom et fermer la boîte de dialogue. La boîte de dialogue se ferme et un nouveau nom de catégorie apparaît dans le panneau.

Création de catégories par la méthode Glisser-déposer

La technique du glisser-déposer est manuelle et n'est pas basée sur des algorithmes. Vous pouvez créer des catégories dans le panneau Catégories en glissant-déposant :

- des concepts, des types ou des patrons extraits du panneau Résultats d'extraction vers le panneau Catégories.
- des concepts extraits du panneau Données vers le panneau Catégories.
- des lignes entières du panneau Données vers le panneau Catégories. Cela permettra de créer une catégorie composée de tous les concepts et les patrons extraits contenus dans cette ligne.

Remarque : Le panneau Résultats d'extraction prend en charge la sélection multiple pour faciliter le glisser-déposer de différents éléments.

Important ! Vous ne pouvez pas faire glisser et déposer des concepts provenant du panneau Données et qui n'ont pas été extraits du texte. Si vous souhaitez forcer l'extraction d'un concept trouvé dans les données, vous devez l'ajouter à un type. Puis réexécutez l'extraction. Les nouveaux résultats de l'extraction contiendront le concept que vous venez d'ajouter. Vous pouvez l'utiliser dans votre catégorie. Pour plus d'informations, reportez-vous à la section Ajout de concepts à des types dans le chapitre 9 sur p. 158.

Pour créer des catégories avec la méthode de glisser-déposer :

▶ Dans le panneau Résultats d'extraction ou le panneau Données, sélectionnez un ou plusieurs concepts, patrons, types, enregistrements ou enregistrements partiels.

▶ Tout en maintenant le bouton de la souris appuyé, faites glisser l'élément vers une catégorie existante ou vers la zone du panneau pour créer une nouvelle catégorie.

▶ Lorsque vous atteignez la zone où vous souhaitez déposer cet élément, relâchez le bouton de la souris. Cet élément est ajouté au panneau Catégories. Les catégories modifiées apparaissent avec une couleur d'arrière-plan spécifique. Cette couleur s'appelle l'arrière-plan des commentaires de catégorie. Pour plus d'informations, reportez-vous à la section Définition des options dans le chapitre 8 sur p. 131.

Remarque : La catégorie résultante a été automatiquement nommée. Si vous le souhaitez, vous pouvez modifier un nom.

Si vous souhaitez voir quels enregistrements sont affectés à une catégorie, sélectionnez cette catégorie dans le panneau Catégories. Le panneau des données est automatiquement actualisé et affiche tous les enregistrements pour cette catégorie.

Utilisation des règles de catégorie

Vous pouvez créer des catégories de différentes façons. Une de ces façons est de définir des règles de catégorie permettant d'exprimer des idées. Les règles de catégorie sont des instructions qui classifient automatiquement les documents ou les enregistrements en une catégorie basée sur une expression logique à l'aide de concepts, de types et de patrons extraits ou d'opérateurs booléens. Par exemple, vous pouvez créer une expression qui signifie *inclure tous les enregistrements contenant le concept extrait* `ambassade` *mais ne pas inclure* `argentine`*dans cette catégorie*.

Alors que certaines règles de catégorie sont produites automatiquement lors de la création des catégories à l'aide de techniques de regroupement, telles que les règles de *cooccurrence* et de *dérivation des racines de concept* (Catégories > Configurer les paramètres > Paramètres avancés : Linguistique), vous pouvez également créer des règles de catégorie manuellement dans l'éditeur de règle en utilisant votre propre compréhension des données et du contexte. Chaque règle est rattachée à une seule catégorie afin que chaque document ou enregistrement correspondant à la règle soit ensuite scoré dans cette catégorie.

Les règles de catégorie permettent d'améliorer la qualité et la productivité de vos résultats de Text Mining et d'autres analyses quantitatives en vous aidant à classer les réponses dans des catégories plus spécifiques. Votre expérience et vos connaissances du marché peuvent vous offrir une compréhension particulière de vos données et du contexte. Cette compréhension peut vous permettre de traduire vos connaissances en règles de catégorie afin de classer vos documents ou vos enregistrements de manière plus efficace et plus précise en combinant des éléments extraits avec la logique booléenne.

La possibilité de créer ces règles améliore la précision, l'efficacité et la productivité du codage, tout en vous permettant d'incorporer vos connaissances du marché à la technologie d'extraction de produit.

Remarque : pour des exemples de correspondances entre les règles et le texte, voir Exemples de règles de catégorie sur p. 208

Syntaxe des règles de catégorie

Alors que certaines règles de catégorie sont produites automatiquement lors de la création des catégories à l'aide de techniques de regroupement, telles que les règles de *cooccurrence* et de *dérivation des racines de concept* (Catégories > Configurer les paramètres > Paramètres avancés : Linguistique), vous pouvez également créer des règles de catégorie manuellement dans l'éditeur de règle. Chaque règle est le descripteur d'une seule catégorie afin que chaque document ou enregistrement correspondant à la règle soit automatiquement scoré dans cette catégorie.

Remarque : pour des exemples de correspondances entre les règles et le texte, voir Exemples de règles de catégorie sur p. 208

Lorsque vous créez ou modifiez une règle, elle doit être ouverte dans le panneau de l'éditeur de règles. Vous pouvez ajouter des concepts, des types ou des patrons et utiliser des caractères génériques pour étendre les correspondances. Lorsque vous utilisez des concepts, des types ou des patrons extraits, vous pouvez bénéficier de la recherche de tous les concepts associés.

Important ! Pour éviter les erreurs courantes, nous vous recommandons de glisser-déposer directement les concepts du panneau Résultats d'extraction, des panneaux Analyse des liens du texte ou du panneau Données dans l'éditeur de règles ou de les ajouter à l'aide des menus contextuels lorsque cela est possible.

Lorsque les concepts, les types et les patrons sont reconnus, une icône apparaît à côté du texte.

 Concept extrait  Type extrait  Patron extrait

Syntaxe et opérateurs de règle

Le tableau suivant contient les caractères avec lesquels vous définissez la syntaxe de votre règle. Utilisez ces caractères avec les concepts, les types et les patrons pour créer votre règle.

Table 10-2
Syntaxe prise en charge

Caractère	Description
`&`	Le "et" booléen. Par exemple, `a & b` contient à la fois `a`*et*`b` comme dans : - `invasion & états-unis` - `2016 & jeux olympiques` - `bonne & pomme`
`|`	Le "ou" booléen est inclusif ce qui signifie que si un ou tous les éléments sont trouvés, une correspondance est établie. Par exemple, `a | b` contient soit `a`*soit*`b` comme dans : - `attaque | france` - `condominium | appartement`
`!()`	Le "pas" booléen. Par exemple, `!(a)` ne contient pas `a` comme dans : `!(bon & hôtel)`, `assassinat & !(autriche)`, ou `!(or) & !(cuivre)`
`*`	Un caractère générique représentant toute expression, d'un caractère unique à un mot entier selon la façon dont il est utilisé. Pour plus d'informations, reportez-vous à la section Utilisation de caractères génériques dans les règles de catégorie sur p. 206.
`()`	Un délimiteur d'expression. Toute expression entre parenthèses est évaluée en premier.
`+`	Le connecteur du patron utilisé pour former un patron avec un ordre spécifique. Vous devez utiliser les crochets lorsqu'ils sont disponibles. Pour plus d'informations, reportez-vous à la section Utilisation des patrons TLA dans les règles de catégorie sur p. 203.
`[]`	Le délimiteur de patrons est nécessaire si vous cherchez à effectuer des correspondances en fonction d'un patron TLA extrait à l'intérieur d'une règle de catégorie. Le contenu entre parenthèses fait référence aux patrons TLA et ne correspondra jamais à des concepts ou des types basés sur une simple cooccurrence. Si vous n'avez pas extrait ce patron TLA, aucune correspondance ne sera possible. Pour plus d'informations, reportez-vous à la section Utilisation des patrons TLA dans les règles de catégorie sur p. 203. N'utilisez pas de crochets si vous cherchez à mettre en correspondance des concepts et des types au lieu des patrons. *Remarque* : Dans les versions précédentes, les règles de cooccurrence et de synonymes générées par les techniques de création de catégories étaient généralement entre crochets. Dans toutes les nouvelles versions, les crochets indiquent désormais la présence d'un patron TLA. Les règles produites par une technique de cooccurrence et des synonymes sont maintenant entourées de parenthèses, comme dans `(enceintes acoustiques|enceintes)`.

Les opérateurs `&` et `|` sont commutatifs par conséquent `a & b = b & a` et `a | b = b | a`.

Ignorer les caractères avec une barre oblique inverse

Si un concept contient un caractère qui est également un caractère de syntaxe, vous devez placer une barre oblique inverse devant ce caractère afin que la règle soit correctement interprétée. La barre oblique inverse (\) permet d'ignorer des caractères qui autrement auraient une signification particulière. Lorsque vous le faites glisser dans l'éditeur, la barre oblique inverse est automatiquement ajoutée.

Les caractères de syntaxe de règle suivants doivent être précédés d'une barre oblique inverse si vous souhaitez qu'ils soient traités tels qu'ils sont plutôt que comme une syntaxe de règle :

& ! | + < > ( ) [ ] *

Par exemple, le concept `r&d` contenant l'opérateur « et » (`&`), la barre oblique inverse est requise lorsque ce concept est saisi dans l'éditeur de règle, sous la forme : `r\&d`.

Utilisation des patrons TLA dans les règles de catégorie

Les patrons d'analyse des liens du texte peuvent être précisément définis en règles de catégorie afin d'obtenir des résultats encore plus précis et contextuels. Lorsque vous définissez un patron dans une règle de catégorie, vous ignorez les résultats d'extraction de concept les plus simples et vous utilisez uniquement les documents et les enregistrements correspondants basés sur les résultats des patrons d'analyse des liens du texte extraits.

Important ! Afin de mettre en correspondance des documents à l'aide de patrons TLA dans vos règles de catégorie, vous devez avoir effectué une extraction avec l'analyse des liens du texte activée. La règle de catégorie recherchera les correspondances trouvées pendant l'extraction. Si vous n'avez pas choisi d'explorer les résultats TLA dans l'onglet Modèle de votre noeud de Text Mining, vous pouvez choisir d'activer l'extraction TLA dans les paramètres d'extraction pendant la session interactive puis effectuer une nouvelle extraction. Pour plus d'informations, reportez-vous à la section Extraction de données dans le chapitre 9 sur p. 143.

Délimitation par crochets. Un patron TLA doit être entouré de crochets `[ ]` si vous l'utilisez à l'intérieur d'une règle de catégorie. Le délimiteur de patrons est nécessaire si vous cherchez à effectuer des correspondances en fonction d'un patron TLA extrait. Puisque les catégories peuvent contenir des types, des concepts ou des patrons, les crochets expliquent à la règle que le contenu entre crochets fait référence au patron TLA extrait. Si vous n'avez pas extrait ce patron TLA, aucune correspondance ne sera possible. Si vous voyez un patron sans crochets, comme `pomme + bonne` dans le panneau Catégories, cela signifie que le patron a été ajouté directement à la catégorie en-dehors de l'éditeur de règle de catégorie. Par exemple, si vous ajoutez un patron de concept directement à une catégorie de la vue d'analyse des liens du texte, il n'apparaîtra pas avec des crochets. Mais, lorsqu'un patron est utilisé à l'intérieur d'une règle de catégorie, vous devez placer le patron entre crochets dans la règle de catégorie, comme dans `[banane + !(bon)]`.

Utilisation du signe + dans les patrons. Dans IBM® SPSS® Modeler Text Analytics , vous pouvez avoir des patrons ayant jusqu'à 6 parties (ou propriétés). Pour indiquer que l'ordre est important, utilisez le signe + pour connecter chaque élément, comme dans `[entreprise1 + a acheté + entreprise2]`. Ici l'ordre est important car il modifierait laquelle des deux entreprises est l'entreprise acquéreuse. L'ordre n'est pas déterminé par la structure de la phrase mais par la structure de la sortie du patron TLA. Par exemple, imaginons le texte "*J'aime Paris*" et que vous souhaitiez extraire cette idée, alors le patron a de fortes chances d'être `[paris + aime]` or `[<Emplacement> + <Positif>]` plutôt que `[<Positif> + <Emplacement>]` car les ressources d'opinion par défaut placent généralement les opinions en deuxième position dans les patrons à deux parties. Afin d'éviter les problèmes, il peut donc être nécessaire d'utiliser directement le patron comme descripteur dans votre catégorie. Mais si vous avez besoin d'utiliser un patron dans une instruction plus complexe, faites particulièrement attention à l'ordre des éléments dans les patrons présentés dans la vue Analyse des liens du texte car l'ordre est très important pour trouver des correspondances.

Par exemple, disons que vous aviez les deux échantillons de texte suivants : "J'*aime l'ananas*" et "Je déteste *l'ananas*. Néanmoins, j'*aime* les fraises". L'expression `aime & l'ananas` correspondrait aux deux textes car il s'agit de l'expression d'un concept et non pas d'une règle des liens du texte (elle n'est pas entre parenthèses). L'expression `ananas + aime` correspond uniquement à "J'*aime l'ananas*" car dans le deuxième texte, le terme *aime* est associé à *fraises* à la place.

Regroupement des patrons. Vous pouvez simplifier les règles avec vos propres patrons. Imaginons que vous souhaitez capturer les trois expressions suivantes, `poivre de cayenne + aime`, `poivre blanc + aime`, et`poivre + aime`. Vous pouvez les regrouper en une seule règle de catégorie comme `[poivre * & aime]`. Si vous avez une autre expression `poivre noir + bon`, vous pouvez regrouper les quatre en une règle comme `[poivre * + <Positive>]`.

Ordre des patrons. Afin de mieux organiser les résultats, les règles d'analyse des liens du texte fournies dans les modèles installés avec votre produits tentent de générer des patrons de base dans le même ordre quel que soit l'ordre des mots dans une phrase. Par exemple, si vous avez un enregistrement contenant le texte, "*Bonnes présentations*." et un autre contenant "*les présentations étaient bonnes*", la même règle correspond à ces deux textes qui sont générés dans le même ordre que `présentation + bonne` dans les résultats des patrons de concept plutôt que `présentation + bonne` et que `bonne + présentation`. Et dans les patrons à deux propriétés comme dans ceux de cet exemple, les concepts affectés aux types dans la bibliothèque Opinions seront présentés en dernier dans les résultats par défaut comme dans `pomme + mauvaise`.

Table 10-3
Syntaxe de patrons et utilisation booléenne

Expression	Correspond à un document ou à un enregistrement qui
`[]`	Contient tous les patrons TLA. Le délimiteur de patrons est nécessaire *dans les règles de catégorie* si vous cherchez à effectuer des correspondances en fonction d'un patron TLA extrait. Le contenu entre crochets fait référence aux patrons TLA et pas à des concepts ou types simples. Si vous n'avez pas extrait ce patron TLA, aucune correspondance ne sera possible. Si vous voulez créer une règle qui ne comprend pas de patrons, vous pouvez utiliser `!([ ])`.
`[a]`	Contient un patron dont au moins un des éléments est `a` quelle que soit sa position dans le patron. Par exemple, `[affaire]` peut renvoyer `[affaire + bonne]` ou simplement `[affaire + .]`
`[a + b]`	Contient un patron de concept. Par exemple, `[affaire + bonne]`. *Remarque :* Si vous souhaitez capturer uniquement ce patron sans ajouter d'autres éléments, nous vous recommandons d'ajouter ce patron directement à votre catégorie plutôt que d'en faire une règle.
`[a + b + c]`	Contient un patron de concept. Le signe + indique que l'ordre des éléments correspondants est important. Par exemple, `[entreprise1 + a acheté + entreprise2]`.
`[<A> + <B>]`	Contient tout patron avec le type `<A>` comme première propriété et le type `<B>` comme deuxième propriété et il y a exactement deux propriétés. Le signe + indique que l'ordre des éléments correspondants est important. Par exemple, `[<Budget> + <Negative>]`. *Remarque :* Si vous souhaitez capturer uniquement ce patron sans ajouter d'autres éléments, nous vous recommandons d'ajouter ce patron directement à votre catégorie plutôt que d'en faire une règle.
`[<A> & <B>]`	Contient tout patron de type avec type `<A>` et type `<B>`. Par exemple, `[<Budget> & <Negative>]`. Ce patron TLA ne sera jamais extrait, mais lorsqu'il est écrit ainsi, il est vaiment égal à `[<Budget> + <Négatif>]|[<Négatif> + <Budget>]`. L'ordre des éléments mis en correspondance n'est pas important. De plus, le patron peut contenir d'autres éléments mais il doit au moins contenir `<Budget>` et `<Negative>`.
`[a + .]`	Contient un patron où `a` est le seul concept et où les autres propriétés de ce patron sont vides. Par exemple : `[affaire + .]` correspond au patron de concept dans lequel le seul résultat est le concept `affaire`. Si vous avez ajouté le concept `affaire` comme descripteur de catégorie, vous obtiendrez tous les enregistrements avec le concept affaire, y compris les instructions positives sur une affaire. Mais, l'utilisation de `[affaire + .]` ne renverra que les résultats de patrons d'enregistrements qui représentent `affaire` et aucune autre relation ou opinion et ne renverra pas `affaire + fantastique`. *Remarque :* Si vous souhaitez capturer uniquement ce patron sans ajouter d'autres éléments, nous vous recommandons d'ajouter ce patron directement à votre catégorie plutôt que d'en faire une règle.
`[<A> + <>]`	Contient un patron où `<A>` est le seul type. Par exemple, `[<Budget> + <>]` correspond au patron dans lequel le seul résultat est un concept de type `<Budget>`. *Remarque :* Vous pouvez utiliser l'opérateur `<>` pour indiquer un type vide uniquement lorsque vous le placez derrière le symbole + dans un patron de type tel que `[<Budget> + <>]` mais pas `[prix + <>]`. *Remarque :* Si vous souhaitez capturer uniquement ce patron sans ajouter d'autres éléments, nous vous recommandons d'ajouter ce patron directement à votre catégorie plutôt que d'en faire une règle.

Expression	Correspond à un document ou à un enregistrement qui
`[a + !(b)]`	Contient au moins un patron qui comprend le concept `a` mais ne comprend pas le concept `b`. Doit contenir au moins un patron. Par exemple, `[prix + !(élevé)]` ou pour les types, `[!(<Fruits>|<Légumes>) + <Positif>]`
`!([<A> & <B>])`	Ne contient pas de patrons spécifiques. Par exemple, `!([<Budget> & <Negative>])`.

Remarque : pour des exemples de correspondances entre les règles et le texte, voir Exemples de règles de catégorie sur p. 208

Utilisation de caractères génériques dans les règles de catégorie

Les caractères génériques peuvent être ajoutés dans les règles afin d'étendre les capacités de mise en correspondance. Le caractère générique * (astérisque) peut être placé avant et/ou après un mot pour indiquer la façon dont les concepts peuvent être mis en correspondance. Deux types de caractères génériques sont disponibles :

- **Les caractères génériques affixes**. Ces caractères génériques sont placés en suffixe ou en préfixe sans espace entre la chaîne et l'astérisque. Par exemple, `opérat*` pourrait renvoyer *opérateur*, *opération*, *opérations*, *opérationnel*, *opérationnelles*, etc.
- **Les caractères génériques nominaux**. Ces caractères génériques sont placés en suffixe ou en préfixe un concept avec un espace entre le concept et l'astérisque. Par exemple, `opération *` pourrait renvoyer *opération*, *opération chirugicale*, *opération mathématique*, etc. De plus, un caractère générique nominal peut être utilisé avec un caractère générique affixe, comme par exemple `* opérat* *`, qui pourrait renvoyer *opération*, *opération chirurgicale*, *opérateur téléphonique*, *post opératoire*, etc. Comme cet exemple l'illustre, nous vous recommandons d'utiliser les caractères génériques avec prudence afin que la recherche ne soit pas trop large et ne capture pas de correspondances indésirables.

Exceptions !

- Un caractère générique ne peut jamais être utilisé seul. Par exemple, `(pomme | * )` n'est pas accepté.
- Un caractère générique ne peut jamais être utilisé pour mettre en correspondance des noms de type. `<Négatif*>` ne sera pas mis en correspondance avec des noms de type.
- Vous ne pouvez pas éviter que certains types soient mis en correspondance avec des concepts à l'aide de caractères génériques. Le type auquel le concept est affecté est automatiquement utilisé.
- Un caractère générique ne peut pas être placé au milieu d'une séquence de mots, qu'il soit ajouté à la fin d'un mot (ouv* compte) ou qu'il soit utilisé seul (ouvrir * compte). Vous ne pouvez pas non plus utiliser les caractères génériques dans les noms de type. Par exemple, `mot* mot`, tel que `pot* recette`, ne correspondra pas à une recette de potage ni à rien d'autre. Toutefois, `pomme* *` correspondra à *pomme de terre*, *pomme vapeur*, *pomme* etc. Dans un autre exemple, `mot * mot`, comme `gâteau * pomme`, ne correspondra pas à *gâteau cannelle pomme* ni à rien d'autre car l'astérisque apparaît entre deux mots. However, `pomme *` correspondra à *pomme de terre*, *pomme*, *pomme vapeur* etc.

Table 10-4
Utilisation des caractères génériques

Expression	Correspond à un document ou à un enregistrement qui
`*pin`	Contient un concept qui se termine par une lettre mais qui peut contenir n'importe quel nombre de lettres comme préfixe. Par exemple : `*pin` se termine par les lettres *pin* mais peut avoir comme préfixe : - `pommes` - `lapin` - `rupin`
`pin*`	Contient un concept qui commence par des lettres mais qui peut contenir n'importe quel nombre de lettres comme suffixe. Par exemple : `pin*` commence par les lettres *pin* mais peut être suivi d'un suffixe ou d'aucun suffixe comme : - `pommes` - `pincement` - `pinceau` Par exemple, `pin* & !(arbre* | pomme)`, qui contient un concept qui débute par les lettres`pin` mais pas un concept qui commence par les lettres *arbre* ni le concept `pomme` ne correspondrait PAS : `pin & pomme` mais correspondrait à : - `pincement` - `pin & forêt`
`*vers*`	Contient un concept qui comprend les lettres vers, mais qui peut avoir n'importe quel nombre de lettres comme préfixe, suffixe ou les deux. Par exemple : `*vers*` pourrait renvoyer : - `inverse` - `pervers` - `versatile`
`* prêt`	Contient un concept qui comprend le mot `prêt` mais peut être un composé avec un autre mot placé devant. Par exemple, `* prêt` pourrait renvoyer : - `prêt` - `banque de prêt` - `remboursement de prêt` Par exemple, `[* argent + <Negative>]` contient un concept qui se termine par le mot `argent` et contient un type `<Negative>` et pourrait renvoyer les patrons de concept suivants : - `transfert d'argent + lent` - `remboursement d'argent + retard`
`voiture *`	Contient un concept qui comprend le mot `voiture` mais qui peut être un composé suivi d'un autre mot. Par exemple, `voiture *` pourrait renvoyer : - `voiture` - `voiture de location` - `voiture de tourisme`
`* pomme *`	Contient un concept qui peut commencer par n'importe quel mot suivi du mot `pomme` et suivi d'un autre mot. `*` signifie 0 ou n caractères, donc cela correspond aussi au mot `pomme`. Par exemple, `* pomme *` pourrait renvoyer : - `pomme de pin` - `tarte aux pommes caramélisée` - `Gratin aux pommes de terre` - `pommes` Par exemple, `[* réservation * * + <Positive>]`, qui contient un concept commençant par le mot `réservation` (sans tenir compte de sa place dans le concept) en première position et contient un type `<Positive>` en deuxième position pourrait renvoyer les patrons de concept : - `système de réservation + bon` - `réservation en ligne + bon`

Remarque : pour des exemples de correspondances entre les règles et le texte, voir Exemples de règles de catégorie sur p. 208

Exemples de règles de catégorie

Pour mieux comprendre en quoi les correspondances entre les règles et les enregistrements sont différentes selon la syntaxe utilisée pour les exprimer, étudiez l'exemple suivant.

Exemples d'enregistrements

Imaginons deux enregistrements :

- **Enregistrement A** : *"lorsque j'ai regardé dans mon porte-monnaie, j'ai vu qu'il y manquait 5 dollars."*
- **Enregistrement B** : *"on a trouvé les 5$ sur l'aire de pique-nique mais la couverture n'était plus là."*

Les deux tableaux suivants montrent ce qui peut être extrait pour les concepts et les types ainsi que les patrons de concept et les patron de type.

Concepts et types extraits de l'exemple

Table 10-5
Exemple de concepts et de types extraits

Concept extrait	Concepts saisis comme
porte-monnaie	<Inconnu>
manquer	<Négatif>
5 dollars	<Devise>
couverture	<Inconnu>
aire de pique-nique	<Inconnu>

Patrons TLA extraits de l'exemple

Table 10-6
Exemple de sortie de patron TLA extraite

Patrons de concepts extraits	Patrons de types extraits	De l'enregistrement
aire de pique-nique + .	<Inconnu> + <>	Enregistrement B
porte-monnaie + .	<Inconnu> + <>	Enregistrement A
couverture + manquer	<Inconnu> + <Négatif>	Enregistrement B
5 dollars + .	<Devise> + <>	Enregistrement B
5 dollars + manquer	<Devise> + <Négatif>	Enregistrement A

Comment mettre en correspondance les règles de catégorie possibles

Le tableau suivant contient une syntaxe qui pourrait être entrée dans l'éditeur de règle de catégorie. Toutes les règles qui se trouvent ici ne fonctionnent pas et toutes ne correspondent pas aux mêmes enregistrements. Regardez la façon dont les différentes syntaxes ont un impact sur les enregistrements mis en correspondance.

Table 10-7
Règles concernant les échantillons

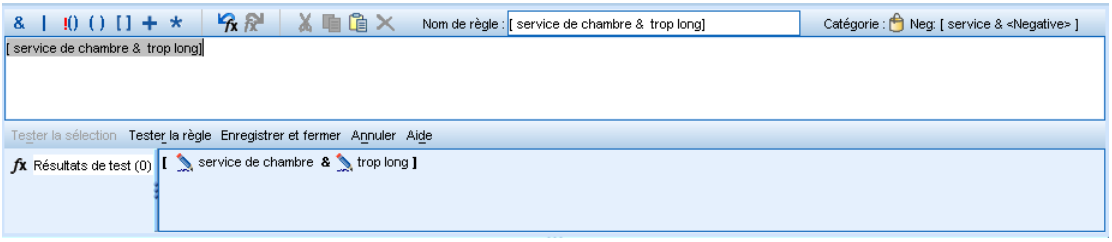
Syntaxe de règle	Résultat
`5 dollars & manquer`	Correspond à la fois à A et à B car ils contiennent tous les deux le concept extrait `manquer` et le concept extrait `5 dollars`. Cela revient à : `(5 dollars & manquer)`
`manquer & 5 dollars`	Correspond à la fois à A et à B car ils contiennent tous les deux le concept extrait `manquer` et le concept extrait `5 dollars`. Cela revient à : `(manquer & 5 dollars)`
`manquer & <Devise>`	Correspond à la fois à A et à B car ils contiennent tous les deux le concept extrait `manquer` et un concept correspondant au type `<Devise>`. Cela revient à : `(manquer & <Devise> )`
`<Devise> & manquer`	Correspond à la fois à A et à B car ils contiennent tous les deux le concept extrait `manquer` et un concept correspondant au type `<Devise>`. Cela revient à : `(<Devise> & manquer)`
`[5 dollars + manquer]`	Correspond à A mais **pas** à B car l'enregistrement B n'a pas produit de sortie de patron B contenant `5 dollars + manquer` (voir le tableau précédent). Cela revient à la sortie de patron TLA : `5 dollars + manquer`
`[manquer + 5 dollars]`	Ne correspond ni à l'enregistrement A ni au B car aucun patron TLA extrait (voir tableau précédent) ne correspond à l'ordre exprimé ici avec `manquer` en première position. Cela revient à la sortie de patron TLA : `5 dollars + manquer`
`[manquer & 5 dollars]`	Correspond à A mais **pas** à B car aucun patron TLA correspondant n'a été extrait de l'enregistrement B. Utiliser le caratère `&` indique que l'ordre n'a pas d'importance lors de la mise en correspondance ; par conséquent, cette règle recherche une correspondance de patron avec `[manquer + 5 dollars]` ou `[5 dollars + manquer]`. Seul `[5 dollars + manquer]` de l'enregistrement A possède une correspondance.
`[manquer + <Devise>]`	Ne correspond ni à l'enregistrement A ni à l'enregistrement B car aucun patron TLA extrait ne correspondait à cet ordre. Cela n'a pas d'équivalent car une sortie TLA n'est basée que sur les termes `(5 dollars + manquer)` ou sur les types `(<Devise> + <Négatif>)`, mais ne mélange pas les types et les concepts.
`[<Devise> + <Négatif>]`	Correspond à l'enregistrement A mais pas à l'enregistrement B car aucun patron TLA n'a été extrait de l'enregistrement B. Cela équivaut à la sortie TLA : `<Devise> + <Négatif>`
`[<Négatif> + <Devise>]`	Ne correspond ni à l'enregistrement A ni à l'enregistrement B car aucun patron TLA extrait ne correspondait à cet ordre. Par défaut, dans le modèle `Opinions` lorsqu'une *rubrique* est trouvée avec une *opinion*, cette *rubrique* (`<Devise>`) occupe la première position et *opinion* (`<Négatif>`) occupe la deuxième.

Création de règles de catégorie

Lorsque vous créez ou modifiez une règle de catégorie, elle doit être ouverte dans le panneau de l'éditeur de règles. Vous pouvez ajouter des concepts, des types ou des patrons et utiliser des caractères génériques pour étendre les correspondances. Lorsque vous utilisez des concepts, des types ou des patrons reconnus, vous pouvez bénéficier de la recherche de tous les concepts associés. Par exemple, lorsque vous utilisez un concept, tous ses termes associés, ses formes plurielles et ses synonymes sont également associés à cette règle. De même, lorsque vous utilisez un type, tous ses concepts sont également capturés par la règle.

Vous pouvez ouvrir l'éditeur de règles en modifiant une règle existante ou en faisant un clic droit sur le nom de la catégorie et en choisissant Créer une règle.

Figure 10-15
Panneau de l'éditeur de règles



Vous pouvez utiliser les menus contextuels, la méthode glisser-déposer ou saisir manuellement les concepts, les types et les patrons dans l'éditeur. Combinez-les ensuite avec les opérateurs booléens (`&`, `!()`, `|`) et les parenthèses pour former vos expressions de règle. Pour éviter les erreurs communes, nous vous recommandons de déplacer les concepts en les faisant glisser directement du panneau Résultats d'extraction ou du panneau Données vers l'éditeur de règle. Soyez attentif à la syntaxe des règles afin d'éviter les erreurs. Pour plus d'informations, reportez-vous à la section Syntaxe des règles de catégorie sur p. 201.

Remarque : pour des exemples de correspondances entre les règles et le texte, voir Exemples de règles de catégorie sur p. 208

Pour créer une règle

- Si vous n'avez pas encore extrait de données ou que votre extraction a expiré, faites-le maintenant. Pour plus d'informations, reportez-vous à la section Extraction de données dans le chapitre 9 sur p. 143.
- Dans le panneau Catégories, sélectionnez la catégorie dans laquelle vous voulez ajouter votre règle.
- Dans les menus, choisissez Catégories > Créer une règle. Le panneau de l'éditeur des règles de catégorie apparaît dans la fenêtre.
- Dans le champ Nom de règle, entrez un nom pour votre règle. Si vous ne donnez pas de nom à la règle, l'expression sera automatiquement utilisée comme nom. Vous pourrez renommer cette règle ultérieurement.

- Dans le plus grand champ de texte de l'expression, vous pouvez effectuer les opérations suivantes :
  - Saisir directement le texte dans le champ ou le faire glisser depuis un autre panneau. Utilisez uniquement des concepts, types et patrons extraits. Par exemple, si vous entrez le mot `chats`, mais que seule la forme au singulier, `chat`, apparaît dans le panneau Résultats d'extraction, l'éditeur ne sera pas en mesure de reconnaître `chats`. Dans ce cas, la forme singulière pourra inclure automatiquement la forme plurielle, ou vous pourrez utiliser un caractère générique. Pour plus d'informations, reportez-vous à la section Syntaxe des règles de catégorie sur p. 201.
  - Sélectionner les concepts, types ou patrons à ajouter aux règles et utiliser les menus.
  - Ajouter des opérateurs booléens pour relier les éléments de votre règle. Utilisez les boutons de la barre d'outils pour ajouter le booléen "et" &, le booléen "ou" |, le booléen "pas" !(), les parenthèses (), et les crochets pour les patrons [ ] à votre règle.
- Cliquez sur le bouton Tester la règle pour vérifier que votre règle est bien constituée. Pour plus d'informations, reportez-vous à la section Syntaxe des règles de catégorie sur p. 201. Le nombre de documents ou d'enregistrements trouvés apparaît entre parenthèses en regard du texte Résultats de test. Les éléments de la règle qui ont été reconnus ou les messages d'erreur éventuels apparaissent à droite de ce texte. Si le graphique à côté du type, du patron ou du concept apparaît avec un point d'interrogation rouge, ceci indique que l'élément ne correspond à aucune extraction connue. S'il n'y a pas de correspondance, la règle ne trouvera aucun enregistrement.
- Pour tester une partie de votre règle, sélectionnez-la, puis cliquez sur Tester la sélection.
- En cas de problème, apportez toutes les modifications nécessaires et testez à nouveau la règle.
- Lorsque vous avez terminé, cliquez sur Enregistrer & Fermer pour enregistrer à nouveau la règle et fermer l'éditeur. Le nouveau nom de la règle apparaît dans la catégorie.

Modification et suppression des règles

Vous pouvez éditer à tout moment une règle que vous avez créée et enregistrée. Pour plus d'informations, reportez-vous à la section Syntaxe des règles de catégorie sur p. 201.

Si une règle ne vous est plus utile, vous pouvez la supprimer.

Pour modifier des règles

- Dans le tableau Descripteurs de la boîte de dialogue Définitions de catégorie, sélectionnez la règle.
- Dans les menus, choisissez Catégories > Editer la règle ou double-cliquez sur le nom de la règle. L'éditeur s'ouvre, avec la règle sélectionnée.
- Modifiez la règle à l'aide des résultats de l'extraction et des boutons de la barre d'outils.
- Testez à nouveau la règle pour vous assurer qu'elle renvoie les résultats attendus.
- Cliquez sur Enregistrer & Fermer pour enregistrer à nouveau la règle et fermer l'éditeur.

Pour supprimer une règle

- Dans le tableau Descripteurs de la boîte de dialogue Définitions de catégorie, sélectionnez la règle.

▶ Dans les menus, sélectionnez Edition > Supprimer. La règle est supprimée de la catégorie.

Import et export de catégories prédéfinies

Si vos propres catégories sont stockées dans un fichier Microsoft Excel (**.xls*, **.xlsx*), vous pouvez les importer dans IBM® SPSS® Modeler Text Analytics .

Vous pouvez également exporter vers un fichier Microsoft Excel (**.xls*, **.xlsx*), les catégories que vous avez dans une session interactive ouverte. Lorsque vous exportez vos catégories, vous pouvez choisir d'inclure ou d'exclure des informations supplémentaires telles que les descripteurs et les scores. Pour plus d'informations, reportez-vous à la section Exporter des catégories sur p. 221.

Si vos catégories prédéfinies n'ont pas de code ou si vous souhaitez de nouveaux codes, vous pouvez générer automatiquement un nouvel ensemble de codes pour l'ensemble de catégories, dans le panneau des catégories en choisissant Catégories > Gestion des catégories > Générer automatiquement des codes à partir des menus. Tous les codes existants seront supprimés et renumérotés automatiquement.

Import de catégories prédéfinies

Vous pouvez importer vos catégories prédéfinies dans IBM® SPSS® Modeler Text Analytics . Avant l'importation, assurez-vous que le fichier de la catégorie prédéfinie est un fichier Microsoft Excel (**.xls*, **.xlsx*) et qu'il est structuré dans un format pris en charge. Vous pouvez également choisir de laisser le produit détecter le format pour vous. Les formats suivants sont pris en charge :

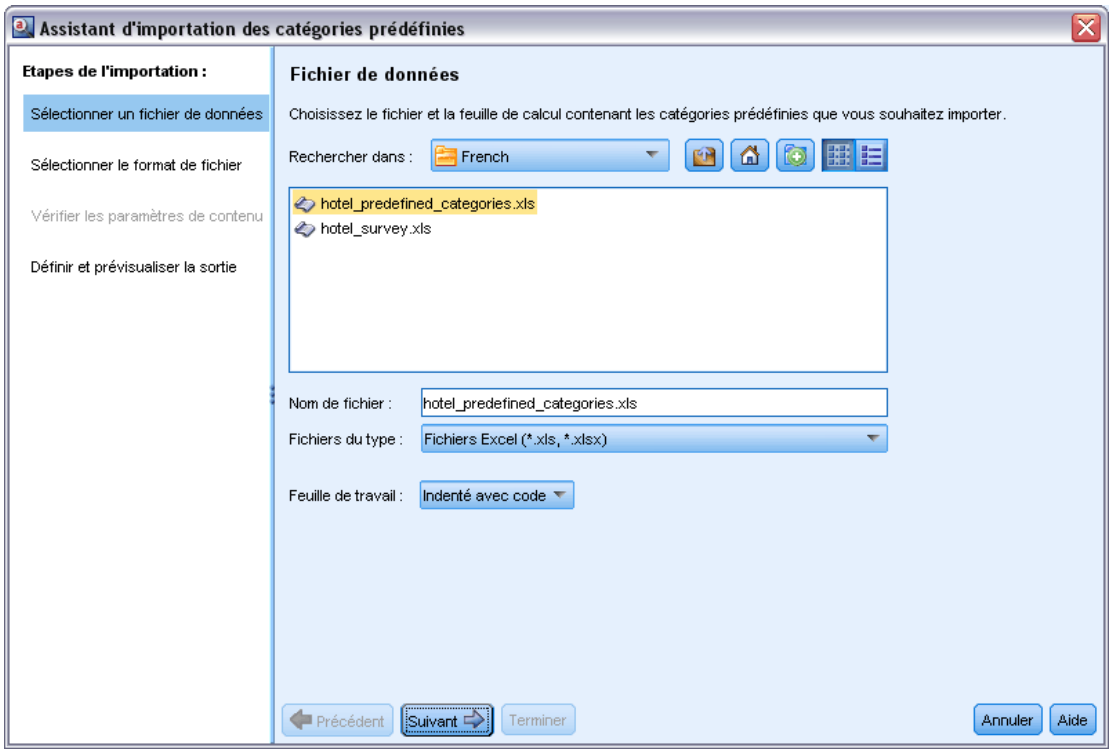
- Format liste plate : Pour plus d'informations, reportez-vous à la section Format liste plate sur p. 217.
- Format compact : Pour plus d'informations, reportez-vous à la section Format compact sur p. 218.
- Format indenté : Pour plus d'informations, reportez-vous à la section Format indenté sur p. 219.

Remarque : Pour la plupart des langues, un flux de démonstration et un fichier de données sont disponibles pour illustrer l'importation des catégories prédéfinies. Recherchez dans le sous-répertoire `<modeler_installation_directory>\Demos\Text_Analytics\` les fichiers correspondant à votre langue.

Pour importer des catégories prédéfinies

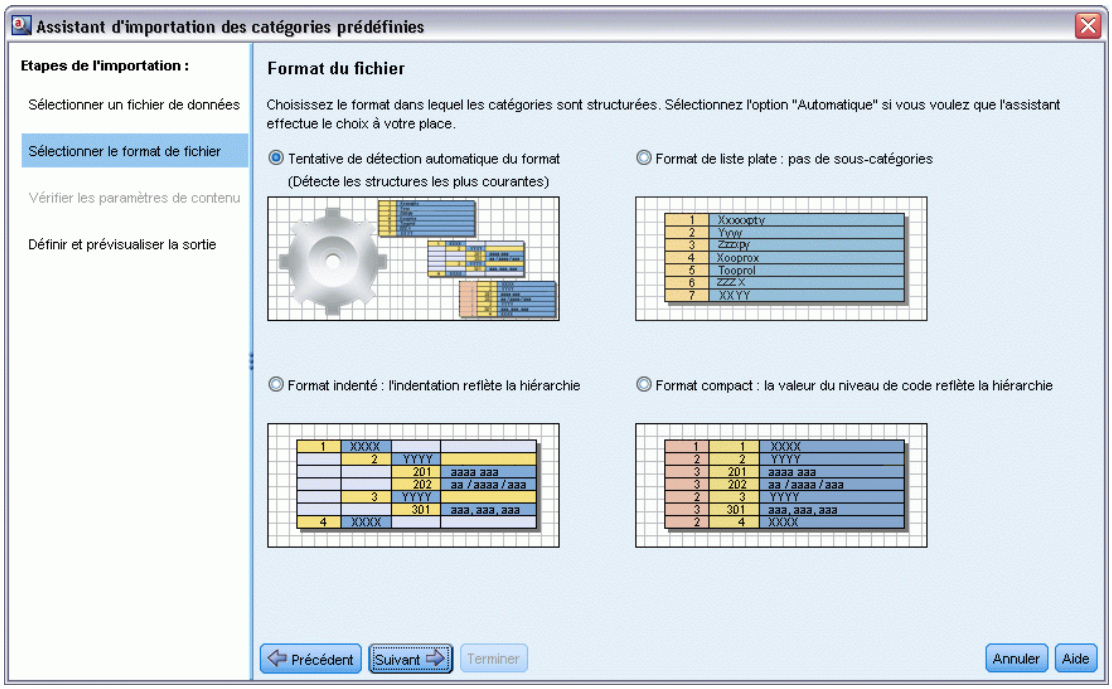
▶ Dans les menus, choisissez Catégories > Gestion des catégories > Importer des catégories prédéfinies. Un assistant d'importation des catégories prédéfinies apparaît.

Figure 10-16
Assistant Import des catégories prédéfinies



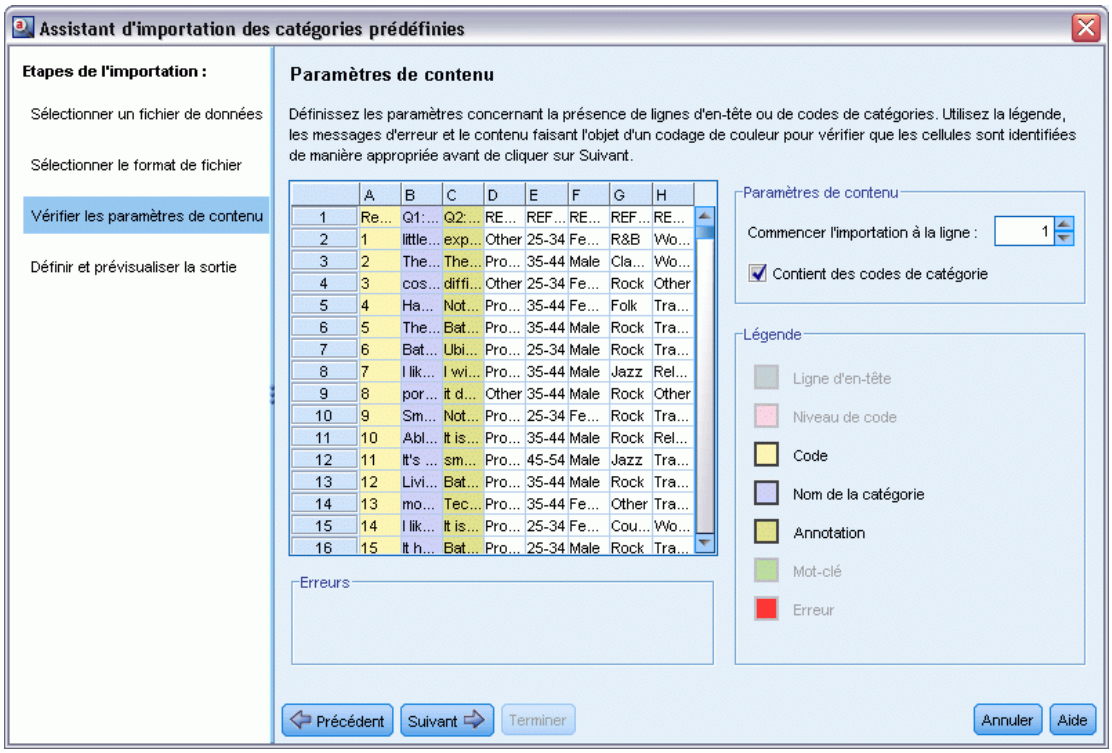
- Dans la liste déroulante Rechercher dans, sélectionnez le pilote et le dossier dans lesquels se trouve le fichier.
- Sélectionnez le fichier dans la liste. Le nom du fichier apparaît dans la zone de texte Nom de fichier.
- Sélectionnez dans la liste, la feuille de calcul contenant les catégories prédéfinies. Le nom de la feuille de calcul apparaît dans le champ Feuille de calcul.
- Cliquez sur Suivant pour commencer à choisir le format de données.

Figure 10-17
Boîte de dialogue Importer des catégories prédéfinies, étape Format de données



- Choisissez le format de votre fichier, ou choisissez l'option permettant au produit de détecter automatiquement le format. La détection automatique est la plus efficace sur les formats les plus courants.
  - Format liste plate : Pour plus d'informations, reportez-vous à la section Format liste plate sur p. 217.
  - Format compact : Pour plus d'informations, reportez-vous à la section Format compact sur p. 218.
  - Format indenté : Pour plus d'informations, reportez-vous à la section Format indenté sur p. 219.
- Cliquez sur Suivant pour définir les options supplémentaires d'import. Si vous avez choisi de détecter automatiquement le format, vous serez orienté directement vers l'étape finale.

Figure 10-18
Importer des catégories prédéfinies, étape Options d'import



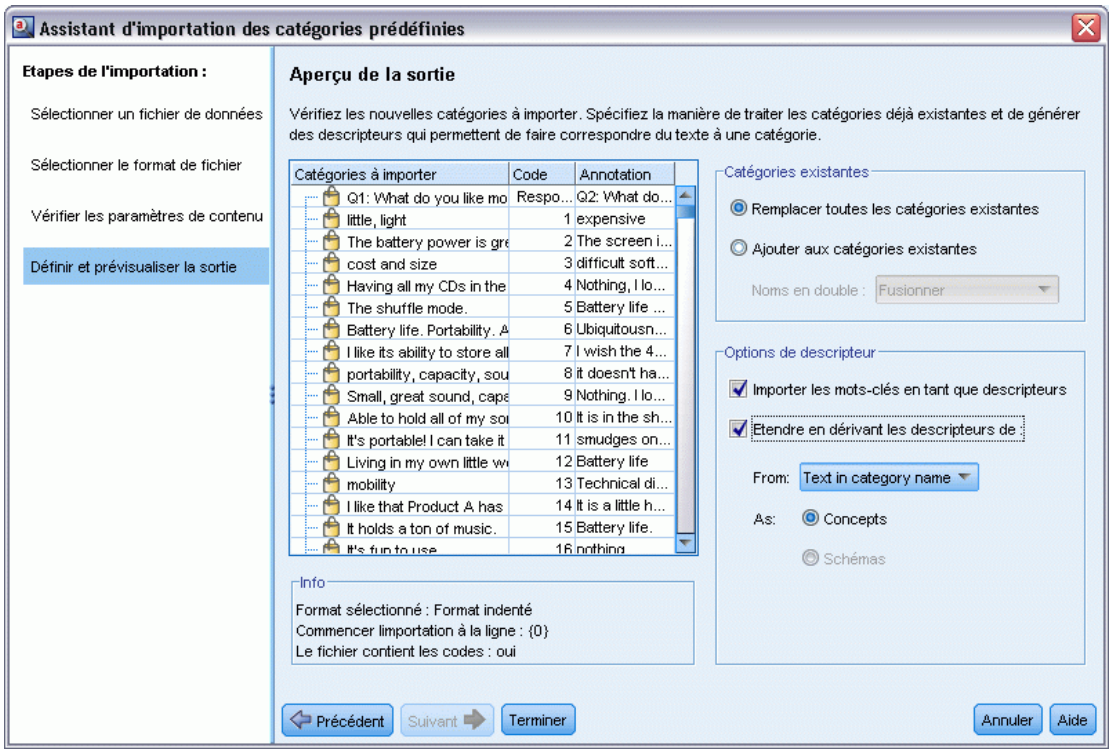
- Si une ou plusieurs lignes contiennent des titres de colonnes ou d'autres types d'information, choisissez le numéro de la ligne à partir de laquelle vous voulez commencer l'importation dans l'option Commencer l'import à la ligne. Par exemple : si vos noms de catégories commencent à la ligne 7, vous devez entrer le chiffre 7 dans cette option pour importer correctement votre fichier.

- Si votre fichier contient des codes de catégories, choisissez l'option Contient des codes de catégories. Ainsi, il sera plus facile à l'assistant de reconnaître correctement vos données.

- Consultez les cellules de codage couleur et la légende afin de vous assurer que les données ont été correctement identifiées. Toute erreur détectée dans le fichier est affichée en rouge et référencée sous le tableau d'aperçu de format. Si le format sélectionné n'est pas le bon, revenez en arrière et choisissez-en un autre. Si vous devez apporter des corrections à votre fichier, effectuez vos changements et redémarrez l'assistant en sélectionnant de nouveau votre fichier. Vous devez corriger toutes les erreurs avant de fermer l'assistant.

- Cliquez sur Suivant pour consulter l'ensemble des catégories et des sous-catégories qui seront importées et pour définir comment créer des descripteurs pour ces catégories.

Figure 10-19
Boîte de dialogue Importer des catégories prédéfinies, étape Prévisualisation



- Consultez l'ensemble des catégories qui seront importées dans le tableau. Si les mots clés que vous avez définis comme descripteurs n'apparaissent pas, il se peut qu'ils n'aient pas été reconnus durant l'import. Assurez-vous qu'ils ont été correctement préfixés et qu'ils apparaissent dans la bonne cellule.

- Choisissez la manière dont vous voulez traiter les catégories pré-existantes dans votre session.
  - Remplacer toutes les catégories existantes. Cette option purge toutes les catégories existantes et les catégories nouvellement importées sont alors utilisées seules à leur place.
  - Ajouter aux catégories existantes. Cette option vous permettra d'importer les catégories et de les fusionner avec les catégories déjà existantes. En ajoutant des catégories à des catégories déjà existantes, vous devez déterminer comment vous voulez que les doublons soient traités. Un des choix (option : Fusionner) est de fusionner toute catégorie importée avec les catégories existantes si elles ont un nom de catégorie commun. Une autre possibilité (option : Exclure de l'import) est d'interdire l'import de catégories si une catégorie portant le même nom existe déjà.

- Import de mot-clés comme descripteurs est une option de descripteur, qui importe les mots-clés identifiés dans vos données en tant que descripteurs pour la catégorie associée.

- Étendre les catégories à partir des descripteurs est une option générant des descripteurs à partir de mots qui représentent le nom de la catégorie ou de la sous-catégorie, et/ou à partir de mots qui forment l'annotation. Si ces mots correspondent aux résultats extraits, alors ils seront ajoutés en tant que descripteurs de la catégorie. Cette option produit de meilleurs résultats lorsque les noms de catégories sont à la fois longs et descriptifs. C'est une méthode rapide pour générer des

descripteurs de catégories, qui à leur tour permettent à la catégorie de capturer les enregistrements contenant ces descripteurs.

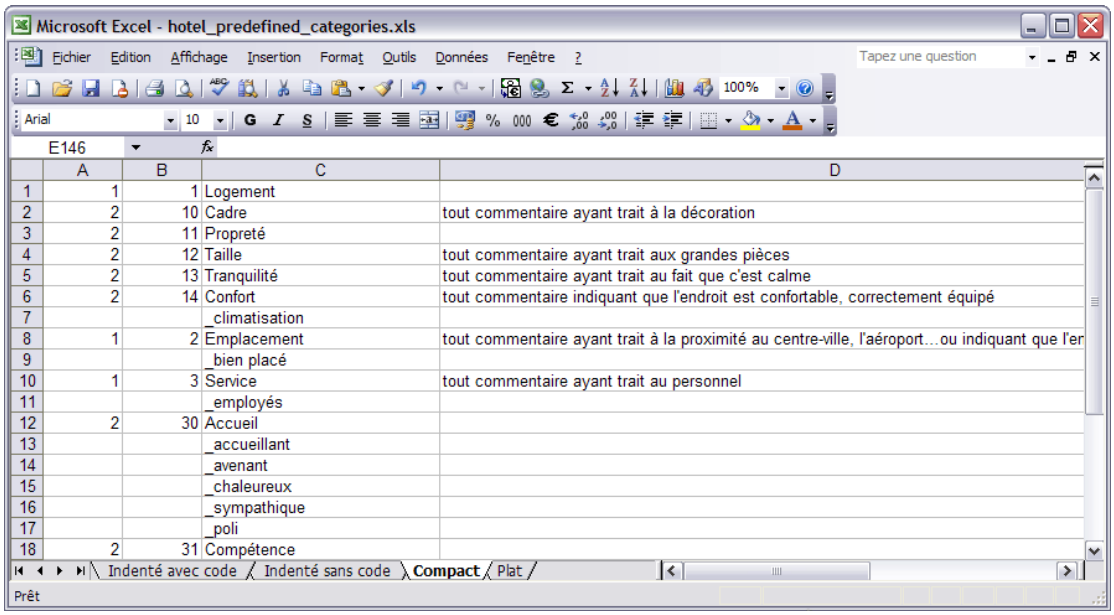
- Le champ À partir de vous permet de choisir depuis quel texte les descripteurs seront choisis, les noms ou les catégories ou sous-catégories, les mots dans les annotations, ou les deux.
- Le champ En tant que vous permet de créer ces descripteurs sous la forme de concepts ou de patrons TLA. Si l'extraction TLA n'a pas eu lieu, les options de patrons de cet assistant sont désactivés.

▶ Cliquez sur Terminer pour importer les catégories prédéfinies dans le panneau Catégories.

Format liste plate

Dans ce format liste plate, il n'y a qu'un niveau de catégories, sans aucune hiérarchie, ce qui signifie qu'il n'y a pas de sous-catégories ou de sous-réseaux. Les noms de catégories sont dans une seule colonne.

Figure 10-20
Exemple de format liste plate



	A	B	C	D
1	1	1	Logement	
2	2	10	Cadre	tout commentaire ayant trait à la décoration
3	2	11	Propreté	
4	2	12	Taille	tout commentaire ayant trait aux grandes pièces
5	2	13	Tranquilité	tout commentaire ayant trait au fait que c'est calme
6	2	14	Confort	tout commentaire indiquant que l'endroit est confortable, correctement équipé
7			_climatisation	
8	1	2	Emplacement	tout commentaire ayant trait à la proximité au centre-ville, l'aéroport…ou indiquant que l'er
9			_bien placé	
10	1	3	Service	tout commentaire ayant trait au personnel
11			_employés	
12	2	30	Accueil	
13			_accueillant	
14			_avenant	
15			_chaleureux	
16			_sympathique	
17			_poli	
18	2	31	Compétence	

Les informations suivantes peuvent être contenues dans un fichier de ce format :

- La colonne des **codes** optionnels contient des valeurs numériques qui identifient chaque catégorie de façon unique. Si vous précisez que les fichiers de données contiennent des codes (option Contient des codes de catégories dans l'étape Paramètres de contenu), alors une colonne contenant des codes uniques pour chaque catégorie doit exister dans la cellule située à gauche du nom de la catégorie. Si vos données ne contiennent pas de codes, mais que vous voulez créer des codes plus tard, vous pourrez toujours le faire en utilisant l'option Catégories > Gestion des catégories > Générer automatiquement des codes.
- Une colonne *obligatoire***nom des catégories** contient tous les noms de catégories. Cette colonne est nécessaire pour importer en utilisant ce format.

- Des **annotations** optionnelles sont disponibles dans la cellule située à droite du nom de catégorie. Cette annotation consiste en un texte décrivant vos catégories ou sous-catégories.
- Des **mots clés** optionnels peuvent être importés en tant que descripteurs de catégories. Afin qu'ils soient reconnus, ces mots clés doivent être placés dans la cellule située directement sous le nom de la catégorie/sous-catégorie associée. La liste des mots clés doit également être préfixée d'un caractère de soulignement (_) par exemple « `_armesàfeu, armes / revolvers` ». La cellule de mot clé peut contenir un ou plusieurs mots pour décrire chaque catégorie. Ces mots seront importés en tant que descripteurs, ou ignorés, en fonction de ce que vous aurez spécifié durant la dernière étape de l'assistant. Plus tard, les descripteurs sont comparés aux résultats extraits du texte. Si une correspondance est détectée, le document ou l'enregistrement est attribué à la catégorie contenant ce descripteur.

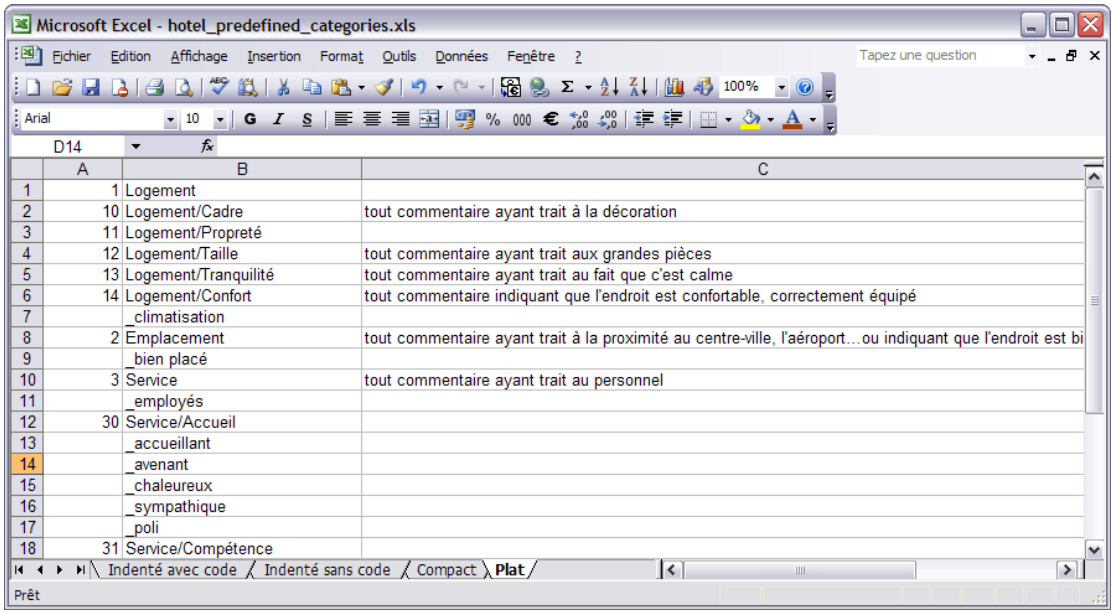
Table 10-8
Format liste plate avec codes, mots clés et annotations

Colonne A	**Colonne B**	**Colonne C**
Code de catégorie (*facultatif*)	Nom de catégorie	Annotation
	_Descripteur/liste des mots-clés (*facultatif*)	

Format compact

Ce format compact est similaire au format liste plate, mis à part le fait qu'il est utilisé avec les catégories hiérarchiques. Par conséquent, une colonne de niveau de code est nécessaire afin de déterminer le niveau hiérarchique de chaque catégorie et sous-catégorie.

Figure 10-21
Exemple de fichier de catégorie prédéfinie compact sous Microsoft Excel



	A	B	C
1	1	Logement	
2	10	Logement/Cadre	tout commentaire ayant trait à la décoration
3	11	Logement/Propreté	
4	12	Logement/Taille	tout commentaire ayant trait aux grandes pièces
5	13	Logement/Tranquilité	tout commentaire ayant trait au fait que c'est calme
6	14	Logement/Confort	tout commentaire indiquant que l'endroit est confortable, correctement équipé
7		_climatisation	
8	2	Emplacement	tout commentaire ayant trait à la proximité au centre-ville, l'aéroport...ou indiquant que l'endroit est bi
9		_bien placé	
10	3	Service	tout commentaire ayant trait au personnel
11		_employés	
12	30	Service/Accueil	
13		_accueillant	
14		_avenant	
15		_chaleureux	
16		_sympathique	
17		_poli	
18	31	Service/Compétence	

Les informations suivantes peuvent être contenues dans un fichier de ce format :

- Une colonne *obligatoire* de **niveau de code** contient les nombres indiquant la position hiérarchique des informations ultérieures pour cette ligne. Par exemple; si les valeurs 1, 2 ou 3 sont spécifiées et que vous avez à la fois des catégories et des sous-catégories, alors 1 correspond aux catégories, 2 aux sous-catégories et 3 aux sous-sous-catégories. Si vous avez uniquement des catégories et des sous-catégories, 1 correspond aux catégories et 2 correspond aux sous-catégories. Et ainsi de suite, jusqu'à la profondeur de catégories souhaitée.
- La colonne optionnelle des **codes** contient des valeurs numériques qui identifient chaque catégorie de façon unique. Si vous précisez que les fichiers de données contiennent des codes (option Contient des codes de catégories dans l'étape Paramètres de contenu), alors une colonne contenant des codes uniques pour chaque catégorie doit exister dans la cellule située à gauche du nom de la catégorie. Si vos données ne contiennent pas de codes, mais que vous voulez créer des codes plus tard, vous pourrez toujours le faire en utilisant l'option Catégories > Gestion des catégories > Générer automatiquement des codes.
- Une colonne *obligatoire***nom des catégories** contient tous les noms de catégories et de sous-catégories. Cette colonne est nécessaire pour importer en utilisant ce format.
- Des **annotations** optionnelles sont disponibles dans la cellule située à droite du nom de catégorie. Cette annotation consiste en un texte décrivant vos catégories ou sous-catégories.
- Des **mots clés** optionnels peuvent être importés en tant que descripteurs de catégories. Afin qu'ils soient reconnus, ces mots clés doivent être placés dans la cellule située directement sous le nom de la catégorie/sous-catégorie associée. La liste des mots clés doit également être préfixée d'un caractère de soulignement (_) par exemple « `_armesàfeu, armes / revolvers` ». La cellule de mot clé peut contenir un ou plusieurs mots pour décrire chaque catégorie. Ces mots seront importés en tant que descripteurs, ou ignorés, en fonction de ce que vous aurez spécifié durant la dernière étape de l'assistant. Plus tard, les descripteurs sont comparés aux résultats extraits du texte. Si une correspondance est détectée, le document ou l'enregistrement est attribué à la catégorie contenant ce descripteur.

Table 10-9
Exemple de format compact avec codes

Colonne A	**Colonne B**	**Colonne C**
Niveau de code hiérarchique	Code de catégorie (*facultatif*)	Nom de la catégorie
Niveau de code hiérarchique	Code de sous-catégorie (*facultatif*)	Nom de sous-catégorie

Table 10-10
Exemple de format compact sans codes

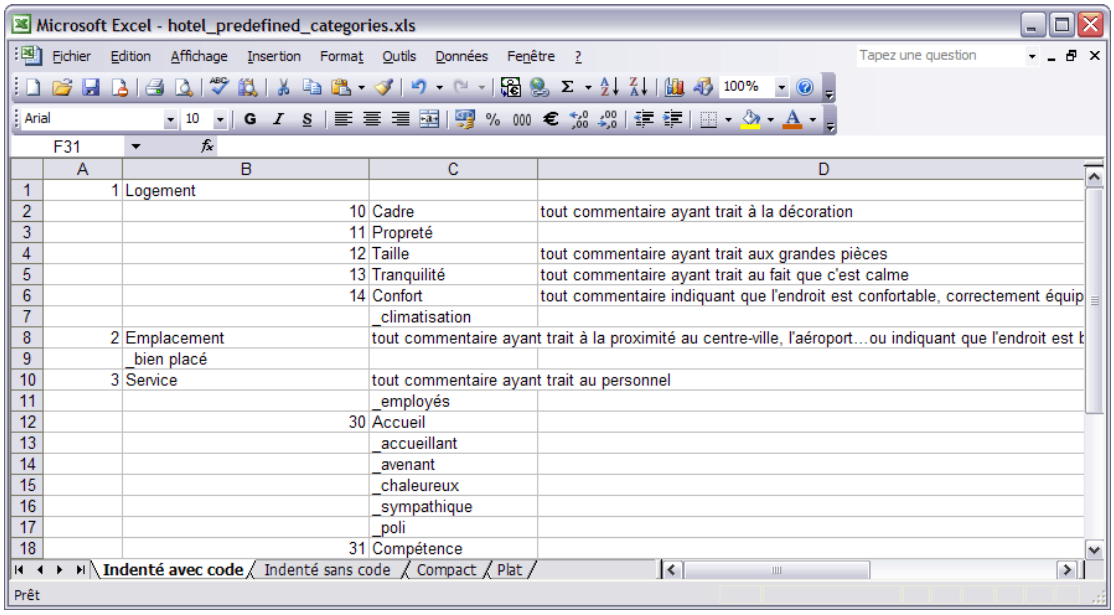
Colonne A	**Colonne B**
Niveau de code hiérarchique	Nom de la catégorie
Niveau de code hiérarchique	Nom de sous-catégorie

Format indenté

Dans le format de fichier indenté, le contenu est organisé de façon hiérarchique ; c'est-à-dire qu'il contient des catégories et un ou plusieurs niveaux de sous-catégories. De plus, sa structure est indentée pour refléter cette hiérarchie. Chaque ligne du fichier contient soit une catégorie, soit une sous-catégorie. Les sous-catégories sont indentées à partir des catégories, et toute sous-sous-catégorie est indentée à partir des sous-catégories, etc. Vous pouvez créer manuellement

cette structure dans Microsoft Excel ou utiliser une structure exportée à partir d'un autre produit et enregistrée sous un format Microsoft Excel.

Figure 10-22
Exemple d'une catégorie indentée dans Microsoft Excel



	A	B	C	D
1	1	Logement		
2		10	Cadre	tout commentaire ayant trait à la décoration
3		11	Propreté	
4		12	Taille	tout commentaire ayant trait aux grandes pièces
5		13	Tranquilité	tout commentaire ayant trait au fait que c'est calme
6		14	Confort	tout commentaire indiquant que l'endroit est confortable, correctement équip
7			_climatisation	
8	2	Emplacement	tout commentaire ayant trait à la proximité au centre-ville, l'aéroport…ou indiquant que l'endroit est b	
9		_bien placé		
10	3	Service	tout commentaire ayant trait au personnel	
11			_employés	
12		30	Accueil	
13			_accueillant	
14			_avenant	
15			_chaleureux	
16			_sympathique	
17			_poli	
18		31	Compétence	

- Les **Codes et les noms de catégories de niveau supérieur** occupent respectivement les colonnes A et B. Ou, si aucun code n'est présent, alors le nom de catégorie occupe la colonne A.
- Les **codes de sous-catégories et les noms de sous-catégories** occupent respectivement les colonnes B et C. Ou, si aucun code n'est présent, les noms de sous-catégories occupent la colonne B. La sous-catégorie est membre d'une catégorie. Vous ne pouvez pas avoir de sous-catégories si vous n'avez pas de catégories.

Table 10-11
Structure indentée avec des codes

Colonne A	**Colonne B**	**Colonne C**	**Colonne D**
Code de catégorie (*facultatif*)	Nom de catégorie		
	Code de sous-catégorie (*facultatif*)	Nom de sous-catégorie	
		Code de sous-sous-catégorie (*facultatif*)	Nom de sous-sous-catégorie

Table 10-12
Structure indentée sans codes

Colonne A	**Colonne B**	**Colonne C**
Nom de la catégorie		
	Nom de sous-catégorie	
		Nom de sous-sous-catégorie

Les informations suivantes peuvent être contenues dans un fichier de ce format :

- Les **codes** optionnels doivent être des valeurs identifiant de manière unique chaque catégorie ou sous-catégorie. Si vous précisez que les fichiers de données contiennent des codes (option Contient des codes de catégories dans l'étape Paramètres de contenu), alors un code unique pour chaque catégorie ou sous-catégorie doit exister dans la cellule située à gauche du nom de la catégorie/sous-catégorie. Si vos données ne contiennent pas de codes, mais que vous voulez créer des codes plus tard, vous pourrez toujours le faire en utilisant l'option Catégories > Gestion des catégories > Générer automatiquement des codes.
- Un **nom** est *obligatoire* pour chaque catégorie et sous-catégorie. Les sous-catégories doivent être indentées des catégories d'une cellule vers la droite et sur une ligne différente.
- Des **annotations** optionnelles sont disponibles dans la cellule située à droite du nom de catégorie. Cette annotation consiste en un texte décrivant vos catégories ou sous-catégories.
- Des **mots clés** optionnels peuvent être importés en tant que descripteurs de catégories. Afin qu'ils soient reconnus, ces mots clés doivent être placés dans la cellule située directement sous le nom de la catégorie/sous-catégorie associée. La liste des mots clés doit également être préfixée d'un caractère de soulignement (_) par exemple « `_armesàfeu, armes / revolvers` ». La cellule de mot clé peut contenir un ou plusieurs mots pour décrire chaque catégorie. Ces mots seront importés en tant que descripteurs, ou ignorés, en fonction de ce que vous aurez spécifié durant la dernière étape de l'assistant. Plus tard, les descripteurs sont comparés aux résultats extraits du texte. Si une correspondance est détectée, le document ou l'enregistrement est attribué à la catégorie contenant ce descripteur.

Important ! Si vous utilisez un code à un niveau, vous devez inclure un code pour chaque catégorie et sous-catégorie. Si cela n'est pas le cas, votre import échouera.

Exporter des catégories

Vous pouvez également exporter vers un fichier Microsoft Excel (**.xls*, **.xlsx*), les catégories que vous avez dans une session interactive ouverte. Les données qui seront exportées viennent essentiellement du contenu actuel du panneau des catégories ou des propriétés des catégories. Toutefois, nous recommandons de scorer de nouveau, si vous prévoyez d'exporter également la valeur de score de Docs..

Toujours exporté..

- S'ils sont présents, les codes de catégories
- Noms de catégories (et de sous-catégories)
- S'ils sont présents, les niveaux de codes (Format *Plat*/*Compact*)
- En-têtes de colonnes (Format *Plat*/*Compact*)

Exporté de manière optionnelle...

- Scores de Docs.
- Annotations de catégories
- Noms des descripteurs
- Nombres de descripteurs

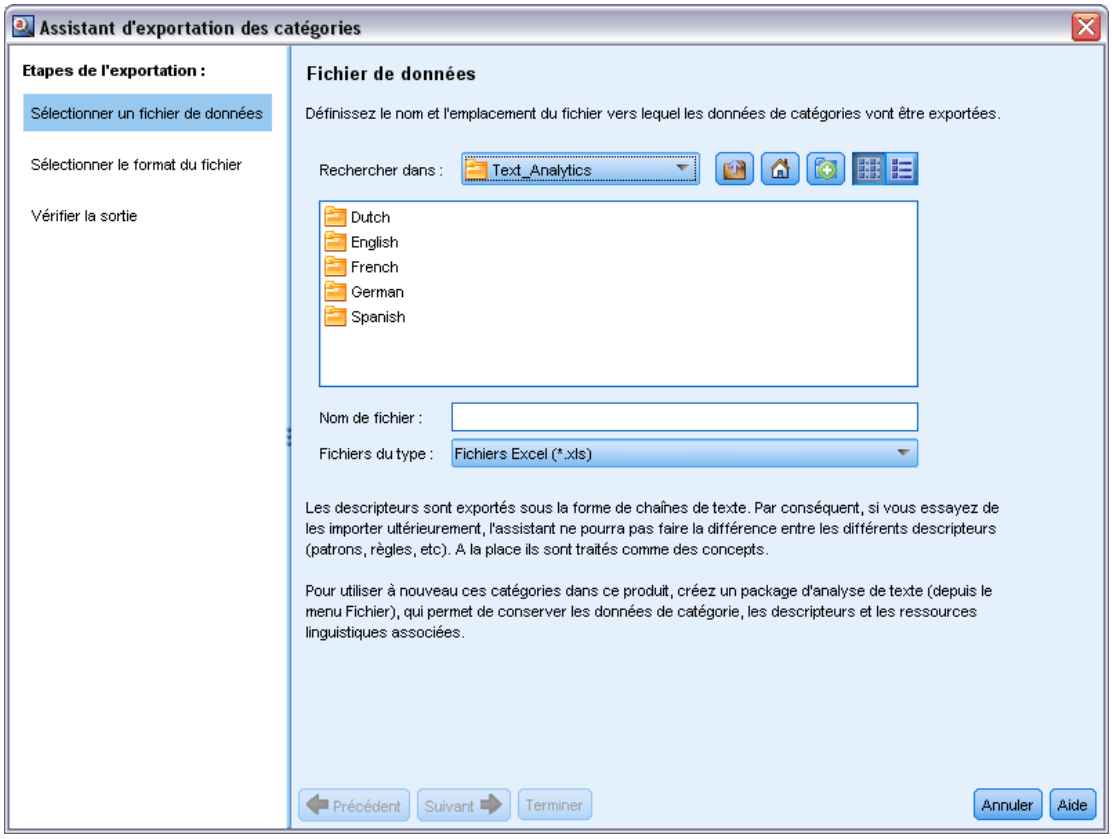
Important ! Lorsque vous exportez des descripteurs, ils sont convertis en chaînes de texte et un caractère de soulignement leur est ajouté comme préfixe. Si vous effectuez de nouveau une importation dans ce produit, la capacité de distinction entre les descripteurs qui sont des patrons, des règles de catégorie ou des concepts bruts est perdue. Si vous souhaitez réutiliser ces catégories dans ce produit, nous vous recommandons fortement de créer un fichier de package d'analyse de texte (TAP), car ce format permet de conserver tous les descripteurs tels qu'ils sont définis, ainsi

que toutes vos catégories, vos codes et les ressources linguistiques utilisées. Les fichiers TAP peuvent être utilisés à la fois dans IBM® SPSS® Modeler Text Analytics et dans IBM® SPSS® Text Analytics for Surveys . Pour plus d'informations, reportez-vous à la section Utilisation des packages d'analyse de texte sur p. 224.

Pour exporter des catégories prédéfinies

▶ Dans les menus, choisissez Catégories > Gestion des catégories > Exporter des catégories. Un assistant d'exportation des catégories apparaît.

Figure 10-23
Assistant d'export de catégories, étape 1

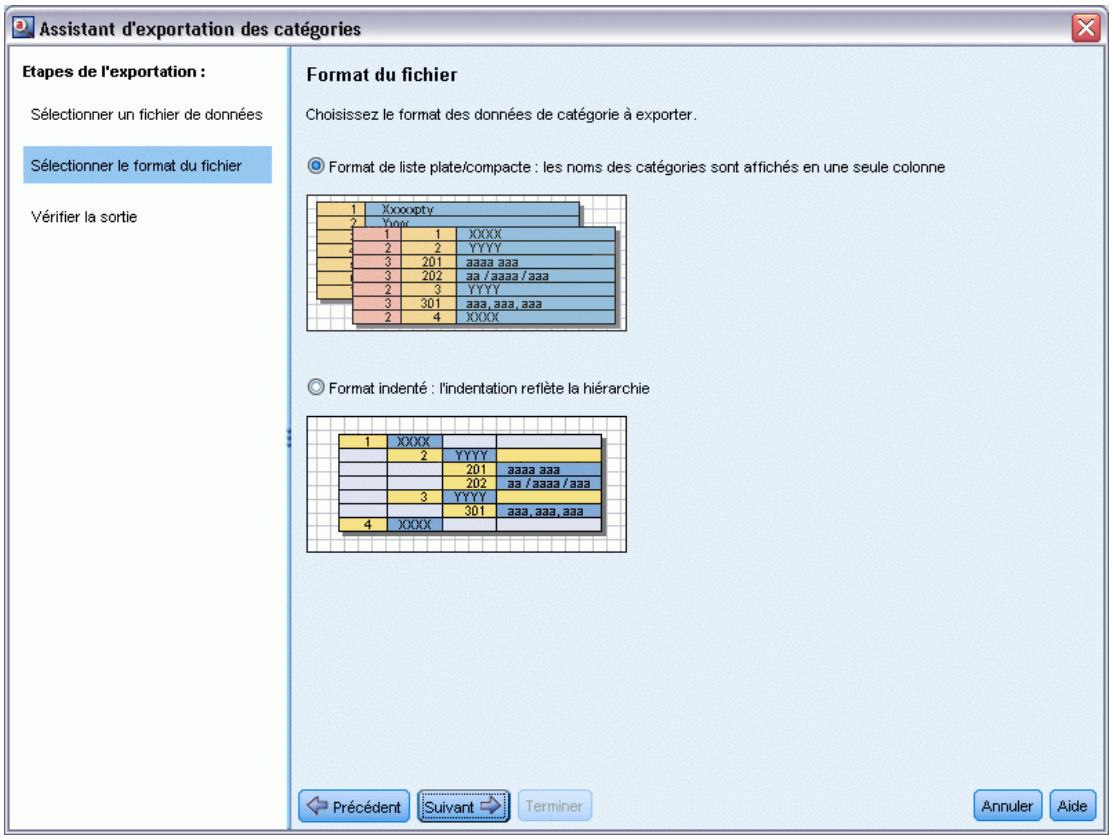


▶ Sélectionnez un emplacement et saisissez le nom du fichier à exporter.

▶ Entrez le nom du fichier de sortie dans la zone de texte Nom du Fichier.

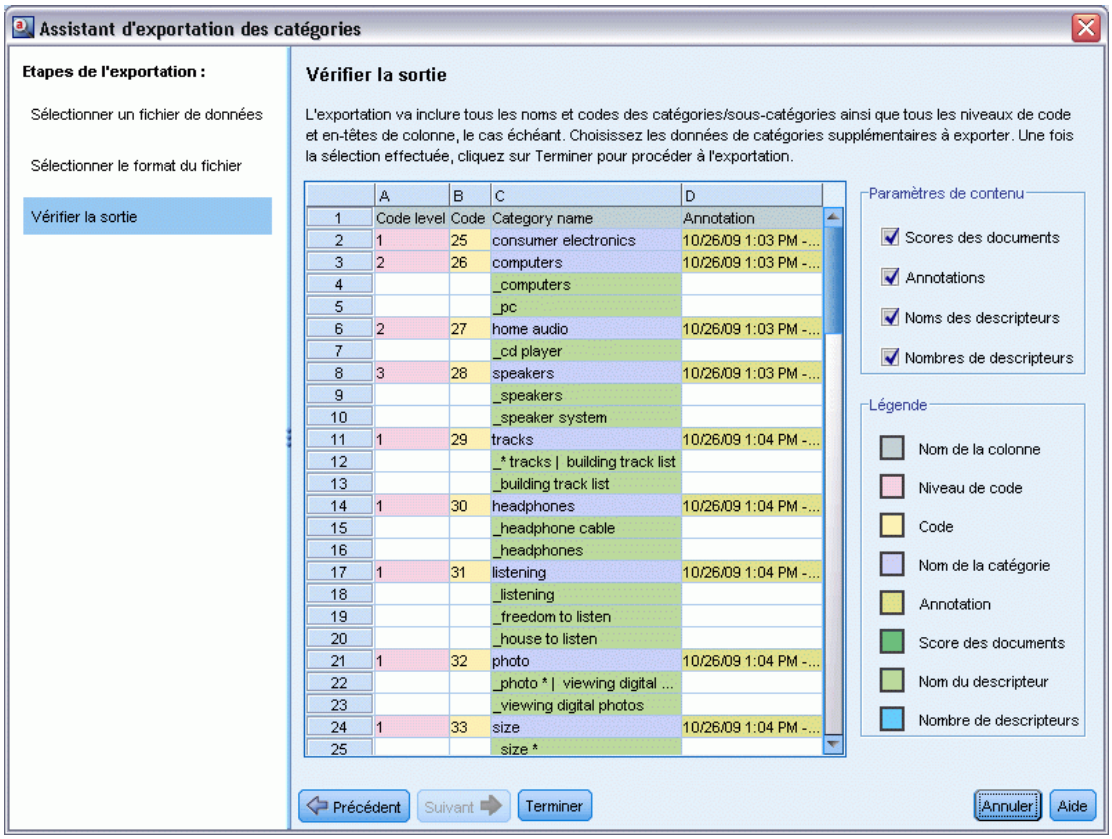
▶ Cliquez sur Suivant pour choisir le format dans lequel vous allez exporter vos données de catégorie.

Figure 10-24
Assistant d'export de catégories, étape 2



- Choisissez le format parmi l'une des options suivantes :
  - Format liste plate ou compacte : Pour plus d'informations, reportez-vous à la section Format liste plate sur p. 217. La liste plate ne contient aucune sous-catégorie. Pour plus d'informations, reportez-vous à la section Format compact sur p. 218. Le format de liste compacte contient des données hiérarchiques.
  - Format indenté : Pour plus d'informations, reportez-vous à la section Format indenté sur p. 219.
- Cliquez sur Suivant pour commencer à choisir le contenu à exporter et pour consulter les données proposées.

Figure 10-25
Assistant d'export de catégories, étape 3



	A	B	C	D
1	Code level	Code	Category name	Annotation
2	1	25	consumer electronics	10/26/09 1:03 PM -...
3	2	26	computers	10/26/09 1:03 PM -...
4			_computers	
5			_pc	
6	2	27	home audio	10/26/09 1:03 PM -...
7			_cd player	
8	3	28	speakers	10/26/09 1:03 PM -...
9			_speakers	
10			_speaker system	
11	1	29	tracks	10/26/09 1:04 PM -...
12			_* tracks | building track list	
13			_building track list	
14	1	30	headphones	10/26/09 1:04 PM -...
15			_headphone cable	
16			_headphones	
17	1	31	listening	10/26/09 1:04 PM -...
18			_listening	
19			_freedom to listen	
20			_house to listen	
21	1	32	photo	10/26/09 1:04 PM -...
22			_photo * | viewing digital ...	
23			_viewing digital photos	
24	1	33	size	10/26/09 1:04 PM -...
25			_size *	

- Consulter le contenu du fichier exporté.
- Sélectionnez ou désélectionnez les paramètres de contenu supplémentaire à exporter tels que les annotations ou les noms de descripteur.
- Cliquez sur Terminer pour exporter les catégories.

Utilisation des packages d'analyse de texte

Un package d'analyse de texte, également appelé TAP (Text Analysis Package), sert en tant que modèle de catégorisations de réponses texte. L'utilisation d'un TAP est un moyen facile de catégoriser vos données textuelles en intervenant le moins possible car il contient les ensembles de catégories prédéfinis et les ressources linguistiques nécessaires au codage d'un grand nombre d'enregistrements rapidement et automatiquement. Les ressources linguistiques permettent d'analyser et d'exploiter les données textuelles pour en extraire des concepts clés. En fonction des concepts clés et des patrons trouvés dans le texte, les enregistrements peuvent être catégorisés dans l'ensemble de catégories sélectionné dans le TAP. Vous pouvez créer votre propre TAP ou enmettre un à jour.

Un TAP est composé des éléments suivants :

- **Ensemble(s) de catégories**. Un ensemble de catégories est principalement composé de catégories prédéfinies, de codes de catégorie, de descripteurs pour chaque catégorie et enfin d'un nom pour l'ensemble en son entier. Les descripteurs sont des éléments linguistiques (concepts, types, patrons et règles) comme le terme *cher* ou le patron *bon prix*. Les descripteurs sont utilisés pour définir une catégorie. Ainsi, lorsque le texte correspond à un descripteur de catégorie, le document ou l'enregistrement est placé dans cette catégorie.
- **Ressources linguistiques**. Les ressources linguistiques sont un ensemble de bibliothèques et de ressources avancées qui permettent d'extraire des concepts et patrons clés. Ces concepts et patrons d'extraction sont utilisés comme descripteurs qui permettent aux enregistrements d'être placés dans une catégorie de l'ensemble de catégories.

Vous pouvez créer votre propre TAP, en mettre un à jour, ou charger des packages d'analyse de texte.

Après avoir sélectionné un TAP et un ensemble de catégories, IBM® SPSS® Modeler Text Analytics peut effectuer une extraction et catégoriser vos documents.

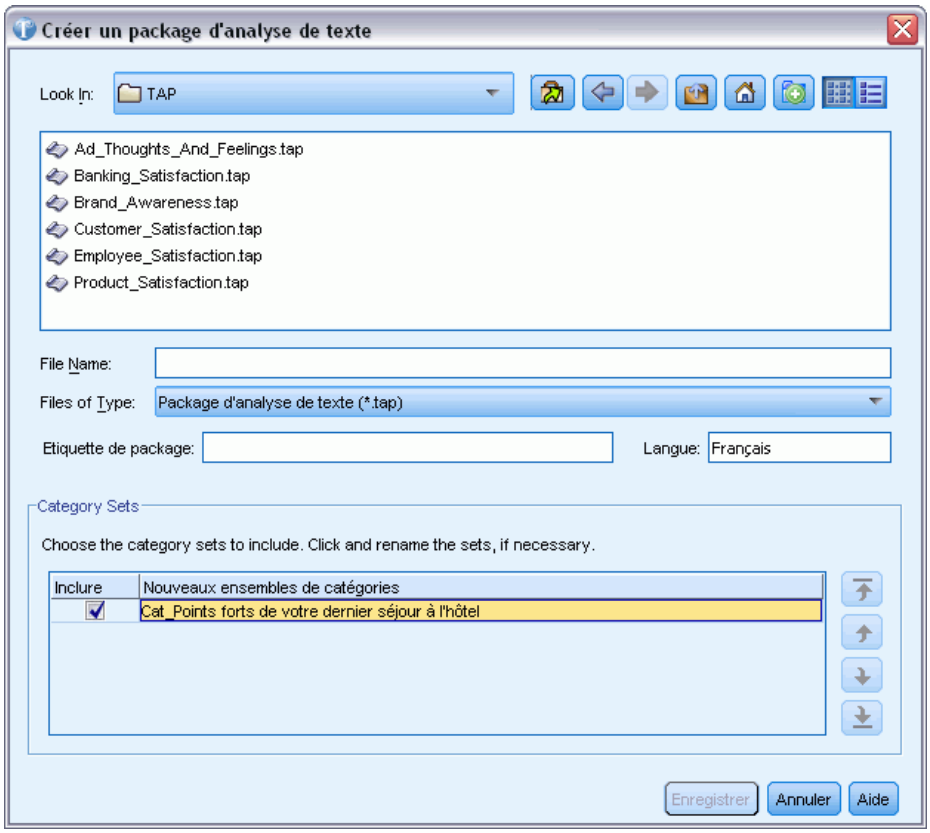
Remarque : Vous pouvez créer et interchanger les TAP entre IBM® SPSS® Text Analytics for Surveys et SPSS Modeler Text Analytics .

Création des packages d'analyse de texte

Lorsque vous avez une session contenant au moins une catégorie et des ressources, vous pouvez créer un package d'analyse de texte (TAP) à partir du contenu de la session interactive ouverte. L'ensemble de catégories et de descripteurs (concepts, types, règles ou résultats de patrons TLA) peut être transformé en TAP utilisant toutes les ressources linguistiques ouvertes dans l'éditeur de ressources.

Vous pouvez voir la langue pour laquelle les ressource ont été créées. La langue est définie dans l'onglet Ressources avancées de Editeur de modèle ou de Editeur de ressources.

Figure 10-26
Boîte de dialogue Créer un package d'analyse de texte

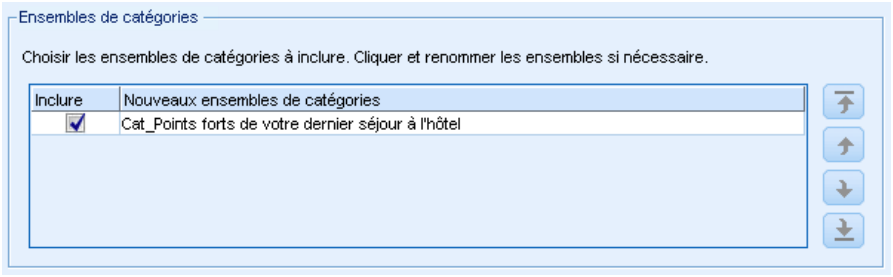


Pour créer un package d'analyse de texte

- Dans les menus, choisissez Fichier > Packages d'analyse de texte > Créer un package. La boîte de dialogue Créer un package apparaît.
- Accédez au répertoire dans lequel vous avez enregistré le TAP. Par défaut, les TAP sont enregistrés dans le sous-répertoire \TAP du répertoire d'installation du produit.
- Entrez un nom pour le TAP dans le champ Nom du fichier.
- Saisissez une étiquette dans le champ Etiquette du package. Lorsque vous saisissez un nom de fichier, ce nom apparaît automatiquement comme étiquette mais vous pouvez la modifier.
- Pour exclure un ensemble de catégories du TAP, désélectionnez la case Inclure. Ainsi, il ne sera pas ajouté à votre package. Par défaut, un ensemble de catégories par question est inclus dans le TAP. Le TAP doit contenir au moins un ensemble de catégories.
- Renommer des ensembles de catégories. La colonne Nouvel ensemble de catégories contient des noms génériques par défaut qui sont générés en ajoutant le préfixe Cat_ au nom de la variable de texte. Un simple clic dans la cellule permet de modifier ce nom. Appuyer sur entrée ou cliquer à un autre droit permet d'appliquer la modification du nom. Si vous renommez un ensemble de

catégories, le nom est uniquement modifié dans le TAP et ne modifie pas le nom de la variable dans la session ouverte.

Figure 10-27
Changement de nom des ensembles de catégories



- Vous pouvez changer l'ordre des ensembles de catégories avec les flèches situées à droite du tableau des ensembles de catégories.
- Cliquez sur Enregistrer pour créer le package d'analyse de texte. La boîte de dialogue se ferme.

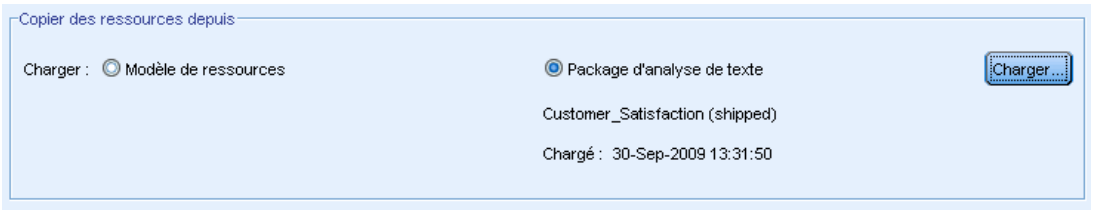
Chargement des packages d'analyse de texte

Lors de la configuration d'un noeud de modélisation Text Mining, vous devez spécifier les ressources à utiliser pendant l'extraction. Plutôt que de choisir un modèle de ressources, vous pouvez sélectionner un TAP (package d'analyse de texte) pour copier non seulement les ressources mais également un ensemble de catégories dans le noeud.

Les TAP sont plus intéressants lors de la création d'un modèle de catégories interactif car vous pouvez utiliser l'ensemble de catégories comme point de départ pour la catégorisation. Lorsque vous exécutez le flux, la session interactive est ouverte et cet ensemble de catégories apparaît dans le panneau Catégories. Ainsi, vous déterminez immédiatement le score de vos documents et de vos enregistrements à l'aide de ces catégories puis vous continuez à affiner, créer et étendre ces catégories jusqu'à ce que vous soyez satisfait. Pour plus d'informations, reportez-vous à la section Stratégies et méthodes de création de catégories sur p. 166.

À partir de la version 14, vous pouvez aussi afficher la langue pour laquelle les ressources de ce TAP ont été définies lorsque vous cliquez sur Charger et que vous choisissez le TAP.

Figure 10-28
Options de l'onglet Modèle pour copier les ressources dans le noeud

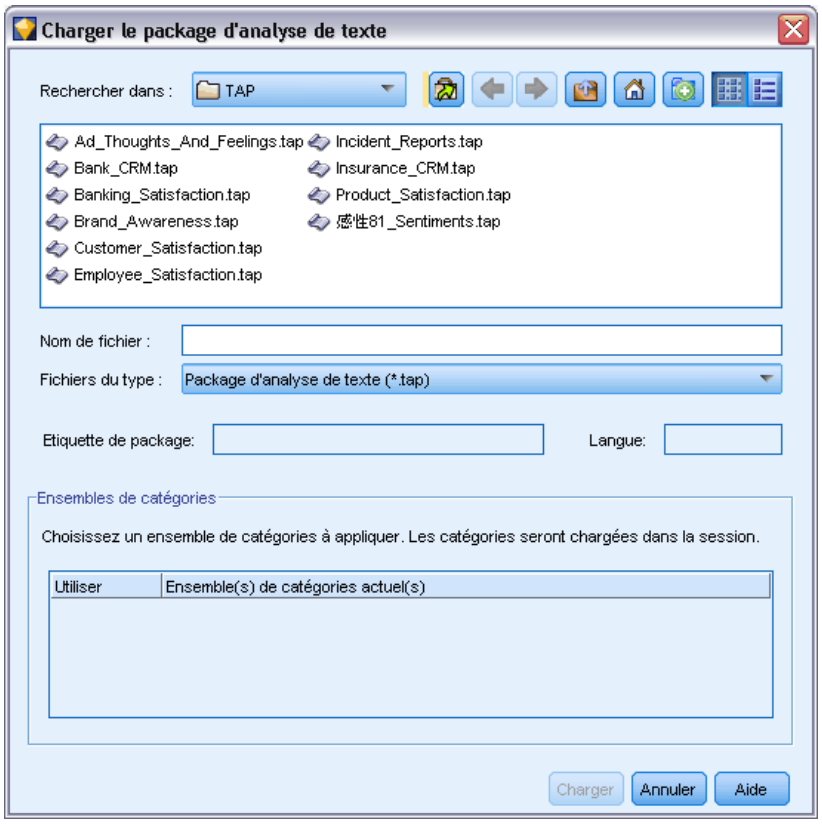


Pour charger un package d'analyse de texte

- Modifiez le noeud de modélisation Text Mining.

- Dans l'onglet Modèle, choisissez *Package d'analyse de texte* dans la section Copier les ressources depuis.
- Cliquez sur Charger. La boîte de dialogue Charger un package d'analyse de texte s'ouvre.

Figure 10-29
Boîte de dialogue Charger un package d'analyse de texte

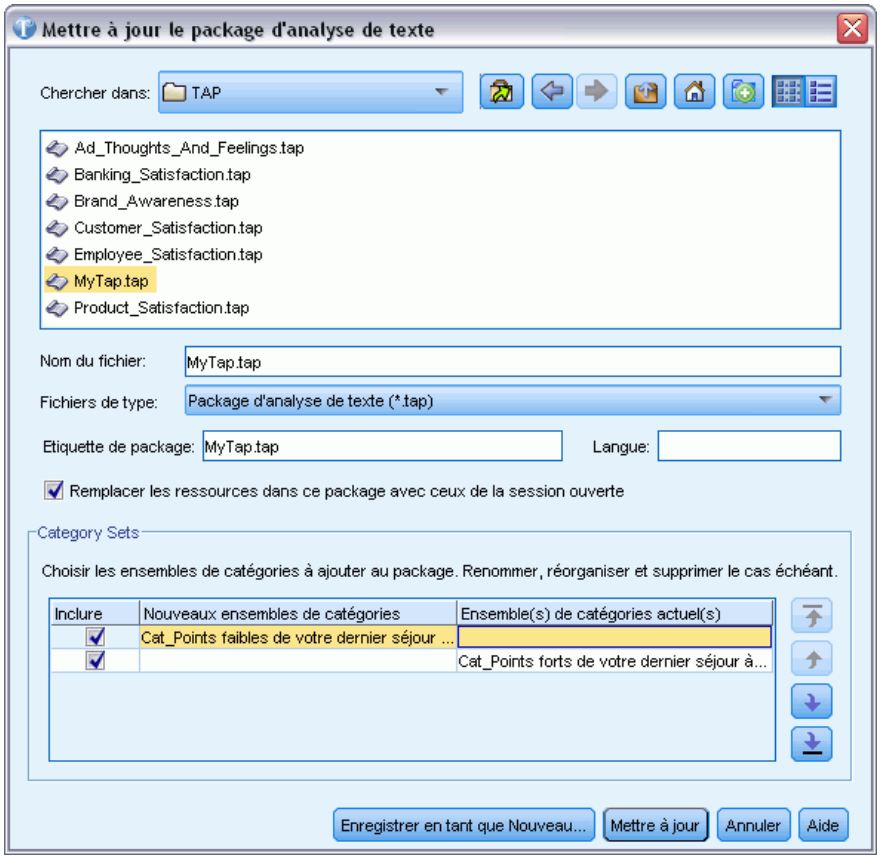


- Recherchez l'emplacement du TAP contenant les ressources et l'ensemble de catégories à copier dans le noeud. Par défaut, les TAP sont enregistrés dans le sous-répertoire \TAP du répertoire d'installation du produit.
- Entrez un nom pour le TAP dans le champ Nom du fichier. Cette étiquette apparaît automatiquement.
- Sélectionnez l'ensemble de catégories à utiliser. Il s'agit de l'ensemble de catégories qui apparaîtra dans la session interactive. Vous pouvez ensuite modifier et améliorer ces catégories manuellement ou en utilisant les options Créer ou étendre des catégories.
- Cliquez sur Charger pour copier le contenu du package d'analyse de texte dans le noeud. La boîte de dialogue se ferme. Lorsqu'un TAP est chargé, une copie de ce TAP est copiée dans le noeud ; ainsi, toutes les modifications effectuées sur les ressources et les catégories ne seront pas reflétées dans le TAP sauf si vous le mettez à jour et le rechargez de manière explicite.

Mise à jour des Packages d'analyse de texte

Si vous apportez des améliorations à un ensemble de catégories, à des ressources linguistiques ou que vous créez un tout nouvel ensemble de catégories, vous pouvez mettre à jour un package d'analyse de texte (TAP) pour que ces améliorations soient plus faciles à réutiliser ultérieurement. Pour ce faire, allez dans la de projet ouverte contenant les informations que vous souhaitez intégrer au TAP. Lorsque vous effectuez une mise à jour, vous pouvez choisir d'ajouter des ensembles de catégories, de remplacer des ressources, de changer l'étiquette du package ou de renommer/réorganiser les ensembles.

Figure 10-30
Boîte de dialogue Mettre à jour un package d'analyse de texte



Pour mettre à jour un package d'analyse de texte

- Dans les menus, choisissez Fichier > Packages d'analyse de texte > Mise à jour d'un package. La boîte de dialogue Mettre à jour un package d'analyse de texte apparaît.
- Accédez au répertoire contenant le package d'analyse de texte à mettre à jour.
- Entrez un nom pour le TAP dans le champ Nom du fichier.
- Pour remplacer les ressources linguistiques à l'intérieur du TAP par celles de la session en cours, sélectionnez l'option Remplacer les ressources de ce package par celles de la session ouverte.

Généralement, il est utile de mettre à jour les ressources linguistiques car elles ont servi à extraire les concepts et patrons clés utilisés pour créer les définitions de catégories. Utiliser les ressources linguistiques les plus récentes permet d'obtenir de meilleurs résultats lors de la catégorisation de vos enregistrements. Si vous ne sélectionnez pas cette option, les ressources linguistiques déjà contenues dans le package sont conservées en l'état.

- Pour modifier uniquement les ressources linguistiques, vérifiez que l'option Remplacer les ressources de ce package par celles de la session ouverte est bien sélectionnée puis sélectionnez uniquement les ensembles de catégories actuels qui se trouvaient dans le TAP.

- Pour inclure le nouvel ensemble de catégories de la session ouverte dans le TAP, cochez la case de chaque ensemble de catégories à ajouter. Vous pouvez ajouter un, plusieurs ou aucun des ensembles de catégories.

- Pour supprimer des ensembles de catégorie du TAP, désélectionnez la case Inclure correspondante. Vous pouvez choisir de supprimer un ensemble de catégories qui se trouvait déjà dans le TAP car vous en ajoutez un ayant été amélioré. Pour ce faire, désélectionnez la case Inclure de l'ensemble de catégories correspondant dans la colonne Ensemble de catégories actuel. Le TAP doit contenir au moins un ensemble de catégories.

- Si nécessaire, modifiez les noms des ensembles de catégories. Un simple clic dans la cellule permet de modifier ce nom. Appuyer sur entrée ou cliquer à un autre droit permet d'appliquer la modification du nom. Si vous renommez un ensemble de catégories, le nom est uniquement modifié dans le TAP et ne modifie pas le nom de la variable dans la session ouverte. Si deux ensembles de catégories ont le même nom, ces noms apparaissent en rouge tant que le doublon n'a pas été corrigé.

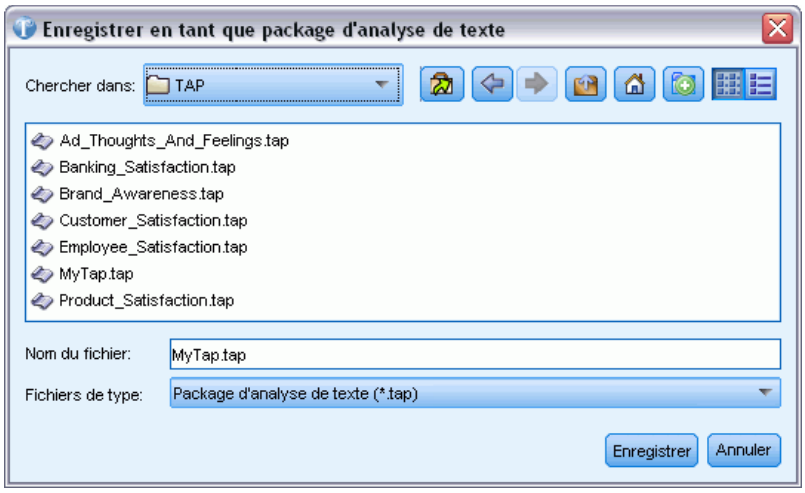
Figure 10-31
Noms en double

Inclure	Nouveaux ensembles de catégories	Ensemble(s) de catégories actuel(s)
☑	Cat_Points forts de votre dernier séjour à ...	
☑	Cat_Points faibles de votre dernier séjour ...	
☑		Cat_Points forts de votre dernier séjour à...
☑		Cat_Points faibles de votre dernier séjour...

- Pour créer un nouveau package en intégrant le contenu de la session au contenu du TAP sélectionné, cliquez sur Enregistrer en tant que Nouveau. La boîte de dialogue Enregistrer comme package d'analyse de texte apparaît. Suivez les instructions suivantes.

- Cliquez sur Mettre à jour pour enregistrer les modifications effectuées dans le TAP sélectionné.

Figure 10-32
Boîte de dialogue Enregistrer comme package d'analyse de texte



Pour enregistrer un package d'analyse de texte

- Accédez au répertoire dans lequel vous avez enregistré le fichier du TAP. Par défaut, les fichiers TAP sont enregistrés dans le sous-répertoire TAP du répertoire d'installation.
- Entrez un nom pour le fichier TAP dans le champ Nom du fichier.
- Saisissez une étiquette dans le champ Etiquette du package. Le nom de fichier choisi est automatiquement utilisé comme étiquette. Mais vous pouvez renommer cette étiquette. Une étiquette est nécessaire.
- Cliquez sur Enregistrer pour créer le nouveau package.

Edition et réglage des catégories

Une fois les catégories créées, vous souhaiterez certainement les analyser et procéder à des ajustements. Outre le réglage des ressources linguistiques, vous devez procéder à l'examen de vos catégories en recherchant des moyens de combiner ou de nettoyer leurs définitions. Vérifiez également certains documents ou enregistrements catégorisés. Vous pouvez également examiner les documents ou les enregistrements d'une catégorie et y opérer des ajustements, de sorte que les catégories puissent collecter les nuances et les distinctions.

Vous pouvez utiliser les techniques intégrées de création de catégories automatisées pour créer vos catégories. Cependant, vous voudrez probablement apporter quelques modifications à ces catégories. Lorsque vous utilisez une ou plusieurs techniques, un certain nombre de nouvelles catégories apparaissent dans la fenêtre. Vous pouvez alors examiner les données d'une catégorie et les ajuster jusqu'à ce que vos définitions de catégorie vous conviennent. Pour plus d'informations, reportez-vous à la section A propos des catégories sur p. 172.

Voici quelques options pour affiner vos catégories, la plupart étant décrites dans les pages suivantes :

- Ajout de descripteurs à vos catégories

- Modification de catégories
- Déplacement de catégories
- Aplatissement des catégories hiérarchiques
- Fusion de catégories
- Suppression de catégories
- Modification et réextraction des ressources linguistiques
- Visualisation de l'interaction de vos catégories et ajustements. Pour plus d'informations, reportez-vous à la section Graphiques et diagrammes de catégorie dans le chapitre 13 sur p. 255.

Ajout de descripteurs aux catégories

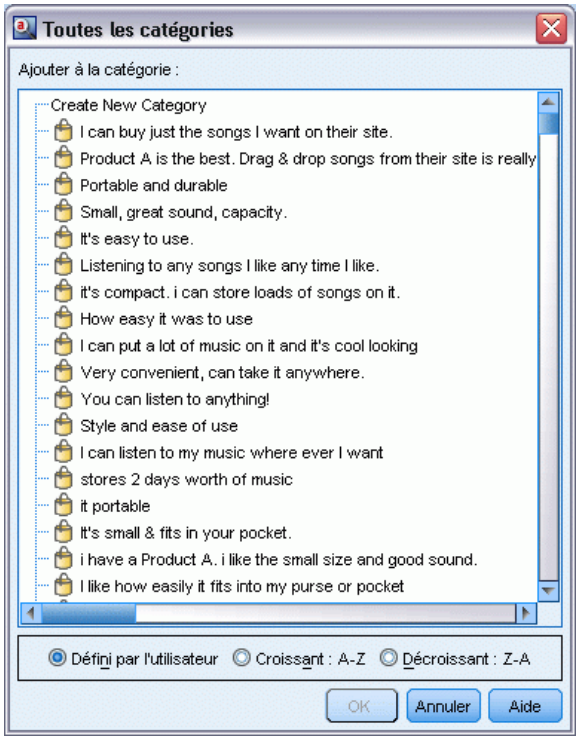
Après avoir utilisé des techniques automatisées, il est possible que vous disposiez encore de résultats d'extraction qui n'ont été utilisés dans aucune des définitions de catégorie. Consultez cette liste dans le panneau des Résultats de l'extraction. Si vous trouvez des éléments que vous souhaitez déplacer vers une catégorie, vous pouvez les ajouter dans une catégorie existante ou nouvelle.

Pour ajouter un concept ou un type à une catégorie

▶ Dans les panneaux Résultats d'extraction et Données, sélectionnez les éléments que vous souhaitez ajouter à une catégorie nouvelle ou existante.

▶ Dans les menus, sélectionnez Catégories > Ajouter à la catégorie. La boîte de dialogue Toutes les catégories apparaît pour présenter l'ensemble des catégories. Sélectionnez la catégorie à laquelle vous voulez ajouter les éléments sélectionnés. Si vous voulez ajouter les éléments à une nouvelle catégorie, sélectionnez Nouvelle catégorie. Une nouvelle catégorie apparaît dans le panneau Catégories, sous le nom du premier élément sélectionné.

Figure 10-33
Boîte de dialogue Toutes les catégories



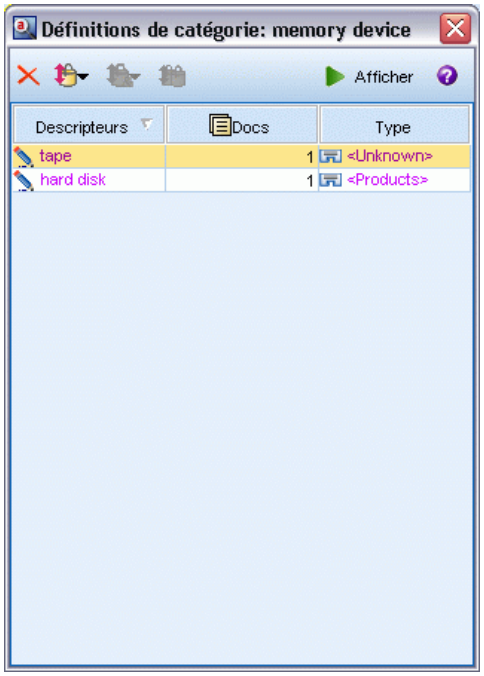
Modification des descripteurs de catégorie

Dès que vous avez créé des catégories, vous pouvez ouvrir chacune d'entre elles pour visualiser l'ensemble des descripteurs qui constituent sa définition. Dans la boîte de dialogue Définitions de catégorie, vous pouvez apporter un certain nombre de modifications à vos descripteurs de catégorie. De plus, si des catégories apparaissent dans l'arborescence des catégories, vous pouvez également les utiliser.

Pour éditer une catégorie

- Sélectionnez la catégorie à éditer dans le panneau Catégories.
- Dans les menus, sélectionnez Affichage > Définitions de catégorie. La boîte de dialogue Définitions de catégorie apparaît.

Figure 10-34
Boîte de dialogue Définitions de catégorie



- Sélectionnez le descripteur que vous souhaitez éditer, puis cliquez sur le bouton correspondant dans la barre d'outils.

Le tableau ci-dessous décrit tous les boutons de la barre d'outils qui permettent de modifier vos définitions de catégorie.

Table 10-13
Boutons de la barre d'outils et descriptions

Icônes	Description
	Supprime les descripteurs sélectionnés de la catégorie.
	Déplace les descripteurs sélectionnés vers une catégorie existante ou nouvelle.
	Déplace les descripteurs sélectionnés vers une catégorie sous la forme d'une règle de catégorie &. Pour plus d'informations, reportez-vous à la section Utilisation des règles de catégorie sur p. 201.
	Déplace chacun des descripteurs sélectionnés dans une nouvelle catégorie qui lui est propre.
 Afficher	Met à jour l'affichage des panneaux Données et Visualisation en fonction des descripteurs sélectionnés.

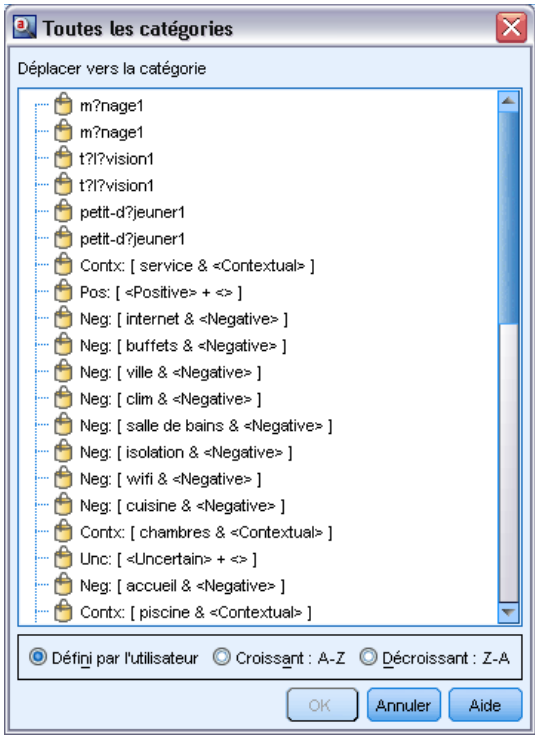
Déplacement de catégories

Si vous voulez placer une catégorie dans une autre catégorie existante ou déplacer des descripteurs dans une autre catégorie, vous pouvez les déplacer.

Pour déplacer une catégorie

- Dans le panneau Catégories, sélectionnez les catégories à déplacer dans une autre catégorie.
- Dans les menus, choisissez Catégories > Déplacer vers la catégorie. Le menu présente un ensemble de catégories, la plus récente figurant en haut de la liste. Sélectionnez le nom de la catégorie vers laquelle vous voulez déplacer les concepts sélectionnés.
  - Si vous voyez le nom que vous recherchez, sélectionnez-le pour l'ajouter à cette catégorie.
  - S'il n'apparaît pas, sélectionnez Plus pour afficher la boîte de dialogue Toutes les catégories et sélectionnez la catégorie dans la liste.

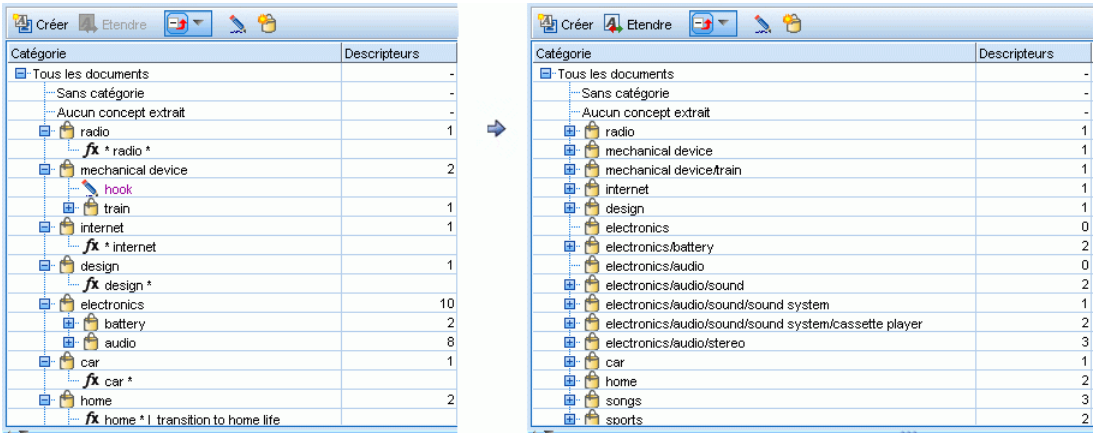
Figure 10-35
Boîte de dialogue Toutes les catégories



Aplatissement des catégories

Lorsque vous avez une structure de catégories hiérarchique avec des catégories et des sous-catégories, vous pouvez aplatir votre structure. Lorsque vous aplatissez une catégorie, tous les descripteurs des sous-catégories de cette catégorie sont déplacés vers la catégorie sélectionnée et les sous-catégories désormais vides sont supprimées. Ainsi, tous les documents qui étaient utilisés pour la mise en correspondance des sous-catégories sont désormais catégorisés dans la catégorie sélectionnée.

Figure 10-36
Catégories aplaties



Pour aplatir une catégorie

- Dans le panneau Catégories, sélectionnez une catégorie (de niveau supérieur ou une sous-catégorie) que vous souhaitez aplatir.
- Dans les menus, sélectionnez Catégories > Aplatir des catégories. Les sous-catégories sont supprimées et les descripteurs sont fusionnés dans la catégorie sélectionnée.

Fusion ou combinaison de catégories

Si vous souhaitez combiner deux catégories ou plus dans une nouvelle catégorie, vous pouvez les fusionner. Quand vous fusionnez des catégories, une nouvelle catégorie est créée avec un nom générique. Tous les concepts, types et patrons utilisés dans les descripteurs de catégorie sont déplacés dans cette nouvelle catégorie. Par la suite, vous pouvez renommer cette catégorie en éditant ses propriétés.

Pour fusionner tout ou partie d'une catégorie

- Dans le panneau Catégories, sélectionnez les éléments que vous souhaitez fusionner.
- Dans les menus, sélectionnez Catégories > Fusionner les catégories. La boîte de dialogue Propriétés des catégories s'affiche et vous permet d'entrer un nom pour la catégorie nouvellement créée. Les catégories sélectionnées sont fusionnées dans la nouvelle catégorie en tant que sous-catégories.

Suppression de catégories

Si vous ne souhaitez pas conserver une catégorie, vous pouvez la supprimer.

Pour supprimer une catégorie

- Dans le panneau Catégories, sélectionnez la ou les catégories à supprimer.
- Dans les menus, sélectionnez Edition > Supprimer.

Chapitre
11

Analyse des clusters

Vous pouvez créer et explorer des clusters de concept dans la vue Clusters (Affichage> Clusters). Un **cluster** est un regroupement de concepts liés, généré par des algorithmes de classification non supervisée qui se fondent sur la fréquence d'apparition de ces concepts dans l'ensemble de documents/enregistrements et la fréquence d'apparition conjointe des concepts dans le même document (ou **cooccurrence**). Chaque concept d'un cluster est cooccurrent avec au moins un autre concept du cluster. Les clusters ont pour objectif de regrouper les concepts apparaissant ensemble, alors que les catégories ont pour objectif de regrouper les documents ou les enregistrements en fonction des correspondances existant entre le texte et les descripteurs (concept, règles, patrons) pour chaque catégorie.

Un cluster adéquat est un cluster présentant des concepts fortement liés et fréquemment cooccurrents, ainsi que dotés de peu de liens vers des concepts d'autres clusters. Lorsque vous travaillez avec des ensembles de données volumineux, cette technique peut aboutir à des temps de traitement beaucoup plus longs.

Remarque : utilisez l'option Nombre maximal de documents à utiliser pour calculer les clusters de la boîte de dialogue Créer des clusters de façon à créer des clusters avec uniquement un sous-ensemble de tous les documents ou enregistrements.

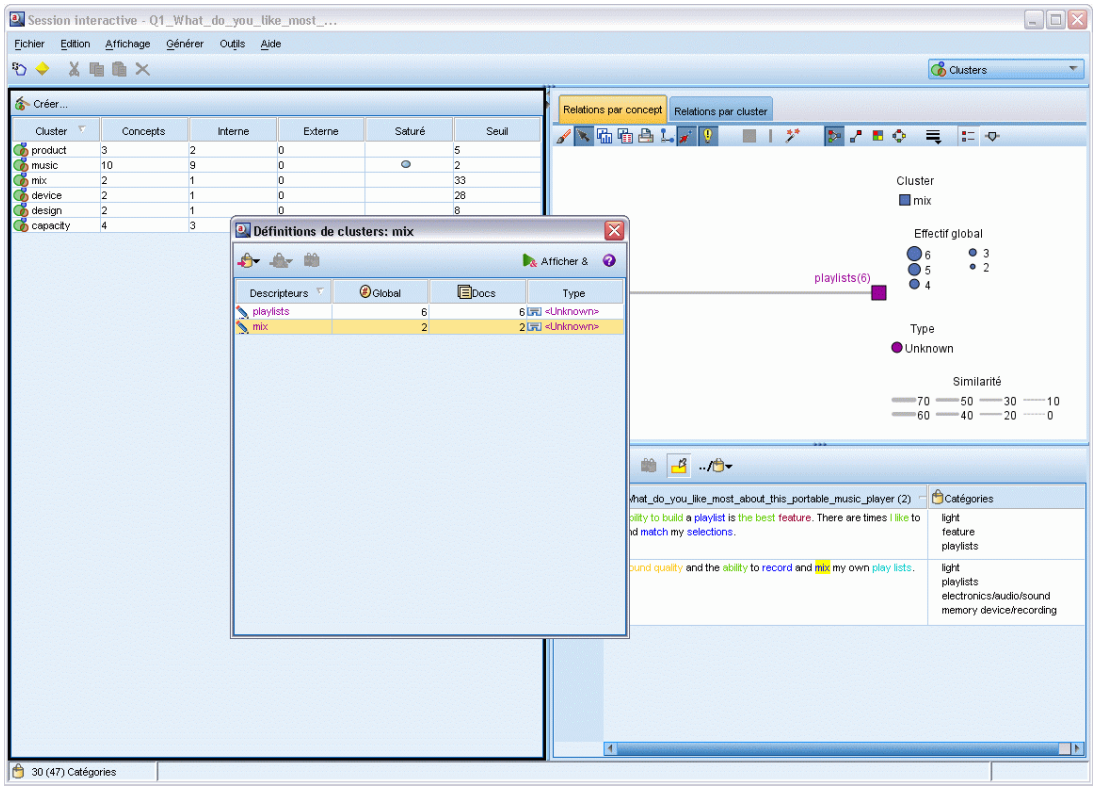
La classification non supervisée est un processus qui commence par l'analyse d'un ensemble de concepts et la recherche des concepts qui sont souvent cooccurrents dans les documents. Deux concepts cooccurrents dans un document sont considérés comme formant une paire de concepts. Ensuite, le processus de classification non supervisée évalue la **valeur de similarité** de chaque paire de concepts en comparant le nombre de documents dans lequel la paire apparaît au nombre de documents dans lequel chaque concept apparaît. Pour plus d'informations, reportez-vous à la section Calcul des valeurs du lien de similarité sur p. 241.

En dernier lieu, le processus de classification non supervisée regroupe les concepts similaires en clusters, par agrégation, et prend en compte la valeur de leurs liens et les paramètres définis dans la boîte de dialogue Créer des clusters. Nous entendons par « agrégation » l'ajout de concepts ou la fusion des clusters plus petits en un cluster plus grand jusqu'à ce que le cluster soit saturé. Un cluster est **saturé** lorsqu'une fusion supplémentaire de concepts ou de petits clusters entraînerait le dépassement des paramètres de la boîte de dialogue Créer des clusters (nombre de concepts, de liens internes ou de liens externes). Un cluster prend le nom du concept qui, en son sein, présente le nombre global le plus élevé de liens vers d'autres concepts du même cluster.

Finalement, toutes les paires de concepts ne se retrouvent pas dans le même cluster puisqu'il peut exister un lien plus fort dans un autre cluster ou la saturation peut empêcher la fusion des clusters dans lesquels elles apparaissent. C'est la raison pour laquelle il existe à la fois des liens internes et des liens externes.

- Les **liens internes** sont des liens entre des paires de concepts au sein d'un cluster. Dans un cluster, tous les concepts ne sont pas liés les uns aux autres. Toutefois, chaque concept est au moins lié à un autre concept de son cluster.
- Les **liens externes** sont des liens entre des paires de concepts dans des clusters distincts (un concept au sein d'un cluster et un concept en dehors, dans un autre cluster).

Figure 11-1
vue Clusters



La vue Clusters est organisée en trois panneaux (vous pouvez masquer ou afficher chaque panneau en sélectionnant son nom dans le menu Affichage) :

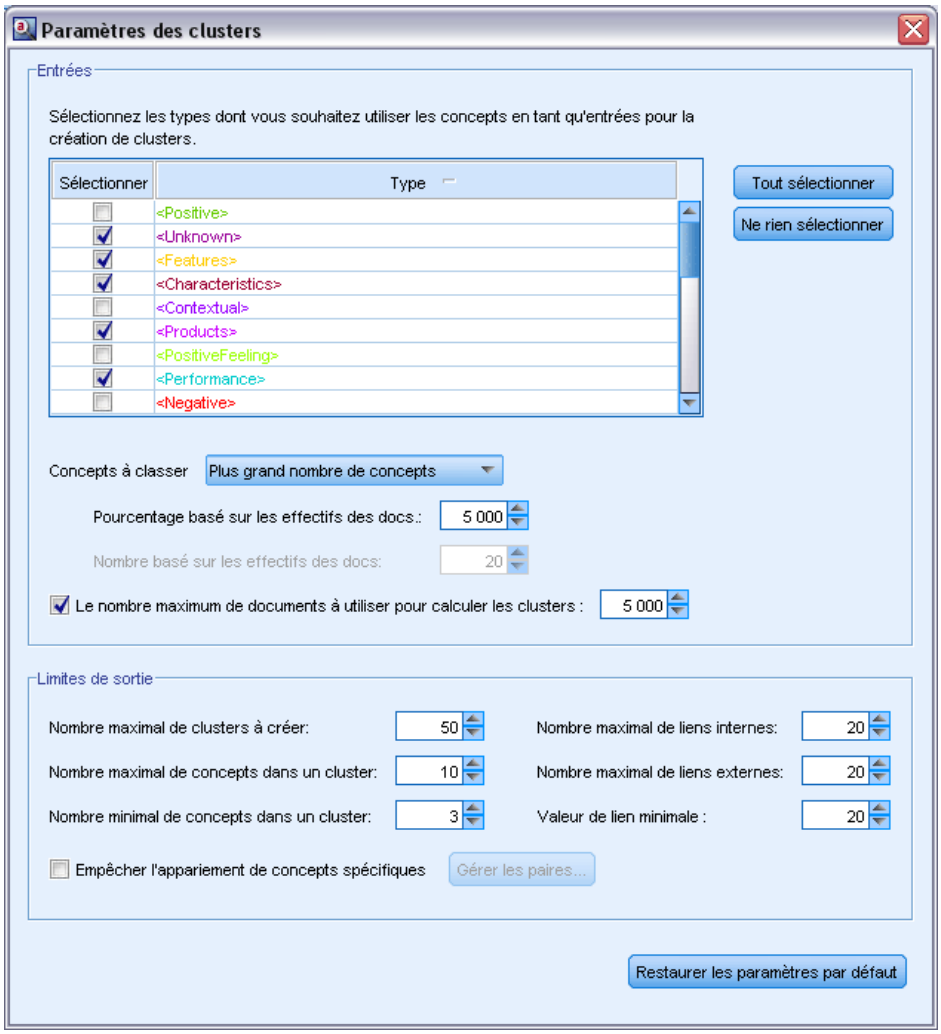
- **Panneau Cluster.** Vous pouvez créer et gérer vos clusters dans ce panneau. Pour plus d'informations, reportez-vous à la section Exploration des Clusters sur p. 242.
- **Panneau Visualisation.** Vous pouvez explorer visuellement vos clusters et analyser la manière dont ils interagissent dans ce panneau. Pour plus d'informations, reportez-vous à la section Graphiques Cluster dans le chapitre 13 sur p. 259.
- **Panneau Données.** Vous pouvez explorer et passer en revue le texte contenu dans les documents et enregistrements qui correspondent aux sélections effectuées dans la boîte de dialogue Définitions du cluster. Pour plus d'informations, reportez-vous à la section Définitions du cluster sur p. 243.

Création de clusters

Lorsque vous accédez pour la première fois à la vue Clusters, aucun cluster n'est visible. Vous pouvez créer des clusters via les menus (Outils > Créer des clusters) ou en cliquant sur le bouton Créer de la barre d'outils. Cette action ouvre la boîte de dialogue Créer des clusters, dans laquelle vous pouvez définir les paramètres et les limites pour la création des clusters.

Remarque : Lorsque les résultats de l'extraction ne correspondent plus aux ressources, ce panneau devient jaune, tout comme le panneau Résultats d'extraction. Vous pouvez procéder à une nouvelle extraction pour obtenir les derniers résultats d'extraction et la couleur jaune disparaîtra. Toutefois, à chaque nouvelle extraction, le panneau Clusters est effacé et vous devez recréer vos clusters. De même, les clusters ne sont pas enregistrés d'une session à l'autre.

Figure 11-2
Boîte de dialogue Créer des clusters



Entrées

Tableau d'**entrées**. Les clusters sont créés à partir de descripteurs dérivés de certains types. Dans le tableau, vous pouvez sélectionner les types à inclure dans le processus de création. Les types qui capturent le plus d'enregistrements ou de documents sont présélectionnés par défaut.

Concepts à classifier : Sélectionnez la méthode de sélection des concepts à utiliser pour la classification. En réduisant le nombre de concepts, vous pouvez accélérer le processus de classification non supervisée. Vous pouvez effectuer une classification à l'aide d'un certain nombre de meilleurs concepts, d'un pourcentage de meilleurs concepts ou en utilisant tous les concepts :

- **Nombre en fonction des effectifs des doc.** Lorsque vous sélectionnez Plus grand nombre de concepts, entrez le nombre de concepts à prendre en compte pour la classification. Les concepts choisis le sont en fonction des plus grands effectifs de documents. Il s'agit du nombre de documents ou d'enregistrements dans lesquels le concept apparaît.
- **Pourcentage en fonction des effectifs des doc.** Lorsque vous sélectionnez Plus grand pourcentage de concepts, entrez le pourcentage de concepts à prendre en compte pour la classification. Les concepts choisis le sont en fonction du pourcentage de concepts qui présentent les plus grands effectifs de documents.

Nombre maximal de documents à utiliser pour calculer les clusters. Par défaut, les valeurs de lien sont calculées à l'aide de l'ensemble complet de documents ou d'enregistrements. Toutefois, dans certains cas, vous pouvez accélérer le processus de classification non supervisée en limitant le nombre de documents ou d'enregistrements utilisés pour calculer les liens. La limitation des documents peut diminuer la qualité des clusters. Pour ce faire, cochez la case à gauche de l'option et entrez le nombre maximal de documents ou d'enregistrements à utiliser.

Limites de sortie

Nombre maximal de clusters à créer. Cette valeur correspond au nombre maximal de clusters à générer et à afficher dans le panneau Clusters. Durant le processus de classification non supervisée, les clusters saturés sont présentés avant les clusters non saturés et, par conséquent, de nombreux clusters obtenus sont saturés. De façon à visualiser davantage de clusters non saturés, vous pouvez régler ce paramètre sur une valeur supérieure au nombre de clusters saturés.

Nombre maximal de concepts dans un cluster. Cette valeur correspond au nombre maximal de concepts que peut contenir un cluster.

Nombre minimal de concepts dans un cluster. Cette valeur correspond au nombre minimal de concepts qui doivent être liés de façon à créer un cluster.

Nombre maximal de liens internes. Cette valeur correspond au nombre maximal de liens internes que peut contenir un cluster. Les liens internes sont des liens entre des paires de concepts au sein d'un cluster.

Nombre maximal de liens externes. Cette valeur correspond au nombre maximal de liens vers des concepts situés en dehors du cluster. Les liens externes sont des liens entre des paires de concepts dans des clusters distinctes.

Valeur de lien minimale. Cette valeur correspond à la plus petite valeur de lien acceptée pour prendre en considération une paire de concepts dans le cadre de la classification non supervisée. La valeur du lien est calculée à l'aide d'une formule de similarité. Pour plus d'informations, reportez-vous à la section Calcul des valeurs du lien de similarité sur p. 241.

Éviter l'appariement de concepts spécifiques. Cochez cette case pour arrêter le processus de regroupement ou d'appariement de deux concepts dans les résultats. Pour créer ou gérer des paires de concepts, cliquez sur Gérer les paires. Pour plus d'informations, reportez-vous à la section Gestion des paires d'exceptions de liens dans le chapitre 10 sur p. 185.

Calcul des valeurs du lien de similarité

La seule connaissance du nombre de documents dans lequel deux concepts sont cooccurrents n'indique pas à elle seule leur degré de similarité. Dans ce cas, il peut être utile de connaître la valeur de similarité. La valeur du lien de similarité est mesurée par le biais de la comparaison entre les effectifs des documents de cooccurrence et les effectifs des documents pour chaque concept de la relation. Lors du calcul de la similarité, l'unité de mesure est le nombre de documents (effectifs des documents) dans lesquels apparaît un concept ou une paire de concepts. Un concept ou une paire de concepts sont « détectés » dans un document s'ils apparaissent *au moins* une fois dans ce dernier. Vous pouvez choisir de représenter graphiquement la valeur du lien de similarité par l'épaisseur de ligne du graphique de concept.

L'algorithme révèle les relations les plus fortes, ce qui signifie que la tendance des concepts à apparaître ensemble dans les données textuelles est largement supérieure à leur tendance à apparaître de façon indépendante. En interne, l'algorithme produit un coefficient de similarité qui s'étend de 0 à 1, où la valeur 1 signifie que les deux concepts apparaissent toujours ensemble et jamais séparément. Le résultat du coefficient de similarité est ensuite multiplié par 100 et arrondi au nombre entier le plus proche. Le coefficient de similarité est calculé à l'aide de la formule présentée dans la figure suivante.

Figure 11-3
Formule du coefficient de similarité

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Où :

- C_I est le nombre de documents ou d'enregistrements dans lequel apparaît le concept `I`.
- C_J est le nombre de documents ou d'enregistrements dans lequel apparaît le concept `J`.
- C_{IJ} est le nombre de documents ou d'enregistrements dans lequel la paire de concepts `I` et `J` est cooccurrente dans l'ensemble de documents.

Par exemple, supposons que vous disposiez de 5 000 documents. Soient `I` et `J` les concepts extraits et `IJ` la cooccurrence de la paire de concepts `I` et `J`. Le tableau suivant propose deux scénarios pour démontrer le calcul du coefficient et de la valeur du lien.

Table 11-1
Exemple de fréquence des concepts

Concept/paire	Scénario A	Scénario B
Concept : `I`	Présence dans 20 documents	Présence dans 30 documents
Concept : `J`	Présence dans 20 documents	Présence dans 60 documents
Paire de concepts : `IJ`	Présence dans 20 documents	Présence dans 20 documents
Coefficient de similarité	1	0.22222
Valeur du lien de similarité	100	22

Dans le scénario A, les concepts I et J ainsi que la paire IJ apparaissent dans 20 documents, ce qui produit un coefficient de similarité de 1. Cela qui signifie que les concepts apparaissent toujours ensemble. La valeur du lien de similarité pour cette paire est de 100.

Dans le scénario B, le concept I apparaît dans 30 documents et le concept J dans 60 documents. En revanche, la paire IJ est seulement présente dans 20 documents. Par conséquent, le coefficient de similarité est de 0,22222. La valeur du lien de similarité pour cette paire est arrondie à 22.

Exploration des Clusters

Après avoir créé des clusters, vous pouvez visualiser un ensemble de résultats dans le panneau Clusters. Pour chaque cluster, le tableau présente les informations suivantes :

- **Cluster.** Il s'agit du nom du cluster. Les clusters sont nommés d'après le concept présentant le plus grand nombre de liens internes.
- **Concepts.** Il s'agit du nombre de concepts dans le cluster. Pour plus d'informations, reportez-vous à la section Définitions du cluster sur p. 243.
- **Interne.** Il s'agit du nombre de liens internes dans le cluster. Les liens internes sont des liens entre des paires de concepts au sein d'un cluster.
- **Externe.** Il s'agit du nombre de liens externes dans le cluster. Les liens externes représentent des liens entre des paires de concepts lorsqu'un concept se situe dans un cluster et l'autre dans un autre cluster.
- **Sat.** La présence de ce symbole indique que ce cluster aurait pu être plus grand, mais que dans ce cas, une ou plusieurs limites auraient été dépassées ; par conséquent, le processus de classification non supervisée a pris fin pour ce cluster, lequel est considéré alors comme étant *saturé*. A la fin du processus de classification non supervisée, les clusters saturées sont présentées avant les clusters non saturés et, par conséquent, de nombreux clusters obtenues sont saturées. De façon à visualiser davantage de clusters non saturés, vous pouvez modifier le paramètre Nombre maximal de clusters à créer sur une valeur supérieure au nombre de clusters saturés ou diminuer le paramètre Valeur de lien minimale. Pour plus d'informations, reportez-vous à la section Création de clusters sur p. 238.
- **Seuil.** Pour toutes les paires de concepts cooccurrents dans le cluster, il s'agit de la valeur la plus faible du lien de similarité de tout le cluster. Pour plus d'informations, reportez-vous à la section Calcul des valeurs du lien de similarité sur p. 241. Un cluster avec une valeur de seuil élevée signifie que les concepts qui le composent présentent une similarité globale plus élevée et sont plus étroitement liés que ceux d'un cluster dont la valeur de seuil est inférieure.

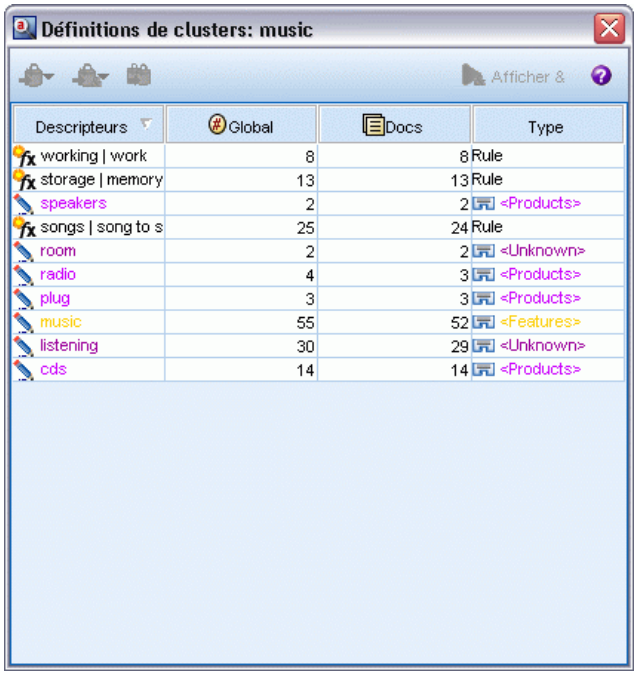
Pour en savoir plus sur un cluster donné, sélectionnez-le et le panneau de visualisation à droite présente alors deux graphiques permettant d'explorer les clusters. Pour plus d'informations, reportez-vous à la section Graphiques Cluster dans le chapitre 13 sur p. 259. Vous pouvez également couper et coller le contenu du tableau dans une autre application.

Lorsque les résultats de l'extraction ne correspondent plus aux ressources, ce panneau devient jaune, tout comme le panneau Résultats d'extraction. Vous pouvez procéder à une nouvelle extraction pour obtenir les derniers résultats d'extraction et la couleur jaune disparaîtra. Toutefois, à chaque nouvelle extraction, le panneau Clusters est effacé et vous devez recréer vos clusters. De même, les clusters ne sont pas enregistrés d'une session à l'autre.

Définitions du cluster

Vous pouvez visualiser tous les concepts d'un cluster en le sélectionnant dans le panneau Clusters et en ouvrant la boîte de dialogue Définitions du cluster (Affichage > Définition du cluster).

Figure 11-4
Boîte de dialogue Définitions du cluster



Tous les concepts du cluster sélectionné apparaissent dans la boîte de dialogue Définitions du cluster. Si vous sélectionnez un ou plusieurs concepts dans la boîte de dialogue Définitions du cluster et que vous cliquez sur Afficher, le panneau Données affiche tous les enregistrements ou documents dans lesquels *tous les concepts sélectionnés apparaissent ensemble*. Toutefois, le panneau Données n'affiche pas d'enregistrements ou de documents texte lorsque vous sélectionnez un cluster dans le panneau Cluster. Pour obtenir des informations générales sur le panneau Données, reportez-vous à Le panneau Données dans le chapitre 10.

La sélection de concepts dans cette boîte de dialogue modifie également le graphique Relations par concept. Pour plus d'informations, reportez-vous à la section Graphiques Cluster dans le chapitre 13 sur p. 259. De la même manière, lorsque vous sélectionnez des concepts dans la boîte de dialogue Définitions du cluster, le panneau Visualisation affiche tous les liens externes et internes issus de ces concepts.

Descriptions de colonne

Des icônes s'affichent pour vous permettre d'identifier facilement chaque descripteur.

Table 11-2
Colonnes et icônes de descripteur

Colonnes	Description
Descripteurs	Nom du concept.
Global	Affiche le nombre de fois que ce descripteur apparaît dans l'ensemble de données, également connu sous le nom de fréquence globale.
Docs	Affiche le nombre de documents ou d'enregistrements dans lesquels ce descripteur apparaît, également connu sous le nom de fréquence du document.
Type	Affiche le ou les types auxquels le descripteur appartient. Si le descripteur est une règle de catégorie, aucun nom de type n'est indiqué dans cette colonne.

Actions de la barre d'outils

Dans cette boîte de dialogue, vous pouvez également sélectionner un ou plusieurs concepts à utiliser dans une catégorie. Il existe plusieurs façons de procéder mais il est plus intéressant de sélectionner des concepts cooccurrents dans un cluster et de les ajouter en tant que règle de catégorie. Pour plus d'informations, reportez-vous à la section Règles de cooccurrence dans le chapitre 10 sur p. 191. Vous pouvez utiliser les boutons de la barre d'outils pour ajouter les concepts aux catégories.

Table 11-3
Boutons de la barre d'outils pour l'ajout de concepts aux catégories

Icônes	Description
	Ajoute les concepts sélectionnés à une nouvelle catégorie ou à une catégorie existante.
	Ajoute les concepts sélectionnés sous la forme d'une règle de catégorie & à une nouvelle catégorie ou à une catégorie existante. Pour plus d'informations, reportez-vous à la section Utilisation des règles de catégorie dans le chapitre 10 sur p. 201.
	Ajoute chacun des concepts sélectionnés sous la forme d'une nouvelle catégorie qui lui est propre.
Afficher &	Met à jour l'affichage des panneaux Données et Visualisation en fonction des descripteurs sélectionnés.

Remarque : Il est également possible d'ajouter des concepts à un type, sous forme de synonymes, ou sous forme d'éléments d'exclusion, à l'aide des menus contextuels.

Chapitre
12

Exploration de l'analyse des liens du texte

Dans la vue Analyse des liens du texte (TLA), vous pouvez explorer des résultats de patrons d'analyse des liens du texte. L'analyse des liens du texte est une technologie de mise en correspondance de patrons qui vous permet de définir des règles de patrons et de les comparer aux concepts extraits et aux relations trouvées dans le texte.

Par exemple, l'extraction d'idées concernant une organisation peut ne présenter qu'un intérêt relatif pour vous. Grâce à l'analyse des liens du texte, vous pouvez en savoir plus sur les liens entre cette organisation et d'autres organisations ou sur les personnes au sein d'une organisation. Vous pouvez également utiliser l'analyse TLA pour extraire des opinions sur des produits ou, pour certaines langues, les relations entre des gènes.

Lorsque vous avez extrait des résultats de patrons TLA, vous pouvez les consulter dans les panneaux Patrons de type et Patrons de concept de la vue Analyse des liens du texte. Pour plus d'informations, reportez-vous à la section Patrons de type et Patrons de concept sur p. 247. Vous pouvez continuer à les explorer dans les panneaux Données ou Visualisations de cette vue. Mais surtout, vous pouvez les ajouter aux catégories.

Si vous n'avez pas choisi cette action, vous pouvez cliquer sur Extraire et choisir Extraction avec analyse des liens du texte dans la boîte de dialogue Paramètres d'extraction. Pour plus d'informations, reportez-vous à la section Extraction des résultats de patrons TLA sur p. 246.

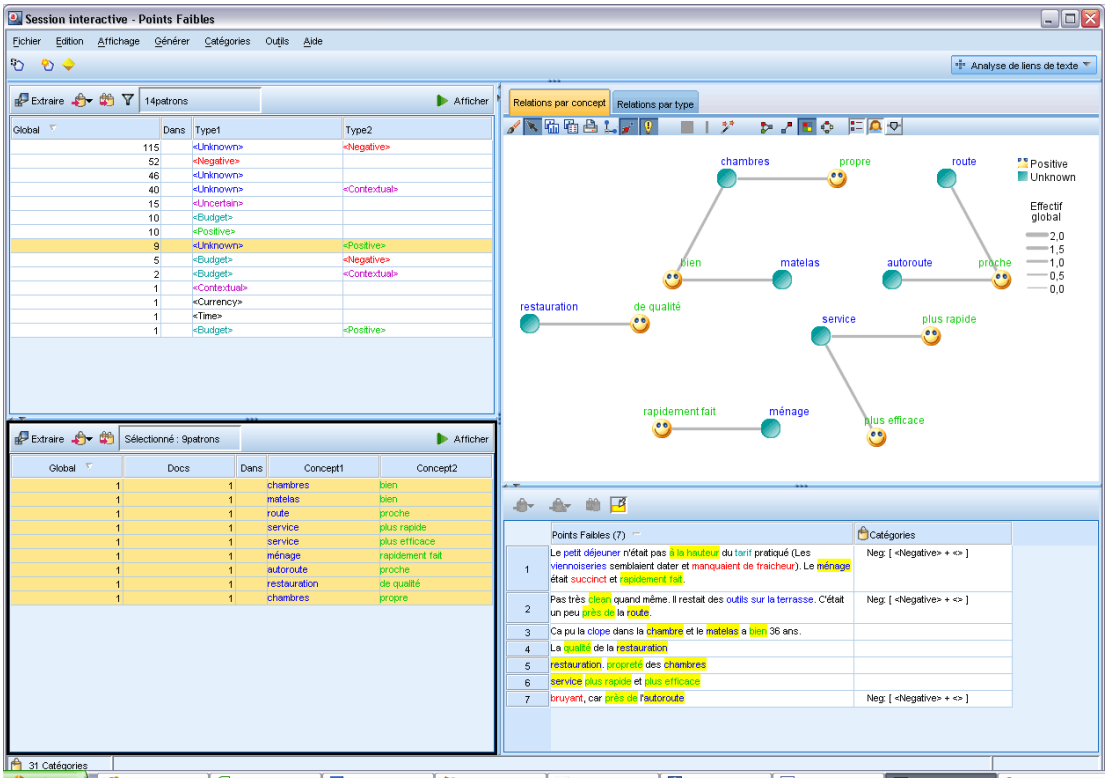
Pour pouvoir extraire les résultats de patrons de l'analyse des liens du texte, des règles de patrons d'analyse des liens du texte doivent être définies dans le modèle de ressources ou dans les bibliothèques que vous utilisez. Vous pouvez utiliser les patrons TLA dans certains modèles de ressources fournis avec IBM® SPSS® Modeler Text Analytics . Le genre de relations et de patrons que vous pouvez extraire dépend entièrement des règles TLA définies dans vos ressources. Vous pouvez définir vos propres règles TLA pour toutes les langues de texte *sauf* le japonais. Les patrons sont constitués de macros, listes de mots et intervalles de mots pour former une requête booléenne, ou règle, qui est comparée à votre texte d'entrée. Pour plus d'informations, reportez-vous à la section A propos des règles des liens du texte dans le chapitre 19 sur p. 338.

Lorsqu'une règle de patron TLA correspond au texte, il est possible d'extraire ce texte sous la forme d'un patron et de le restructurer sous la forme de données de sortie. Les résultats sont alors visibles dans les panneaux de la vue Analyse des liens du texte. Chaque panneau peut être masqué ou affiché en sélectionnant son nom dans le menu Affichage :

- **Panneaux Patrons de type et Patrons de concept.** Ces deux panneaux permettent de créer et d'explorer des patrons. Pour plus d'informations, reportez-vous à la section Patrons de type et Patrons de concept sur p. 247.

- **Panneau Visualisation.** Ce panneau permet d'explorer visuellement la façon dont les concepts et les types des patrons interagissent. Pour plus d'informations, reportez-vous à la section Graphiques Analyse des liens du texte dans le chapitre 13 sur p. 261.
- **Panneau Données.** Vous pouvez explorer et passer en revue le texte contenu dans les documents et enregistrements qui correspondent aux sélections effectuées dans un autre panneau. Pour plus d'informations, reportez-vous à la section Panneau Données sur p. 251.

Figure 12-1
Vue Analyse des liens du texte



Extraction des résultats de patrons TLA

Le résultat du processus d'extraction correspond à un ensemble de concepts, de types et de patrons d'analyse des liens du texte (TLA) si la fonction a été activée. Si vous avez extrait des patrons TLA, vous pouvez les visualiser dans la vue Analyse des liens du texte. Lorsque les résultats de l'extraction ne sont pas synchronisés avec les ressources, les panneaux de patrons deviennent jaunes et indiquent ainsi qu'une nouvelle extraction produirait des résultats différents.

Vous devez choisir d'extraire ces patrons dans le paramètre de noeud ou dans la boîte de dialogue Extraire à l'aide de l'option Extraction avec analyse des liens du texte. Pour plus d'informations, reportez-vous à la section Extraction de données dans le chapitre 9 sur p. 143.

Remarque : Il existe un rapport entre la taille de votre ensemble de données et la durée nécessaire à l'exécution du processus d'extraction. Reportez-vous aux instructions d'installation pour obtenir des statistiques et des recommandations sur les performances. Vous pouvez toujours envisager d'insérer un noeud Echantillonner en amont ou d'optimiser la configuration de votre machine.

Pour extraire des données

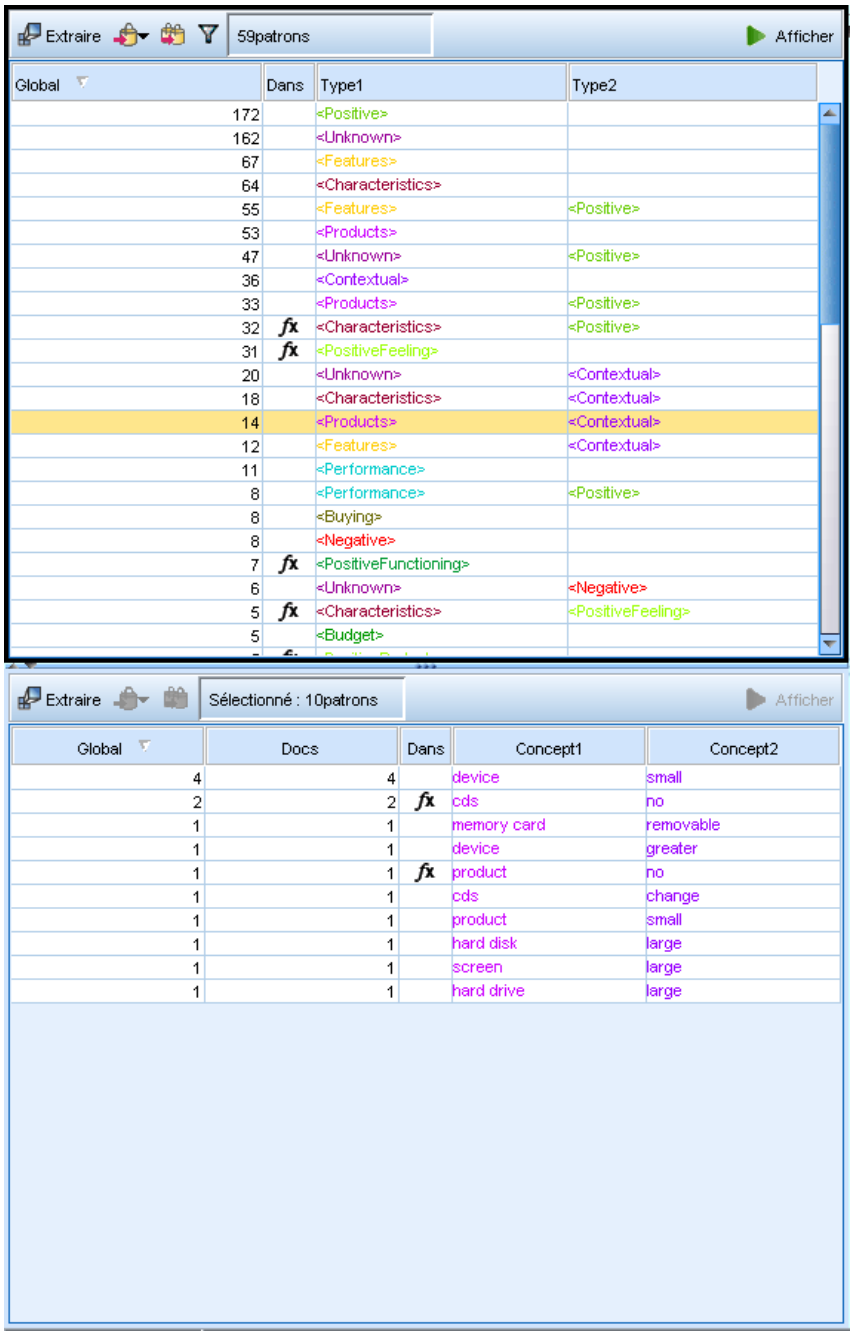
- A partir des menus, sélectionnez Outils> Extraire. Vous pouvez également cliquer sur le bouton Extraire de la barre d'outils.
- Modifiez les options à utiliser. Gardez à l'esprit que l'option Extraction avec analyse des liens du texte doit être sélectionnée dans cet onglet et que votre modèle doit comporter des règles TLA pour que des résultats de patrons TLA puissent être extraits. Pour plus d'informations, reportez-vous à la section Extraction de données dans le chapitre 9 sur p. 143.
- Cliquez sur Extraire pour lancer le processus d'extraction.

  Dès le début de l'extraction, une boîte de dialogue indique la progression du processus. Si vous voulez annuler l'extraction, cliquez sur Annuler. Lorsque l'extraction est terminée, la boîte de dialogue se ferme et les résultats apparaissent dans le panneau. Pour plus d'informations, reportez-vous à la section Patrons de type et Patrons de concept sur p. 247.

Patrons de type et Patrons de concept

Les patrons sont constitués de deux éléments : une combinaison de concepts et de types. Les patrons s'avèrent particulièrement utiles lorsque vous tentez de découvrir des opinions sur un sujet donné ou des relations entre des concepts. Pour vous, l'intérêt de l'extraction du nom du produit de votre concurrent peut être limité. Dans ce cas, vous pouvez examiner les patrons extraits pour éventuellement découvrir des exemples de document ou d'enregistrement contenant du texte indiquant que le produit est bon, mauvais ou cher.

Figure 12-2
Vue Analyse des liens du texte : panneaux Patrons de type et Patrons de concept



Un patron peut inclure jusqu'à six types ou six concepts. C'est la raison pour laquelle les lignes des deux panneaux de patrons contiennent jusqu'à six propriétés, ou positions. Chaque propriété correspond à la position spécifique d'un élément dans la règle de patron TLA telle qu'elle est définie dans les ressources linguistiques. Dans la session interactive, si une propriété ne contient pas de valeur, elle n'apparaît pas dans le tableau. Par exemple, si les résultats de patrons les

plus longs ne contiennent pas plus de quatre propriétés, les deux dernières n'apparaissent pas. Pour plus d'informations, reportez-vous à la section A propos des règles des liens du texte dans le chapitre 19 sur p. 338.

Lorsque vous extrayez des résultats de patrons, ils sont d'abord regroupés au niveau du type, puis divisés en patrons de concept. C'est la raison pour laquelle il existe deux panneaux de résultats différents : Patrons de type (en haut à gauche) et Patrons de concept (en bas à gauche). Pour afficher tous les patrons de concepts retournés, sélectionnez tous les patrons de types. Le panneau inférieur des concepts affichera alors tous les patrons de concepts jusqu'à la valeur de rang maximal (tel que défini dans la boîte de dialogue Filtrer).

Patrons de type. Ce panneau présente les résultats de patrons comportant au moins deux types associés correspondant à une règle de patron TLA. Les patrons de type se présentent sous la forme `<Organization>` + `<Location>` + `<Positive>`, ce qui permet d'obtenir un commentaire positif concernant une organisation située dans une location particulière. La syntaxe est la suivante :

```
<Type1> + <Type2> + <Type3> + <Type4> + <Type5> + <Type6>
```

Patrons de concept. Ce panneau présente les résultats de patrons au niveau du concept pour tous les patrons de type actuellement sélectionnés dans le panneau Patrons de type situé au-dessus. Les patrons de concept suivent une structure de type `hôtel` + `paris` + `merveilleux`. La syntaxe est la suivante :

```
concept1 + concept2 + concept3 + concept4 + concept5 + concept6
```

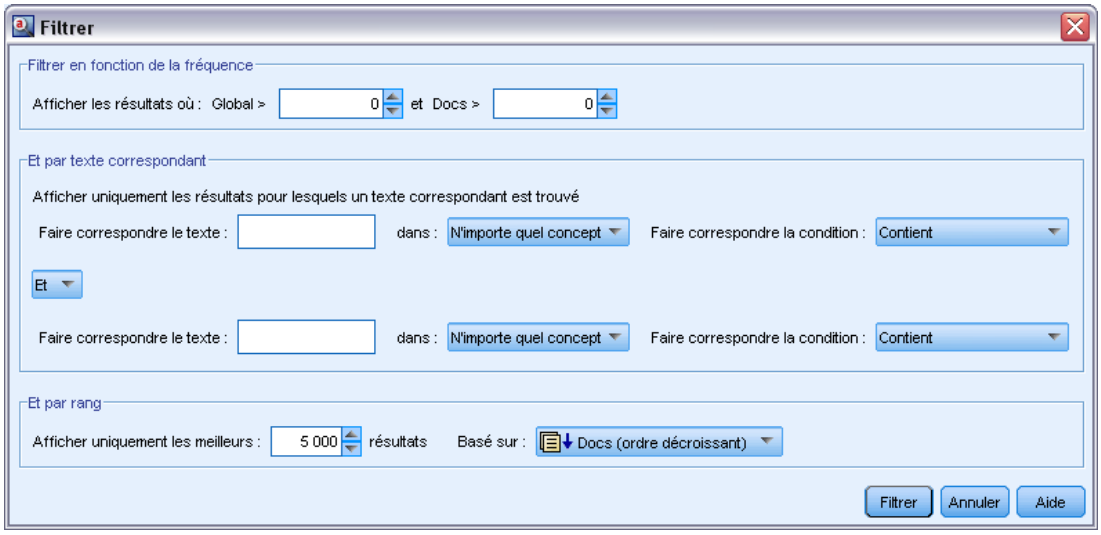
Lorsque les résultats de patrons utilisent moins de six propriétés (valeur maximale), seul le nombre de propriétés (ou colonnes) nécessaire apparaît. Toute propriété vide détectée entre deux propriétés renseignées est ignorée de façon à ce que le patron `<Type1>+<>+<Type2>+<>+<>+<>` puisse être représenté par `<Type1>+<Type3>` . Pour un patron de concept, nous obtenons `concept1+.+concept2` (où `.` représente une valeur nulle).

Comme avec les résultats d'extraction dans la vue Catégories et concepts,vous pouvez vérifier les résultats ici. Si vous souhaitez affiner les types et concepts qui constituent ces patrons, procédez aux modifications dans le panneau Résultats d'extraction de la vue Catégories et concepts ou directement dans l'éditeur de ressources, puis exécutez une nouvelle extraction des patrons. Chaque fois qu'un concept, type ou patron est utilisé tel quel dans une définition de catégorie ou comme partie de règle, une icône de catégorie ou de règle apparaît dans la colonne In du tableau des patrons ou des résultats d'extraction.

Filtrage des résultats TLA

Lorsque vous travaillez sur des ensembles de données très volumineux, le processus d'extraction peut renvoyer des millions de résultats. Pour de nombreux utilisateurs, cette quantité peut compliquer l'examen des résultats. Vous avez cependant la possibilité de filtrer ces résultats, afin de mettre en évidence les plus intéressants. Pour limiter les patrons à afficher, vous pouvez modifier les paramètres dans la boîte de dialogue Filtrer. Tous ces paramètres sont utilisés ensemble.

Figure 12-3
Boîte de dialogue Filtrer (dans la vue Analyse des liens du texte)



Filtrer en fonction de la fréquence. Vous pouvez appliquer un filtre afin de n'afficher que les résultats présentant une certaine valeur de fréquence globale ou de documents.

- La **fréquence globale** est le nombre total de fois où un patron apparaît dans l'ensemble complet des documents ou enregistrements. Elle est indiquée dans la colonne Global.
- La **fréquence de documents** est le nombre total de documents ou d'enregistrements dans lesquels un patron apparaît. Elle est indiquée dans la colonne Docs.

Par exemple, si un patron apparaît 300 fois dans 500 enregistrements, nous déclarons que ce patron a une fréquence globale de 300 et une fréquence de documents de 500.

Et par texte correspondant. Vous pouvez également appliquer un filtre n'affichant que les résultats correspondant à la règle que vous définissez ici. Entrez l'ensemble de caractères qui doit être mis en correspondance dans le champ Texte correspondant, puis indiquez si la recherche de ce texte doit porter sur les noms de concept ou de type (en identifiant le numéro de propriété) ou si elle doit porter sur l'ensemble. Ensuite, sélectionnez la condition à laquelle appliquer la correspondance (il n'est pas nécessaire d'utiliser des chevrons pour marquer le début ou la fin d'un nom de type). Sélectionnez Et ou Ou dans la liste déroulante de sorte que la règle corresponde aux deux instructions ou à une seule, puis définissez la seconde instruction de correspondance du texte de la même manière que la première.

Table 12-1
Conditions de correspondance de texte

Condition	Description
Contient	Texte mis en correspondance si la chaîne apparaît n'importe où. (Option par défaut)
Commence par	Le texte est seulement mis en correspondance si le concept ou le type commence par le texte entré.
Se termine par	Le texte est seulement mis en correspondance si le concept ou le type se termine par le texte entré.
Correspondance exacte	Toute la chaîne doit concorder avec le nom du concept ou du type.

Et par rang. Vous pouvez également procéder à un filtrage pour afficher uniquement les premiers patrons en fonction de la fréquence globale (Global) ou de la fréquence de documents (Docs), dans l'ordre croissant ou décroissant. Cette valeur de rang maximale limite le nombre total de patrons renvoyés pour l'affichage.

Lorsque le filtre est appliqué, le produit ajoute des patrons de type jusqu'à ce que le nombre total maximal de patrons de concept (rang maximal) soit dépassé. Il commence par examiner le patron de type de premier rang, puis relève la somme des patrons de concept correspondants. Si cette somme ne dépasse pas le rang maximal, les patrons sont affichés dans la vue. Ensuite, les patrons de concept du patron de type suivant sont totalisés. Si ce nombre ajouté au nombre total de patrons de concept du patron de type précédent est inférieur au rang maximal, ces patrons sont également affichés dans la vue. Ce processus continue jusqu'à ce qu'un nombre aussi élevé que possible de patrons soit affiché, dans la limite du rang maximal.

Résultats affichés dans le panneau de patrons

Voici des exemples d'affichage des résultats dans la barre d'outils du panneau des patrons, sur la base des filtres.

Figure 12-4
Résultats du filtre, exemple 1

Dans cet exemple, la barre d'outils montre que le nombre de patrons renvoyé est limité en raison du rang maximal spécifié dans le filtre. La présence d'une icône violette indique que le nombre maximal de patrons est atteint. Placez le curseur sur l'icône pour obtenir plus d'informations. Reportez-vous ci-dessus à l'explication du filtre Et par rang.

Figure 12-5
Résultats du filtre, exemple 2

Dans cet exemple, la barre d'outils montre que les résultats ont été limités par un filtre de correspondance de texte (voir l'icône loupe). Vous pouvez pointer sur l'icône pour visualiser la correspondance de texte.

Pour filtrer les résultats

- A partir des menus, sélectionnez Outils> Filtrer. La boîte de dialogue Filtrer apparaît.
- Sélectionnez et affinez les filtres à utiliser.
- Cliquez sur OK pour appliquer les filtres et visualiser les nouveaux résultats.

Panneau Données

Lorsque vous extrayez et explorez des patrons d'analyse des liens du texte, vous pouvez passer en revue certaines des données avec lesquelles vous travaillez. Par exemple, vous pouvez visualiser les enregistrements réels dans lesquels un groupe de patrons a été découvert. Vous pouvez consulter les enregistrements ou les documents dans le panneau Données, situé dans l'angle

inférieur droit. S'il n'apparaît pas par défaut, sélectionnez Affichage> Panneaux > Données dans les menus.

Le panneau Données présente une ligne par document ou enregistrement correspondant à une sélection dans la vue, jusqu'à une certaine limite d'affichage. Par défaut, le nombre de documents ou d'enregistrements affichés dans le panneau Données est limité pour vous permettre de consulter vos données plus rapidement. Cependant, vous pouvez modifier cette limite dans la boîte de dialogue Options. Pour plus d'informations, reportez-vous à la section Options : onglet Session dans le chapitre 8 sur p. 131.

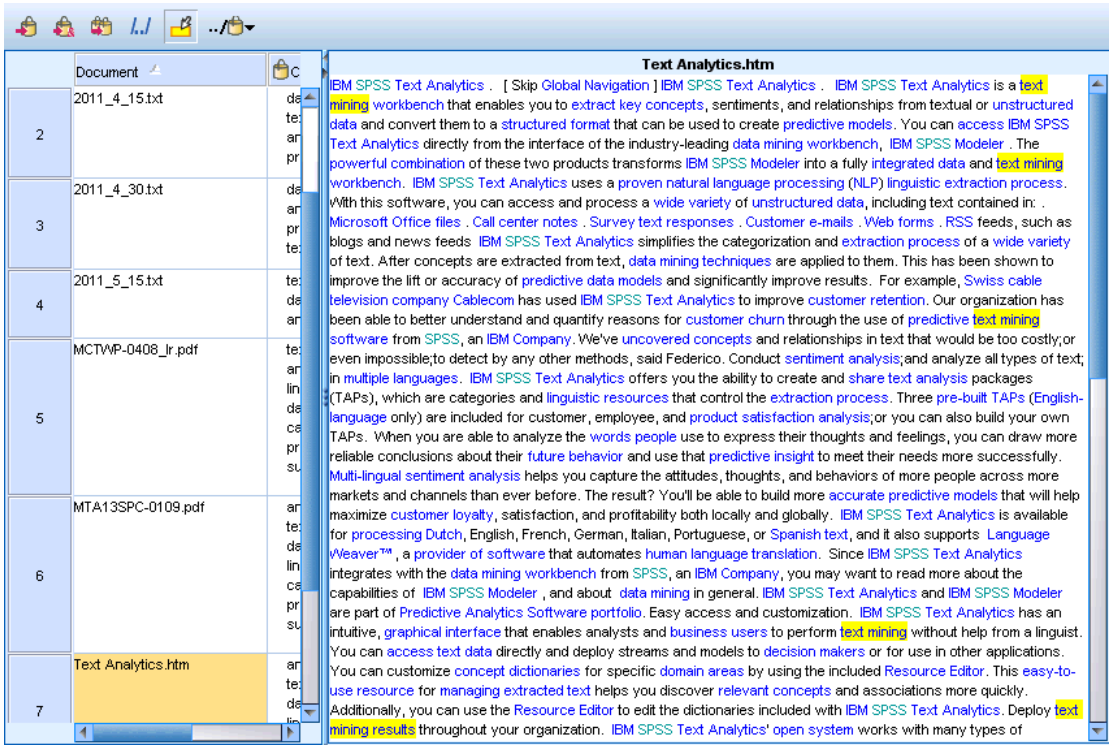
Affichage et actualisation du panneau Données

L'affichage du panneau Données n'est pas automatiquement actualisé car en présence d'ensembles de données volumineux, l'opération prendrait trop de temps. Par conséquent, lorsque vous sélectionnez des patrons de type ou de concept dans cette vue, vous pouvez cliquer sur Afficher pour actualiser le contenu du panneau Données.

Documents texte ou enregistrements

Si vos données textuelles sont sous la forme d'enregistrements et que le texte est relativement bref, le champ de texte du panneau Données affiche les informations dans leur intégralité. Cependant, si vous utilisez des enregistrements et de grands ensembles de données, la colonne du champ de texte affiche une petite partie du texte et ouvre un panneau Aperçu du texte à droite qui permet de consulter une plus grande partie du texte de l'enregistrement sélectionné dans la table, voire son intégralité. Si vos données textuelles se présentent sous la forme de documents, le panneau Données affiche le nom de fichier du document. Lorsque vous sélectionnez un document, le panneau Aperçu du texte s'ouvre et affiche le texte du document sélectionné.

Figure 12-6
Panneau de données avec panneau Aperçu du texte



Couleurs et mise en surbrillance

Chaque fois que vous affichez des données, des concepts et des descripteurs trouvés dans ces documents ou enregistrements, ils apparaissent en couleur pour vous permettre de les identifier facilement dans le texte. Le code couleur correspond aux types auxquels les concepts appartiennent. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher le concept sous lequel il a été extrait et le type auquel il a été affecté. Tout texte n'ayant pas été extrait apparaît en noir. En règle générale, ces mots non extraits sont souvent des connecteurs (*et* ou *avec*), des pronoms (*me* ou *ils*), et des verbes (*être*, *avoir* ou *prendre*).

Colonnes du panneau Données

Alors que la colonne de champ du texte est toujours visible, il est possible d'afficher également d'autres colonnes. Pour afficher d'autres colonnes, cliquez sur Affichage > Panneau Données dans les menus, puis sélectionnez la colonne que vous souhaitez afficher dans le panneau de données. Les colonnes pouvant être affichées sont les suivantes :

- **"Nom du champ de texte" (#)/Documents.** Ajoute une colonne pour les données textuelles à partir desquelles des concepts et des types ont été extraits. Si vos données sont contenues dans des documents, la colonne est appelée Documents, et seul le nom de fichier du document ou son chemin d'accès complet est visible. Pour examiner le texte de ces documents, vous devez consulter le panneau Aperçu du texte. Le nombre de lignes du panneau Données est indiqué entre parenthèses après le nom de cette colonne. Il peut arriver que les documents ou

les enregistrements ne soient pas tous affichés en raison d'une limite définie dans la boîte de dialogue Options pour optimiser la vitesse de chargement. Si la limite est atteinte, le nombre sera suivi de - Max. Pour plus d'informations, reportez-vous à la section Options : onglet Session dans le chapitre 8 sur p. 131.

- **Catégories.** Répertorie chacune des catégories à laquelle appartient un enregistrement. Lorsque cette colonne est affichée, l'actualisation du panneau Données peut prendre plus de temps afin d'afficher les informations les plus récentes.
- **Rang de pertinence.** Donne un rang pour chaque enregistrement dans une seule catégorie. Ce rang montre dans quelle mesure l'enregistrement correspond à la catégorie par rapport aux autres enregistrements dans cette catégorie. Sélectionnez une catégorie dans le panneau Catégories (panneau supérieur gauche) pour voir le rang. Pour plus d'informations, reportez-vous à la section Pertinence des catégories dans le chapitre 10 sur p. 176.
- **Effectifs de catégories** Répertorie le nombre de catégories auxquelles appartient un enregistrement.

Chapitre
13

Visualisation des graphiques

La vue Catégories et concepts, la vue Clusters et la vue Analyse des liens du texte présentent toutes un panneau de visualisation dans l'angle supérieur droit de la fenêtre. Vous pouvez utiliser ce panneau pour explorer visuellement les données. Les graphiques et diagrammes suivants sont disponibles.

- **Vue Catégories et concepts.** Cette vue inclut trois graphiques et diagrammes : *Barre Catégorie*, *Relations de catégorie* et *Tableau des relations de catégorie*. Dans cette vue, les graphiques sont mis à jour uniquement lorsque vous cliquez sur Afficher. Pour plus d'informations, reportez-vous à la section Graphiques et diagrammes de catégorie sur p. 255.
- **Vue Clusters.** Cette vue inclut deux graphiques Relations : le graphique *Relations par concept* et le graphique *Relations par cluster*. Pour plus d'informations, reportez-vous à la section Graphiques Cluster sur p. 259.
- **Vue Analyse des liens du texte.** Cette vue inclut deux graphiques Relations : le graphique *Relations par concept* et le graphique *Relations par type*. Pour plus d'informations, reportez-vous à la section Graphiques Analyse des liens du texte sur p. 261.

Pour des informations supplémentaires sur l'ensemble des barres d'outils et des palettes générales utilisées pour l'édition des graphiques, consultez la section sur l'Edition des graphiques dans l'aide en ligne ou dans le fichier *SourceProcessOutputNodes.pdf*, disponible dans le dossier *\Documentation\en* du DVDIBM® SPSS® Modeler.

Graphiques et diagrammes de catégorie

Lors de la création des catégories, il est important de prendre le temps de passer en revue les définitions de catégorie, les documents ou enregistrements qu'elles contiennent, et les chevauchements de catégorie. Le panneau de visualisation offre plusieurs perspectives sur les catégories. Le panneau Visualisation est situé dans l'angle supérieur droit de la vue Catégories et concepts. S'il est encore masqué, vous pouvez y accéder à partir du menu Affichage (Affichage > Panneaux > Visualisation).

Dans cette vue, le panneau de visualisation offre trois perspectives sur les similarités de catégorisation des documents ou des enregistrements. Vous pouvez vous appuyer sur les graphiques et diagrammes de ce panneau pour analyser les résultats de la catégorisation, et affiner les catégories ou générer des rapports. Lorsque vous affinez des catégories, vous pouvez utiliser ce panneau pour examiner les définitions de catégorie afin de découvrir les catégories trop similaires (par exemple, les catégories qui partagent plus de 75 % de leurs documents ou enregistrements) ou trop différentes. Si deux catégories sont trop similaires, vous pouvez décider de les combiner. Vous pouvez également décider d'affiner les définitions des catégories en supprimant certains descripteurs d'une catégorie.

En fonction des éléments sélectionnés dans le panneau Résultats d'extraction , dans le panneau Catégories ou dans la boîte de dialogue Définitions de catégorie, vous pouvez visualiser les interactions correspondantes entre les documents/enregistrements et les catégories de chaque onglet de ce panneau. Chacun présente des informations similaires, mais d'une façon différente ou avec un niveau de détail différent. Néanmoins, afin d'actualiser un graphique pour la sélection actuelle, cliquez sur Afficher dans la barre d'outils du panneau ou de la boîte de dialogue dans lequel vous avez effectué votre sélection.

Le panneau Visualisation de la vue Catégories et Concepts propose les graphiques et diagrammes suivants :

- **Diagramme Barre Catégorie.** Un tableau et un diagramme en bâton présentent le chevauchement entre les documents/enregistrements correspondant à votre sélection et les catégories associées. Le diagramme en bâton présente également le rapport entre les documents/enregistrements dans les catégories et le nombre total de documents/enregistrements. Pour plus d'informations, reportez-vous à la section Diagramme Barre Catégorie sur p. 256.
- **Graphique Relations de catégorie.** Ce graphique présente le chevauchement des documents/enregistrements pour les catégories auxquelles appartiennent les documents/enregistrements, en fonction de la sélection dans les autres panneaux. Pour plus d'informations, reportez-vous à la section Graphique Relations de catégorie sur p. 257.
- **Tableau des relations de catégorie.** Cet onglet présente les mêmes informations que l'onglet Relations de catégorie, mais sous forme de tableau. Il contient trois colonnes qu'il est possible de trier en cliquant sur leur en-tête. Pour plus d'informations, reportez-vous à la section Tableau des relations de catégorie sur p. 258.

Pour plus d'informations, reportez-vous à la section Catégorisation des données textuelles dans le chapitre 10 sur p. 162.

Diagramme Barre Catégorie

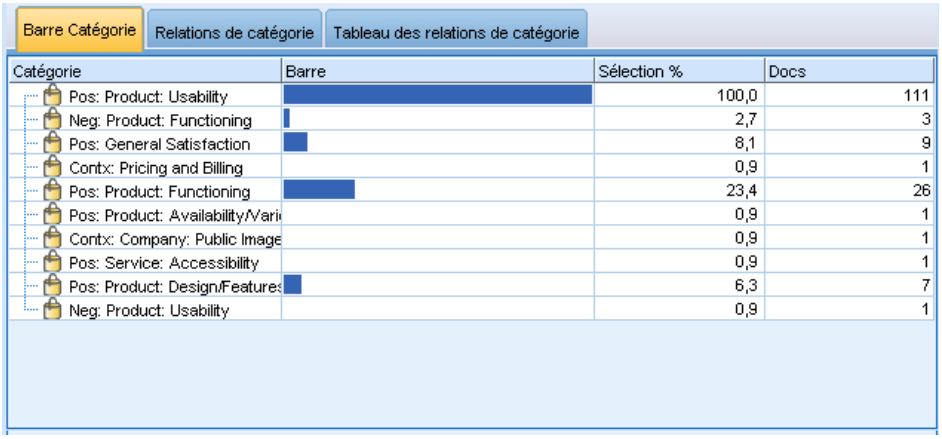
Cet onglet affiche un tableau et un diagramme en bâton qui présentent le chevauchement entre les documents/enregistrements correspondant à votre sélection et les catégories associées. Le diagramme en bâton présente également le rapport entre les documents/enregistrements dans les catégories et le nombre total de documents ou d'enregistrements. Vous ne pouvez pas modifier la présentation de ce diagramme. Toutefois, vous pouvez trier les colonnes en cliquant sur leur en-tête.

Le diagramme contient les colonnes suivantes :

- **Catégorie.** Cette colonne présente le nom des catégories de votre sélection. Par défaut, la catégorie la plus courante de votre sélection est répertoriée en premier.
- **Barre.** Cette colonne présente, de manière visuelle, le rapport entre les documents ouenregistrements d'une catégorie donnée et le nombre total de documents ou d'enregistrements.
- **Sélection %.** Cette colonne présente un pourcentage basé sur le rapport entre le nombre total de documents ou d'enregistrements pour une catégorie et le nombre total de documents ou d'enregistrements représentés dans la sélection.

- **Docs.** Cette colonne présente le nombre de documents ou d'enregistrements d'une sélection pour la catégorie donnée.

Figure 13-1
Diagramme Barre Catégorie

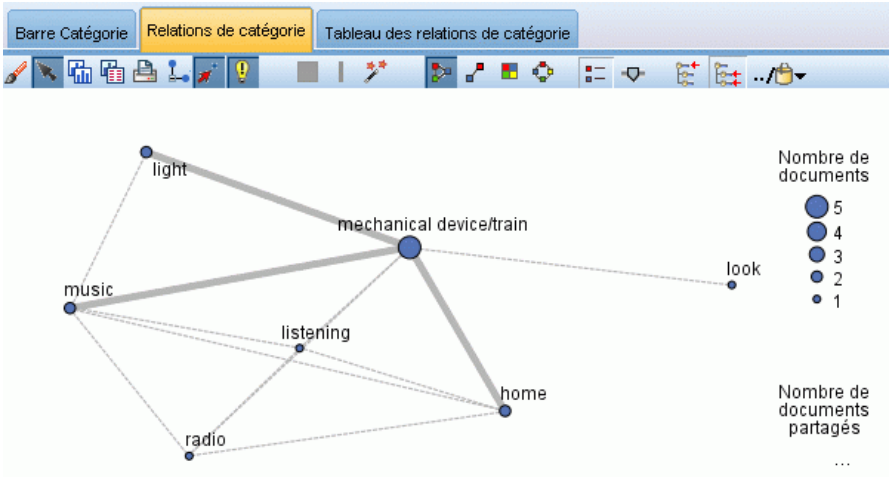


Catégorie	Barre	Sélection %	Docs
Pos: Product: Usability		100,0	111
Neg: Product: Functioning		2,7	3
Pos: General Satisfaction		8,1	9
Contx: Pricing and Billing		0,9	1
Pos: Product: Functioning		23,4	26
Pos: Product: Availability/Vari		0,9	1
Contx: Company: Public Image		0,9	1
Pos: Service: Accessibility		0,9	1
Pos: Product: Design/Features		6,3	7
Neg: Product: Usability		0,9	1

Graphique Relations de catégorie

Cet onglet affiche un graphique Relations de catégorie. Les relations présentent le chevauchement des documents ou des enregistrements pour les catégories auxquelles les documents ou les enregistrements appartiennent, en fonction de la sélection dans les autres panneaux. S'il existe des étiquettes de catégorie, celles-ci apparaissent dans le graphique. Vous pouvez sélectionner la présentation du graphique (réseau, cercle, orientée ou grille) à l'aide des boutons de la barre d'outils de ce panneau.

Figure 13-2
Graphique Relations de catégorie, présentation de grille



Dans les relations, chaque noeud représente une catégorie. A l'aide de la souris, vous pouvez sélectionner et déplacer les noeuds au sein du panneau. La taille du noeud représente la taille relative basée sur le nombre de documents ou d'enregistrements pour cette catégorie dans votre

sélection. L'épaisseur et la couleur de la ligne entre deux catégories représentent le nombre de documents ou d'enregistrements communs qu'elles contiennent. Si vous pointez sur un nœud en mode d'interaction, une info-bulle affiche le nom (ou l'étiquette) de la catégorie et le nombre global de documents ou enregistrements dans la catégorie.

Remarque : Par défaut, le mode d'interaction est activé pour les graphiques dans lesquels vous pouvez déplacer des noeuds. Toutefois, vous pouvez passer en mode d'édition pour modifier la présentation de vos graphiques, notamment les couleurs, polices et légendes. Pour plus d'informations, reportez-vous à la section Utilisation des palettes et des barres d'outils de graphiques sur p. 264.

Tableau des relations de catégorie

Cet onglet présente les mêmes informations que l'onglet Relations de catégorie, mais sous forme de tableau. Le tableau contient trois colonnes qu'il est possible de trier en cliquant sur leur en-tête :

- **Effectif.** Cette colonne présente le nombre de documents ou d'enregistrements que les deux catégories partagent ou ont en commun.
- **Catégorie 1.** Cette colonne présente le nom de la première catégorie, suivi entre parenthèses du nombre total de documents ou d'enregistrements qu'elle contient.
- **Catégorie 2.** Cette colonne présente le nom de la deuxième catégorie, suivi entre parenthèses du nombre total de documents ou d'enregistrements qu'elle contient.

Figure 13-3
Tableau des relations de catégorie

Barre Catégorie | Relations de catégorie | Tableau des relations de catégorie

Effectifs	Catégorie 1	Catégorie 2
1	Contx: Company: Public Image/Reputation(1)	Neg: Product: Functioning(3)
1	Contx: Company: Public Image/Reputation(1)	Neg: Product: Usability(1)
1	Contx: Company: Public Image/Reputation(1)	Pos: Product: Functioning(26)
1	Contx: Company: Public Image/Reputation(1)	Pos: Product: Usability(111)
1	Neg: Product: Functioning(3)	Pos: General Satisfaction(9)
1	Neg: Product: Functioning(3)	Pos: Product: Availability/Variety/Size(1)
1	Neg: Product: Functioning(3)	Neg: Product: Usability(1)
1	Neg: Product: Functioning(3)	Pos: Product: Functioning(26)
1	Neg: Product: Usability(1)	Pos: Product: Functioning(26)
1	Neg: Product: Usability(1)	Pos: Product: Usability(111)
1	Pos: General Satisfaction(9)	Pos: Product: Availability/Variety/Size(1)
1	Pos: Product: Design/Features(7)	Pos: General Satisfaction(9)
1	Pos: Product: Functioning(26)	Pos: General Satisfaction(9)
1	Pos: Product: Usability(111)	Contx: Pricing and Billing(1)
1	Pos: Product: Usability(111)	Pos: Product: Availability/Variety/Size(1)
1	Pos: Product: Usability(111)	Pos: Service: Accessibility(1)
3	Neg: Product: Functioning(3)	Pos: Product: Usability(111)
7	Pos: Product: Design/Features(7)	Pos: Product: Usability(111)
9	Pos: Product: Usability(111)	Pos: General Satisfaction(9)
26	Pos: Product: Functioning(26)	Pos: Product: Usability(111)

Graphiques Cluster

Après avoir créé vos clusters, vous pouvez les explorer visuellement dans les graphiques Relations du panneau Visualisation. Ce panneau offre deux perspectives sur la classification non supervisée : un graphique Relations par concept et un graphique Relations par cluster. Il est possible d'utiliser les graphiques Relations de ce panneau pour analyser les résultats de classification non supervisée et découvrir certains concepts et règles que vous pouvez ajouter à vos catégories. Le panneau Visualisation est situé dans l'angle supérieur droit de la vue Clusters. S'il est masqué, vous pouvez y accéder à partir du menu Affichage (Affichage > Panneaux > Visualisation). En sélectionnant un cluster dans le panneau Cluster, vous pouvez afficher automatiquement les graphiques correspondants dans le panneau Visualisation.

Remarque : par défaut, les graphiques se trouvent dans le mode d'interaction/de sélection (dans lequel vous pouvez déplacer des noeuds). Toutefois, vous pouvez modifier la présentation de vos graphiques en mode d'édition, notamment les couleurs, polices et légendes. Pour plus d'informations, reportez-vous à la section Utilisation des palettes et des barres d'outils de graphiques sur p. 264.

La vue Clusters inclut deux graphiques Relations.

- **Graphique Relations par concept**. Ce graphique présente tous les concepts des clusters sélectionnés, ainsi que des concepts liés en dehors du cluster. Ce graphique peut vous aider à visualiser la façon dont les concepts sont liés au sein d'un cluster, ainsi que d'éventuels liens externes. Pour plus d'informations, reportez-vous à la section Graphique Relations par concept sur p. 259.
- **Graphique Relations par cluster**. Ce graphique présente sous la forme de lignes en pointillé les clusters sélectionnés avec tous les liens externes entre eux. Pour plus d'informations, reportez-vous à la section Graphique Relations par cluster sur p. 260.

Pour plus d'informations, reportez-vous à la section Analyse des clusters dans le chapitre 11 sur p. 237.

Graphique Relations par concept

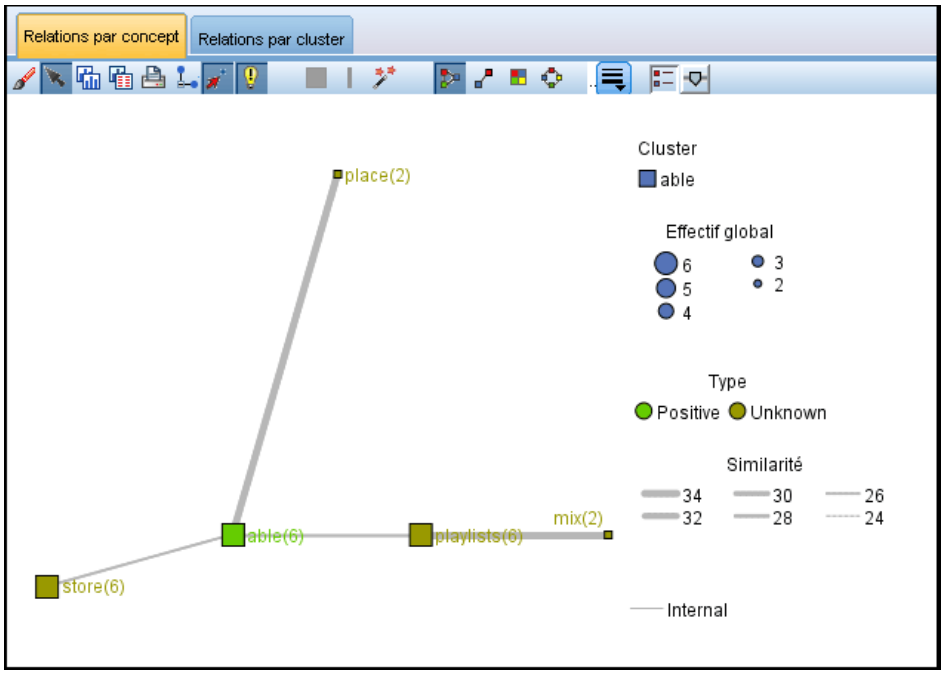
Cet onglet affiche un graphique Relations représentant tous les concepts des clusters sélectionnés, ainsi que les concepts liés en dehors du cluster. Ce graphique peut vous aider à visualiser la façon dont les concepts sont liés au sein d'un cluster, ainsi que d'éventuels liens externes. Dans un cluster, chaque concept est représenté sous la forme d'un noeud avec un code de couleur en fonction de la couleur de type. Pour plus d'informations, reportez-vous à la section Création de types dans le chapitre 17 sur p. 306.

Les liens internes entre les concepts d'un cluster sont tracés et l'épaisseur des lignes représentant chaque lien est directement liée aux effectifs des documents pour la cooccurrence de chaque paire de concepts ou à la valeur du lien de similarité, selon votre choix dans la barre d'outils du graphique. Les liens externes entre les concepts d'un cluster et les concepts situés en dehors du cluster sont également représentés.

Si des concepts sont sélectionnés dans la boîte de dialogue Définitions du cluster, le graphique Relations par concept affiche ces concepts et tout lien interne et externe associé à ces concepts. Les liens entre d'autres concepts, qui n'incluent pas l'un des concepts sélectionnés, n'apparaissent pas sur le graphique.

Remarque : par défaut, les graphiques se trouvent dans le mode d'interaction/de sélection (dans lequel vous pouvez déplacer des noeuds). Toutefois, vous pouvez modifier la présentation de vos graphiques en mode d'édition, notamment les couleurs, polices et légendes. Pour plus d'informations, reportez-vous à la section Utilisation des palettes et des barres d'outils de graphiques sur p. 264.

Figure 13-4
Graphique Relations par concept



Graphique Relations par cluster

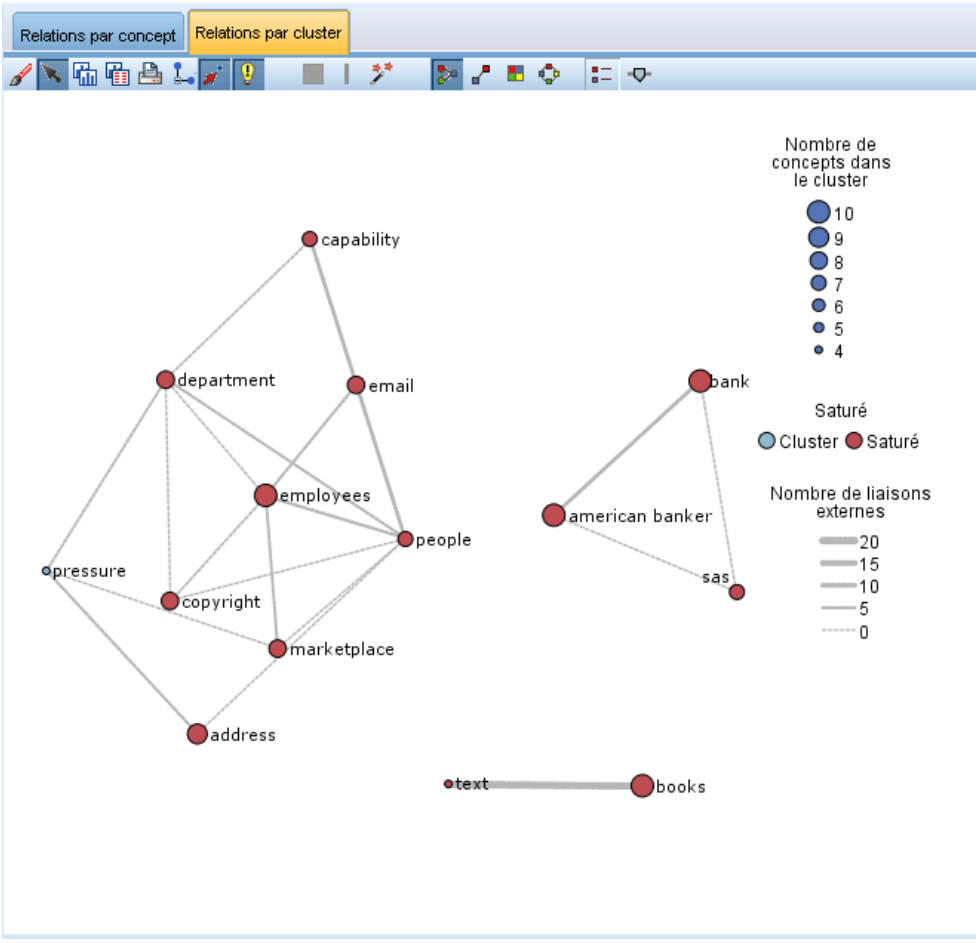
Cet onglet affiche un graphique Relations représentant les clusters sélectionnés. Les liens externes entre les clusters sélectionnés, ainsi que les liens entre d'autres clusters apparaissent tous sous la forme de lignes en pointillé. Dans un graphique Relations par cluster, chaque noeud représente la totalité d'un cluster et l'épaisseur des lignes tracées entre eux représente le nombre de liens externes entre deux clusters.

Important ! Pour pouvoir afficher un graphique Relations par cluster, vous devez avoir créé des clusters présentant des liens externes. Les liens externes sont des liens entre des paires de concepts dans des clusters distincts (un concept au sein d'un cluster et un concept en dehors, dans un autre cluster).

Par exemple, prenons deux clusters. La `Cluster A` inclut trois concepts : `A1`, `A2` et `A3`. La `Cluster B` inclut deux concepts : `B1` et `B2`. Les concepts suivants sont liés : `A1-A2`, `A1-A3`, `A2-B1` (externe), `A2-B2` (externe), `A1-B2` (externe) et `B1-B2`. En d'autres termes, dans le graphique Relations par cluster, l'épaisseur de ligne représente les trois liens externes.

Remarque : par défaut, les graphiques se trouvent dans le mode d'interaction/de sélection (dans lequel vous pouvez déplacer des noeuds). Toutefois, vous pouvez modifier la présentation de vos graphiques en mode d'édition, notamment les couleurs, polices et légendes. Pour plus d'informations, reportez-vous à la section Utilisation des palettes et des barres d'outils de graphiques sur p. 264.

Figure 13-5
Graphique Relations par cluster



Graphiques Analyse des liens du texte

Après avoir extrait vos patrons Analyse des liens du texte (TLA), vous pouvez les explorer visuellement dans les graphiques Relations du panneau Visualisation. Ce panneau offre deux perspectives sur les patrons TLA : un graphique Relations de concept (patron) et un graphique Relations de type (patron). Il est possible d'utiliser les graphiques Relations de ce panneau pour représenter visuellement des patrons. Le panneau Visualisation est situé dans l'angle supérieur droit de la vue Analyse des liens du texte. S'il est masqué, vous pouvez y accéder à partir du menu Affichage (Affichage > Panneaux > Visualisation). En l'absence de sélection, la zone de graphique est vide.

Remarque : par défaut, les graphiques se trouvent dans le mode d'interaction/de sélection (dans lequel vous pouvez déplacer des noeuds). Toutefois, vous pouvez modifier la présentation de vos graphiques en mode d'édition, notamment les couleurs, polices et légendes. Pour plus d'informations, reportez-vous à la section Utilisation des palettes et des barres d'outils de graphiques sur p. 264.

La vue Analyse des liens du texte inclut deux graphiques Relations.

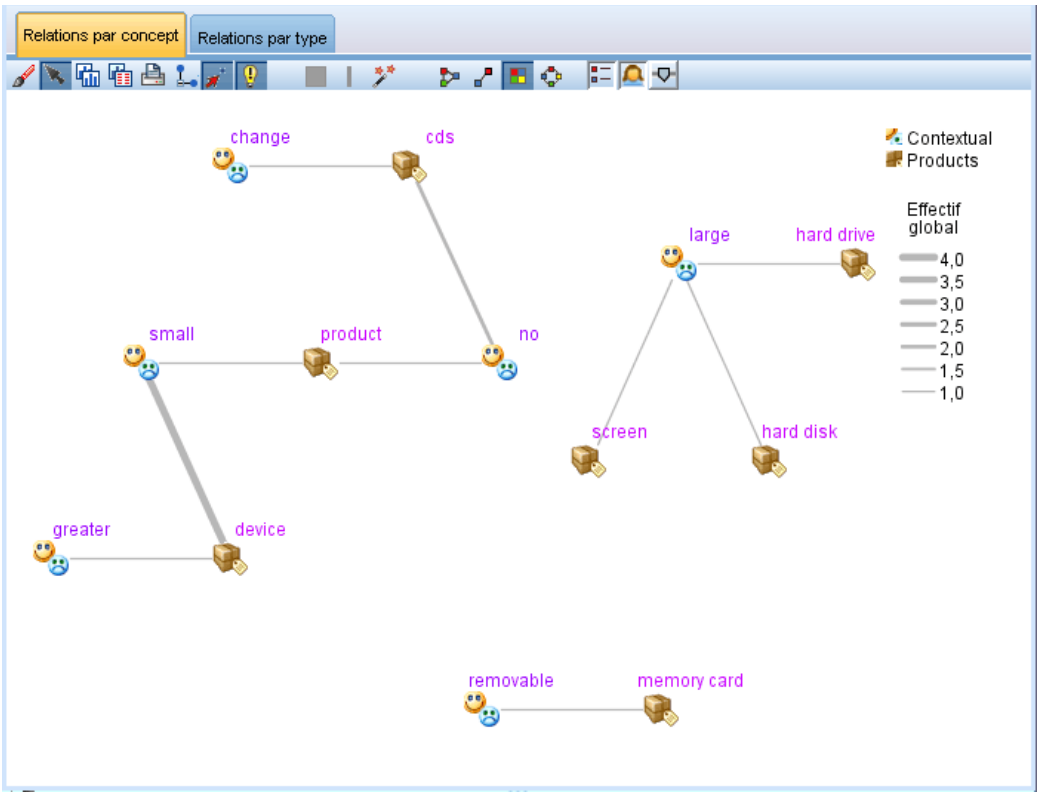
- **Graphique Relations par concept**. Ce graphique présente tous les concepts figurant dans les patrons sélectionnés. Dans un graphique de concept, l'épaisseur des lignes et la taille des noeuds (lorsque les icônes de type ne sont pas représentées) indiquent le nombre d'occurrences globales figurant dans la table sélectionnée. Pour plus d'informations, reportez-vous à la section Graphique Relations par concept sur p. 262.
- **Graphique Relations par type**. Ce graphique présente tous les types figurant dans les patrons sélectionnés. Dans un graphique de type, l'épaisseur des lignes et la taille des noeuds (lorsque les icônes de type ne sont pas représentées) indiquent le nombre d'occurrences globales figurant dans la table sélectionnée. Les noeuds sont représentés soit par une couleur de type, soit par une icône. Pour plus d'informations, reportez-vous à la section Graphique Relations par type sur p. 263.

Pour plus d'informations, reportez-vous à la section Exploration de l'analyse des liens du texte dans le chapitre 12 sur p. 245.

Graphique Relations par concept

Ce graphique Relations présente tous les concepts représentés dans la sélection actuelle. Par exemple, si vous avez sélectionné un patron de type avec trois patrons de concept correspondants, le graphique affiche trois ensembles de concepts liés. L'épaisseur de ligne et la taille des noeuds d'un graphique de concept représentent les fréquences globales. Le graphique représente visuellement des informations identiques aux éléments sélectionnés dans les panneaux de patrons. Le type de chaque concept est présenté par une couleur ou une icône, selon ce que vous avez sélectionné dans la barre d'outils du graphique. Pour plus d'informations, reportez-vous à la section Utilisation des palettes et des barres d'outils de graphiques sur p. 264.

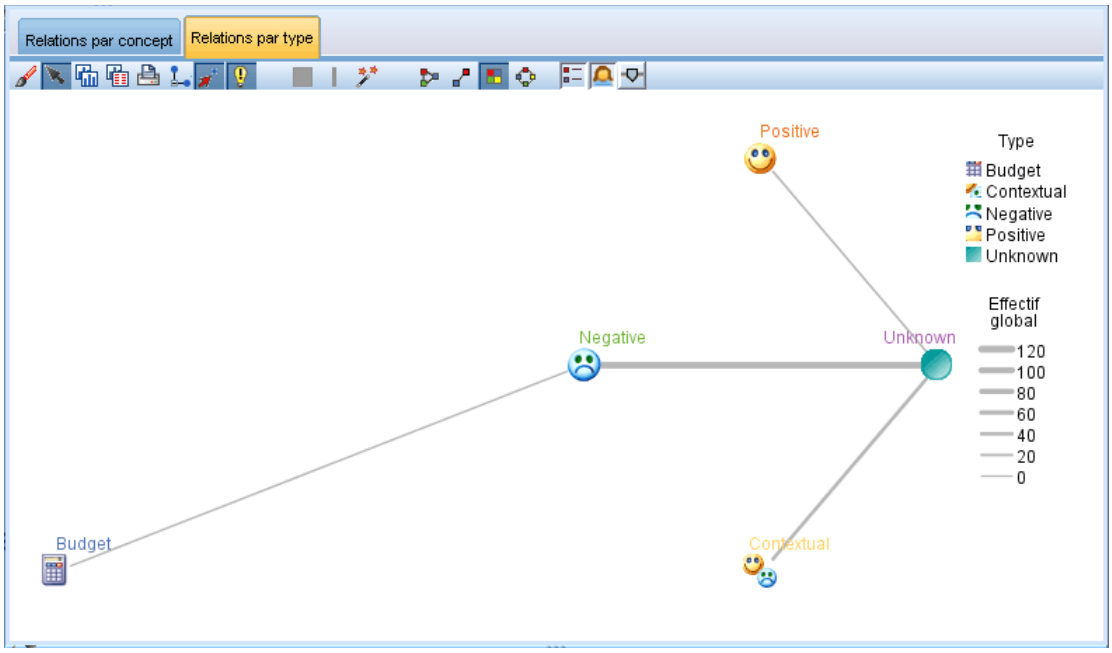
Figure 13-6
Graphique Relations par concept



Graphique Relations par type

Ce graphique Relations représente chaque patron de type pour la sélection actuelle. Par exemple, si vous avez sélectionné deux patrons de concept, le graphique affiche un noeud par type dans les patrons sélectionnés et les liens entre eux détectés dans le même patron. L'épaisseur de ligne et la taille des noeuds représentent les fréquences globales pour l'ensemble. Le graphique représente visuellement des informations identiques aux éléments sélectionnés dans les panneaux de patrons. Outre les noms de type qui apparaissent dans le graphique, les types sont également identifiés grâce à leur couleur ou à une icône de type, selon ce que vous avez sélectionné dans la barre d'outils du graphique. Pour plus d'informations, reportez-vous à la section Utilisation des palettes et des barres d'outils de graphiques sur p. 264.

Figure 13-7
Graphique Relations par type



Utilisation des palettes et des barres d'outils de graphiques

Pour chaque graphique, une barre d'outils fournit un accès rapide à certaines palettes courantes avec lesquelles vous pouvez effectuer un certain nombre d'actions sur vos graphiques. Chaque vue (Catégories et concepts, Clusters et Analyse des liens du texte) présente une barre d'outils légèrement différente. Vous pouvez choisir entre le mode d'affichage *Sélection/Interaction* ou le mode d'affichage *Edition*.

Tandis que le mode d'interaction vous permet d'explorer de manière analytique les données et les valeurs représentées par la visualisation, le mode d'édition vous permet de modifier la mise en forme et l'aspect de la visualisation. Vous pouvez par exemple modifier les polices et les couleurs pour respecter le guide de style de votre organisation. Pour sélectionner ce mode, sélectionnez Affichage > Panneau Visualisation > Mode d'édition dans les menus (ou cliquez sur l'icône de la barre d'outils).

En mode d'édition, plusieurs barres d'outils ont un impact sur différents aspects de la présentation de la visualisation. Si vous n'utilisez pas certaines barres d'outils, vous pouvez les masquer afin d'augmenter l'espace disponible dans la boîte de dialogue dans laquelle le graphique est affiché. Pour sélectionner ou désélectionner des barres d'outils, cliquez sur le nom de la barre d'outils ou de la palette souhaitée dans le menu Affichage.

Pour des informations supplémentaires sur l'ensemble des barres d'outils et des palettes générales utilisées pour l'édition des graphiques, consultez la section sur l'Edition des graphiques dans l'aide en ligne ou dans le fichier *SourceProcessOutputNodes.pdf*, disponible dans le dossier *\Documentation\en* du DVDIBM® SPSS® Modeler.

Table 13-1
Boutons de la barre d'outil de Text Analytics

Bouton/liste	Description
	Active le mode d'édition. Basculez vers le mode d'édition pour modifier l'apparence du graphique, par exemple pour agrandir la police, modifier les couleurs pour respecter le guide de style de l'entreprise, ou supprimer des étiquettes et des légendes.
	Active le mode d'interaction. Par défaut, le mode d'interaction est activé, ce qui signifie que vous pouvez déplacer et faire glisser des noeuds dans le graphique et pointer sur des objets du graphique pour obtenir des informations supplémentaires via une info-bulle.
	Sélectionnez un type d'affichage des relations pour les graphiques dans les vues Catégories et concepts et Analyse des liens du texte. ■ Présentation en cercle. Présentation générale qu'il est possible d'appliquer à n'importe quel graphique. Pour créer un graphique, elle suppose que les liens ne sont pas orientés et traite tous les noeuds de la même façon. Les noeuds sont placés uniquement autour du périmètre d'un cercle. ■ Présentation en réseau. Présentation générale qu'il est possible d'appliquer à n'importe quel graphique. Pour créer un graphique, elle suppose que les liens ne sont pas orientés et traite tous les noeuds de la même façon. Les noeuds sont placés librement au sein de la présentation. ■ Présentation orientée. Il convient d'utiliser cette présentation uniquement pour les graphiques orientés. Cette présentation produit des structures ressemblant à un arbre, des noeuds racine aux noeuds feuille, et fournit une organisation par couleurs. Les données hiérarchiques ont un aspect attrayant avec cette présentation. ■ Présentation de grille. Présentation générale qu'il est possible d'appliquer à n'importe quel graphique. Pour créer un graphique, elle suppose que les liens ne sont pas orientés et traite tous les noeuds de la même façon. Les noeuds sont placés uniquement aux points de grille au sein de l'espace.
	Représentation de la taille des liens. Choisissez ce que l'épaisseur de la ligne représente dans le graphique. S'applique uniquement à la vue Clusters. Le graphique Relations par cluster affiche uniquement le nombre de liens externes entre les clusters. Vous pouvez choisir entre : ■ Similarité. L'épaisseur indique le nombre de liens externes entre deux clusters ■ Cooccurrence. L'épaisseur indique le nombre de documents dans lesquels une cooccurrence de descripteurs a lieu.
	Bouton bascule qui, lorsqu'il est pressé, affiche la légende. Lorsque le bouton est désactivé, la légende n'apparaît pas.
	Bouton bascule qui, lorsqu'il est activé, affiche les icônes de type dans le graphique plutôt que les couleurs de type. S'applique uniquement à la vue Analyse des liens du texte.
	Bouton bascule qui, lorsqu'il est pressé, affiche le curseur des liens sous le graphique. Vous pouvez filtrer les résultats en faisant glisser la flèche.
	Il affichera le graphique pour le niveau de catégories sélectionné le plus élevé plutôt que celui de leurs sous-catégories.

Bouton/liste	Description
	Il affichera le graphique pour le niveau de catégories sélectionné le plus bas.
	Cette option contrôle le mode d'affichage des noms de sous-catégories dans les résultats. ■ **Chemin d'accès complet aux catégories.** Cette option va générer le nom de la catégorie et le chemin d'accès complet aux catégories parents le cas échéant en utilisant des barres obliques pour séparer les noms de catégories des noms de sous-catégories. ■ **Chemin d'accès court aux catégories.** Cette option va générer seulement le nom de la catégorie mais utilise des point de suspension pour afficher le nombre de catégories parents pour la catégorie en question. ■ **Catégorie du niveau le plus bas.** Cette option va générer seulement le nom de la catégorie sans afficher le chemin d'accès complet ou les catégories parents.

Chapitre
14

Editeur de ressources de session

IBM® SPSS® Modeler Text Analytics capture et extrait rapidement et avec précision des concepts-clés de données texte. Ce processus d'extraction repose principalement sur les ressources linguistiques afin de déterminer la façon d'extraire des informations des données textuelles. Par défaut, les ressources proviennent de modèles de ressources.

SPSS Modeler Text Analytics est fourni avec des **modèles de ressources** spécialisés qui contiennent des ressources linguistiques et non linguistiques sous forme de bibliothèques et de ressources avancées, pour vous permettre de définir le traitement et l'extraction des données. Pour plus d'informations, reportez-vous à la section Modèles et ressources dans le chapitre 15 sur p. 273.

Dans la boîte de dialogue de nœud, vous pouvez charger une copie des ressources du modèle vers le nœud. Dans une session interactive, vous pouvez personnaliser ces ressources en fonction des données du nœud si vous le désirez. Au cours d'une session interactive, vous pouvez utiliser vos ressources dans la vue de Editeur de ressources. Lorsque vous lancez une session interactive, une extraction est effectuée en utilisant les ressources chargées dans la boîte de dialogue du nœud, si vous n'avez pas placé les données et les résultats d'extraction dans le nœud.

Modification des ressources dans l'éditeur de ressources

L'Editeur de ressources permet d'accéder aux ressources utilisées pour produire les résultats de l'extraction (concepts, types et patrons) pour une session interactive. Cet éditeur est très similaire à l'Editeur de modèle, sauf que dans l'Editeur de ressources vous modifiez les ressources de la session. Une fois que vous avez terminé d'utiliser les ressources et toute autre tâche, vous pouvez mettre à jour le nœud de modélisation pour enregistrer ce travail pour pouvoir le restaurer dans une session interactive suivante. Pour plus d'informations, reportez-vous à la section Mise à jour des nœuds de modélisation et enregistrement dans le chapitre 8 sur p. 136.

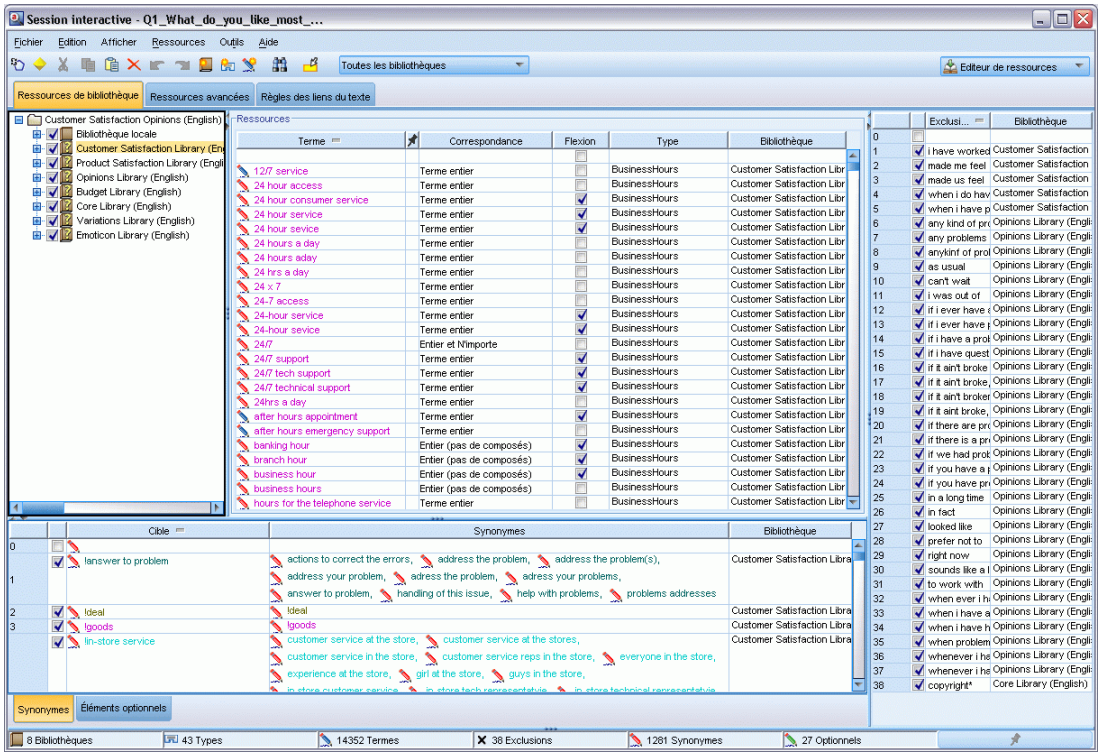
Si vous voulez travailler directement dans les modèles utilisés pour charger les ressources dans les nœuds, il est recommandé d'utiliser l'Editeur de modèle. La plupart des tâches que vous pouvez exécuter dans l'Editeur de ressources s'exécutent de la même manière dans l'Editeur de modèle, à savoir :

- **Utilisation de bibliothèques**Pour plus d'informations, reportez-vous à la section Utilisation des bibliothèques dans le chapitre 16 sur p. 290.
- **Création de déclarations de types**Pour plus d'informations, reportez-vous à la section Création de types dans le chapitre 17 sur p. 306.
- **Ajout de termes à la déclaration**Pour plus d'informations, reportez-vous à la section Ajout de termes dans le chapitre 17 sur p. 308.
- **Création de synonymes**Pour plus d'informations, reportez-vous à la section Définition de synonymes dans le chapitre 17 sur p. 317.

- **Import et export des modèles**Pour plus d'informations, reportez-vous à la section Import et export des modèles dans le chapitre 15 sur p. 283.
- **Publication de bibliothèques**. Pour plus d'informations, reportez-vous à la section Publication de bibliothèques dans le chapitre 16 sur p. 300.

Pour le texte en néerlandais, en anglais, en français, en allemand, en italien, en portugais, et en espagnol

Figure 14-1
Vue de l'éditeur de ressources pour les langues autres que le japonais

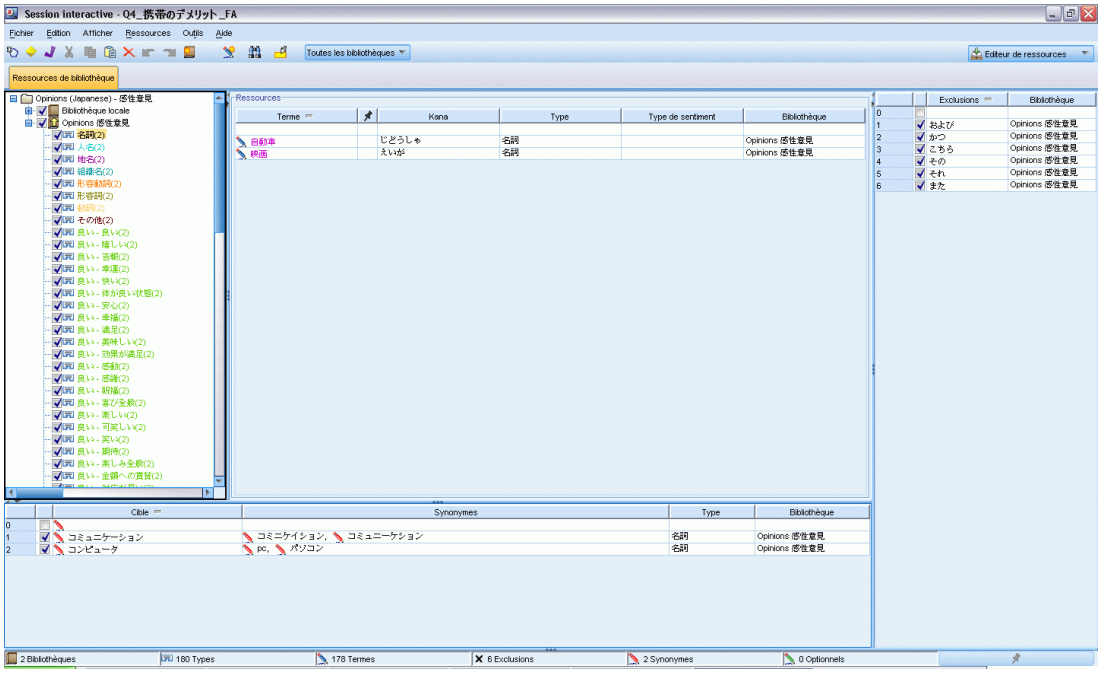


Pour les textes en japonais

L'interface de l'éditeur pour la langue japonaise est différente de celle des autres langues. Pour plus d'informations, reportez-vous à la section Modification des ressources pour du texte en japonais dans l'annexe A sur p. 374.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Figure 14-2
Vue de l'éditeur de ressources pour le texte en japonais

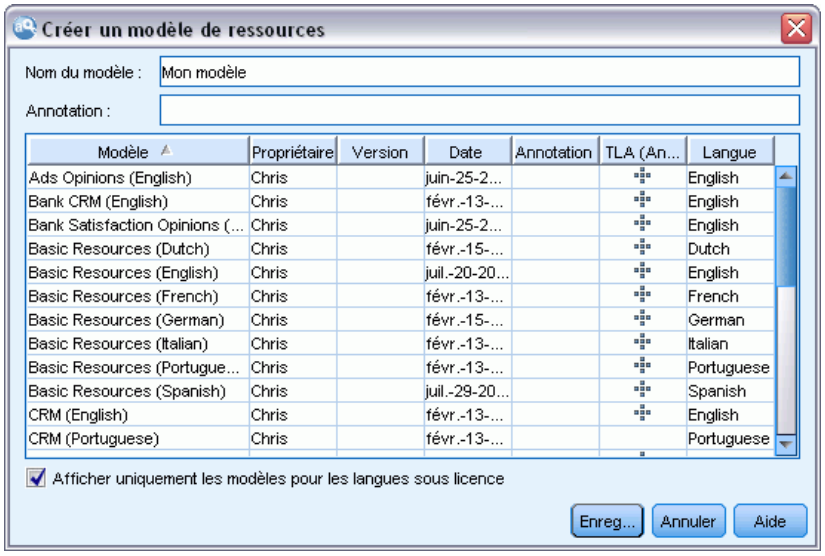


Création et mise à jour de modèles

Chaque fois que vous apportez des modifications à vos ressources et que vous souhaitez les réutiliser ultérieurement, enregistrez-les en tant que modèle. Lorsque vous effectuez cette opération, vous pouvez choisir d'enregistrer vos ressources en utilisant le nom d'un modèle existant ou en indiquant un nouveau nom. Ensuite, chaque fois que vous chargerez ce modèle, vous obtiendrez les mêmes ressources. Pour plus d'informations, reportez-vous à la section Copie des ressources à partir de modèles et de TAP dans le chapitre 3 sur p. 42.

Remarque : Vous avez également la possibilité de publier et de partager vos bibliothèques. Pour plus d'informations, reportez-vous à la section Partage de bibliothèques dans le chapitre 16 sur p. 298.

Figure 14-3
Boîte de dialogue Créer un modèle de ressource



Pour créer (ou mettre à jour) un modèle

- Dans les menus de la vue Editeur de ressources, choisissez Ressources > Créer un modèle de ressource. La boîte de dialogue Créer un modèle de ressource apparaît.
- Entrez un nouveau nom dans le champ Nom du modèle, si vous créez un modèle. Sélectionnez un modèle dans le tableau si vous souhaitez remplacer un modèle existant par les ressources chargées.
- Cliquez sur Enregistrer pour créer le modèle.

Important ! Etant donné que les modèles sont chargés lorsque vous les sélectionnez dans le nœud et non pas lorsque le flux est exécuté, veillez à recharger le modèle de ressources dans tous les autres nœuds dans lesquels il est utilisé pour obtenir les dernières modifications apportées. Pour plus d'informations, reportez-vous à la section Mise à jour des ressources d'un nœud après le chargement dans le chapitre 15 sur p. 281.

Changement des modèles de ressources

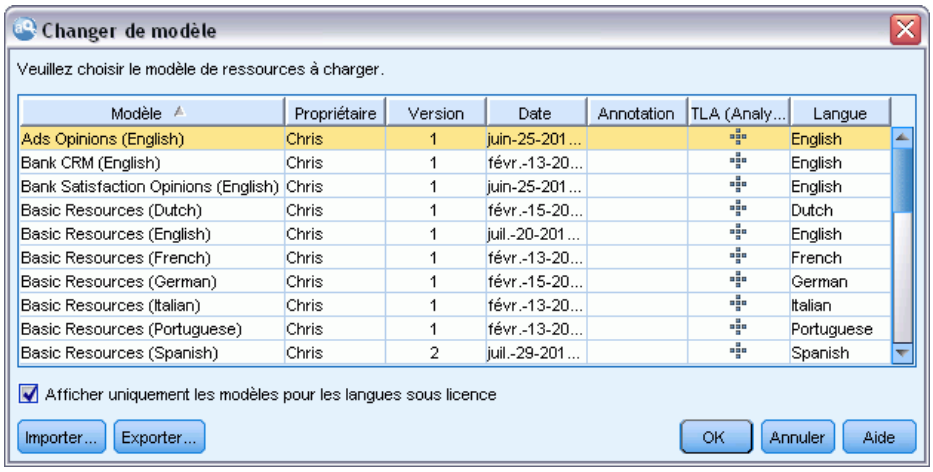
Si vous voulez remplacer les ressources chargées dans la session par une copie de celles d'un autre modèle, vous pouvez changer ces modèles. Cette action remplacera toutes les ressources chargées dans la session. Si vous remplacez des ressources pour disposer de règles de patrons Analyse des liens du texte (TLA) prédéfinies, veillez à sélectionner un modèle qui les contient dans la colonne TLA.

Important ! Vous ne pouvez pas basculer d'un modèle en japonais vers un modèle ne l'étant pas et vice-versa. *Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Remplacer des ressources est particulièrement utile lorsque vous souhaitez restaurer le travail d'une session (catégories, patrons et ressources) tout en chargeant une copie mise à jour des ressources à partir d'un modèle pour ne pas perdre le reste du travail de votre session.Vous pouvez sélectionner le modèle dont vous souhaitez copier le contenu dans l'Editeur de ressources puis cliquez sur OK. Vous remplacez ainsi les ressources contenues dans cette session. Veillez à mettre à jour le nœud de modélisation à la fin de la session si vous voulez conserver ces modifications lorsque vous lancez de nouveau la session interactive.

Remarque : Si vous utilisez le contenu d'un autre modèle au cours d'une session interactive, le nom du modèle figurant dans le nœud correspond toujours au dernier modèle chargé et copié. Pour pouvoir tirer parti de ces ressources ou du travail d'une autre session, mettez à jour le nœud de modélisation avant de quitter la session et sélectionnez l'option Utiliser le travail d'une session dans le nœud. Pour plus d'informations, reportez-vous à la section Mise à jour des nœuds de modélisation et enregistrement dans le chapitre 8 sur p. 136.

Figure 14-4
Boîte de dialogue Changer de modèle



Pour Changer de modèle

- Dans les menus de la vue Editeur de ressources, choisissez Ressources > Changer de modèles de ressource. La boîte de dialogue Changer de modèles de ressource apparaît.
- Sélectionnez le modèle à utiliser parmi ceux répertoriés dans le tableau.
- Cliquez sur OK pour abandonner les ressources chargées et charger une copie de celles contenues dans le modèle sélectionné à la place. Si vous avez apporté des modifications à vos ressources et souhaitez enregistrer vos bibliothèques pour un usage ultérieur, vous pouvez les publier, les mettre à jour et les partager avant de passer à un autre modèle. Pour plus d'informations, reportez-vous à la section Partage de bibliothèques dans le chapitre 16 sur p. 298.

Partie III:
Modèles et ressources

Chapitre
15

Modèles et ressources

IBM® SPSS® Modeler Text Analytics capture et extrait rapidement et avec précision des concepts-clés de données texte. Ce processus d'extraction repose principalement sur les ressources linguistiques afin de déterminer la façon d'extraire des informations des données textuelles. Pour plus d'informations, reportez-vous à la section Fonctionnement de l'extraction dans le chapitre 1 sur p. 7. Vous pouvez affiner ces ressources dans la vue de l'Editeur de ressources.

Lorsque vous installez le logiciel, vous obtenez aussi un ensemble de ressources spécialisées. Ces ressources fournies vous permettent de bénéficier d'années de recherche et d'un réglage pour des langues et applications spécifiques. Etant donné que les ressources fournies ne sont pas toujours parfaitement adaptées au contexte de vos données, vous pouvez modifier ces modèles de ressources, ou même créer et utiliser des bibliothèques personnalisées adaptées aux données de votre entreprise. Ces ressources sont disponibles sous diverses formes et chacune peut être utilisée dans votre session. Les ressources se trouvent aux endroits suivants :

- **Modèles de ressources**. Les modèles sont constitués d'un ensemble de bibliothèques, de types et de ressources avancées qui forment un ensemble spécialisé de ressources adapté à un domaine ou contexte particulier comme les opinions sur des produits.
- **Packages d'analyses de texte (TAP)**. En plus des ressources stockées dans un modèle, les TAP regroupent également un ou plusieurs ensembles de catégories générés à l'aide de ces ressources afin que les catégories et les ressources soient stockées ensemble et puissent être réutilisées. Pour plus d'informations, reportez-vous à la section Utilisation des packages d'analyse de texte dans le chapitre 10 sur p. 224.
- **Bibliothèques**. Les bibliothèques sont utilisées comme blocs de construction pour les TAP et les modèles. Elles peuvent aussi être ajoutées individuellement aux ressources dans votre session. Chaque bibliothèque est composée de plusieurs déclarations utilisées pour définir et gérer les types, les synonymes et les exclusions. Alors que les bibliothèques sont également fournies individuellement, elles sont prégroupées dans les modèles et les TAP. Pour plus d'informations, reportez-vous à la section Utilisation des bibliothèques dans le chapitre 16 sur p. 290.

Remarque : Au cours de l'extraction, certaines ressources internes compilées sont aussi utilisées. Ces ressources compilées contiennent de nombreuses définitions qui complètent les types de Core Library. Ces ressources compilées ne peuvent pas être éditées.

L'Editeur de ressources permet d'accéder aux ressources utilisées pour produire les résultats de l'extraction (concepts, types et patrons). Vous pouvez effectuer de nombreuses tâches dans l'Editeur de ressources, dont :

- **Utilisation de bibliothèques**Pour plus d'informations, reportez-vous à la section Utilisation des bibliothèques dans le chapitre 16 sur p. 290.
- **Création de déclarations de types**Pour plus d'informations, reportez-vous à la section Création de types dans le chapitre 17 sur p. 306.
- **Ajout de termes à la déclaration**Pour plus d'informations, reportez-vous à la section Ajout de termes dans le chapitre 17 sur p. 308.

- **Création de synonymes**Pour plus d'informations, reportez-vous à la section Définition de synonymes dans le chapitre 17 sur p. 317.
- **Mise à jour des ressources dans les TAP**. Pour plus d'informations, reportez-vous à la section Mise à jour des Packages d'analyse de texte dans le chapitre 10 sur p. 229.
- **Création de modèles**. Pour plus d'informations, reportez-vous à la section Création et mise à jour de modèles dans le chapitre 14 sur p. 269.
- **Import et export des modèles**Pour plus d'informations, reportez-vous à la section Import et export des modèles sur p. 283.
- **Publication de bibliothèques**. Pour plus d'informations, reportez-vous à la section Publication de bibliothèques dans le chapitre 16 sur p. 300.

Editeur de modèle et éditeur de ressource

Vous pouvez utiliser et modifier les modèles, les bibliothèques et les ressources principalement de deux manières. Vous pouvez travailler sur les ressources linguistiques dans l'Editeur de modèle ou l'Editeur de ressources.

Editeur de modèle

L'Editeur de modèle vous permet de créer et modifier des modèles de ressources sans session interactive et indépendamment d'un nœud ou flux spécifique. Vous pouvez utiliser cet éditeur pour créer ou modifier des modèles de ressources avant de les charger dans le nœud d'analyse des liens du texte et le nœud de modélisation Text Mining.

Editeur de modèle est accessible via la barre d'outils principale IBM® SPSS® Modeler ou le menu Outils > Editeur de modèle.

Editeur de ressources

L'Editeur de ressources qui est accessible dans une session interactive, permet d'utiliser les ressources dans le contexte d'un nœud ou d'un ensemble de données. Lorsque vous ajoutez un nœud de modélisation Text Mining à un flux, vous pouvez charger une copie du contenu du modèle de ressources ou une copie d'un package d'analyse de texte (ensembles de catégories *et* ressources) pour contrôler l'extraction de texte pour l'opération de Text Mining. Lorsque vous lancez une session interactive, vous pouvez créer des catégories, extraire des patrons d'analyse des liens du texte et créer des modèles, mais également ajuster les ressources des données de la session dans une vue de l'Editeur de ressources intégrée. Pour plus d'informations, reportez-vous à la section Modification des ressources dans l'éditeur de ressources dans le chapitre 14 sur p. 267.

Lorsque vous utilisez les ressources dans une session interactive, les modifications s'appliquent uniquement à la session. Si vous voulez enregistrer le travail (ressources, catégorie, patrons, etc.) pour continuer dans une session suivante, vous devez mettre à jour le nœud de modélisation. Pour plus d'informations, reportez-vous à la section Mise à jour des nœuds de modélisation et enregistrement dans le chapitre 8 sur p. 136.

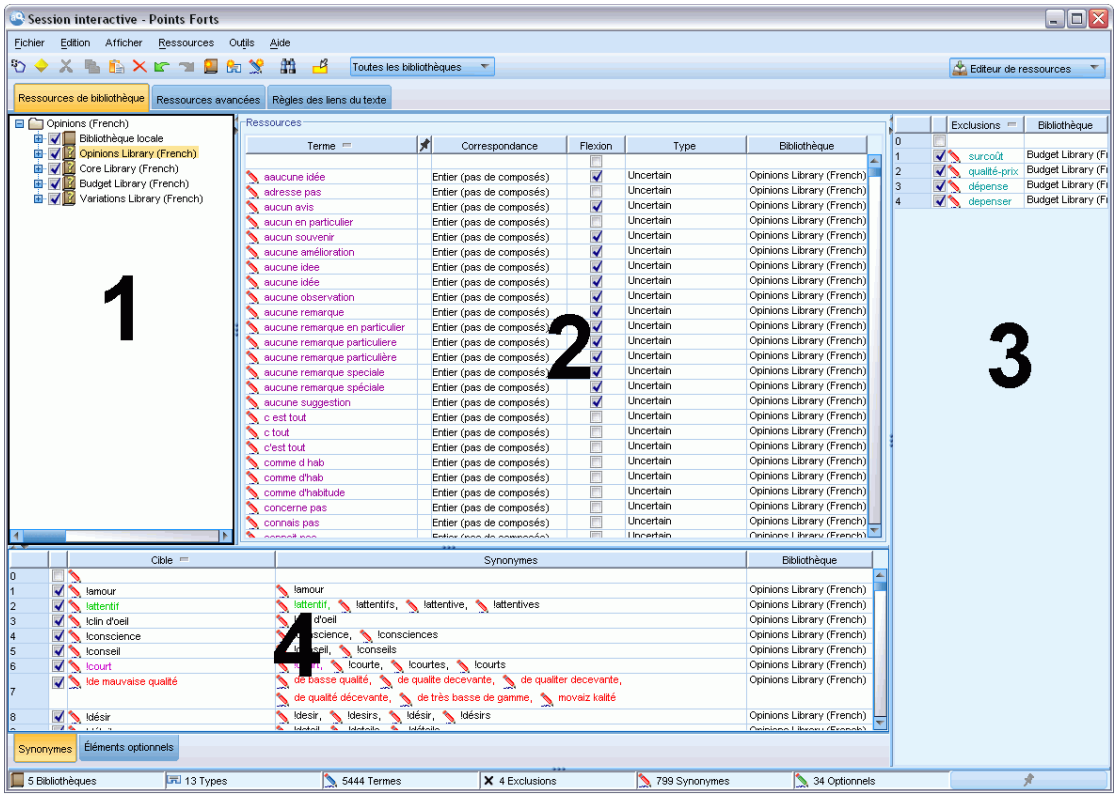
Si vous voulez enregistrer les modifications dans le modèle d'origine, dont le contenu a été copié vers le nœud de modélisation, pour pouvoir charger le modèle mis à jour dans d'autres nœuds, vous pouvez créer un modèle depuis les ressources. Pour plus d'informations, reportez-vous à la section Création et mise à jour de modèles dans le chapitre 14 sur p. 269.

Interface de l'éditeur

Les opérations que vous effectuez dans l'Editeur de modèles ou l'Editeur de ressources concernent la gestion et l'adaptation des ressources linguistiques. Ces ressources sont stockées sous la forme de modèles et de bibliothèques. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

Onglet Ressources de bibliothèque

Figure 15-1
Editeur de modèle de Text Mining



L'interface comporte quatre parties :

1. Panneau Arborescence des bibliothèques. Situé dans l'angle supérieur gauche, ce plan présente l'arborescence des bibliothèques. Vous pouvez activer et désactiver les bibliothèques figurant dans cette arborescence. Vous pouvez également filtrer les vues des autres panneaux en sélectionnant une bibliothèque dans l'arborescence. Vous pouvez effectuer plusieurs opérations dans cette arborescence à l'aide des menus contextuels. Lorsque vous développez une bibliothèque figurant

dans l'arborescence, vous pouvez consulter l'ensemble des types qu'elle comporte. Vous pouvez également filtrer cette liste avec le menu Affichage si vous souhaitez vous concentrer sur une bibliothèque spécifique.

2. Listes des termes du panneau Déclarations de types. Ce panneau, situé à droite de l'arborescence des bibliothèques, affiche les listes de termes des déclarations de types correspondant aux bibliothèques sélectionnées dans l'arborescence des bibliothèques. Une **déclaration de types** est un ensemble de termes devant être regroupés sous une même étiquette ou sous un même nom de type. Lorsque le moteur du programme d'extraction lit vos données textuelles, il compare les mots trouvés dans le texte aux termes figurant dans les déclarations de types. Si un concept extrait figure sous la forme d'un terme dans une déclaration de types, le nom du type en question lui est alors affecté. Vous pouvez considérer la déclaration de types comme étant un dictionnaire distinct qui regroupe les termes présentant des points communs. Par exemple, le type `<Location>` qui figure dans Core Library comporte des concepts tels que `new orleans`, `great britain` et `new york`. Ces termes désignent tous des lieux géographiques. Une bibliothèque peut comporter une ou plusieurs déclarations de types. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

3. Panneau Dictionnaire d'exclusions. Ce panneau, situé sur le côté droit, affiche la collection de termes qui seront exclus des résultats finaux de l'extraction. Les termes apparaissant dans ce dictionnaire d'exclusions n'apparaissent pas dans le panneau Résultats d'extraction. Les termes exclus peuvent être stockés dans la bibliothèque de votre choix. Mais, le panneau Dictionnaire d'exclusions affiche tous les termes exclus de toutes les bibliothèques visibles dans l'arborescence des bibliothèques.Pour plus d'informations, reportez-vous à la section Dictionnaires d'exclusions dans le chapitre 17 sur p. 321.

4. Panneau Dictionnaire de substitutions. Ce panneau, situé en bas à gauche, affiche les synonymes et les éléments optionnels, chacun dans leur propre onglet. Les synonymes et les élements optionnels permettent de grouper les termes similaires sous un concept principal ou cible dans les résultats d'extraction finaux. Ce dictionnaire peut comporter des synonymes connus, des synonymes et des éléments définis par l'utilisateur, ainsi que les fautes d'orthographe courantes associées à leur correction. Les définitions des synonymes et les éléments optionnels peuvent être stockés dans la bibliothèque de votre choix. Mais, le panneau Dictionnaire de substitutions affiche tous les contenus de toutes les bibliothèques visibles dans l'arborescence des bibliothèques. Pendant que ce panneau affiche tous les synonymes et éléments facultatifs de toutes les bibliothèques, les substitutions de toutes les bibliothèques de l'arborescence apparaissent ensemble dans ce panneau. Une bibliothèque ne peut comporter qu'un seul dictionnaire de substitutions. Pour plus d'informations, reportez-vous à la section Dictionnaires des substitutions/synonymes dans le chapitre 17 sur p. 315. Veuillez noter que l'onglet des éléments optionnels ne s'applique pas pour les ressources linguistiques du texte en japonais

Remarques :

- pour procéder à un filtrage visant à n'afficher que les informations propres à une seule bibliothèque, vous pouvez modifier l'affichage de la bibliothèque à l'aide de la liste déroulante figurant dans la barre d'outils. Elle comporte une entrée de niveau supérieur appelée Toutes les bibliothèques, ainsi qu'une autre entrée supplémentaire pour chacune des bibliothèques.

Pour plus d'informations, reportez-vous à la section Affichage des bibliothèques dans le chapitre 16 sur p. 294.

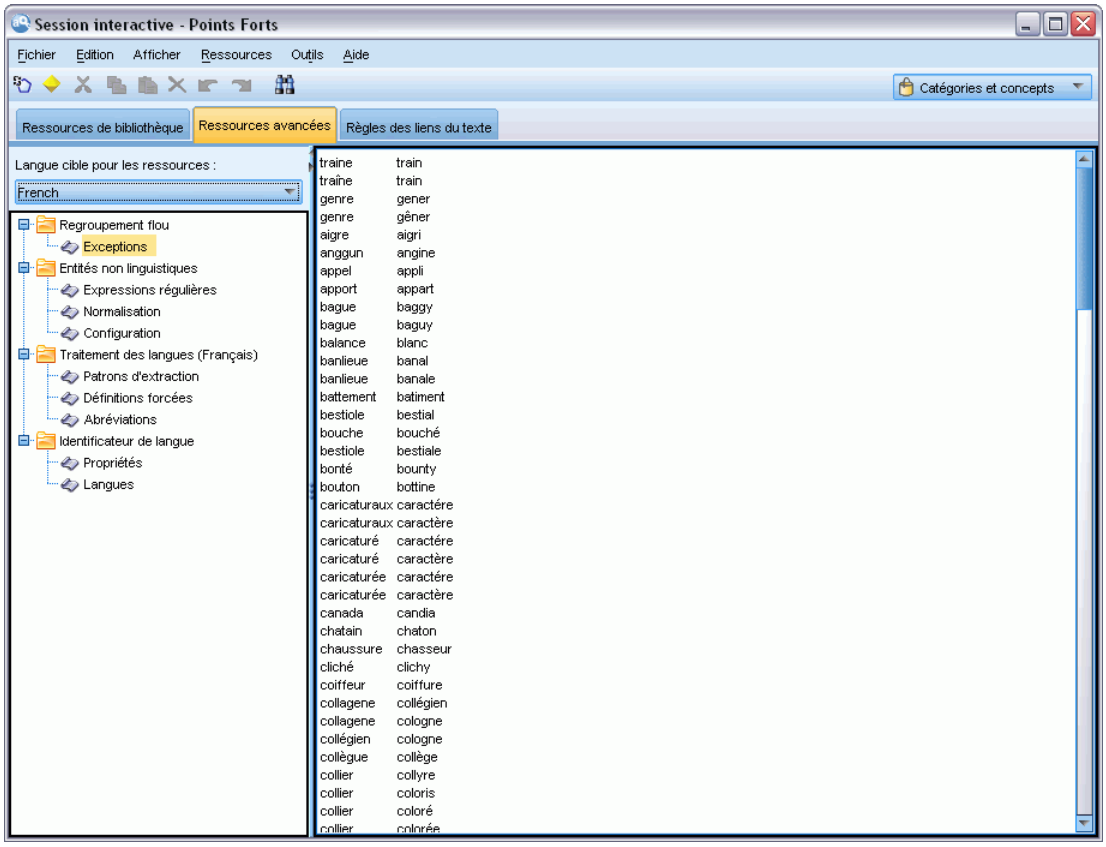
- L'interface de l'éditeur pour la langue japonaise est différente de celle des autres langues. Pour plus d'informations, reportez-vous à la section Modification des ressources pour du texte en japonais dans l'annexe A sur p. 374.*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Onglet Ressources avancées

Les ressources avancées sont disponibles dans le deuxième onglet de la vue de l'éditeur. Vous pouvez consulter et modifier les ressources avancées dans cet onglet. Pour plus d'informations, reportez-vous à la section À propos des ressources avancées dans le chapitre 18 sur p. 324.

Important ! Cet onglet n'est pas disponible pour les ressources adaptées au texte japonais.

Figure 15-2
Editeur de modèles de Text Mining - Onglet Ressources avancées



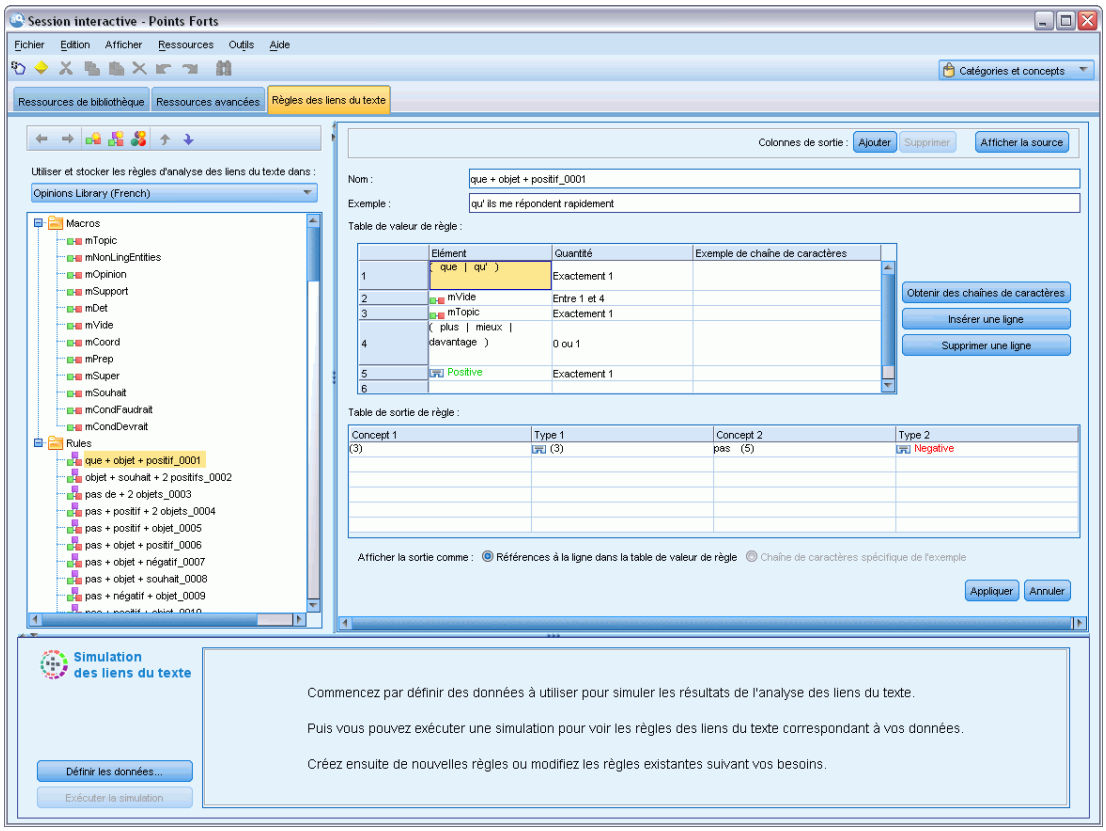
Onglet Règles des liens du texte

Depuis la version 14, les règles d'analyse des liens du texte peuvent être modifiées dans leur propre onglet de la vue de l'éditeur. Vous pouvez utiliser l'éditeur de règles, créer vos propres règles et même effectuer des simulations pour connaître l'influence de vos règles sur les résultats

TLA. Pour plus d'informations, reportez-vous à la section A propos des règles des liens du texte dans le chapitre 19 sur p. 338.

Important ! Cet onglet n'est pas disponible pour les ressources adaptées au texte japonais.

Figure 15-3
Editeur de modèles de Text Mining - Onglet Règles des liens du texte

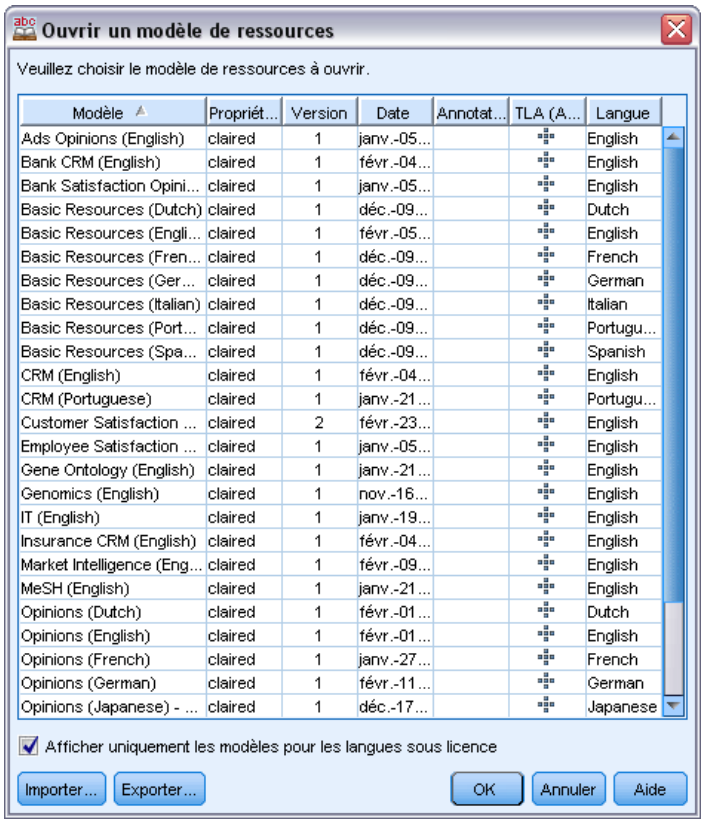


Ouverture des modèles

Lorsque vous lancez l'Editeur de modèle, un message vous demande d'ouvrir un modèle. Vous pouvez également ouvrir un modèle depuis le menu Fichier. Si vous voulez un modèle contenant des règles d'analyse des liens du texte (TLA), sélectionnez un modèle ayant une icône dans la colonne TLA. La langue pour laquelle le modèle est créé figure dans la colonne Langue.

Si vous voulez importer un modèle qui ne figure pas dans le tableau ou exporter un modèle, vous pouvez utiliser les boutons de la boîte de dialogue Ouvrir un modèle. Pour plus d'informations, reportez-vous à la section Import et export des modèles sur p. 283.

Figure 15-4
Boîte de dialogue Ouvrir un modèle de ressources



Pour ouvrir un modèle

- Dans les menus de l'Editeur de modèle, choisissez Fichier > Ouvrir un modèle de ressources. La boîte de dialogue Ouvrir un modèle de ressources apparaît.
- Sélectionnez le modèle à utiliser parmi ceux répertoriés dans le tableau.
- Cliquez sur OK pour ouvrir le modèle. Si un modèle est ouvert dans l'éditeur, cliquez sur OK pour abandonner ce modèle et afficher le modèle que vous avez sélectionné. Si vous avez apporté des modifications à vos ressources et souhaitez enregistrer vos bibliothèques pour un usage ultérieur, vous pouvez les publier, les mettre à jour et les partager avant d'ouvrir un autre modèle. Pour plus d'informations, reportez-vous à la section Partage de bibliothèques dans le chapitre 16 sur p. 298.

Enregistrement des modèles

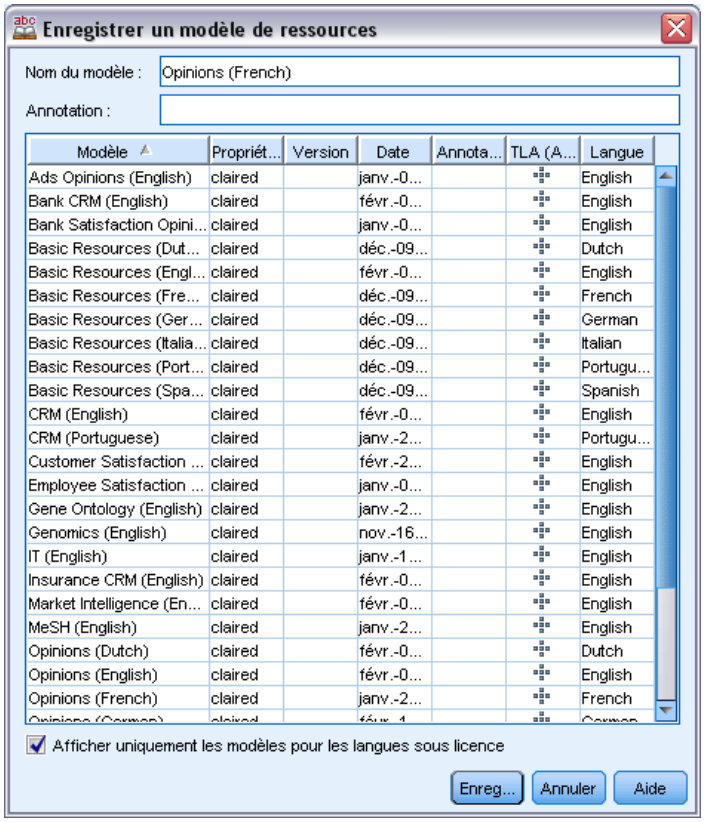
Dans l'Editeur de modèle, vous pouvez enregistrer les modifications d'un modèle. Vous pouvez choisir d'enregistrer vos ressources en utilisant le nom d'un modèle existant ou en indiquant un nouveau nom.

Si vous modifiez un modèle que vous avez chargé dans un nœud précédemment, vous devez recharger le contenu du modèle dans le nœud pour disposer des toutes dernières modifications. Pour plus d'informations, reportez-vous à la section Copie des ressources à partir de modèles et de TAP dans le chapitre 3 sur p. 42.

Ou bien, si vous utilisez l'option Utiliser les informations interactives enregistrées dans l'onglet Modèle du nœud Text Mining, ce qui implique que vous utilisez les ressources d'une session interactive précédente, vous devez utiliser les ressources de ce modèle dans la session interactive. Pour plus d'informations, reportez-vous à la section Changement des modèles de ressources dans le chapitre 14 sur p. 270.

Remarque : Vous avez également la possibilité de publier et de partager vos bibliothèques. Pour plus d'informations, reportez-vous à la section Partage de bibliothèques dans le chapitre 16 sur p. 298.

Figure 15-5
Boîte de dialogue Enregistrer un modèle de ressources



Pour enregistrer un modèle

- Dans les menus de l'Editeur de modèle, choisissez Fichier > Enregistrer un modèle de ressources. La boîte de dialogue Enregistrer un modèle de ressources s'ouvre.
- Insérez un nouveau nom dans le champ Nom du modèle, si vous souhaitez enregistrer ce modèle en tant que nouveau modèle. Sélectionnez un modèle dans le tableau si vous souhaitez remplacer un modèle existant par les ressources chargées.

- Si vous le souhaitez, entrez une description pour afficher un commentaire ou une annotation dans le tableau.
- Cliquez sur Enregistrer pour enregistrer le modèle.

Important ! Parce que les ressources des modèles ou des TAP sont chargées/copiées dans le nœud, vous devez mettre à jour ces ressources en les rechargeant si vous effectuez des modifications sur un modèle et que vous souhaitez retrouver ces modifications dans le flux existant. Pour plus d'informations, reportez-vous à la section Mise à jour des ressources d'un nœud après le chargement sur p. 281.

Mise à jour des ressources d'un nœud après le chargement

Par défaut, lorsque vous ajoutez un nœud à un flux, un ensemble de ressources d'un modèle par défaut est chargé et incorporé à votre nœud. Et si vous modifiez des modèles ou utilisez un TAP, lorsque vous les chargez, une copie de ces ressources remplace alors les ressources. Parce que les modèles et les TAP ne sont pas directement liés au nœud, les modifications effectuées sur un modèle ou un TAP ne sont pas automatiquement disponibles dans un nœud préexistant. Afin de pouvoir utiliser ces modifications, vous devez mettre à jour les ressources dans ce nœud. Les ressources peuvent être mises à jour de deux manières.

Méthode 1 : Rechargement des ressources dans l'onglet Modèle

Pour mettre à jour les ressources dans le nœud en utilisant un modèle ou un TAP nouveau ou mis à jour, vous pouvez le recharger dans l'onglet Modèle du nœud. En le rechargeant, vous remplacez la copie des ressources dans le nœud par une copie plus récente. Par souci pratique, l'heure et la date de mise à jour apparaissent dans l'onglet Modèle avec le nom du modèle d'origine. Pour plus d'informations, reportez-vous à la section Copie des ressources à partir de modèles et de TAP dans le chapitre 3 sur p. 42.

Toutefois, si vous utilisez les données d'une session interactive dans un nœud de modélisation Text Mining et avez sélectionné l'optionUtiliser le travail d'une session dans l'onglet Modèle, le travail et les ressources de la session sauvegardée sont utilisés et le bouton Charger est désactivé. Il est désactivé, car au cours d'une session interactive, vous avez choisi l'option Mettre à jour le nœud de modélisation et vous avez conservé les catégories, les ressources et un autre travail de session. Dans ce cas, si vous voulez utiliser ou mettre à jour ces ressources, vous pouvez essayer d'utiliser la méthode suivante de changement des ressources dans l'Editeur de ressources.

Méthode 2 : Changement des ressources dans l'Editeur de ressources

Lorsque vous voulez utiliser des ressources différentes au cours d'une session interactive, vous pouvez changer ces ressources en utilisant la boîte de dialogue Changer de modèle. Cela est particulièrement pratique si vous voulez réutiliser un travail de catégorie existant et remplacer les ressources. Dans ce cas, vous pouvez sélectionner l'option Utiliser le travail d'une session dans l'onglet Modèle d'un nœud de modélisation Text Mining. Cette action désactive l'option de rechargement d'un modèle en utilisant la boîte de dialogue du nœud et conserve les paramètres et modifications effectuées pendant votre session. Vous pouvez alors lancer la session interactive en exécutant le flux et en remplaçant les ressources dans l'Editeur de ressources. Pour plus

d'informations, reportez-vous à la section Changement des modèles de ressources dans le chapitre 14 sur p. 270.

Pour pouvoir conserver le travail d'une session dans les sessions suivantes, y compris les ressources, vous devez mettre à jour le nœud de modélisation sans la session interactive pour que les ressources (et les autres données) soient enregistrées dans le nœud. Pour plus d'informations, reportez-vous à la section Mise à jour des nœuds de modélisation et enregistrement dans le chapitre 8 sur p. 136.

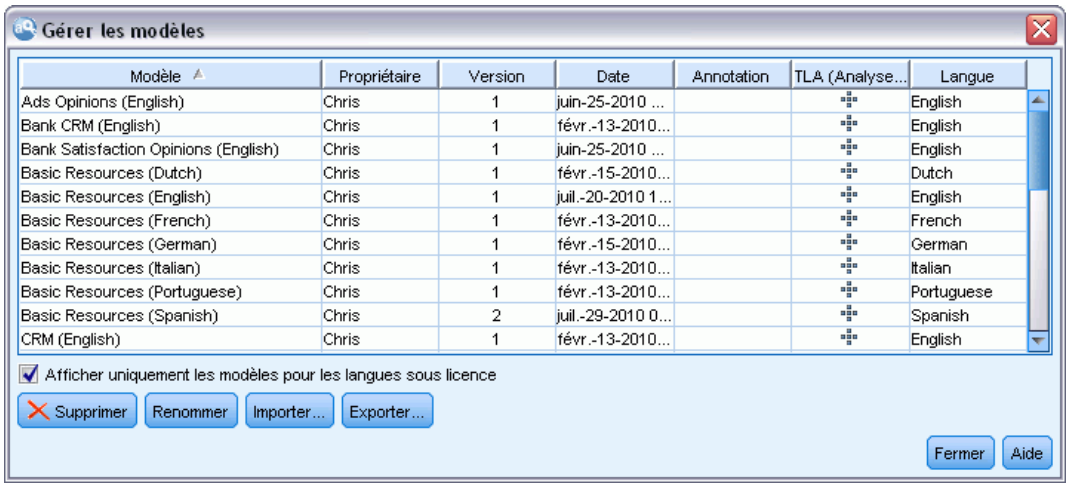
Remarque : Si vous utilisez le contenu d'un autre modèle au cours d'une session interactive, le nom du modèle figurant dans le nœud correspond toujours au dernier modèle chargé et copié. Pour pouvoir tirer parti de ces ressources ou du travail d'une autre session, mettez à jour le nœud de modélisation avant de quitter la session.

Gestion des modèles

Il existe également certaines tâches de gestion de base que vous pouvez effectuer occasionnellement sur vos modèles telles que l'attribution de nouveau nom, l'import et l'export, ou la suppression des modèles obsolètes. Ces tâches sont réalisées dans la boîte de dialogue Gérer les modèles. L'import et l'export des modèles permettent de les partager avec d'autres utilisateurs. Pour plus d'informations, reportez-vous à la section Import et export des modèles sur p. 283.

Remarque : Vous ne pouvez pas renommer ou supprimer les modèles installés (ou fournis) avec ce produit. Si vous voulez renommer, vous pouvez ouvrir le modèle installé et en créer un nouveau avec le nom de votre choix. Vous pouvez supprimer vos modèles personnalisés ; néanmoins, si vous essayez de supprimer un modèle fourni, il sera réinitialisé à la version installée à l'origine.

Figure 15-6
Boîte de dialogue Gérer les modèles



Pour renommer un modèle

- Dans les menus, sélectionnez Ressources > Gérer les modèles de ressources. La boîte de dialogue Gérer les modèles apparaît.

- Sélectionnez le modèle à renommer et cliquez sur Renommer. La zone de nom devient un champ modifiable dans le tableau.
- Tapez un nouveau nom et appuyez sur la touche Entrée. Un message de confirmation apparaît.
- Si le nouveau nom vous satisfait, cliquez sur Oui. Sinon, cliquez sur Non.

Pour supprimer un modèle

- Dans les menus, sélectionnez Ressources > Gérer les modèles de ressources. La boîte de dialogue Gérer les modèles apparaît.
- Dans la boîte de dialogue Gérer les modèles, sélectionnez le modèle à supprimer.
- Choisissez Supprimer. Un message de confirmation apparaît.
- Cliquez sur Oui pour supprimer ou sur Non pour annuler la demande de suppression. Si vous cliquez sur Oui, le modèle est supprimé.

Import et export des modèles

Vous pouvez partager les modèles avec d'autres utilisateurs ou ordinateurs en les important et en les exportant. Les modèles sont stockés dans une base de données interne, mais ils ne peuvent pas être exportés sous forme de fichiers **.lrt* sur votre disque dur.

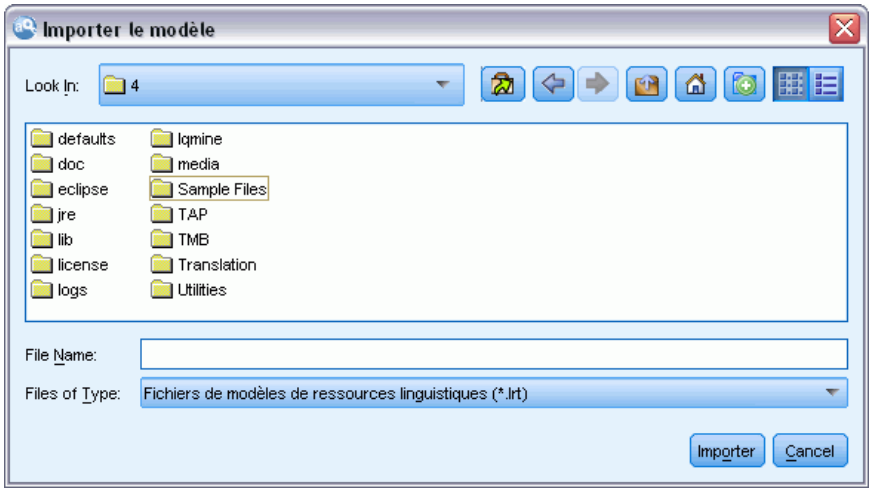
Etant donné que dans certains cas, vous voulez importer et exporter des modèles, il existe des boîtes de dialogue qui offrent ces fonctions.

- Ouvrez la boîte de dialogue dans l'Editeur de modèle
- Boîte de dialogue Charger des ressources dans le nœud de modélisation Text Mining et le nœud Analyse des liens du texte.
- Boîte de dialogue Gérer les modèles dans l'Editeur de modèle et l'Editeur de ressources.

Pour importer un modèle

- Dans la boîte de dialogue, cliquez sur Importer. La boîte de dialogue Importer un modèle apparaît.

Figure 15-7
Boîte de dialogue Importer un modèle

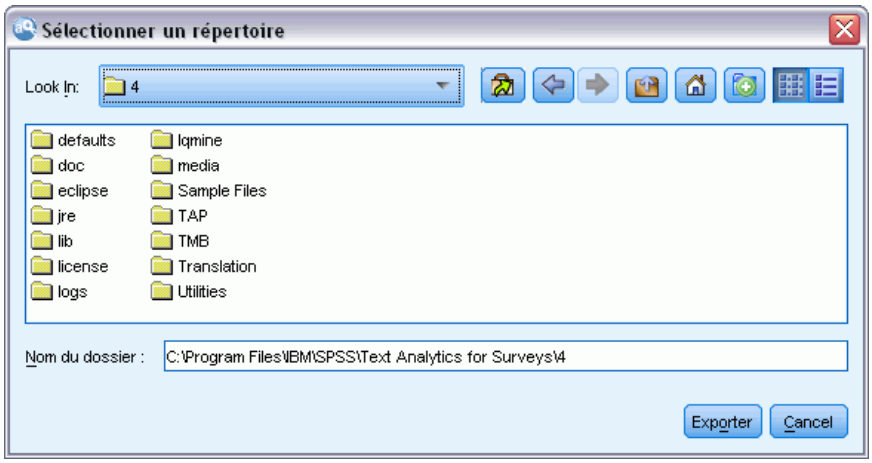


- Sélectionnez le fichier du modèle de ressources (*.*lrt*) à importer et cliquez sur Importer. Enregistrez le modèle importé en lui attribuant un nouveau nom ou en remplaçant le modèle existant. La boîte de dialogue se ferme et le modèle apparaît dans le tableau.

Pour exporter un modèle

- Dans la boîte de dialogue, sélectionnez le modèle à exporter et cliquez sur Exporter. La boîte de dialogue Sélectionner un répertoire apparaît.

Figure 15-8
Boîte de dialogue Sélectionner un répertoire



- Sélectionnez le répertoire vers lequel vous souhaitez exporter et cliquez sur Exporter. Cette boîte de dialogue se ferme et le modèle est exporté et porte l'extension de fichier (*.*lrt*).

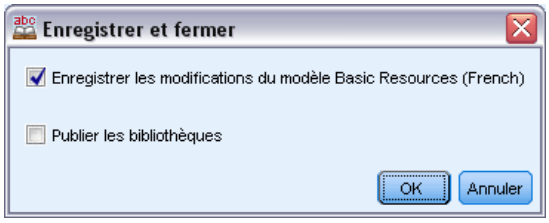
Sortie de l'Editeur de modèle

Lorsque vous avez terminé de travailler dans l'Editeur de modèle, vous pouvez enregistrer votre travail et quitter l'éditeur.

Pour quitter l'Editeur de modèle

▶ Dans les menus, sélectionnez Fichier > Fermer. La boîte de dialogue Enregistrer et fermer s'ouvre.

Figure 15-9
Boîte de dialogue Enregistrer et fermer



▶ Sélectionnez Enregistrer les modifications dans un modèle pour enregistrer le modèle ouvert avant de fermer l'éditeur.

▶ Sélectionnez Publier les bibliothèques pour publier les bibliothèques dans le modèle ouvert avant de fermer l'éditeur. Si vous sélectionnez cette option, un message vous demande de sélectionner les bibliothèques à publier. Pour plus d'informations, reportez-vous à la section Publication de bibliothèques dans le chapitre 16 sur p. 300.

Sauvegarde des ressources

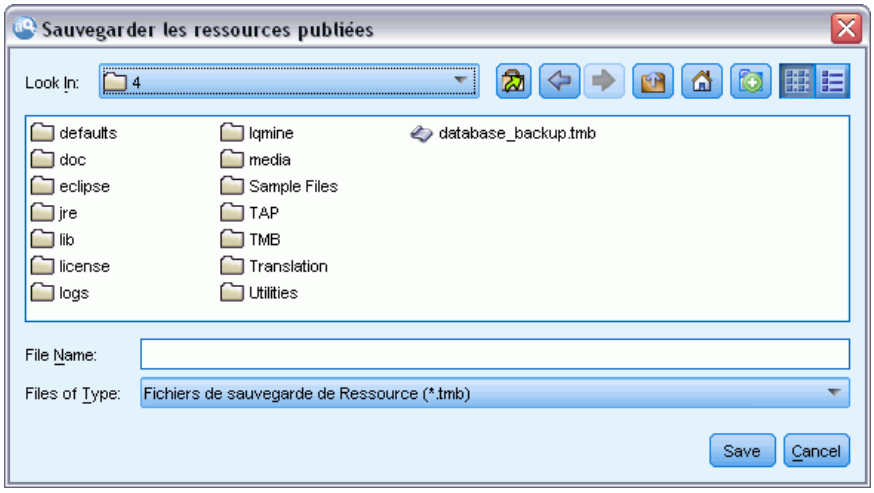
Il se peut que vous ayez besoin, de temps à autre, de sauvegarder vos ressources par mesure de sécurité.

Important ! Lorsque vous procédez à une restauration, le contenu intégral de vos ressources est supprimé et seul le contenu du fichier de sauvegarde est accessible dans le produit. Ceci inclut tout travail en cours.

Pour sauvegarder les ressources

▶ Dans les menus, sélectionnez Ressources > Outils de sauvegarde > Sauvegarder les ressources. La boîte de dialogue Sauvegarder apparaît.

Figure 15-10
Boîte de dialogue Sauvegarder les ressources

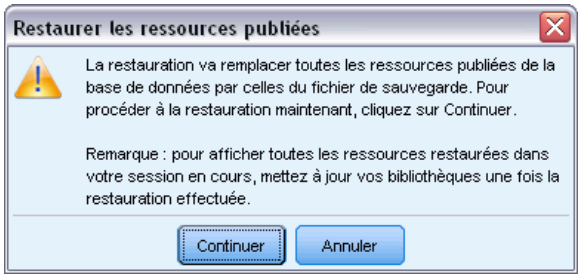


▶ Nommez votre fichier de sauvegarde et cliquez sur Enregistrer. La boîte de dialogue se ferme et le fichier de sauvegarde est créé.

Pour restaurer les ressources

▶ Dans les menus, sélectionnez Ressources > Outils de sauvegarde > Restaurer les ressources. Une alerte vous avertit que la restauration va écraser le contenu actuel de votre base de données.

Figure 15-11
Message d'alerte d'écrasement



▶ Cliquez sur Oui pour continuer. La boîte de dialogue apparaît.

Figure 15-12
Boîte de dialogue Restaurer les ressources

▶ Sélectionnez le fichier de sauvegarde à restaurer et cliquez sur Ouvrir. La boîte de dialogue se ferme et les ressources sont restaurées dans l'application.

Important ! Lorsque vous procédez à une restauration, le contenu intégral de vos ressources est supprimé et seul le contenu du fichier de sauvegarde est accessible dans le produit. Ceci inclut tout travail en cours.

Import des fichiers de ressources

Si vous avez effectué des modifications directement dans les fichiers de ressources en-dehors de ce produit, vous pouvez les importer dans une bibliothèque donnée, en sélectionnant cette bibliothèque et en procédant à l'importation. Lorsque vous importez un répertoire, vous pouvez également importer l'ensemble des fichiers pris en charge dans une bibliothèque ouverte spécifique. Vous ne pouvez importer que les fichiers **.txt*.

Important ! Pour le texte en japonais, les fichiers .txt à importer doivent être codés en UTF8. En outre, vous ne pouvez pas importer de listes d'exclusion pour le japonais. *Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

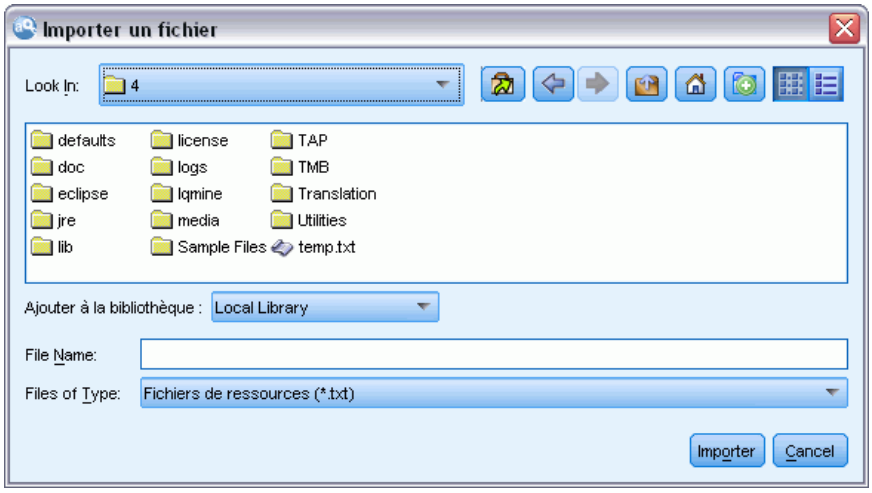
Chaque fichier importé doit contenir seulement une entrée par fichier, et si le contenu est structuré comme :

- Une liste de mots ou d'expressions (une par ligne). Le fichier est importé comme liste de termes d'une déclaration de types, où la déclaration de types prend le nom du fichier sans l'extension. Pour le texte en japonais, le nom du fichier doit correspondre à un type japonais connu. Si le nom du fichier correspond au nom de type d'un sentiment plutôt qu'à un type japonais de base, les termes apparaissant dans le fichier seront attribués au type de sentiment et au type de base par défaut 名詞.
- Une liste d'entrées telle que *terme1*<TAB>*terme2*, est ensuite importée comme liste de synonymes, avec *terme1* comme ensemble des termes sous-jacents et *terme2* comme terme cible. Pour le texte en japonais, la valeur par défaut 名詞 est affectée au terme cible.

Pour importer un unique fichier de ressources

▶ Dans les menus, choisissez Ressources > Importer des fichiers > Importer un fichier. La boîte de dialogue Importer un fichier apparaît.

Figure 15-13
Boîte de dialogue Importer un fichier

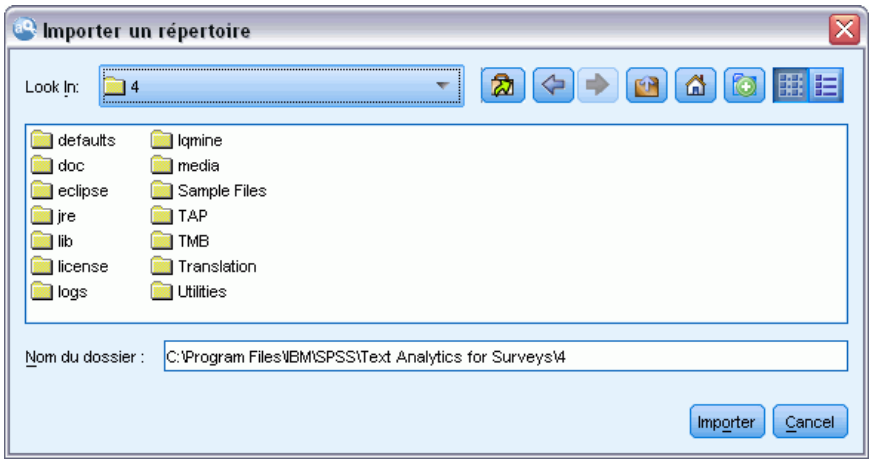


▶ Sélectionnez le fichier que vous souhaitez importer et cliquez sur Importer. Un format interne est appliqué au contenu du fichier qui est ensuite ajouté à votre bibliothèque.

Pour importer l'ensemble des fichiers d'un répertoire

▶ Dans les menus, choisissez Ressources > Importer des fichiers > Importer un répertoire entier. La boîte de dialogue Importer un répertoire apparaît.

Figure 15-14
Boîte de dialogue Importer un répertoire



▶ Dans la liste Importer, sélectionnez la bibliothèque dans laquelle vous souhaitez importer l'ensemble des fichiers de ressources. Si vous sélectionnez l'option Par défaut, une bibliothèque est créée. Elle porte le nom du répertoire.

- Sélectionnez le répertoire à partir duquel les fichiers doivent être importés. Les sous-répertoires ne seront pas lus.
- Cliquez sur Importer. La boîte de dialogue se ferme et le contenu des fichiers de ressources importés apparaît dans l'éditeur sous forme de dictionnaires et de fichiers de ressources avancés.

Chapitre
16

Utilisation des bibliothèques

Les ressources utilisées par le moteur d'extraction pour extraire et regrouper les termes des données textuelles contiennent une ou plusieurs bibliothèques. Vous pouvez visualiser l'ensemble des bibliothèques dans l'arborescence des bibliothèques située dans la partie supérieure gauche de l'Editeur de modèle et de l'Editeur de ressources. Les bibliothèques comportent trois types de dictionnaires : Types, Substitutions et Exclusions. Pour plus d'informations, reportez-vous à la section À propos des dictionnaires de bibliothèque dans le chapitre 17 sur p. 304.

Le modèle de ressources ou les ressources du TAP choisi comprennent plusieurs bibliothèques afin de vous permettre de procéder immédiatement à l'extraction des concepts de vos données textuelles. Cependant, vous pouvez également créer vos propres bibliothèques et les publier pour pouvoir les réutiliser. Pour plus d'informations, reportez-vous à la section Publication de bibliothèques sur p. 300.

Supposons, par exemple, que vous utilisez fréquemment des données textuelles relatives au secteur automobile. Après avoir analysé vos données, vous décidez de créer des ressources personnalisées pour gérer les termes propres à ce secteur. A l'aide d'Editeur de modèle, vous pouvez créer un nouveau modèle et, dans ce modèle, une bibliothèque permettant d'extraire et de regrouper les termes correspondant au secteur automobile. Etant donné que vous aurez à nouveau besoin de ces informations dans cette bibliothèque, publiez votre bibliothèque dans un référentiel central, accessible à partir de la boîte de dialogue **Gérer les bibliothèques**, afin qu'elle puisse être réutilisée indépendamment dans différentes sessions de flux.

Supposons que vous souhaitez également regrouper les termes propres à différents sous-secteurs, tels que les dispositifs électroniques, les moteurs, les systèmes de refroidissement, voire un fabricant ou un marché particulier. Vous pouvez créer une bibliothèque pour chaque groupe, puis publier ces bibliothèques de manière à ce qu'elles puissent être utilisées avec plusieurs ensembles de données textuelles. De cette manière, vous pouvez ajouter les bibliothèques qui correspondent le mieux au contenu de vos données textuelles.

Remarque : les ressources supplémentaires peuvent être configurées et gérées dans l'onglet Ressources avancées. Certaines s'appliquent à toutes les bibliothèques et gèrent les entités non linguistiques, les exceptions de regroupement flou, etc. De plus, vous pouvez également modifier les règles de patrons d'analyse des liens du texte, qui sont spécifiques à la bibliothèque, dans l'onglet Règles des liens du texte. Pour plus d'informations, reportez-vous à la section À propos des ressources avancées dans le chapitre 18 sur p. 324.

Bibliothèques fournies

Par défaut, plusieurs bibliothèques sont installées avec IBM® SPSS® Modeler Text Analytics . Vous pouvez utiliser ces bibliothèques préformatées pour accéder à des milliers de termes et synonymes prédéfinis, ainsi qu'à plusieurs types différents. Ces bibliothèques fournies sont affinées pour être utilisées dans plusieurs domaines différents et sont disponibles dans plusieurs langues.

Remarque : Pour des informations spécifiques sur les ressources de texte japonais et les exceptions, consultez Exceptions pour le texte en japonais sur p. 368*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

De nombreuses bibliothèques existent mais les plus couramment utilisées sont les suivantes :

- **Bibliothèque locale.** Utilisée pour stocker les dictionnaires définis par l'utilisateur. Il s'agit d'une bibliothèque vide ajoutée par défaut à toutes les ressources. Elle contient également une déclaration de types vide. Elle est particulièrement utile pour effectuer des modifications ou des améliorations aux ressources directement (comme ajouter un mot à un type) depuis la vue Catégories et concepts, la vue Clusters et la vue Analyse des liens du texte . Dans ce cas, ces modifications et ces améliorations sont automatiquement stockées dans la première bibliothèque répertoriée dans l'arborescence de bibliothèques d' Editeur de ressources ; par défaut, il s'agit de la *bibliothèque locale*. Vous ne pouvez pas publier cette bibliothèque car elle est spécifique aux données du projet de . Pour publier son contenu, renommez tout d'abord la bibliothèque.
- **Core library.** Utilisée dans la plupart des cas, puisqu'elle comprend les cinq types de base intégrés représentant les personnes, les lieux, les organisations, les produits et les éléments inconnus. Bien que seuls quelques termes soient répertoriés dans l'une des déclarations de types, les types représentés dans Core library viennent en fait compléter les types fiables détectés dans les ressources internes, compilées et livrées avec votre produit de Text Mining. Ces ressources internes et compilées contiennent des milliers de termes pour chaque type. Pour cette raison, même si vous ne voyez pas un terme dans une liste de termes des déclarations de types, il peut quand même être extrait et saisi avec un type Core. Cela explique la façon dont les noms comme *George* peuvent être extraits et recevoir le type `<Person>`, alors que seul *John*> apparaît dans la déclaration de types `<Person>` de Core library. De même, si vous n'incluez pas Core Library, il se peut que vous aperceviez toujours ces types dans vos résultats d'extraction puisque les ressources compilées qui les contiennent sont toujours utilisées par le moteur du programme d'extraction.
- **Opinions Library.** Utilisé le plus souvent pour extraire des opinions et des sentiments des données textuelles. Cette bibliothèque comprend des milliers de mots qui représentent des attitudes, des qualificatifs et des préférences et qui — lorsqu'ils sont utilisés avec d'autres termes — indiquent une opinion sur un sujet. Cette bibliothèque contient de nombreux types intégrés, des synonymes et des exclusions. Elle comprend également un grand nombre de règles de patrons utilisées pour l'analyse des liens du texte. Pour pouvoir utiliser les règles d'analyse des liens du texte dans cette bibliothèque et les résultats de patrons qu'elles génèrent, cette bibliothèque doit être spécifiée dans l'onglet Règles des liens du texte. Pour plus d'informations, reportez-vous à la section A propos des règles des liens du texte dans le chapitre 19 sur p. 338.
- **Budget Library.** Utilisée pour extraire les termes faisant référence au coût d'un objet ou d'un service. La bibliothèque comprend plusieurs mots et expressions représentant des adjectifs, des qualificatifs et des jugements concernant le prix ou la qualité d'un objet ou d'un service.
- **Variations Library.** Utilisée pour inclure les observations dans lesquelles certaines variations de langue requièrent des définitions de synonyme pour être correctement regroupées. Cette bibliothèque ne comporte que des définitions de synonyme.

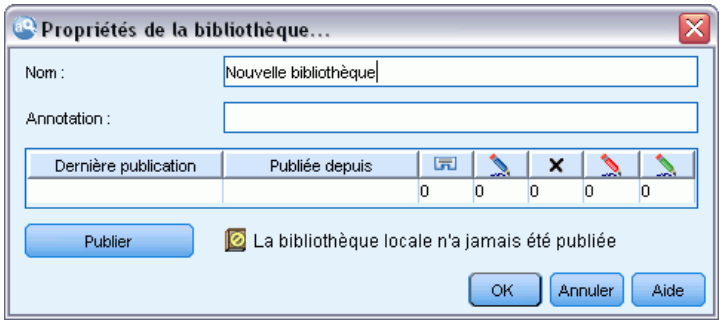
Bien que certaines bibliothèques fournies en dehors des modèles ressemblent au contenu de certains modèles, les modèles ont été spécifiquement adaptés à certaines applications et contiennent des ressources avancées supplémentaires. Nous vous recommandons d'essayer d'utiliser un modèle conçu pour le genre de données textuelles que vous utilisez et d'effectuer les modifications sur ces ressources plutôt que de simplement ajouter des bibliothèques individuelles à un modèle plus générique.

Les ressources compilées sont également fournies avec SPSS Modeler Text Analytics . Elles sont systématiquement utilisées au cours du processus d'extraction et contiennent un grand nombre de définitions complémentaires dans les déclarations de types intégrées aux bibliothèques par défaut. Etant donné que ces ressources sont compilées, il est impossible de les visualiser ou de les modifier. Vous pouvez toutefois forcer n'importe quel autre dictionnaire à accepter un terme classé par type par les ressources compilées. Pour plus d'informations, reportez-vous à la section Ajout des termes forcés dans le chapitre 17 sur p. 312.

Création de bibliothèques

Vous pouvez créer autant de bibliothèques que vous le souhaitez. Après avoir créé une nouvelle bibliothèque, vous pouvez commencer à y créer des déclarations de types et insérer des termes, des synonymes et des exclusions.

Figure 16-1
Boîte de dialogue Propriétés de bibliothèque



Pour créer une bibliothèque

- A partir des menus, sélectionnez Ressources > Nouvelle bibliothèque. La boîte de dialogue Propriétés de bibliothèque s'ouvre.
- Nommez la bibliothèque dans le champ Nom.
- Si vous le souhaitez, insérez un commentaire dans le champ Annotation.
- Cliquez sur Publier si vous souhaitez publier cette bibliothèque maintenant, sans y insérer quoi que ce soit. Pour plus d'informations, reportez-vous à la section Partage de bibliothèques sur p. 298. Vous pouvez également la publier plus tard à n'importe quel moment.
- Cliquez sur OK pour créer la bibliothèque. La boîte de dialogue se ferme et la bibliothèque apparaît dans l'arborescence. Si vous développez la bibliothèque dans l'arborescence, vous vous apercevez qu'une déclaration de types vide y a été incluse automatiquement. Vous pouvez

ajouter des termes immédiatement. Pour plus d'informations, reportez-vous à la section Ajout de termes dans le chapitre 17 sur p. 308.

Ajout de bibliothèques publiques

Si vous souhaitez réutiliser une bibliothèque contenant les données d'une autre session, vous pouvez l'ajouter à vos ressources en cours à condition qu'il s'agisse d'une bibliothèque publique. Une **bibliothèque publique** est une bibliothèque qui a été publiée. Pour plus d'informations, reportez-vous à la section Publication de bibliothèques sur p. 300.

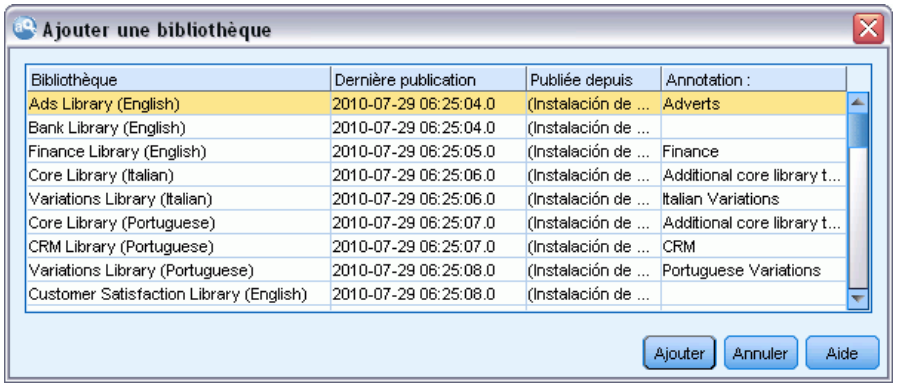
Important ! Vous ne pouvez pas ajouter de bibliothèque japonaise à des ressources non japonaises et vice versa. *Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Lorsque vous ajoutez une bibliothèque publique, une copie **locale** est incorporée à vos données de session. Il est possible d'apporter des modifications à cette bibliothèque, cependant, vous êtes tenu de publier à nouveau sa version publique si vous souhaitez partager les modifications effectuées.

Lorsque vous ajoutez une bibliothèque publique, une boîte de dialogue Résoudre les conflits peut apparaître s'il existe des conflits entre les termes et les types d'une bibliothèque et les autres bibliothèques locales. Il est nécessaire de résoudre ces conflits ou d'accepter les résolutions proposées afin de pouvoir réaliser cette opération. Pour plus d'informations, reportez-vous à la section Résolution des conflits sur p. 302.

Remarque : Si vous mettez systématiquement à jour vos bibliothèques lorsque vous lancez une session interactive ou si vous publiez une bibliothèque lorsque vous la fermez , vous avez moins de risques d'avoir des bibliothèques qui ne sont pas synchronisées. Pour plus d'informations, reportez-vous à la section Partage de bibliothèques sur p. 298.

Figure 16-2
Boîte de dialogue Ajouter une bibliothèque



Pour ajouter une bibliothèque

- A partir des menus, sélectionnez Ressources > Ajouter une bibliothèque. La boîte de dialogue Ajouter une bibliothèque apparaît.
- Sélectionnez la ou les bibliothèques dans la liste.

- Cliquez sur Ajouter. En cas de conflits entre les bibliothèques qui viennent d'être ajoutées et les anciennes, vous êtes invité à vérifier les résolutions de conflit ou à les modifier avant de réaliser l'opération. Pour plus d'informations, reportez-vous à la section Résolution des conflits sur p. 302.

Recherche de termes et de types

Vous pouvez rechercher dans les nombreux panneaux de l'éditeur à l'aide de la fonction Rechercher. Dans l'éditeur, sélectionnez Edition > Rechercher dans les menus. La barre d'outils Rechercher apparaît. Utilisez cette barre d'outils pour rechercher une occurrence à la fois. En cliquant à nouveau sur Rechercher, vous pouvez rechercher les occurrences suivantes du terme recherché.

Lors de la recherche, l'éditeur effectue la recherche uniquement dans les bibliothèques répertoriées dans la liste déroulante de la barre d'outils Rechercher. Si l'option Toutes les bibliothèques est sélectionnée, le programme effectue la recherche dans tous les supports dans l'éditeur.

Lorsque vous démarrez une recherche, elle débute dans la zone active. La recherche continue dans chaque section, formant une boucle jusqu'à revenir à la cellule active. Il est possible d'inverser l'ordre de la recherche en utilisant les flèches de direction. Vous pouvez également choisir si votre recherche est sensible ou non à la casse.

Pour rechercher des chaînes dans la vue

- Dans les menus, sélectionnez Edition > Rechercher. La barre d'outils Rechercher apparaît.
- Insérez la chaîne de caractères que vous souhaitez rechercher.
- Cliquez sur le bouton Rechercher pour commencer la recherche. La prochaine occurrence du terme ou du type est ensuite mise en surbrillance.
- Cliquez à nouveau sur le bouton pour passer d'une occurrence à une autre.

Affichage des bibliothèques

Vous pouvez afficher le contenu d'une bibliothèque donnée ou celui de toutes les bibliothèques. Cela peut s'avérer utile lorsque vous travaillez avec plusieurs bibliothèques ou lorsque vous souhaitez passer en revue le contenu d'une bibliothèque spécifique avant sa publication. Le fait de modifier l'affichage n'a un impact que sur ce que vous voyez dans l'onglet Ressources de bibliothèque mais cette modification n'empêche pas l'utilisation des bibliothèques au cours de l'extraction. Pour plus d'informations, reportez-vous à la section Désactivation des bibliothèques locales sur p. 296.

L'affichage par défaut est Toutes les bibliothèques. Cette option affiche toutes les bibliothèques de l'arborescence et leur contenu dans d'autres panneaux. Vous pouvez modifier cette sélection en utilisant la liste déroulante de la barre d'outils ou via une sélection de menus (Affichage > Bibliothèques). Si une seule bibliothèque est affichée, tous les éléments des autres bibliothèques disparaissent de la vue mais continuent d'être lus pendant l'extraction.

Pour afficher la vue Bibliothèque

- Dans les menus de l'onglet Ressources de bibliothèque, choisissez Vue > Bibliothèques. Un menu comprenant toutes les bibliothèques locales apparaît.
- Sélectionnez la bibliothèque que vous souhaitez afficher ou sélectionnez l'option Toutes les bibliothèques pour afficher le contenu de toutes les bibliothèques. Le contenu de la vue est filtré en fonction de votre sélection.

Gestion des bibliothèques locales

Les bibliothèques locales sont les bibliothèques qui se trouvent dans votre projet de session interactiveou dans un modèle, contrairement aux bibliothèques publiques. Pour plus d'informations, reportez-vous à la section Gestion des bibliothèques publiques sur p. 296. Vous souhaiterez peut-être exécuter certaines tâches de gestion de bibliothèques locales, notamment : renommer, désactiver ou supprimer une bibliothèque locale.

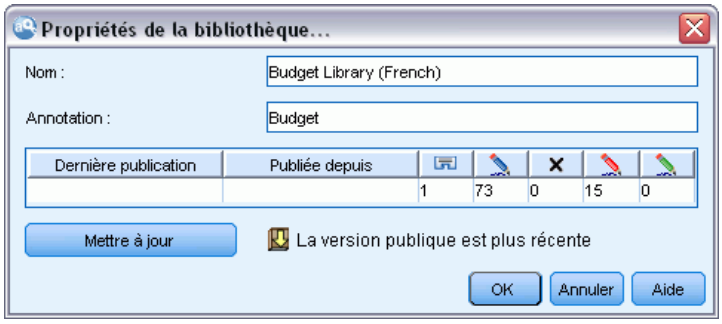
Attribution d'un nouveau nom à une bibliothèque locale

Vous pouvez renommer les bibliothèques locales. Lorsque vous renommez une bibliothèque locale, vous la dissociez de la version publique éventuelle. Cela signifie que les modifications suivantes ne peuvent plus être partagées avec la version publique. Vous pouvez republier cette bibliothèque locale sous son nouveau nom. Cela signifie également que vous ne pouvez pas incorporer les modifications apportées à cette version locale dans la version publique originale.

Remarque : Il est impossible de renommer une bibliothèque publique.

- Dans les menus, sélectionnez Edition> Propriétés de bibliothèque. La boîte de dialogue Propriétés de bibliothèque apparaît.

Figure 16-3
Boîte de dialogue Propriétés de bibliothèque



Pour renommer une bibliothèque locale

- Dans l'arborescence, sélectionnez la bibliothèque à renommer.
- Attribuez un nouveau nom à la bibliothèque dans le champ Nom.

- Cliquez sur OK pour accepter le nouveau nom de la bibliothèque. La boîte de dialogue se ferme et son nom apparaît dans l'arborescence.

Désactivation des bibliothèques locales

Si vous souhaitez exclure temporairement une bibliothèque du processus d'extraction, désélectionnez la case à cocher à gauche du nom de la bibliothèque dans l'arborescence. Vous indiquez ainsi que vous souhaitez conserver la bibliothèque, mais en ignorer le contenu pendant la recherche de conflits et le processus d'extraction.

Pour désactiver une bibliothèque

- Dans l'arborescence de bibliothèques, sélectionnez la bibliothèque à désactiver.
- Cliquez sur la barre d'espacement. La coche disparaît de la case figurant à gauche du nom.

Suppression des bibliothèques locales

Vous pouvez supprimer une bibliothèque sans supprimer la version publique de la bibliothèque et inversement. La suppression d'une bibliothèque locale supprime la bibliothèque et tout son contenu de la session uniquement. Le fait de supprimer la version locale d'une bibliothèque ne supprime pas la bibliothèque des autres sessions et ne supprime pas non plus la version publique. Pour plus d'informations, reportez-vous à la section Gestion des bibliothèques publiques sur p. 296.

Pour supprimer une bibliothèque locale

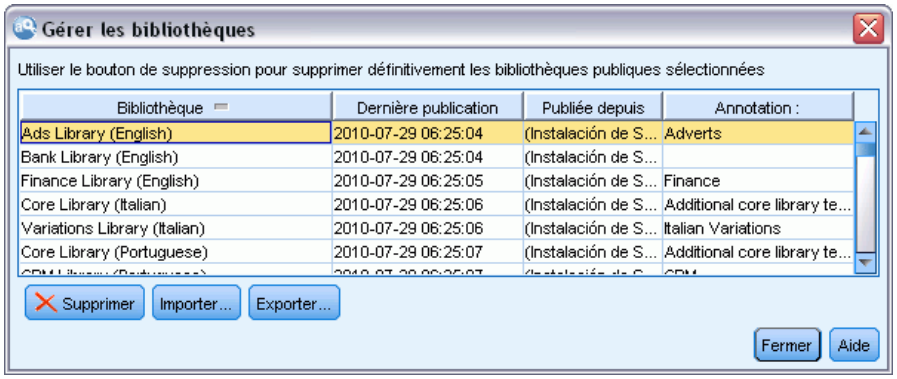
- Dans l'arborescence, sélectionnez la bibliothèque à supprimer.
- Dans les menus, choisissez Edition > Supprimer pour supprimer la bibliothèque. La bibliothèque est supprimée.
- Si vous n'avez jamais publié cette bibliothèque auparavant, un message vous demande si vous souhaitez supprimer ou conserver la bibliothèque. Cliquez sur Supprimer pour poursuivre ou sur Conserver pour conserver cette bibliothèque.

  Remarque : Vous devez toujours conserver au moins une bibliothèque.

Gestion des bibliothèques publiques

Afin de réutiliser les bibliothèques locales, vous pouvez les publier et ensuite les utiliser à partir de la boîte de dialogue Gérer les bibliothèques (Ressources > Gérer les bibliothèques). Pour plus d'informations, reportez-vous à la section Partage de bibliothèques sur p. 298. Vous pouvez effectuer des tâches de base de gestion des bibliothèques publiques : import, export ou suppression d'une bibliothèque publique. Il est impossible de renommer une bibliothèque publique.

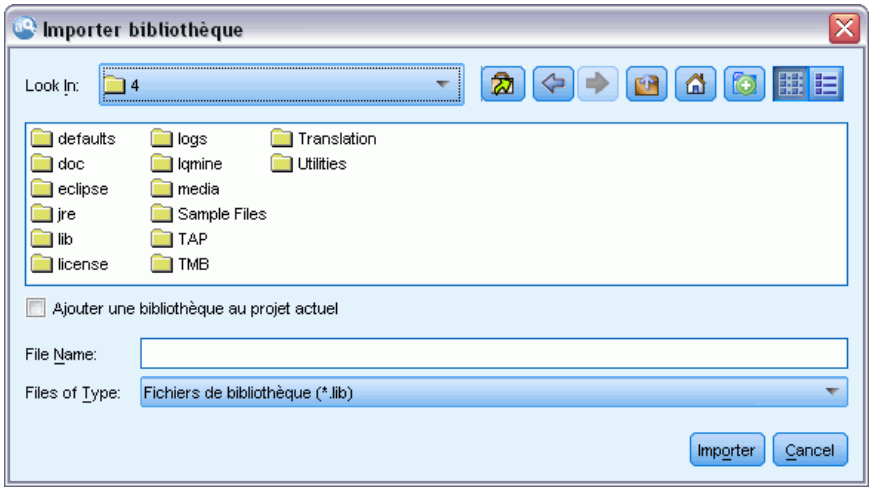
Figure 16-4
Boîte de dialogue Gérer les bibliothèques



Import de bibliothèques publiques

▶ Dans la boîte de dialogue Gérer les bibliothèques, cliquez sur Importer. La boîte de dialogue Importer une bibliothèque apparaît.

Figure 16-5
Boîte de dialogue Importer une bibliothèque



▶ Sélectionnez le fichier de bibliothèque (**.lib*) à importer et, si vous souhaitez également ajouter cette bibliothèque localement, sélectionnez l'option Ajouter une bibliothèque au projet actuel.

▶ Cliquez sur Importer. La boîte de dialogue se ferme. Si une bibliothèque publique de même nom existe déjà, vous êtes invité à renommer la bibliothèque en cours d'import ou à remplacer la bibliothèque publique en cours.

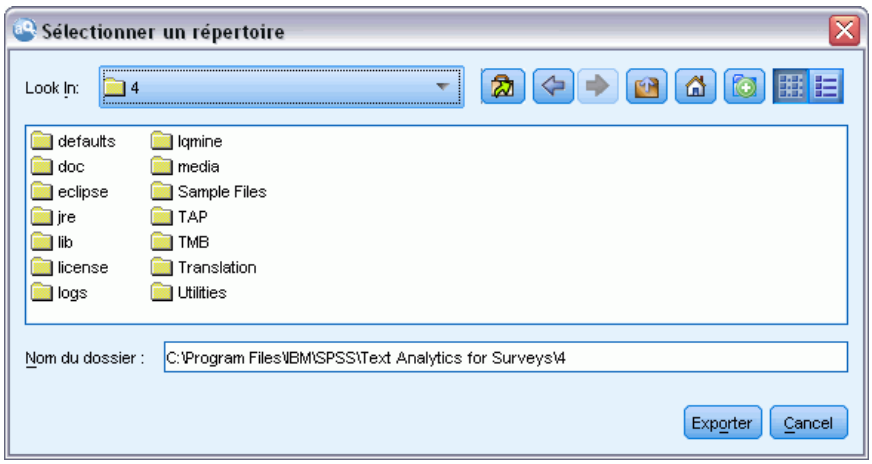
Export de bibliothèques publiques

Vous pouvez exporter les bibliothèques publiques au format *.lib* afin de pouvoir les partager.

▶ Dans la boîte de dialogue Gérer les bibliothèques, sélectionnez dans la liste la bibliothèque à exporter.

▶ Cliquez sur Export. La boîte de dialogue Sélectionner un répertoire apparaît.

Figure 16-6
Boîte de dialogue Sélectionner un répertoire



▶ Sélectionnez le répertoire vers lequel vous souhaitez exporter et cliquez sur Exporter. La boîte de dialogue disparaît et le fichier de bibliothèque (**.lib*) est exporté.

Suppression de bibliothèques publiques

Vous pouvez supprimer une bibliothèque locale sans supprimer la version publique de la bibliothèque et inversement. Cependant, si la bibliothèque est supprimée de cette boîte de dialogue, elle ne peut plus être ajoutée aux ressources de session jusqu'à ce qu'une version locale soit à nouveau publiée.

Si vous supprimez une bibliothèque qui a été installée avec le produit, la version originale installée est restaurée.

▶ Dans la boîte de dialogue Gérer les bibliothèques, sélectionnez la bibliothèque à supprimer. Vous pouvez trier la liste en cliquant sur l'en-tête approprié.

▶ Cliquez sur Supprimer pour supprimer la bibliothèque. IBM® SPSS® Modeler Text Analytics vérifie si la version locale de la bibliothèque est la même que la bibliothèque publique. Si c'est le cas, la bibliothèque est supprimée sans message d'alerte. Si les versions de la bibliothèque diffèrent, un message d'alerte apparaît pour vous demander si vous souhaitez conserver ou supprimer la version publique.

Partage de bibliothèques

Les bibliothèques permettent d'utiliser des ressources de manière à pouvoir les partager facilement parmi plusieurs sessions interactives. Les bibliothèques existent en deux états ou deux versions. Les bibliothèques modifiables dans l'éditeur de ressources et une partie de la session interactive s'appellent des **bibliothèques locales**. Lorsque vous travaillez dans une session interactive, vous pouvez effectuer de nombreux changements, par exemple dans la bibliothèque *Vegetables* (Légumes). Si vos changements peuvent s'avérer utiles avec d'autres données, vous pouvez

rendre ces ressources disponibles en créant une **version publique** de la bibliothèque *Vegetables* (Légumes). Une bibliothèque publique, comme son nom l'indique, est mise à la disposition de ressources dans n'importe quelle session interactive.

Vous pouvez visualiser les bibliothèques publiques dans la boîte de dialogue Gérer les bibliothèques. Une fois que cette version de bibliothèque publique existe, vous pouvez l'ajouter aux ressources dans d'autres contextes de manière à ce que ces ressources linguistiques personnalisées puissent être partagées.

Les bibliothèques fournies sont initialement des bibliothèques publiques. Il est possible de modifier les ressources de ces bibliothèques avant de créer une version publique. Ces nouvelles versions sont ensuite accessibles dans d'autres sessions interactives.

Lorsque vous utilisez vos bibliothèques et y apportez des modifications, les différentes versions ne sont plus synchronisées. Dans certains cas, une version locale peut être plus récente que la version publique et, dans d'autres cas, la version publique peut être plus récente que la version locale. La version publique peut contenir des changements que la version locale n'a pas et inversement. C'est le cas si la version publique a été mise à jour depuis une autre session interactive. Si les versions de vos bibliothèques ne sont plus synchronisées, vous pouvez rétablir la synchronisation. La synchronisation des versions de bibliothèque consiste à publier une nouvelle fois et/ou mettre à jour les bibliothèques locales.

Chaque fois que vous lancez ou fermez une session interactive, vous êtes invité à synchroniser toutes les bibliothèques qui doivent être mises à jour ou republiées. De plus, vous pouvez facilement identifier l'état de synchronisation de votre bibliothèque locale à l'aide de l'icône qui apparaît à côté du nom de la bibliothèque dans l'arborescence ou en affichant la boîte de dialogue Propriétés de bibliothèque. Vous pouvez également effectuer cette opération à tout moment en sélectionnant les options de menu correspondantes. Le tableau ci-dessous décrit les cinq états possibles et leurs icônes associées.

Table 16-1
Etats de synchronisation des bibliothèques locales

Icône	Description de l'état des bibliothèques locales
	Non publiée — La bibliothèque locale n'a jamais été publiée.
	Synchronisé — Les versions locale et publique de la bibliothèque sont identiques. Cela s'applique également à la *bibliothèque locale*, qui ne peut pas être publiée car elle est censée contenir uniquement des ressources propres à la session.
	Obsolète — La version publique de la bibliothèque est plus récente que la version locale. Vous pouvez mettre à jour votre version locale avec les modifications.
	Plus récent — La version locale de la bibliothèque est plus récente que la version publique. Vous pouvez republier votre version locale pour que la version publique soit actualisée.
	Désynchronisé — Les bibliothèques locale et publique contiennent chacune des modifications que l'autre bibliothèque n'a pas. Vous devez choisir soit de mettre à jour votre bibliothèque locale, soit de la publier. Si vous la mettez à jour, vous perdez les modifications que vous avez effectuées depuis votre dernière mise à jour ou publication. Si vous optez pour la publication, vous supprimez les changements effectués dans la version publique.

Remarque : Si vous mettez systématiquement à jour vos bibliothèques lorsque vous lancez une session interactive ou si vous publiez une bibliothèque lorsque vous la fermez , vous avez moins de risques d'avoir des bibliothèques qui ne sont pas synchronisées.

Vous pouvez republier une bibliothèque si vous pensez que les changements apportés peuvent être utiles à d'autres flux pouvant également contenir cette bibliothèque. Par conséquent, si vos changements peuvent être utiles à d'autres flux, vous pouvez mettre à jour les versions locales dans ces flux. De cette manière, vous pouvez créer des flux pour chaque contexte ou domaine qui s'applique à vos données : vous pouvez créer de nouvelles bibliothèques et/ou ajouter des bibliothèques publiques à vos ressources.

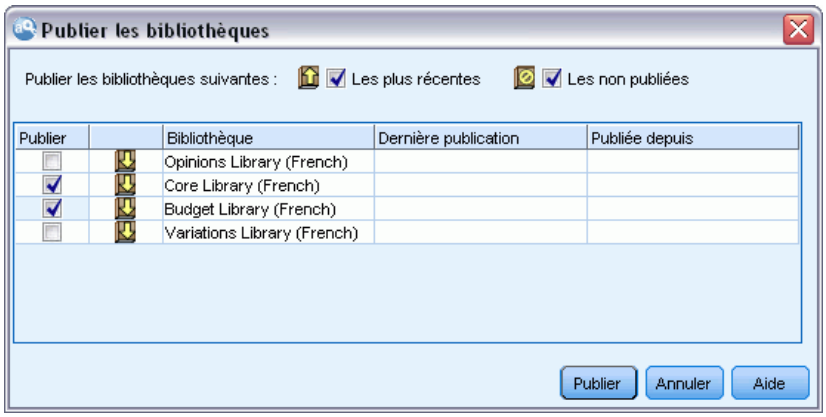
Lorsqu'une version publique de bibliothèque est partagée, les risques de différence entre les versions locale et publique sont plus importants. Chaque fois que vous lancez ou fermez et publiez depuis une session interactive, ouvrez ou fermez un modèle depuis l'Editeur de modèle, un message apparaît pour vous permettre de publier et/ou de mettre à jour les bibliothèques dont la version n'est pas synchronisée avec celles contenues dans la boîte de dialogue Gérer les bibliothèques. Si la version publique de la bibliothèque est plus récente que la version locale, une boîte de dialogue vous demande si vous souhaitez procéder à une mise à jour. Vous pouvez choisir de conserver la version locale en l'état au lieu de la mettre à jour pour qu'elle contienne les changements de la version publique. Vous pouvez également décider de la mettre à jour pour qu'elle reflète les changements insérés dans la version publique.

Publication de bibliothèques

Si vous n'avez jamais publié de bibliothèques, sachez que la publication consiste à créer une copie publique de votre bibliothèque locale dans la base de données. Si vous republiez une bibliothèque, le contenu de la bibliothèque locale remplace le contenu de la version publique existante. Après une nouvelle publication, vous pouvez mettre à jour cette bibliothèque dans n'importe quelle session de flux de manière à ce que leurs versions locales soient synchronisées avec la version publique. Même si vous pouvez publier une bibliothèque, une version locale est toujours stockée dans la session.

Important ! Si vous apportez des modifications à votre bibliothèque locale et si, dans le même temps, la version publique de la bibliothèque est également modifiée, votre bibliothèque est considérée comme désynchronisée. Nous vous recommandons de commencer par mettre à jour la version locale avec les modifications de la version publique, d'effectuer tous les changements nécessaires et ensuite, de publier à nouveau la version locale pour que les deux versions soient identiques. Si vous effectuez des changements et que vous commencez par une publication, vous remplacerez les changements apportés à la version publique.

Figure 16-7
Boîte de dialogue Publier les bibliothèques



Pour publier des bibliothèques locales dans la base de données

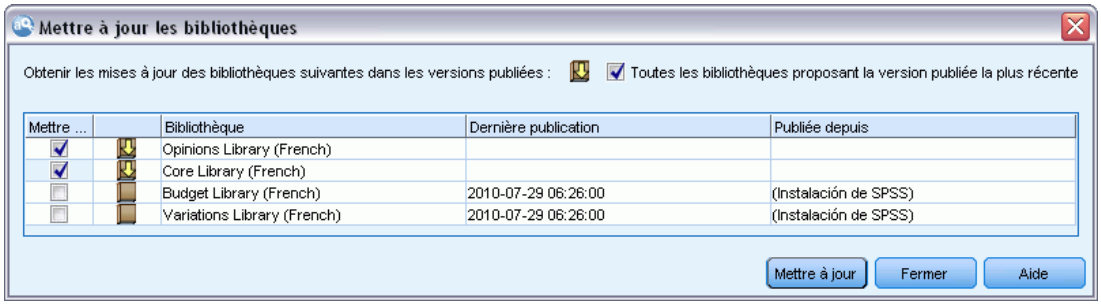
- A partir des menus, sélectionnez Ressources > Publier les bibliothèques. La boîte de dialogue Publier les bibliothèques apparaît. Toutes les bibliothèques devant être publiées sont sélectionnées par défaut.
- Cochez la case située à gauche de chaque bibliothèque à publier ou republier.
- Cliquez sur Publier pour publier les bibliothèques dans la base de données Gérer les bibliothèques.

Mise à jour des bibliothèques

Chaque fois que vous lancez ou fermez une session interactive, vous pouvez mettre à jour ou publier toutes les bibliothèques qui ne sont plus synchronisées avec les versions publiques. Si la version publique de la bibliothèque est plus récente que la version locale, une boîte de dialogue vous demande si vous souhaitez mettre à jour la bibliothèque. Vous pouvez choisir soit de conserver la version locale au lieu de la mettre à jour avec le contenu de la version publique, soit de remplacer la version locale par la publique. Si la version publique d'une bibliothèque est plus récente que la version locale, vous pouvez mettre à jour la version locale pour synchroniser son contenu avec celui de la version publique. La mise à jour consiste à incorporer dans votre version locale les changements trouvés dans la version publique.

Remarque : Si vous mettez systématiquement à jour vos bibliothèques lorsque vous lancez une session interactive ou si vous publiez une bibliothèque lorsque vous la fermez , vous avez moins de risques d'avoir des bibliothèques qui ne sont pas synchronisées. Pour plus d'informations, reportez-vous à la section Partage de bibliothèques sur p. 298.

Figure 16-8
Boîte de dialogue Mettre à jour les bibliothèques



Pour mettre à jour les bibliothèques locales

- A partir des menus, sélectionnez Ressources > Mettre à jour les bibliothèques. La boîte de dialogue Mettre à jour les bibliothèques apparaît. Toutes les bibliothèques devant être mises à jour sont sélectionnées par défaut.
- Cochez la case située à gauche de chaque bibliothèque à publier ou republier.
- Cliquez sur Mettre à jour pour mettre à jour les bibliothèques locales.

Résolution des conflits

Conflits entre bibliothèque locale et bibliothèque publique

Chaque fois que vous lancez une session de flux, IBM® SPSS® Modeler Text Analytics compare les bibliothèques locales et celles répertoriées dans la boîte de dialogue Gérer les bibliothèques. Si une bibliothèque locale de votre session n'est pas synchronisée avec les versions publiées, la boîte de dialogue Avertissement de synchronisation de bibliothèques apparaît. Vous pouvez choisir parmi les options suivantes afin de sélectionner les versions des bibliothèques que vous souhaitez utiliser :

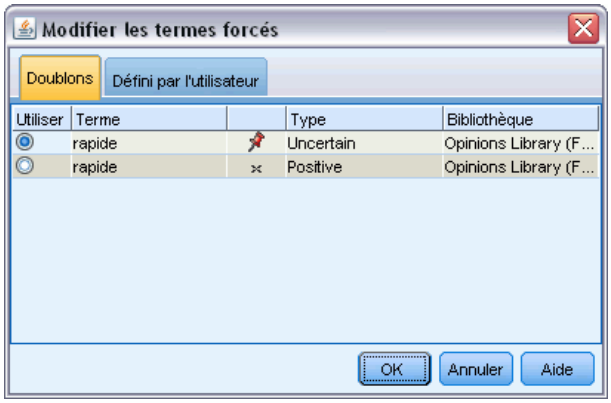
- **Toutes les bibliothèques locales vers un fichier.** Cette option conserve toutes vos bibliothèques en l'état. Vous pouvez toujours les republier ou les mettre à jour ultérieurement.
- **Toutes les bibliothèques publiées sur cet ordinateur.** Cette option remplace les bibliothèques locales affichées par les versions de la base de données.
- **Toutes les bibliothèques les plus récentes.** Cette option remplace toutes les anciennes bibliothèques locales par les versions publiques plus récentes contenues dans la base de données.
- **Autre.** Cette option permet de sélectionner manuellement les versions de votre choix en les sélectionnant dans le tableau.

Conflits de termes forcés

Lorsque vous ajoutez une bibliothèque publique ou mettez à jour une bibliothèque locale, vous pouvez détecter des conflits et des doublons entre les termes et les types de cette bibliothèque et ceux des autres bibliothèques stockées dans vos ressources. Si cela se produit, vous êtes invité à

vérifier, dans la boîte de dialogue Editer les termes forcés, les résolutions de conflit proposées ou à les modifier avant de réaliser l'opération. Pour plus d'informations, reportez-vous à la section Ajout des termes forcés dans le chapitre 17 sur p. 312.

Figure 16-9
Boîte de dialogue Editer les termes forcés



La boîte de dialogue Editer les termes forcés contient toutes les paires de termes et types contradictoires. L'alternance des couleurs de l'arrière-plan permet de distinguer chaque paire contradictoire. Ces couleurs peuvent être modifiées dans la boîte de dialogue Options. Pour plus d'informations, reportez-vous à la section Options : Onglet Affichage dans le chapitre 8 sur p. 133. La boîte de dialogue Éditer les termes forcés contient deux onglets :

- **Doublons.** Cet onglet contient les doublons rencontrés dans les bibliothèques. Si une icône représentant une punaise apparaît après un terme, cela signifie que l'occurrence du terme a été forcée. Si une icône X apparaît en noir, cela signifie que cette occurrence du terme sera ignorée pendant l'extraction, car elle a été imposée ailleurs.
- **Défini par l'utilisateur.** Cet onglet contient la liste de tous les termes qui ont été forcés manuellement dans le panneau des termes de la déclaration de types et non pas via les conflits.

Remarque : La boîte de dialogue Editer les termes forcés apparaît après l'ajout ou la mise à jour d'une bibliothèque. Si vous quittez la boîte de dialogue, vous n'annulez pas la mise à jour ou l'ajout de la bibliothèque.

Pour résoudre des conflits

▶ Dans la boîte de dialogue Editer les termes forcés, sélectionnez le bouton radio de la colonne Utiliser du terme que vous souhaitez forcer.

▶ Lorsque vous avez terminé, cliquez sur OK pour appliquer les termes forcés et fermer la boîte de dialogue. Si vous cliquez sur Annuler, vous annulerez les changements effectués dans cette boîte de dialogue.

Chapitre
17

À propos des dictionnaires de bibliothèque

Les ressources servant à extraire des données textuelles sont stockées sous la forme de modèles et de bibliothèques. Une bibliothèque peut être constituée de trois dictionnaires.

- La **déclaration de types** comprend un ensemble de termes regroupés sous une même étiquette ou sous un même nom de type. Lorsque le moteur d'extraction lit vos données textuelles, il compare les mots rencontrés dans le texte aux termes définis dans vos déclarations de types. Au cours de l'extraction, les formes fléchies des termes et synonymes d'un type sont regroupées sous un terme cible nommé concept. Les concepts extraits sont affectés à la déclaration de types dans laquelle ils apparaissent en tant que termes. Vous pouvez gérer vos déclarations de types dans les panneaux situés en haut à gauche et au centre de l'éditeur — l'arborescence de la bibliothèque et le panneau des termes. Pour plus d'informations, reportez-vous à la section Déclarations de types sur p. 304.
- Le **dictionnaire de substitutions** contient un ensemble de mots définis comme synonymes ou comme éléments optionnels qui sont utilisés pour regrouper des termes similaires sous un seul terme cible, appelé concept dans les résultats d'extraction finaux. Vous pouvez gérer vos dictionnaires de substitutions dans le panneau inférieur gauche de l'éditeur à l'aide des onglets Synonymes et Optionnels. Pour plus d'informations, reportez-vous à la section Dictionnaires des substitutions/synonymes sur p. 315.
- Le **dictionnaire d'exclusions** comprend un ensemble de termes et de types qui seront supprimés des résultats finaux de l'extraction. Vous pouvez gérer vos dictionnaires d'exclusion dans le panneau le plus à droite de l'éditeur. Pour plus d'informations, reportez-vous à la section Dictionnaires d'exclusions sur p. 321.

Pour plus d'informations, reportez-vous à la section Utilisation des bibliothèques dans le chapitre 16 sur p. 290.

Déclarations de types

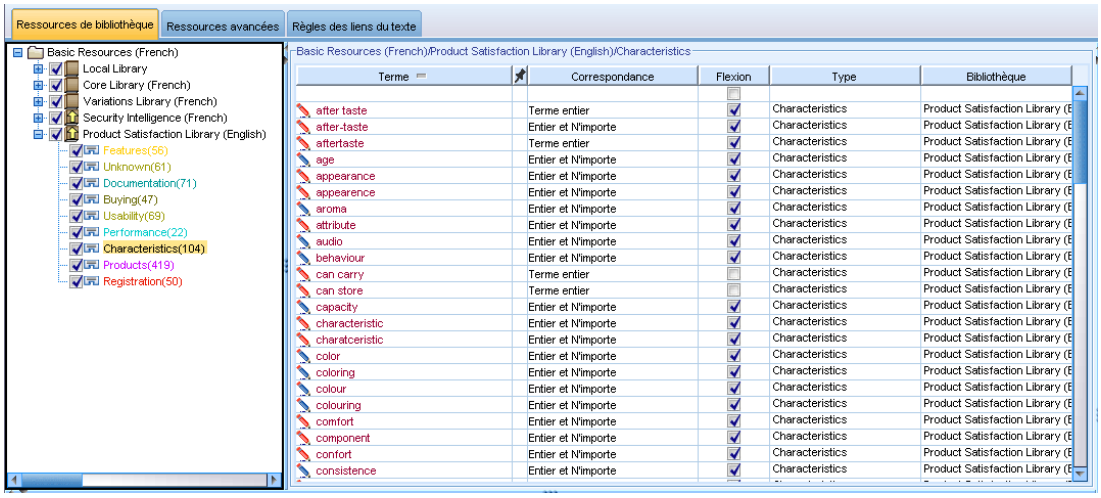
Une **déclaration de types** est constituée d'un nom de type, d'une étiquette et d'une liste de termes. La gestion des déclarations de types s'effectue dans les panneaux supérieurs gauche et central de l'onglet Ressources de bibliothèque dans l'éditeur. Vous pouvez accéder à cette vue avec Vue > Editeur de ressources dans les menus lors d'une session interactive. Autrement, vous pouvez modifier les dictionnaires d'un modèle spécifique dans l'Editeur de modèle.

Lorsque le moteur du programme d'extraction lit vos données textuelles, il compare les mots rencontrés dans le texte aux termes définis dans vos déclarations de types. Les termes sont des mots ou des expressions dans les déclarations de types dans vos ressources linguistiques.

Lorsqu'un mot correspond à un terme, il est attribué au nom de type de ce terme. Lorsque les ressources sont lues au cours de l'extraction, les termes trouvés dans le texte traversent ensuite plusieurs étapes de traitement avant de devenir des concepts dans le panneau Résultats d'extraction. Si plusieurs termes appartenant à la même déclaration de types sont déterminés

comme étant des synonymes par le moteur du programme d'extraction, ils sont regroupés sous le terme le plus fréquent et nommés un **concept** dans le panneau Résultats d'extraction. Par exemple, si les termes `question` et `requête` peuvent apparaître dans le nom de concept `question` à la fin.

Figure 17-1
Arborescence de bibliothèques et panneau de termes



La liste des déclarations de types est affichée dans le panneau qui contient l'arborescence de bibliothèques, à gauche. Le contenu de chaque déclaration de types apparaît dans le panneau central. Les déclarations de types offrent bien plus qu'une simple liste de termes. La manière dont les mots de vos données textuelles sont mis en correspondance avec les concepts affectés aux déclarations de types est déterminée par l'option de mise en correspondance. Une **option de mise en correspondance** définit le positionnement d'un terme par rapport à une expression ou un mot candidat dans les données textuelles. Pour plus d'informations, reportez-vous à la section Ajout de termes sur p. 308.

Remarque : Certaines options, telles que l'option de mise en correspondance ou les formes fléchies, ne s'appliquent pas au texte en japonais. Pour plus d'informations, reportez-vous à la section Modification des propriétés de types japonais dans l'annexe A sur p. 382.*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Par ailleurs, vous pouvez étendre les termes de votre déclaration de types en indiquant si vous souhaitez générer et ajouter automatiquement des formes fléchies aux termes de la déclaration. En générant les formes fléchies, vous ajoutez automatiquement les formes plurielles de termes au singulier, les formes au singulier de termes au pluriel et les adjectifs dans la déclaration de types. Pour plus d'informations, reportez-vous à la section Ajout de termes sur p. 308.

Remarque : Dans la plupart des langues, les concepts qui ne se trouvent dans aucune déclaration de type mais sont extraits du texte sont automatiquement dotés du type `<Unknown>` ; cependant, pour le texte en japonais ils sont automatiquement dotés du type <名詞>*Remarque* : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Types intégrés

IBM® SPSS® Modeler Text Analytics est fourni avec un ensemble de ressources linguistiques sous la forme de bibliothèques et de ressources compilées. Les bibliothèques fournies contiennent un ensemble de déclarations de types intégrées parmi lesquelles `<Location>`, `<Organization>`, `<Person>` et `<Product>`.

Remarque : L'ensemble de types intégrés par défaut est différent pour le texte japonais. Pour des informations supplémentaires sur les types fournis avec les ressources japonaises, consultez Types disponibles pour du texte en japonais. *Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Ces déclarations de types sont utilisées par le moteur du programme d'extraction pour affecter des types aux concepts qu'il extrait (par exemple, le type `<Location>` est affecté au concept `paris`). Bien qu'un grand nombre de termes aient été définis dans les déclarations de types intégrées, ceux-ci ne couvrent pas toutes les possibilités. Par conséquent, vous pouvez les compléter ou en créer qui vous soient propres. Pour obtenir la description du contenu d'une déclaration de types fournie en particulier, reportez-vous à l'annotation dans la boîte de dialogue Propriétés de type. Sélectionnez le type dans l'arborescence et choisissez Edition > Propriétés dans le menu contextuel.

Remarque : Outre les bibliothèques fournies, les ressources compilées (également utilisées par le moteur d'extraction) contiennent un grand nombre de définitions complémentaires aux déclarations de types intégrées, mais leur contenu n'est pas visible dans le produit. Vous pouvez toutefois forcer n'importe quel autre dictionnaire à accepter un terme classé par type par les déclarations compilées. Pour plus d'informations, reportez-vous à la section Ajout des termes forcés sur p. 312.

Création de types

Vous pouvez créer des déclarations de types pour faciliter le regroupement de termes similaires. Quand des termes qui figurent dans cette déclaration sont découverts au cours du processus d'extraction, ils sont affectés à ce nom de type et extraits sous un nom de concept. Lorsque vous créez une bibliothèque, une bibliothèque de types vide est toujours incluse, afin de vous permettre de commencer immédiatement à entrer des termes.

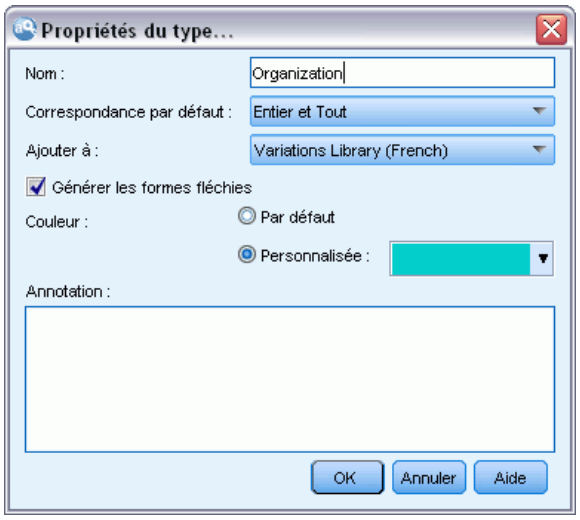
Important ! : Vous ne pouvez pas créer de nouveaux types pour des ressources en japonais. Pour des informations supplémentaires sur les déclarations de types des ressources japonaises, consultez Panneau Arborescence des bibliothèques japonaises, panneau des types et panneau des termes. *Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Si vous analysez un texte concernant des aliments et souhaitez regrouper les termes relatifs aux légumes, vous pouvez créer votre propre déclaration de types `<Légumes>`. Vous pouvez ensuite y ajouter des termes tels que `carotte`, `brocoli` et `épinard` si vous estimez qu'il s'agit de termes importants qui vont apparaître dans le texte. Ensuite, au cours de l'extraction, si l'un de ces termes est identifié, il est extrait en tant que concept et affecté au type `<Vegetables>` (Légumes).

Il n'est pas nécessaire de définir toutes les formes d'un mot ou d'une expression, car vous pouvez choisir de générer toutes les formes fléchies des termes. Lorsque vous choisissez cette option, le moteur d'extraction reconnaît automatiquement les formes au singulier ou au pluriel des

termes parmi les autres formes et les associe à ce type. Cette option s'avère particulièrement utile quand votre type contient principalement des noms, dans la mesure où il est peu probable que vous cherchiez à obtenir les formes fléchies de verbes ou d'adjectifs.

Figure 17-2
Boîte de dialogue Propriétés de type



Nom. Nom que vous attribuez à la déclaration de types que vous créez. Nous vous recommandons de ne pas utiliser d'espaces dans les noms de types, en particulier si deux noms de types ou plus commencent par le même mot.

Remarque : il existe certaines contraintes sur les noms de type et l'utilisation des symboles. Par exemple, n'utilisez pas de symboles tels que « @ » or « ! » dans le nom.

Mise en correspondance par défaut. L'attribut Mise en correspondance par défaut indique au moteur d'extraction la manière dont il doit mettre en correspondance ce terme et les données textuelles. Dès que vous ajoutez un terme à cette déclaration de types, cet attribut de mise en correspondance lui est associé automatiquement. Vous avez toujours la possibilité de modifier manuellement le choix de correspondance dans la liste des termes. Vous avez le choix entre les options suivantes : Terme entier, Début, Fin, Tout, Début ou Fin, Entier et Début, Entier et Fin et Entier et (Début ou Fin) et Entier (pas de composés). Pour plus d'informations, reportez-vous à la section Ajout de termes sur p. 308. Cette option ne s'applique pas aux ressources en japonais.

Ajouter à. Ce champ indique la bibliothèque dans laquelle vous allez créer votre nouvelle déclaration de types.

Générer les formes fléchies. Cette option indique au moteur d'extraction d'utiliser la fonction de morphologie grammaticale pour collecter et regrouper les formes similaires des termes que vous ajoutez à cette déclaration (le singulier et le pluriel d'un terme, par exemple). Cette fonction est particulièrement utile lorsque votre type contient surtout des noms. Quand vous sélectionnez cette option, elle est appliquée par défaut à tous les nouveaux termes ajoutés au type, mais vous pouvez la modifier manuellement dans la liste. Cette option ne s'applique pas aux ressources en japonais.

Couleur. Ce champ vous permet de distinguer les résultats de ce type dans l'interface. Si vous sélectionnez Par défaut, la couleur par défaut du type est également utilisée pour cette déclaration de types. La couleur par défaut se configure dans la boîte de dialogue des options. Pour plus d'informations, reportez-vous à la section Options : Onglet Affichage dans le chapitre 8 sur p. 133. Si vous sélectionnez Personnalisé, sélectionnez une couleur dans la liste déroulante.

Annotations. Ce champ est facultatif et sert à entrer des commentaires ou des descriptions.

Pour créer une déclaration de types

- Sélectionnez la bibliothèque dans laquelle vous souhaitez créer une déclaration de types.
- A partir des menus, sélectionnez Outils > Nouveau type. La boîte de dialogue Propriétés de type apparaît.
- Entrez le nom de votre déclaration de types dans la zone de texte Nom et choisissez les options souhaitées.
- Cliquez sur OK pour créer la déclaration de types. Le nouveau type apparaît dans l'arborescence de bibliothèques et s'affiche dans le panneau central. Vous pouvez commencer immédiatement à ajouter des termes. Pour plus d'informations, reportez-vous à Ajout de termes.

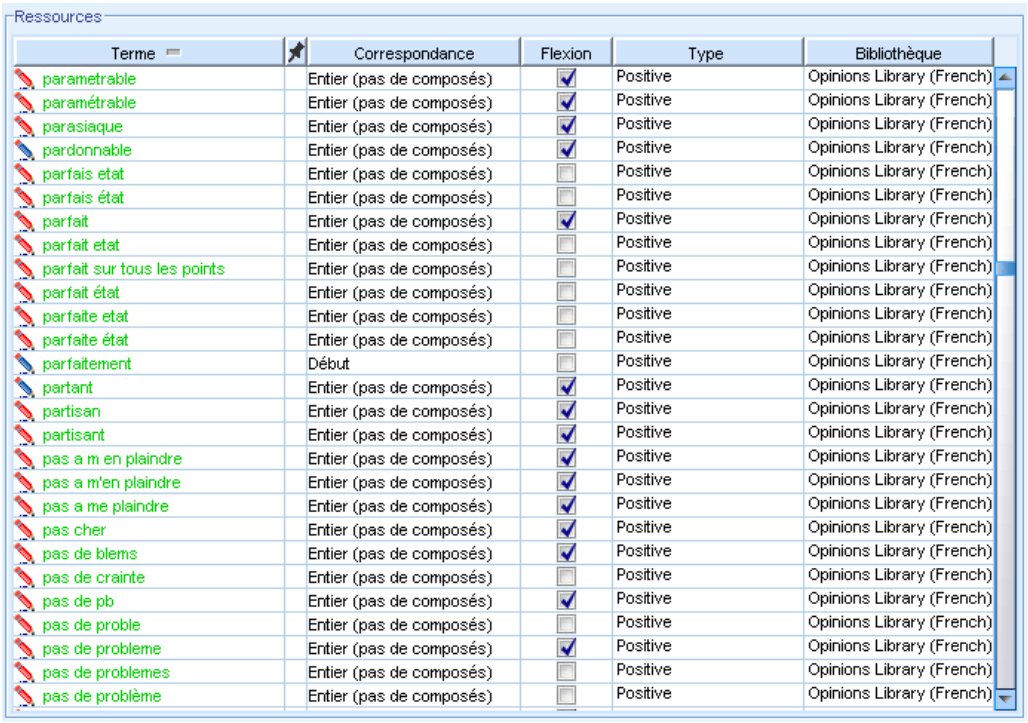
  Remarque : Ces instructions indiquent la façon d'effectuer des modifications dans la vue Editeur de ressources ou dans la vue Editeur de modèle. N'oubliez pas que vous pouvez également réaliser ce type de réglage directement à partir du panneau Résultats d'extraction , du panneau Données, du panneau Catégorie ou de la boîte de dialogue Définitions de clusters dans les autres vues. Pour plus d'informations, reportez-vous à la section Affinage des résultats de l'extraction dans le chapitre 9 sur p. 154.

Ajout de termes

L'arborescence de bibliothèques contient les bibliothèques. Il peut être développé pour présenter les déclarations de types qu'elles contiennent. Dans le panneau central, une liste affiche les termes de la bibliothèque ou de la déclaration de types sélectionnée, en fonction de l'élément sélectionné dans l'arborescence.

Important ! Les termes sont définis autrement pour les ressources en japonais. Pour plus d'informations, reportez-vous à la section Panneau Arborescence des bibliothèques japonaises, panneau des types et panneau des termes dans l'annexe A sur p. 376.*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Figure 17-3
Panneau des termes


Ressources

Terme	Correspondance	Flexion	Type	Bibliothèque
parametrable	Entier (pas de composés)	☑	Positive	Opinions Library (French)
paramétrable	Entier (pas de composés)	☑	Positive	Opinions Library (French)
parasiaque	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pardonnable	Entier (pas de composés)	☑	Positive	Opinions Library (French)
parfais etat	Entier (pas de composés)	☐	Positive	Opinions Library (French)
parfais état	Entier (pas de composés)	☐	Positive	Opinions Library (French)
parfait	Entier (pas de composés)	☑	Positive	Opinions Library (French)
parfait etat	Entier (pas de composés)	☐	Positive	Opinions Library (French)
parfait sur tous les points	Entier (pas de composés)	☐	Positive	Opinions Library (French)
parfait état	Entier (pas de composés)	☐	Positive	Opinions Library (French)
parfaite etat	Entier (pas de composés)	☐	Positive	Opinions Library (French)
parfaite état	Entier (pas de composés)	☐	Positive	Opinions Library (French)
parfaitement	Début	☐	Positive	Opinions Library (French)
partant	Entier (pas de composés)	☑	Positive	Opinions Library (French)
partisan	Entier (pas de composés)	☑	Positive	Opinions Library (French)
partisant	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pas a m en plaindre	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pas a m'en plaindre	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pas a me plaindre	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pas cher	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pas de blems	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pas de crainte	Entier (pas de composés)	☐	Positive	Opinions Library (French)
pas de pb	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pas de proble	Entier (pas de composés)	☐	Positive	Opinions Library (French)
pas de probleme	Entier (pas de composés)	☑	Positive	Opinions Library (French)
pas de problemes	Entier (pas de composés)	☐	Positive	Opinions Library (French)
pas de problème	Entier (pas de composés)	☐	Positive	Opinions Library (French)

Dans Editeur de ressources, vous pouvez ajouter des termes directement à une déclaration de types : soit directement dans le panneau de termes, soit par l'intermédiaire de la boîte de dialogue Ajouter des nouveaux termes. Les termes que vous ajoutez peuvent être des mots simples ou composés. La ligne vide, située en haut de la liste, vous permet d'ajouter de nouveaux termes à tout moment.

Remarque : Ces instructions indiquent la façon d'effectuer des modifications dans la vue Editeur de ressources ou dans la vue Editeur de modèle. N'oubliez pas que vous pouvez également réaliser ce type de réglage directement à partir du panneau Résultats d'extraction , du panneau Données, du panneau Catégorie ou de la boîte de dialogue Définitions de clusters dans les autres vues. Pour plus d'informations, reportez-vous à la section Affinage des résultats de l'extraction dans le chapitre 9 sur p. 154.

Colonne Terme

Dans cette colonne, entrez des mots simples ou composés dans la cellule. La couleur d'affichage du terme dépend de la couleur du type dans lequel le terme est enregistré ou a été imposé. Vous pouvez changer les couleurs des types dans la boîte de dialogue Propriétés de type. Pour plus d'informations, reportez-vous à la section Création de types sur p. 306.

Colonne Forcer

Dans cette colonne, sélectionnez cette cellule pour afficher une icône de punaise qui indique au moteur d'extraction d'ignorer les autres occurrences de ce même terme dans les autres bibliothèques. Pour plus d'informations, reportez-vous à la section Ajout des termes forcés sur p. 312.

Colonne Correspondance

Dans cette colonne, sélectionnez une option de mise en correspondance pour indiquer au moteur d'extraction comment il doit faire correspondre ce terme aux données textuelles. Consultez le tableau pour voir des exemples. Pour changer la valeur par défaut, éditez les propriétés du type. Pour plus d'informations, reportez-vous à la section Création de types sur p. 306. Dans les menus, choisissez Edition > Modifier la mise en correspondance. Ci-dessous figurent les options de mise en correspondance de base, car des combinaisons de celles-ci sont également possibles :

- Début. Si le terme dans la déclaration correspond au premier mot d'un concept extrait du texte, ce type est attribué. Par exemple, si vous entrez `tarte`, le terme `tarte aux pommes` est renvoyé.
- Fin. Si le terme dans la déclaration correspond au dernier mot d'un concept extrait du texte, ce type est attribué. Par exemple, si vous entrez `tarte`, `moule à tarte` est renvoyé.
- N'importe. Si le terme dans la déclaration correspond à tout mot d'un concept extrait du texte, ce type est attribué. Par exemple, si vous entrez `tarte`, l'option Tout classe `tarte aux pommes`, `moule à tarte` et `moule pour tarte aux pommes` sous un même type.
- Terme entier. Ce type est attribué si l'intégralité du concept extrait du texte correspond exactement avec le terme de la déclaration. L'ajout d'un terme comme Terme entier, Entier et début, Entier et fin, Entier et n'importe ou Entier (pas de composés) force l'extraction d'un terme.

  De plus, comme le type `<Person>` extrait uniquement les noms en deux parties, comme *edith piaf* ou *mohandas gandhi*, vous voudrez peut-être ajouter explicitement les prénoms à cette déclaration de types si vous essayez d'extraire un prénom lorsqu'aucun nom de famille n'est mentionné. Par exemple, si vous voulez extraire toutes les instances du nom *edith*, vous devez ajouter `edith` au type `<Person>` à l'aide des options Terme entier ou Entier et début.
- Entier (pas de composés). Si le concept complet extrait du texte correspond au terme exact dans la déclaration, ce type est attribué et l'extraction s'arrête afin d'empêcher l'extraction d'associer le terme aux composés plus longs. Par exemple, si vous saisissez `pomme`, l'option Entier (pas de composé) produira `pomme` et n'extraira pas le composé `jus de pomme`, à moins que ce ne soit forcé ailleurs.

Dans le tableau suivant, on peut supposer que le terme `pomme` est dans une déclaration de types. En fonction de l'option de mise en correspondance, ce tableau indique les concepts qui seraient extraits et entrés s'ils étaient trouvés dans le texte.

Table 17-1
Exemples de correspondance

Options de mise en correspondance pour le terme :  `pommes`	**Concepts extraits**			
	`pommes`	`tarte` ***pommes***	***pommes*** `mûres`	***tarte*** **aux** `pommes` ***maison***
Terme entier	✔			
Démarrer		✔		
Fin			✔	
Début ou fin		✔	✔	
Entier et début	✔	✔		
Entier et fin	✔		✔	
Entier et (début ou fin)	✔	✔	✔	
N'importe		✔	✔	✔
Entier et n'importe	✔	✔	✔	✔
Entier (pas de composés)	✔	*jamais extrait*	*jamais extrait*	*jamais extrait*

Colonne Flexion

Dans cette colonne, sélectionnez si le moteur du programme d'extraction doit générer les formes fléchies de ce terme au cours de l'extraction afin qu'elles soient regroupées. La valeur par défaut de cette colonne est définie dans les propriétés du type, mais vous pouvez modifier cette option au cas par cas directement dans la colonne. Dans les menus, sélectionnez Edition > Modifier la flexion.

Colonne Type

Dans cette colonne, sélectionnez une déclaration de types dans la liste déroulante. La liste de types est filtrée en fonction de votre sélection dans l'arborescence de la bibliothèque. Le premier type de la liste est toujours celui sélectionné par défaut dans l'arborescence de la bibliothèque. Dans les menus, sélectionnez Edition > Modifier le type.

Colonne Bibliothèque

Dans cette colonne, la bibliothèque dans laquelle votre terme est stocké s'affiche. Pour transférer un terme dans un autre type de l'arborescence, faites-le glisser et déposez-le sur l'autre bibliothèque.

Pour ajouter un terme simple à une déclaration de types

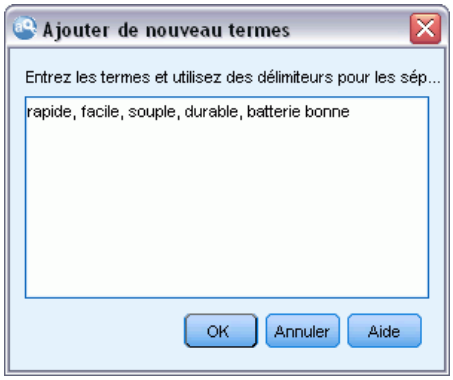
- Dans l'arborescence de la bibliothèque, sélectionnez la déclaration de types à laquelle vous voulez ajouter le terme.
- Dans la liste de termes du panneau central, entrez votre terme dans la première cellule vide disponible et définissez les options souhaitées pour ce terme.

Pour ajouter plusieurs termes à une déclaration de types

- Dans l'arborescence de la bibliothèque, sélectionnez la déclaration de types à laquelle vous voulez ajouter des termes.

- A partir des menus, sélectionnez Outils> Nouveaux termes. La boîte de dialogue Ajouter des nouveaux termes apparaît.

Figure 17-4
Boîte de dialogue Ajouter des nouveaux termes



- Entrez les termes que vous souhaitez ajouter à la déclaration de types sélectionnée. Pour ce faire, saisissez les termes au clavier, ou copiez et collez un ensemble de termes. Si vous entrez plusieurs termes, séparez-les au moyen du séparateur défini dans la boîte de dialogue Options ou ajoutez un terme par nouvelle ligne. Pour plus d'informations, reportez-vous à la section Définition des options dans le chapitre 8 sur p. 131.

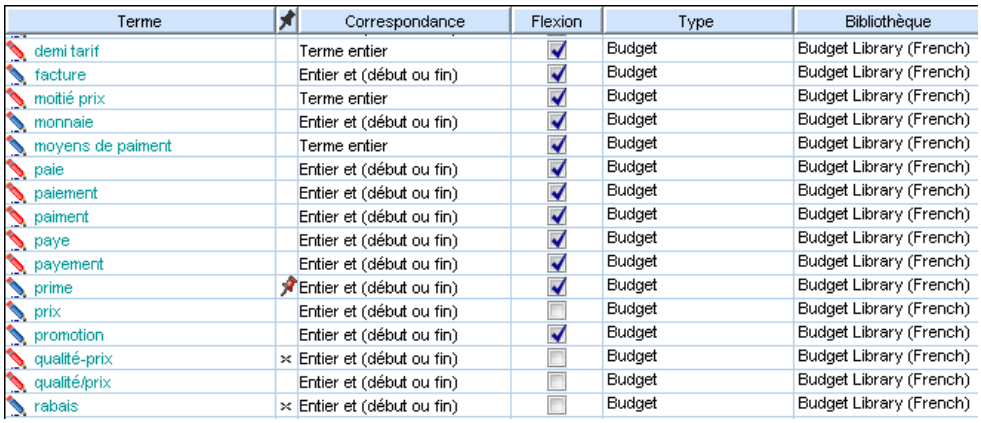
- Cliquez sur OK pour ajouter les termes à la déclaration. L'option de mise en correspondance est définie automatiquement sur la valeur par défaut de cette bibliothèque de types. La boîte de dialogue se ferme et les nouveaux termes apparaissent dans la déclaration.

Ajout des termes forcés

Si vous souhaitez affecter un terme à un type particulier, vous pouvez l'ajouter à la déclaration de types correspondante. Toutefois, si plusieurs termes ont le même nom, le moteur d'extraction doit savoir quel type utiliser. Par conséquent, vous êtes invité à sélectionner le type à utiliser. Cette opération est appelée **ajout des termes forcés** un terme dans un type. Cette option est particulièrement utile lorsque vous remplacez le type attribué d'une déclaration compilée (interne, non modifiable). En général, nous recommandons d'éviter les termes doubles.

L'ajout des termes forcés ne *supprime* pas les autres occurrences du terme, mais elles sont ignorées par le moteur d'extraction. Vous pouvez ensuite désigner l'occurrence qui doit être utilisée en activant ou en désactivant l'ajout des termes forcés. Vous pouvez également imposer un terme dans une déclaration de types lorsque vous ajoutez une bibliothèque publique ou la mettez à jour.

Figure 17-5
Icônes de statut d'imposition

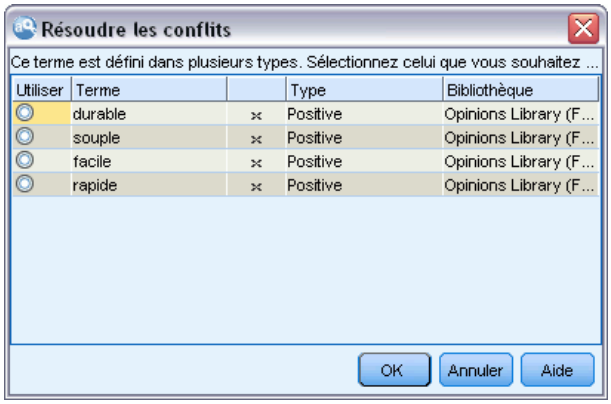


Terme		Correspondance	Flexion	Type	Bibliothèque
demi tarif		Terme entier	☑	Budget	Budget Library (French)
facture		Entier et (début ou fin)	☑	Budget	Budget Library (French)
moitié prix		Terme entier	☑	Budget	Budget Library (French)
monnaie		Entier et (début ou fin)	☑	Budget	Budget Library (French)
moyens de paiment		Terme entier	☑	Budget	Budget Library (French)
paie		Entier et (début ou fin)	☑	Budget	Budget Library (French)
paiement		Entier et (début ou fin)	☑	Budget	Budget Library (French)
paiment		Entier et (début ou fin)	☑	Budget	Budget Library (French)
paye		Entier et (début ou fin)	☑	Budget	Budget Library (French)
payement		Entier et (début ou fin)	☑	Budget	Budget Library (French)
prime	📌	Entier et (début ou fin)	☑	Budget	Budget Library (French)
prix		Entier et (début ou fin)	☐	Budget	Budget Library (French)
promotion		Entier et (début ou fin)	☑	Budget	Budget Library (French)
qualité-prix	✕	Entier et (début ou fin)	☐	Budget	Budget Library (French)
qualité/prix		Entier et (début ou fin)	☐	Budget	Budget Library (French)
rabais	✕	Entier et (début ou fin)	☐	Budget	Budget Library (French)

Vous pouvez voir les termes qui sont imposés ou ignorés dans la colonne Forcer, en deuxième position dans le panneau des termes. Si une icône de punaise apparaît, cela signifie que cette occurrence du terme a été imposée. Si une icône X apparaît en noir, cela signifie que cette occurrence du terme sera ignorée pendant l'extraction, car elle a été imposée ailleurs. Par ailleurs, quand vous imposez un terme, celui-ci apparaît dans la couleur du type dans lequel il a été imposé. Cela signifie que si vous avez imposé dans `Type 1` un terme présent à la fois dans `Type 1` et `Type 2`, il apparaît dans la fenêtre dans la couleur définie pour `Type 1`.

Vous pouvez double-cliquer sur l'icône pour modifier le statut. Si le terme apparaît ailleurs, la boîte de dialogue Résoudre les conflits qui apparaît vous permet de sélectionner l'occurrence à utiliser.

Figure 17-6
Boîte de dialogue Résoudre les conflits



Attribution de nouveaux noms aux types

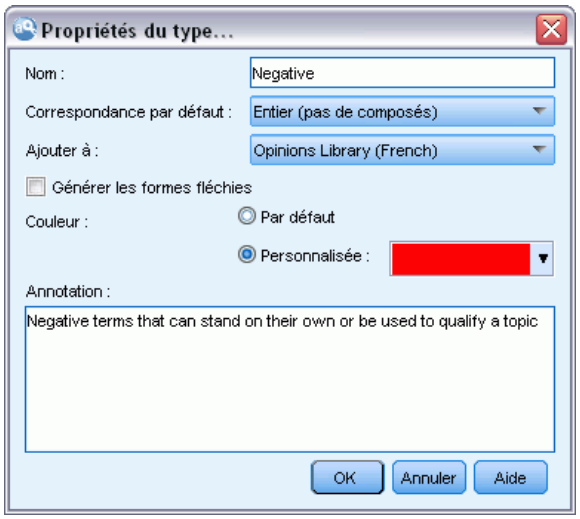
Vous pouvez renommer une déclaration de types ou configurer ses autres paramètres en éditant les propriétés du type.

Important ! Nous vous recommandons de ne pas utiliser d'espaces dans les noms de types, en particulier si deux noms de types ou plus commencent par le même mot. Nous vous recommandons également de ne pas renommer les types dans les bibliothèques Principale ou Opinions ni de modifier les attributs de correspondance par défaut.

Pour renommer un type

- Dans l'arborescence de bibliothèques, sélectionnez la déclaration de types que vous souhaitez renommer.
- Cliquez avec le bouton droit de la souris, puis choisissez Propriétés de type dans le menu contextuel. La boîte de dialogue Propriétés de type apparaît.

Figure 17-7
Boîte de dialogue Propriétés de type



- Entrez le nouveau nom de votre déclaration de types dans la zone de texte Nom.
- Cliquez sur OK pour accepter le nouveau nom. Le nouveau nom du type apparaît dans l'arborescence de bibliothèques.

Déplacement de types

Vous pouvez faire glisser une déclaration de types vers un autre emplacement d'une même bibliothèque ou vers une autre bibliothèque de l'arborescence.

Pour réorganiser un type au sein d'une même bibliothèque

- Dans l'arborescence de bibliothèques, sélectionnez la déclaration de types à déplacer.
- Dans les menus, choisissez Edition > Monter d'un niveau pour remonter la déclaration de types d'une position dans l'arborescence de la bibliothèque ou Edition > Descendre d'un niveau pour le descendre d'une position.

Pour déplacer un type vers une autre bibliothèque

- Dans l'arborescence de bibliothèques, sélectionnez la déclaration de types à déplacer.
- Cliquez avec le bouton droit de la souris, puis choisissez Propriétés de type dans le menu contextuel. La boîte de dialogue Propriétés de type apparaît. (Vous pouvez également faire glisser et déposer le type dans une autre bibliothèque).
- Dans la zone de liste Ajouter à, sélectionnez la bibliothèque dans laquelle vous voulez déplacer la déclaration de types.
- Cliquez sur OK. La boîte de dialogue se referme et le type figure maintenant dans la bibliothèque sélectionnée.

Désactiver et supprimer des types

Si vous souhaitez supprimer temporairement une déclaration de types, vous pouvez la désactiver en décochant la case située à gauche de son nom dans l'arborescence de bibliothèques. Vous indiquez ainsi que vous souhaitez conserver la déclaration dans votre bibliothèque, mais en ignorer le contenu pendant la recherche de conflits et le processus d'extraction.

Vous pouvez aussi supprimer définitivement des déclarations de types d'une bibliothèque.

Pour désactiver une déclaration de types

- Dans l'arborescence de bibliothèques, sélectionnez la déclaration de types à désactiver.
- Cliquez sur la barre d'espacement. La coche disparaît de la case figurant à gauche du nom du type.

Pour supprimer une déclaration de types

- Dans l'arborescence de bibliothèques, sélectionnez la déclaration de types à supprimer.
- Dans les menus, choisissez Edition > Supprimer pour supprimer la déclaration de types.

Dictionnaires des substitutions/synonymes

Un **dictionnaire de substitutions** est un ensemble de termes qui permet de regrouper des termes similaires sous un terme cible. Les dictionnaires de substitutions sont gérés dans le panneau du bas de l'onglet Ressources de bibliothèque. Vous pouvez accéder à cette vue avec Vue > Editeur de ressources dans les menus lors d'une session interactive. Autrement, vous pouvez modifier les dictionnaires d'un modèle spécifique dans l'Editeur de modèle.

Vous pouvez définir deux formes de substitution dans ce dictionnaire : **synonymes** et **éléments optionnels**. Vous pouvez cliquer sur les onglets de ce panneau pour passer de l'un à l'autre.

Après avoir exécuté une extraction sur vos données textuelles, vous pouvez trouver plusieurs concepts qui sont des synonymes ou des formes fléchies d'autres concepts. En identifiant les éléments optionnels et les synonymes, vous pouvez forcer le moteur du programme d'extraction à les faire correspondre à un terme cible unique.

La substitution à l'aide de synonymes et d'éléments optionnels réduit le nombre de concepts dans le panneau Résultats d'extraction en les combinant pour former des concepts plus importants et représentatifs avec une fréquence Doc. plus élevée.

Remarque : Pour les ressources de texte japonais, les éléments optionnels ne s'appliquent pas et ne sont pas disponibles. De plus, les synonymes sont gérés un peu différemment pour le texte japonais. Pour plus d'informations, reportez-vous à la section Utilisation du dictionnaire des synonymes pour du texte japonais dans l'annexe A sur p. 383.*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Synonymes

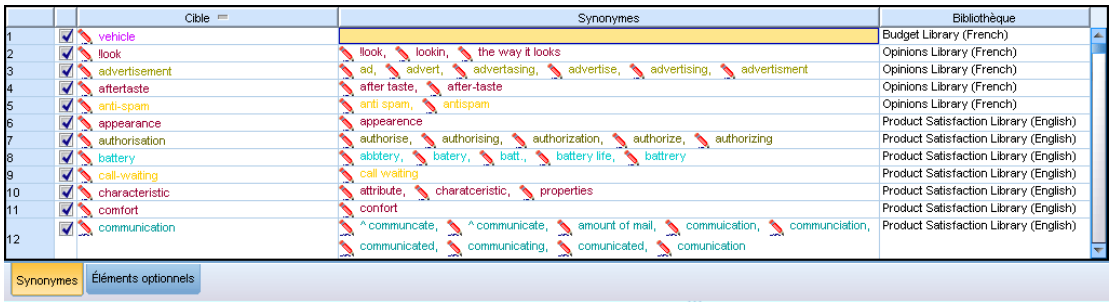
Les synonymes sont des mots ayant le même sens. Vous pouvez également utiliser des synonymes pour regrouper des termes et leurs abréviations, ou pour réunir les mots fréquemment mal orthographiés sous la version correcte du mot. Vous pouvez définir ces synonymes dans l'onglet Synonymes.

Une définition de synonyme est constituée de deux parties. La première est un terme cible, qui est le terme sous lequel vous souhaitez que le moteur du programme d'extraction regroupe tous les termes synonymes. Ce terme cible deviendra probablement le concept apparaissant dans le panneau Résultats d'extraction, sauf s'il est utilisé comme synonyme d'un autre terme cible ou s'il est exclu. La deuxième est la liste de synonymes qui sera regroupée sous le terme cible.

Par exemple, si vous voulez que `automobile` soit remplacé par `véhicule`, `automobile` sera le synonyme et `véhicule` le terme cible.

Vous pouvez entrer n'importe quel mot dans la colonne Synonyme, mais si le mot n'est pas trouvé au cours de l'extraction et si le terme avait une option de mise en correspondance de `Entier`, aucune substitution ne peut être effectuée. En revanche, il n'est pas nécessaire que le terme cible soit extrait pour que les synonymes soient regroupés sous ce terme.

Figure 17-8
Dictionnaire de substitutions, onglet Synonymes



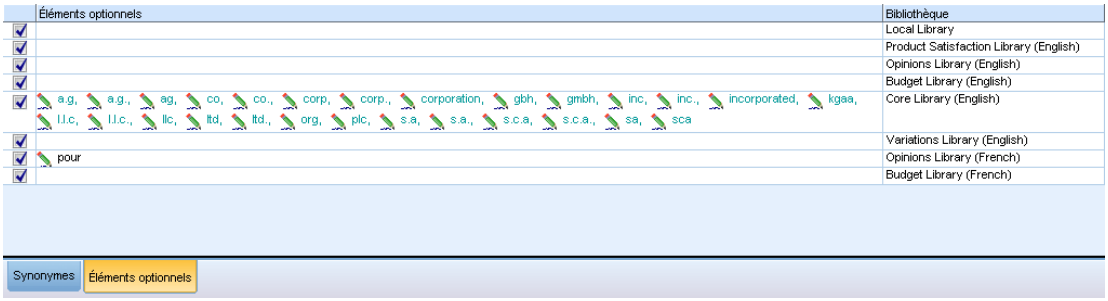
Éléments optionnels

Les éléments optionnels désignent les mots optionnels d'un terme composé qui peuvent être ignorés pendant l'extraction afin de conserver un regroupement de termes similaires, même s'ils apparaissent légèrement différents dans le texte. Les éléments optionnels sont des mots simples dont la suppression d'un terme composé peut créer une correspondance avec un autre terme. Ces mots simples peuvent apparaître à n'importe quel endroit du terme composé, à savoir au début, au milieu ou à la fin. Vous pouvez définir les éléments optionnels dans l'onglet Optionnels.

Par exemple, pour regrouper les termes `ibm` et `ibm corp`, vous devez déclarer que `corp` doit ici être traité comme élément optionnel. Dans un autre exemple, si vous signalez que le terme `accès` est un élément optionnel et que l'extraction renvoie `vitesse d'accès internet` et `vitesse internet`, ils sont regroupés sous le terme qui revient le plus fréquemment.

Remarque : Pour des ressources de texte japonais, il n'existe pas d'onglet Eléments optionnels car les éléments optionnels ne s'appliquent pas. l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Figure 17-9
Dictionnaire de substitutions - Onglet Optionnels



Définition de synonymes

Dans l'onglet Synonymes, vous pouvez entrer une définition de synonyme dans la ligne vide qui figure en haut du tableau. Commencez par définir le terme cible et ses synonymes. Vous pouvez également sélectionner la bibliothèque dans laquelle vous souhaitez stocker la définition. Toutes les occurrences des synonymes sont ensuite regroupées sous le terme cible lors de l'extraction finale. Pour plus d'informations, reportez-vous à la section Ajout de termes sur p. 308.

Par exemple, si les données textuelles fournissent diverses informations sur les télécommunications, vous pouvez rencontrer ces termes : `téléphone cellulaire`, `téléphone portable` et `téléphone mobile`. Vous pouvez alors décider d'indiquer que `cellulaire` et `mobile` sont synonymes de `portable`. Si vous définissez ces synonymes, chaque occurrence extraite de `téléphone cellulaire` et `téléphone mobile` sera considérée comme le même terme que `téléphone portable` et apparaîtra avec elle dans la liste des termes.

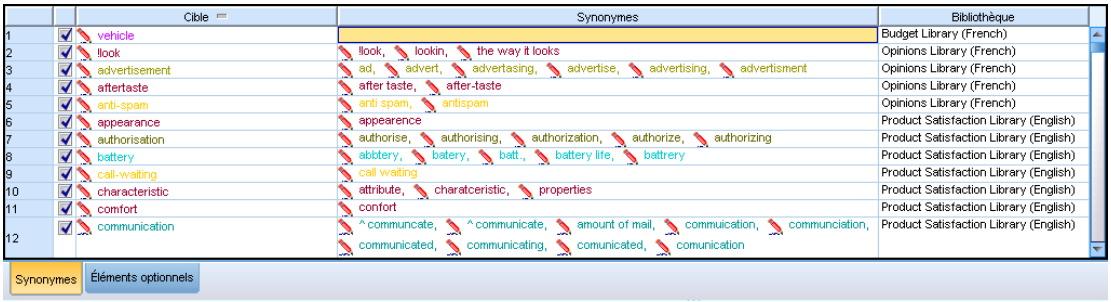
Quand vous créez vos déclarations de types, vous pouvez entrer un terme puis penser à trois ou quatre synonymes de ce terme. Dans ce cas, vous pouvez saisir tous les termes puis votre terme cible dans le dictionnaire de substitution avant de faire glisser les synonymes.

Remarque : Les synonymes sont gérés un peu différemment pour le texte japonais. Pour plus d'informations, reportez-vous à la section Utilisation du dictionnaire des synonymes pour du texte japonais dans l'annexe A sur p. 383.*Remarque* : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

La substitution de synonymes est également appliquée aux formes fléchies (le pluriel, par exemple) du synonyme. En fonction du contexte, il est conseillé de définir des limites régissant le processus de substitution des termes. Des caractères sont alors utilisés pour restreindre le traitement des synonymes :

- **point d'exclamation (!)** Lorsque le point d'exclamation précède directement le synonyme `!synonyme`, cela indique qu'aucune forme fléchie du synonyme ne sera substituée par le terme cible. Néanmoins, un point d'exclamation placé directement avant le terme cible `!target-term` signale qu'aucune partie du mot composé cible ou les variantes ne doit faire l'objet d'une substitution supplémentaire.
- **Astérisque (*).** Un astérisque placé directement après un synonyme, comme `synonyme*`, signifie que vous souhaitez remplacer ce mot par le terme cible. Par exemple, si vous définissez `manage*` comme synonyme et `management` comme terme cible, le mot `managers` est remplacé par le terme cible `management`. Vous pouvez également ajouter un espace et un astérisque après le mot (`synonyme *`) comme `internet *`. Si vous définissez le terme cible `internet` et les synonymes `internet * *` et `web *`, les termes `internet haut débit` et `web indépendant` sont alors remplacés par `internet`. Vous ne pouvez pas utiliser le caractère générique de l'astérisque au début d'un mot ou d'une chaîne dans ce dictionnaire.
- **Caret (^).** Un caret et un espace placés avant le synonyme, comme `^ synonyme`, indiquent que le regroupement de synonymes ne s'applique que lorsque le terme commence par le synonyme. Par exemple, si vous définissez `^ salaire` comme synonyme et `revenus` comme cible et que ces deux termes sont extraits, ils sont regroupés sous le terme `revenus`. En revanche, si « `bulletin de salaire` » et « `revenus` » sont extraits, ils ne sont pas réunis car « `bulletin de salaire` » ne commence pas par « `salaire` ». Un espace doit séparer ce symbole du synonyme.
- **Symbole dollar ($).** Un espace et un symbole dollar placés après le synonyme (`synonyme $`, par exemple) signalent que le regroupement de synonymes ne s'applique que lorsque le synonyme figure à la fin du terme. Par exemple, si vous définissez le synonyme `capital $` et le terme cible `argent` et que ces deux termes sont extraits, ils sont tous deux regroupés sous le terme `argent`. En revanche, si `capital monétaire` et `argent` sont extraits, ils ne sont pas réunis car `capital monétaire` ne se termine pas par `capital`. Un espace doit séparer ce symbole du synonyme.
- **Caret (^) et symbole dollar ($).** Si le caret et le dollar sont utilisés ensemble (par exemple `^ synonym $`), seules les concordances exactes sont mises en correspondance avec le synonyme. Cela signifie qu'aucun mot ne peut apparaître avant ou après le synonyme dans le terme extrait pour que le regroupement s'effectue. Par exemple, vous pouvez définir `^ van $` comme synonyme et `fourgon` comme cible pour que `van` soit regroupé avec `fourgon`, mais pas `ludwig van beethoven`. Par ailleurs, dès lors que vous définissez un synonyme à l'aide des symboles caret et dollar et que ce mot apparaît n'importe où dans le texte source, le synonyme est automatiquement extrait.

Figure 17-10
Dictionnaire de substitutions, onglet Synonymes avec exemple



Remarque : Ces caractères spéciaux et caractères génériques ne sont pas pris en charge pour le texte japonais. Pour plus d'informations, reportez-vous à la section Utilisation du dictionnaire des synonymes pour du texte japonais dans l'annexe A sur p. 383.*Remarque* : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Pour ajouter une entrée de synonyme

- Affichez le panneau de substitution et cliquez sur l'onglet Synonymes dans l'angle inférieur gauche.
- Dans la ligne vide en haut du tableau, entrez le terme cible dans la colonne Cible. Le terme cible que vous entrez apparaît en couleur. Cette couleur représente le type dans lequel le terme apparaît ou a été imposé, le cas échéant. Si le terme apparaît en noir, il n'est alors inclus dans aucune déclaration de types.
- Cliquez dans la deuxième cellule à droite de la cible et entrez l'ensemble de synonymes. Séparez chaque entrée à l'aide du séparateur global défini dans la boîte de dialogue Options. Pour plus d'informations, reportez-vous à la section Définition des options dans le chapitre 8 sur p. 131. Les termes que vous entrez apparaissent en couleur. Cette couleur représente le type dans lequel le terme apparaît. Si le terme apparaît en noir, il n'est alors inclus dans aucune déclaration de types.
- Cliquez dans la dernière cellule pour sélectionner la bibliothèque dans laquelle vous voulez stocker cette définition de synonyme.

  Remarque : Ces instructions indiquent la façon d'effectuer des modifications dans la vue Editeur de ressources ou dans la vue Editeur de modèle. N'oubliez pas que vous pouvez également réaliser ce type de réglage directement à partir du panneau Résultats d'extraction , du panneau Données, du panneau Catégorie ou de la boîte de dialogue Définitions de clusters dans les autres vues. Pour plus d'informations, reportez-vous à la section Affinage des résultats de l'extraction dans le chapitre 9 sur p. 154.

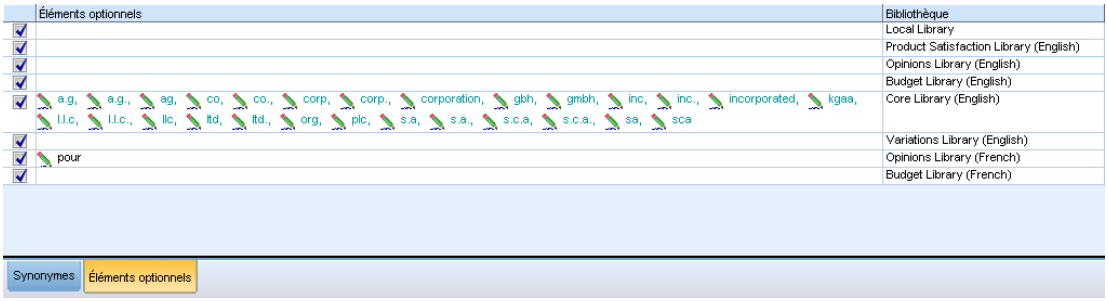
Définition des éléments optionnels

Dans l'onglet Optionnels, vous pouvez définir des éléments optionnels pour n'importe quelle bibliothèque. Ces entrées sont regroupées pour chaque bibliothèque. Dès que vous ajoutez une bibliothèque dans l'arborescence de bibliothèques, une ligne vide d'élément optionnel est ajoutée dans l'onglet Optionnels.

Toutes les entrées sont transformées automatiquement en mots en minuscules. Le moteur d'extraction met indifféremment en correspondance les mots du texte en majuscules et en minuscules.

Remarque : Pour les ressources de texte japonais, les éléments optionnels ne s'appliquent pas et ne sont pas disponibles.

Figure 17-11
Dictionnaire de substitutions - Onglet Optionnels



Remarque : Les termes sont délimités à l'aide du séparateur défini dans la boîte de dialogue Options. Pour plus d'informations, reportez-vous à la section Définition des options dans le chapitre 8 sur p. 131. Si l'élément optionnel que vous entrez contient le même séparateur qu'une partie du terme, faites-le précéder d'une barre oblique inversée.

Pour ajouter une entrée

- Le panneau de substitution étant affiché, cliquez sur l'onglet Optionnels dans l'angle inférieur gauche de l'éditeur.
- Cliquez dans la cellule de la colonne Eléments optionnels de la bibliothèque à laquelle vous souhaitez ajouter cette entrée.
- Entrez l'élément optionnel. Séparez chaque entrée à l'aide du séparateur global défini dans la boîte de dialogue Options. Pour plus d'informations, reportez-vous à la section Définition des options dans le chapitre 8 sur p. 131.

Désactiver et supprimer des substitutions

Vous pouvez supprimer une entrée de façon temporaire. Pour cela, vous devez la désactiver dans votre dictionnaire. Lorsque vous désactivez une entrée, celle-ci est ignorée au cours des extractions.

Vous pouvez aussi supprimer toutes les entrées obsolètes de votre dictionnaire de substitutions.

Pour désactiver une entrée

- Dans votre dictionnaire, sélectionnez l'entrée à désactiver.

- Cliquez sur la barre d'espacement. La coche disparaît de la case figurant à gauche de l'entrée.

  Remarque : vous pouvez également désactiver la case à cocher située à gauche de l'entrée.

Pour supprimer une entrée de synonyme

- Dans votre dictionnaire, sélectionnez l'entrée à supprimer.
- Dans le menu, sélectionnez Edition > Supprimer ou appuyez sur la touche Suppr sur votre clavier. L'entrée ne figure plus dans le dictionnaire.

Pour supprimer une entrée d'élément optionnel

- Dans votre dictionnaire, double-cliquez sur l'entrée à supprimer.
- Supprimez manuellement le terme.
- Appuyez sur Entrée pour appliquer la modification.

Dictionnaires d'exclusions

Un **dictionnaire d'exclusions** est une liste de mots, de phrases ou de chaînes partielles. Tout terme correspondant à ou contenant une entrée dans le dictionnaire d'exclusions sera ignoré ou exclu de l'extraction. La gestion des dictionnaires d'exclusions s'effectue dans le panneau de droite de l'éditeur. En règle générale, vous entrez dans cette liste les mots ou expressions de liaison utilisés pour la continuité du texte, mais qui ne lui apportent rien d'important et risquent d'encombrer les résultats de l'extraction. En ajoutant ces termes au dictionnaire d'exclusions, vous êtes assuré qu'ils ne seront jamais extraits.

La gestion des dictionnaires d'exclusion s'effectue dans le panneau supérieur droit de l'onglet Ressources de bibliothèque dans l'éditeur. Vous pouvez accéder à cette vue avec Vue > Editeur de ressources dans les menus lors d'une session interactive. Autrement, vous pouvez modifier les dictionnaires d'un modèle spécifique dans l'Editeur de modèle.

Figure 17-12
Panneau Dictionnaire d'exclusions

		Exclusions	Bibliothèque
0	☐		
1	☑	any kind of problem	Opinions Library (English)
2	☑	any problems i have	Opinions Library (English)
3	☑	anykinf of problem	Opinions Library (English)
4	☑	as usual	Opinions Library (English)
5	☑	can't wait	Opinions Library (English)
6	☑	1 conseil	Opinions Library (French)
7	☑	bien souvent	Opinions Library (French)
8	☑	en interet	Opinions Library (French)
9	☑	grand h	Opinions Library (French)
10	☑	mais qualité	Opinions Library (French)
11			

Vous pouvez entrer un mot, une expression ou une chaîne partielle dans la ligne vide figurant dans le haut du tableau du dictionnaire d'exclusions. Vous pouvez ajouter des chaînes de caractères à votre dictionnaire d'exclusions sous la forme d'un ou de plusieurs mots, voire de mots partiels à l'aide du caractère générique astérisque. Les entrées déclarées dans le dictionnaire d'exclusions

servent à empêcher l'extraction de ces concepts. Si une entrée est également déclarée ailleurs dans l'interface, par exemple dans une déclaration de types, elle apparaît barrée dans les autres dictionnaires, ce qui indique qu'elle est actuellement exclue. Il n'est pas nécessaire que cette chaîne apparaisse dans les données textuelles ou qu'elle fasse partie d'une déclaration de types pour être appliquée.

Remarque : Si un concept ajouté au dictionnaire d'exclusions sert également de cible dans une entrée de synonyme, la cible et tous ses synonymes sont également exclus. Pour plus d'informations, reportez-vous à la section Définition de synonymes sur p. 317.

Utilisation de caractères génériques (*)

Pour toutes les langues de texte, en-dehors du japonais, vous pouvez utiliser le caractère générique astérisque pour signaler que vous souhaitez que l'entrée d'exclusion soit considérée comme une chaîne partielle. Tous les mots détectés par le moteur du programme d'extraction, et qui commencent ou finissent par l'une des chaînes entrées ici, sont exclus de l'extraction finale. Toutefois, il est interdit d'utiliser les caractères génériques dans deux cas de figure :

- Tiret (-) précédé d'un astérisque, comme *-
- Apostrophe (') précédée d'un astérisque, comme *'s

Table 17-2
Exemples d'entrées d'exclusion

Entrée	Exemple	Résultats
mot	*suivant*	Aucun concept (ou ses termes) n'est extrait s'il contient le mot *suivant*.
expression	*par exemple*	Aucun concept (ou ses termes) n'est extrait s'il contient l'expression *par exemple*.
partiel	*final**	Exclut tous les concepts (ou leurs termes) concordants ou contenant des variations du mot *final*, par exemple `finale`, `finaliste`, `finalement` ou encore `finale 2010`.
partiel	**piste*	Exclut tous les concepts (ou leurs termes) concordants ou contenant des variations du mot *piste*, tels que `trappiste`, `monotypiste`, `copiste`, `endoscopiste`, `alpiste` ou `avant-piste`.

Pour ajouter des entrées

▶ Dans la ligne vide qui figure dans le haut du tableau, entrez un terme. Le terme que vous entrez apparaît en couleur. Cette couleur représente le type dans lequel le terme apparaît. Si le terme apparaît en noir, il n'est alors inclus dans aucune déclaration de types.

Pour désactiver des entrées

Vous pouvez supprimer temporairement une entrée en la désactivant dans votre dictionnaire d'exclusions. Lorsque vous désactivez une entrée, celle-ci est ignorée au cours des extractions.

▶ Dans votre dictionnaire d'exclusions, sélectionnez l'entrée à désactiver.

▶ Cliquez sur la barre d'espacement. La coche disparaît de la case figurant à gauche de l'entrée.

Remarque : vous pouvez également désactiver la case à cocher située à gauche de l'entrée.

Pour supprimer des entrées

Vous pouvez supprimer les entrées inutiles de votre dictionnaire d'exclusions.

- Dans votre dictionnaire d'exclusions, sélectionnez l'entrée à supprimer.
- Dans les menus, sélectionnez Edition > Supprimer. L'entrée ne figure plus dans le dictionnaire.

Chapitre
18

À propos des ressources avancées

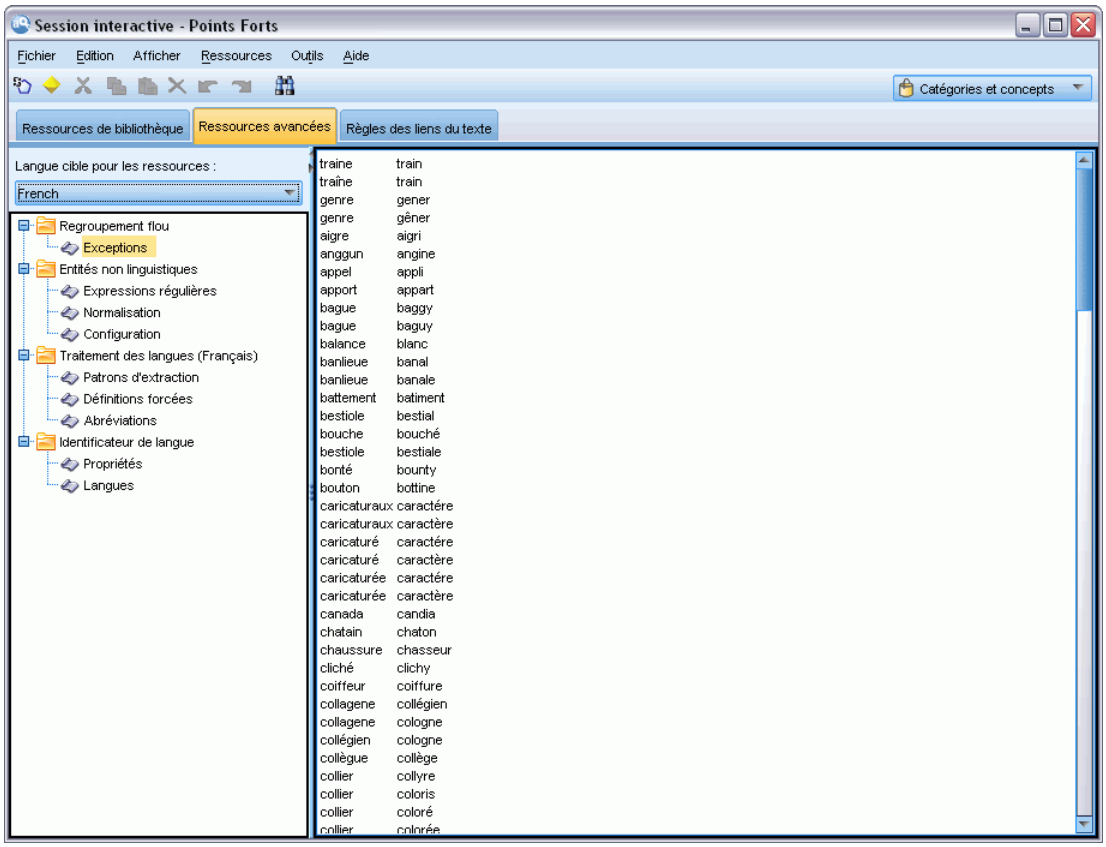
Outre les déclarations de types, les dictionnaires d'exclusions et de substitutions, vous pouvez également utiliser divers paramètres de ressources avancées tels que les paramètres de regroupements flous ou les définitions de type non linguistiques. Vous pouvez trouver ces ressources dans l'onglet Ressources avancées dans la vue Editeur de modèle ou la vue Editeur de ressources.

Important ! Cet onglet n'est pas disponible pour les ressources adaptées au texte japonais.

Lorsque vous allez dans l'onglet Ressources avancées, vous pouvez modifier les informations suivantes :

- **Langue cible pour les ressources.** Permet de sélectionner la langue pour laquelle les ressources seront créées et affinées. Pour plus d'informations, reportez-vous à la section Langue cible pour les ressources sur p. 327.
- **Regroupement flou (Exceptions).** Permet d'exclure des paires de mots de l'algorithme de regroupement flou (correction des fautes d'orthographe). Pour plus d'informations, reportez-vous à la section Regroupement flou sur p. 328.
- **Entités non linguistiques.** Permet d'activer et de désactiver les entités non linguistiques pouvant être extraites, ainsi que les expressions régulières et les règles de normalisation qui sont appliquées lors de leur extraction. Pour plus d'informations, reportez-vous à la section Entités non linguistiques sur p. 329.
- **Gestion des langues.** Permet de déclarer les méthodes spéciales de structuration des phrases (patrons d'extraction et définitions forcées) et d'utilisation des abréviations pour la langue sélectionnée. Pour plus d'informations, reportez-vous à la section Traitement des langues sur p. 334.
- **Identificateur de langue.** Permet de configurer l'identificateur de langue automatique activé lorsque la langue est définie sur Toutes. Pour plus d'informations, reportez-vous à la section Identificateur de langue sur p. 336.

Figure 18-1
Editeur de modèles de Text Mining - Onglet Ressources avancées



Remarque : vous pouvez utiliser la barre d'outils Rechercher/Remplacer pour rechercher des informations rapidement ou pour apporter des modifications globales à une section. Pour plus d'informations, reportez-vous à la section Remplacement sur p. 326.

Pour modifier les ressources avancées

- Localisez et sélectionnez la section de ressources dans laquelle vous souhaitez effectuer l'édition. Le contenu apparaît dans le panneau de droite.
- Si nécessaire, à l'aide du menu ou des boutons de la barre d'outils, coupez, copiez ou collez des éléments.
- Editez les fichiers que vous souhaitez modifier en utilisant les règles de formatage de la section. Les modifications sont enregistrées dès que vous les apportez. Utilisez la flèche d'annulation ou de rétablissement de la barre d'outils pour rétablir vos précédentes modifications.

Recherche

Il peut s'avérer nécessaire de localiser rapidement des informations dans une section particulière. Par exemple, si vous procédez à l'analyse des liens du texte, vous disposez de centaines de macros et définitions de patrons. La fonction Rechercher vous permet de rechercher rapidement une

règle spécifique. Pour rechercher des informations dans une section, utilisez la barre d'outils Rechercher.

Figure 18-2
Barre d'outils Rechercher

Pour utiliser la fonction de recherche

- Localisez et sélectionnez la section de ressources dans laquelle vous souhaitez effectuer la recherche. Son contenu apparaît dans le panneau de droite de l'éditeur.
- Dans les menus, sélectionnez Edition > Rechercher. La barre d'outils Rechercher apparaît dans la partie supérieure droite de la boîte de dialogue Editer les ressources avancées.
- Saisissez la chaîne de mots que vous recherchez dans la zone de texte. Vous pouvez contrôler la casse, les correspondances partielles et le sens de la recherche à l'aide des boutons de la barre d'outils.
- Cliquez sur Rechercher pour lancer la recherche. Si une correspondance est détectée, le texte est mis en surbrillance dans la fenêtre.
- Cliquez de nouveau sur Rechercher pour rechercher la correspondance suivante.

Remarque : Lorsque vous utilisez l'onglet Règles des liens du texte, l'option Rechercher est uniquement disponible lorsque vous consultez le code source.

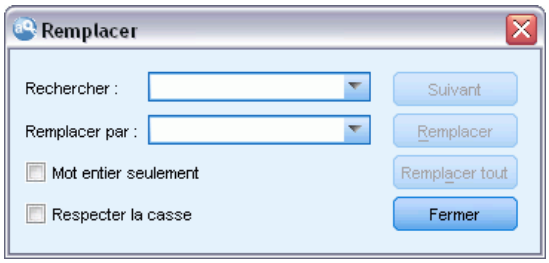
Remplacement

Dans certains cas, il peut s'avérer nécessaire d'effectuer des mises à jour globales à vos ressources avancées. La fonction Remplacer peut vous aider à effectuer des mises à jour homogènes sur vos informations.

Pour utiliser la fonction de remplacement

- Localisez et sélectionnez la section de ressources dans laquelle vous souhaitez effectuer la recherche et le remplacement. Son contenu apparaît dans le panneau de droite de l'éditeur.
- Dans les menus, sélectionnez Edition > Remplacer. La boîte de dialogue Remplacer apparaît.

Figure 18-3
Boîte de dialogue Remplacer



- Dans la zone de texte Rechercher, saisissez la chaîne de mots que vous souhaitez rechercher.

- Dans la zone de texte Remplacer par, saisissez la chaîne que vous souhaitez utiliser à la place du texte trouvé.
- Sélectionnez Rechercher uniquement les mots entiers si vous ne souhaitez rechercher ou remplacer que des mots complets.
- Sélectionnez Respecter la casse si vous ne souhaitez rechercher ou remplacer que les mots dont la casse correspond exactement.
- Cliquez sur Suivant pour rechercher une correspondance. Si une correspondance est détectée, le texte est mis en surbrillance dans la fenêtre. Si vous ne voulez pas remplacer cette correspondance, cliquez de nouveau sur Suivant jusqu'à ce que vous ayez trouvé une correspondance que vous souhaitez remplacer.
- Cliquez sur Remplacer pour remplacer la correspondance sélectionnée.
- Cliquez sur Remplacer pour remplacer toutes les correspondances de la section. Un message indique le nombre de remplacements effectués.
- Lorsque les remplacements sont terminés, cliquez sur Fermer. La boîte de dialogue se ferme.

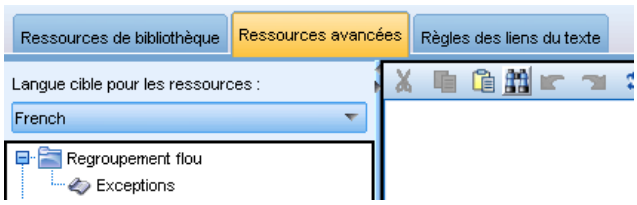
  Remarque : En cas d'erreur de remplacement, vous pouvez annuler l'opération. Pour cela, fermez la boîte de dialogue et choisissez Edition >Annuler dans les menus. Vous devez suivre cette procédure pour chaque modification à annuler.

Langue cible pour les ressources

Les ressources sont créées pour une langue de texte spécifique. La langue pour laquelle ces ressources sont adaptées est définie dans l'onglet Ressources avancées. Si nécessaire, vous pouvez passer à une autre langue en sélectionnant cette langue dans la zone de liste déroulante Langue cible pour les ressources. En outre, la langue répertoriée ici sera la langue utilisée pour tout package d'analyse de texte que vous créez avec ces ressources.

Important ! Vous aurez rarement besoin de modifier la langue de vos ressources. Si vous la modifiez, cela peut générer des problèmes, dans le cas où vos ressources ne correspondraient plus à la langue d'extraction. Toutefois, vous pouvez avoir besoin de modifier une langue si vous projetez d'utiliser l'option `TOUTES` les langues lors de l'extraction, dans le cas où votre texte comporte plusieurs langues. En modifiant la langue, vous pouvez accéder, par exemple, aux ressources de traitement de la langue des patrons d'extraction, et aux abréviations et forcer des définitions pour la langue secondaire qui vous intéresse. Toutefois, gardez à l'esprit que vous devez à nouveau définir la langue sur la langue principale de votre extraction, avant de publier, d'enregistrer des modifications apportées aux ressources ou d'exécuter une nouvelle extraction.

Figure 18-4
Définition de la langue cible



Regroupement flou

Dans le nœud Text Mining et dans Paramètres d'extraction, si vous sélectionnez Traitement des fautes de frappe. Nombre de caractères minimum requis :, vous avez activé l'algorithme de regroupement flou.

Le regroupement flou permet de regrouper les mots dont l'orthographe est souvent incorrecte ou proche en supprimant temporairement toutes les voyelles (à l'exception de la première voyelle) et les consonnes doubles ou triples des mots extraits, et en les comparant ensuite afin de déterminer s'ils sont identiques. Pendant le processus d'extraction, la fonction de regroupement flou est appliquée aux termes extraits et les résultats sont comparés afin de déterminer s'il existe des correspondances. Si tel est le cas, les termes initiaux sont regroupés dans la liste d'extraction finale. Ils sont placés sous le terme qui compte le plus d'occurrences dans les données.

Remarque : Si les deux termes comparés sont attribués à différents types, à l'exclusion du type `<Unknown>`, la méthode de regroupement flou ne doit pas être appliquée à cette paire. Autrement dit, les termes doivent appartenir au même type ou au type `<Unknown>` afin d'appliquer la méthode.

Si vous avez activé cette fonction et remarqué que deux mots orthographiés de façon similaire étaient regroupés de façon erronée, vous pouvez les exclure du regroupement flou. Pour cela, entrez les paires dont la mise en correspondance est incorrecte dans la section Exceptions de l'onglet Ressources avancées. Pour plus d'informations, reportez-vous à la section À propos des ressources avancées sur p. 324.

L'exemple suivant montre le fonctionnement du regroupement flou. Si le regroupement flou est activé, ces mots apparaissent similaires et sont mis en correspondance comme suit :

```
color   -> colr                mountain -> montn
colour  -> colr                montana  -> montn

modeling  -> modlng            furniture -> furntr
modelling -> modlng            furnature -> furntr
```

Dans l'exemple précédent, vous souhaiterez vraisemblablement que les termes `paysage` et `passage` ne soient pas regroupés. Par conséquent, vous pouvez les entrer dans la section Exceptions comme suit :

```
paysage      passage
```

Important ! Dans certains cas, les exceptions de regroupement flou n'empêchent pas 2 mots d'être mis en paire car certaines règles de synonymes s'appliquent. Dans ce cas, vous pouvez essayer d'entrer des synonymes à l'aide du caractère générique du point d'exclamation (`!`) pour que les mots ne soient pas des synonymes dans les résultats. Pour plus d'informations, reportez-vous à la section Définition de synonymes dans le chapitre 17 sur p. 317.

Règles de formatage pour le Regroupement flou (Exceptions)

- Définissez une seule paire d'exception par ligne.
- Utilisez des mots simples ou composés.
- N'utilisez que des minuscules dans les mots. Les mots en majuscules seront ignorés.
- Utilisez le caractère `TAB` pour séparer les deux mots de chaque paire.

Entités non linguistiques

Lorsque vous travaillez avec certains types de données, vous pouvez souhaiter extraire des dates, des numéros de Sécurité sociale, des pourcentages ou toute autre entité non linguistique. Ces entités sont déclarées explicitement dans le fichier de configuration, dans lequel vous pouvez activer ou désactiver les entités. Pour plus d'informations, reportez-vous à la section Configuration sur p. 332. Pour optimiser la sortie dans le moteur du programme d'extraction, l'entrée issue du traitement non linguistique est normalisée pour regrouper les entités semblables en fonction de formats prédéfinis. Pour plus d'informations, reportez-vous à la section Normalisation sur p. 332.

Remarque : vous pouvez activer et désactiver l'extraction des entités non linguistiques dans les paramètres d'extraction.

Entités non linguistiques disponibles

Les entités non linguistiques du tableau suivant peuvent être extraites. Le nom de type est entre parenthèses.

Adresses (`<Address>`)	Organisations (`<Organization>`)
Acides aminés (`<Aminoacid>`)	Pourcentages (`<Percent>`)
Devises (`<Currency>`)	Produits (`<Product>`)
Dates (`<Date>`)	Protéines (`<Gene>`)
Délai (`<Delay>`)	Numéros de téléphone (`<PhoneNumber>`)
Chiffres (`<Digit>`)	Temps (`<Time>`)
Adresses électroniques (`<email>`)	Sécurité sociale (Etats-Unis) (`<SocialSecurityNumber>`)
Adresses HTTP/URL (`<url>`)	Poids et mesures (`<Weights-Measures>`)
Adresse IP (`<IP>`)	

Nettoyage du texte pour traitement

Avant l'extraction des entités non linguistiques, le texte d'entrée est nettoyé. Durant cette étape, les modifications temporaires suivantes sont effectuées afin que les entités non linguistiques puissent être identifiées et extraites ainsi :

- Toute séquence de deux espaces ou plus est remplacée par un seul espace.
- Les tabulations sont remplacées par un espace.
- Les caractères de fin de ligne uniques ou les caractères de séquence sont remplacés par un espace, tandis que les séquences de fin de ligne multiples sont marquées comme la fin d'un paragraphe. Une fin de ligne peut être indiquée par des retours chariot (CR) et des sauts de ligne (LF) ou les deux à la fois.
- Les balises HTML et XML sont temporairement supprimées et ignorées.

Expressions régulières

Lors de l'extraction d'entités non linguistiques, vous pouvez modifier ou ajouter des définitions d'expressions régulières à utiliser pour identifier des expressions régulières. Pour cela, rendez-vous dans la section Expressions régulières dans l'onglet Ressources avancées. Pour plus d'informations, reportez-vous à la section À propos des ressources avancées sur p. 324.

Le fichier est divisé en plusieurs sections distinctes. La première section s'intitule `[macros]`. Une autre section peut venir s'y ajouter, pour chaque entité non linguistique. Vous pouvez ajouter des sections à ce fichier. Dans chaque section, les règles sont numérotées (*regexp1*, *regexp2*, etc.). Ces règles doivent être numérotées de manière séquentielle de 1 à *n*. Toute interruption dans la numérotation entraîne la suspension du traitement de ce fichier.

Dans certains cas, une entité varie en fonction de la langue. Une entité est considérée comme variant en fonction de la langue si son paramètre de langue a une valeur autre que 0 dans le fichier de configuration. Pour plus d'informations, reportez-vous à la section Configuration sur p. 332. Lorsqu'une entité varie en fonction de la langue, la langue doit être utilisée comme préfixe du nom de la section (par exemple, `[english/PhoneNumber]`). Si l'entité `PhoneNumber` adopte la valeur de langue 2, cette section contient des règles applicables uniquement aux numéros de téléphone anglais.

Important ! Si vous modifiez ce fichier (ou un autre) dans l'éditeur et que le moteur du programme d'extraction ne fonctionne plus comme vous le souhaitez, sélectionnez l'option Rétablir les valeurs d'origine dans la barre d'outils pour rétablir le fichier d'origine livré. Ce fichier requiert un certain niveau de connaissance des expressions régulières. Pour obtenir une assistance dans ce domaine, contactez IBM Corp..

Caractères spéciaux. [] {} () \ * + ? | ^ $

Tous les caractères s'auto-correspondent à l'exception des caractères spéciaux suivants, qui sont utilisés dans un but spécifique dans les expressions : `.[{()\*+?|^$` Pour pouvoir être utilisés tels quels, ces caractères doivent être précédés d'une barre oblique inverse (\) dans la définition.

Par exemple, si vous essayez d'extraire des adresses Web, le point est très important pour l'entité, par conséquent, vous devez le faire précéder d'une barre oblique inverse :

```
www\.[a-z]+\.[a-z]+
```

Opérateurs de répétition et quantifieurs ? + * {}

Pour rendre les définitions plus flexibles, vous pouvez utiliser plusieurs caractères génériques qui sont standard dans les expressions régulières. Il s'agit de `*` `?` `+`

- *L'astérisque* * indique qu'il y a *zéro occurrence ou plus* de la chaîne précédente. Par exemple, `ab*c` correspond à « *ac* », « *abc* », « *abbbc* », etc.
- *Le signe plus* + indique qu'il y a *une ou plusieurs occurrences* de la chaîne précédente. Par exemple, `ab+c` correspond à « *abc* », « *abbc* », « *abbbc* », mais pas à « *ac* ».
- *Le point d'interrogation ?* indique qu'il y a *zéro ou une occurrence* de la chaîne précédente.

Par exemple, `modell?ing` correspond à la fois à « *modeling* » et à « *modeling* ».

- *Limiter la répétition avec des accolades {}* indique les limites de la répétition. Par exemple :
  - `[0-9]{n}` correspond à un chiffre répété exactement *n* fois.
    Par exemple, `[0-9]{4}` correspondra à "*1998*", mais ni à "*33*" ni à "*19983*".
  - `[0-9]{n,}` correspond à un chiffre répété *n fois ou plus*.
    Par exemple, `[0-9]{3,}` correspondra à "*199*" ou à "*1998*", mais pas à "*19*".
  - `[0-9]{n,m}` correspond à un chiffre répété entre *n et m fois inclus*.
    Par exemple, `[0-9]{3,5}` correspondra à "*199*", "*1998*" ou à "*19983*", mais pas à "*19*" ni à "*199835*".

Espaces facultatifs et traits d'union

Dans certains cas, vous voulez inclure un espace facultatif dans une définition. Par exemple, si vous vouliez extraire des devises telles que « *pesos uruguayens* », « *peso uruguayen* », « *pesos Uruguay* », « *peso Uruguay* », « *pesos* » ou « *peso* », vous devriez tenir compte du fait qu'il peut y avoir deux mots séparés par un espace. Dans ce cas, cette définition devrait être écrite sous la forme `?pesos?( uruguayens| Uruguay )`. Puisque *uruguayen* ou *Uruguay* sont précédés par un espace lorsqu'ils sont utilisés avec *pesos*/*peso*, l'espace facultatif doit être défini à l'intérieur de la séquence facultative `( uruguayens| Uruguay)`. Si l'espace ne figurait pas dans la séquence facultative (par exemple `?pesos ?(uruguayens|Uruguay)`), il ne trouverait pas de correspondance avec "*pesos*" ou "*peso*" puisque l'espace serait requis.

Si vous cherchez des séries d'éléments incluant des traits d'union (-) dans une liste, le trait d'union doit être défini en dernier. Par exemple, si vous cherchez une virgule (,) ou un trait d'union (-), utilisez `[,-]` et non pas `[-,]`.

Ordre des chaînes dans les listes et les macros

Vous devez toujours définir la séquence la plus longue avant une plus courte, sinon la plus longue ne sera jamais lue puisque la correspondance se fera sur la plus courte. Par exemple, si vous cherchiez des chaînes "*janvier*" ou "*janv*", vous devez définir "*janvier*" avant "*janv*". Ainsi, par exemple `(janvier|janv)` et non `(janv|janvier)`. Cela s'applique aussi aux macros, puisque les macros sont des listes de chaînes.

Ordre des règles dans la section des définitions

Définissez une seule règle par ligne. Dans chaque section, les règles sont numérotées (*regexp1*, *regexp2*, etc.). Ces règles doivent être numérotées de manière séquentielle de 1 à *n*. Toute interruption dans la numérotation entraîne la suspension du traitement de ce fichier. Pour désactiver une entrée, placez le symbole # au début de la ligne utilisée pour définir l'expression régulière. Pour activer une entrée, supprimez le caractère # en début de ligne.

Dans chaque section, les règles les plus spécifiques doivent être définies avant les plus générales afin de garantir un traitement correct. Par exemple, si vous cherchiez une date au format "*mois année*" et au format "*mois*", la règle "*mois année*" doit être définie avant la règle "*mois*". Voici comment elle doit être définie :

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}
```

```
#@# January
regexp2=$(MONTH)
```

et non pas

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

Utilisation des macros dans les règles

Chaque fois qu'une séquence spécifique est utilisée dans plusieurs règles, vous pouvez utiliser une macro. Ensuite, si vous devez modifier la définition de cette séquence, vous aurez à la modifier une seule fois et non dans toutes les règles y faisant référence. Par exemple, en supposant que vous ayez la macro suivante :

```
MOIS=((janvier|février|mars|avril|juin|juillet|août|septembre|octobre|
novembre|décembre)|(jan|fév|mar|avr|mai|juin|juil|aoû|sep|oct|nov|déc)(\.)?)
```

Chaque fois que vous faites référence au nom de la macro, il doit être mis au format `$()`, comme dans : `regexp1=$(MOIS)`

Toutes les macros doivent être définies dans la section `[macros]`.

Normalisation

Lors de l'extraction d'entités non linguistiques, les entités rencontrées sont normalisées pour regrouper les entités semblables en fonction de formats prédéfinis. Par exemple, les symboles de devise et leur équivalent en toutes lettres sont considérés comme étant identiques. Les entrées de normalisation sont stockées dans la section Normalisation de l'onglet Ressources avancées. Pour plus d'informations, reportez-vous à la section À propos des ressources avancées sur p. 324. Le fichier est divisé en plusieurs sections distinctes.

Important ! Ce fichier s'adresse à des utilisateurs avertis uniquement. Il est peu probable que vous ayez à modifier ce fichier. Pour obtenir une assistance dans ce domaine, contactez IBM Corp..

Règles de formatage pour Normalisation

- Ajoutez uniquement une entrée de normalisation par ligne.
- Respectez bien les sections de ce fichier. Vous ne pouvez pas ajouter de nouvelles sections.
- Pour désactiver une entrée, entrez le symbole # au début de la ligne concernée. Pour activer une entrée, supprimez le caractère # en début de ligne.

Configuration

Vous pouvez activer et désactiver les types d'entité non linguistique que vous souhaitez extraire dans le fichier de configuration des entités non linguistiques. Désactivez les entités dont vous n'avez pas besoin pour diminuer le temps de traitement nécessaire. Pour cela, rendez-vous dans la section Configuration de l'onglet Ressources avancées. Pour plus d'informations, reportez-vous à la section À propos des ressources avancées sur p. 324. Si l'extraction non linguistique est activée,

le moteur du programme d'extraction lit ce fichier de configuration lors du processus d'extraction afin de déterminer les types d'entité non linguistique à extraire.

La syntaxe de ce fichier est la suivante :

#name<TAB>*Langue*<TAB>*Code*

Table 18-1
Syntaxe du fichier de configuration

Etiquette de colonne	Description
`#name`	Libellé selon lequel les entités non linguistiques seront référencées dans les deux autres fichiers nécessaires à l'extraction d'entités non linguistiques. Les mots utilisés ici distinguent les majuscules des minuscules.
`Langue`	Langue des documents. Il est préférable de sélectionner une langue précise. Toutefois, vous pouvez également choisir l'option N'importe. Les options suivantes sont proposées : `0` = N'importe quelle langue étant utilisée chaque fois qu'une expression régulière n'est pas spécifique à une langue et pourrait être utilisée dans plusieurs modèles avec différentes langues, par exemple une adresse IP/URL/e-mail ; `1` = Français ; `2` = Anglais ; `4` = Allemand ; `5` = Espagnol ; `6` = Néerlandais ; 8 = Portugais ; `10` = Italien.
`Code`	Code de catégories grammaticales. La plupart des entités prennent la valeur "s", sauf rares exceptions. Valeurs possibles : `s` = mot vide ; `a` = adjectif ; `n` = nom. Lorsque cette option est activée, les entités non linguistiques sont d'abord extraites, puis les patrons d'extraction sont appliqués afin d'identifier le rôle dans un contexte plus large. Par exemple, les pourcentages se voient attribuer la valeur "a." Imaginons que la valeur 30 % soit extraite en tant qu'entité non linguistique. Elle sera identifiée comme étant un adjectif. Par conséquent, si votre texte contient les mots "30%" augmentation salaire, l'entité non linguistique "30%" est incluse dans le patron de catégorie grammaticale "ann" (adjectif nom nom).

Ordre dans la définition des entités

L'ordre dans lequel les entités sont déclarées dans ce fichier est important car il affecte leur mode d'extraction. Ces entrées sont appliquées dans l'ordre dans lequel elles sont répertoriées. Toute modification de l'ordre se répercute sur les résultats. Les entités non linguistiques les plus spécifiques doivent être définies avant les plus générales.

Par exemple, l'entité non linguistique "`Acide aminé`" est définie par :

```
regexp1=($(AA)-?$(NUM))
```

où `$(AA)` correspond à "`(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)`", qui sont des séquences de trois lettres spécifiques correspondant à des acides aminés particuliers.

D'autre part, l'entité non linguistique « `Gène` » est plus générale et elle est définie par :

```
regexp1=p[0-9]{2,3}
regexp2=[a-z]{2,4}-?[0-9]{1,3}-?[r]
regexp3=[a-z]{2,4}-?[0-9]{1,3}-?p?
```

Si « `Gène` » est défini avant « `Acide aminé` » dans la section Configuration, alors « `Acide aminé` » ne sera jamais mis en correspondance, puisque `regexp3` de « `Gène` » sera toujours mis en correspondance en premier.

Règles de formatage pour la configuration

- Utilisez le caractère `TAB` pour séparer les entrées d'une colonne.
- Ne supprimez aucune ligne.
- Respectez la syntaxe indiquée dans le tableau précédent.
- Pour désactiver une entrée, entrez le symbole # au début de la ligne concernée. Pour activer une entité, supprimez le caractère # en début de ligne.

Traitement des langues

Chaque langue moderne exprime des idées, structure des phrases et utilise des abréviations d'une façon particulière. Dans la section Gestion des langues, vous pouvez éditer les patrons d'extraction, forcer des définitions pour ces patrons et indiquer des abréviations pour la langue que vous avez sélectionnée dans la liste déroulante Langue.

- Patrons d'extraction
- Définitions forcées
- Abréviations

Patrons d'extraction

Lors de l'extraction d'informations à partir de vos documents, le moteur du programme d'extraction applique une série de patrons d'extraction à un « ensemble » de mots dans le texte afin d'identifier ceux (mots et expressions) susceptibles d'être extraits. Vous pouvez ajouter ou modifier les patrons d'extraction.

Les catégories grammaticales incluent les éléments grammaticaux tels que substantifs, adjectifs, participes passés, déterminants, prépositions, coordonnants, prénoms, sigles et particules. Une série de ces éléments constitue un patron d'extraction. Dans les produits de Text Mining IBM Corp., chaque catégorie grammaticale est représentée par un caractère unique afin de faciliter la définition des patrons. Par exemple, un adjectif est représenté par la lettre (minuscule) *a*. L'ensemble de codes pris en charge apparaît par défaut en haut de chaque fichier de patron d'extraction par défaut, conjointement avec une série de patrons et des exemples de chaque patron pour illustrer chaque code utilisé.

Règles de formatage applicables aux patrons d'extraction

- Un patron par ligne.
- Utilisez le caractère `#` en début de ligne pour désactiver un patron.

L'ordre dans lequel vous répertoriez les patrons d'extraction est très important car une séquence de mots donnée est lue une seule fois par le moteur du programme d'extraction et est affectée au premier patron d'extraction pour lequel le moteur détecte une correspondance.

Définitions forcées

Lors de l'extraction d'informations à partir de vos documents, le moteur du programme d'extraction analyse le texte et identifie la catégorie grammaticale pour chaque mot rencontré. Dans certains cas, un mot peut présenter différents rôles en fonction du contexte. Si vous forcez un mot à prendre un rôle grammatical particulier ou si vous excluez complètement le mot du traitement, vous pouvez le faire dans la section Définition forcée dans l'onglet Ressources avancées. Pour plus d'informations, reportez-vous à la section À propos des ressources avancées sur p. 324.

Pour imposer une catégorie grammaticale pour un mot donné, vous devez ajouter une ligne à cette section à l'aide de la syntaxe suivante :

```
terme:code
```

Table 18-2
Description de la syntaxe

Entrée	Description
`terme`	Un nom de terme.
`code`	Code à caractère unique représentant la catégorie grammaticale. Vous pouvez répertorier jusqu'à six codes de catégorie grammaticale différents par expression uniterme. En outre, vous pouvez arrêter l'extraction d'un mot en mots ou phrases composés en utilisant le code `s` (en minuscule), par exemple : `additional:s`.

Règles de formatage applicables aux définitions imposées.

- Une ligne par mot.
- Les termes ne peuvent pas contenir le signe « deux-points ».
- Utilisez le code `s` (lettre minuscule) en tant que code de catégorie grammaticale pour arrêter définitivement l'extraction d'un mot.
- Utilisez jusqu'à six codes de catégorie grammaticale par ligne. Les codes de catégories grammaticales pris en charge sont affichés dans la section Patrons d'extraction. Pour plus d'informations, reportez-vous à la section Patrons d'extraction sur p. 334.
- Utilisez l'astérisque (*) en tant que caractère générique à la fin d'une chaîne pour les correspondances partielles. Par exemple, si vous saisissez `add*:s`, des mots tels que `additionner`, `addition`, `additionnel`, `addenda` et `additif` ne sont jamais extraits en tant que mot ou dans le cadre d'un mot composé. Toutefois, si une correspondance de mot est explicitement déclarée en tant que terme dans un dictionnaire compilé ou dans les définitions imposées, le mot sera extrait. Par exemple, si vous saisissez à la fois `add*:s` et `addendum:n`, le mot `addendaum` sera extrait s'il est détecté dans le texte.

Abréviations

Lorsque le moteur du programme d'extraction traite le texte, il interprète généralement tout point rencontré comme marquant la fin d'une phrase. La plupart du temps, cela s'avère correct ; toutefois, cette gestion des signes de ponctuation que sont les points ne s'applique pas lorsque le texte contient des abréviations.

Si vous extrayez des termes de votre texte et que vous vous apercevez que certaines abréviations ont été mal gérées, vous devez déclarer ces dernières de manière explicite dans cette section.

Remarque : si l'abréviation apparaît déjà dans une définition de synonyme ou est définie en tant que terme dans une déclaration de types, il n'est pas nécessaire d'ajouter l'entrée d'abréviation ici.

Règles de formatage applicables aux abréviations

- Définissez uniquement une abréviation par ligne.

Identificateur de langue

Bien qu'il soit préférable de sélectionner les langues spécifiques pour les données textuelles que vous analysez, vous pouvez également activer l'option Toutes lorsque le texte peut être dans plusieurs langues différentes ou inconnues. L'option de langue Toutes utilise un moteur de reconnaissance automatique de la langue, appelé Identificateur de langue. L'identificateur de langue analyse les documents afin d'identifier ceux dont la langue est prise en charge et applique automatiquement les dictionnaires internes les mieux adaptés à chaque fichier lors de l'extraction. L'option Toutes est régie par les paramètres des sections Propriétés.

Propriétés

La configuration de l'identificateur de langue s'effectue à l'aide des paramètres de cette section. Le tableau suivant décrit les paramètres que vous pouvez définir dans la section Identificateur de langue - Propriétés dans l'onglet Ressources avancées. Pour plus d'informations, reportez-vous à la section À propos des ressources avancées sur p. 324.

Table 18-3
Description des paramètres

Paramètre	Description
NUM_CHARS	Indique le nombre de caractères que doit lire le moteur du programme d'extraction afin de pouvoir déterminer la langue du texte. Plus le nombre est petit, plus la langue est identifiée rapidement. En revanche, un paramètre élevé garantit une plus grande fiabilité de l'identification. Si vous paramétrez cette valeur sur 0, le programme d'extraction lit l'ensemble du document.
USE_FIRST_SUPPORTED_LANGUAGE	Indique si le moteur du programme d'extraction utilise la première langue prise en charge détectée par l'identificateur de langue. Si vous paramétrez cette valeur sur 1, le système utilise la première langue prise en charge. Si vous paramétrez cette valeur sur 0, le système utilise la langue par défaut.
FALLBACK_LANGUAGE	Indique la langue à utiliser si l'identificateur détecte une langue non prise en charge. Les valeurs possibles sont l'anglais, le français, l'allemand, l'espagnol, le néerlandais, l'italien et l'option ignore. Si vous choisissez la valeur ignore, les documents dont la langue n'est pas prise en charge sont ignorés.

Langues

L'identificateur de langue prend en charge de nombreuses langues. Vous pouvez éditer la liste des langues dans la section Identificateur de langue - Langues de l'onglet Ressources avancées.

Vous pouvez considérer d'éliminer les langues de cette liste qui sont peu susceptibles d'être utilisées car plus la liste compte de langues, plus grandes sont les chances d'obtenir des faux positifs et une plus médiocre performance. Il n'est toutefois pas possible d'ajouter des langues à ce fichier. Placez les langues les plus probables en haut de la liste, afin d'aider l'identificateur de langue à trouver plus rapidement des correspondances à vos documents.

Chapitre
19

A propos des règles des liens du texte

L'analyse des liens du texte (TLA) est une technologie de mise en correspondance de patrons, utilisée pour extraire les relations trouvées dans votre texte à l'aide d'un ensemble de règles. Lorsque l'analyse des liens du texte est activée pour l'extraction, les données du texte sont évaluées en fonction de ces règles. Lorsqu'une correspondance est trouvée, le patron de l'analyse des liens du texte est extrait et présenté. Ces règles sont définies dans l'onglet Règles des liens du texte.

Par exemple, extraire des informations sur une organisation peut ne pas représenter assez d'intérêt pour vous, mais grâce à l'analyse des liens du texte, vous pouvez également vous familiariser avec les liens existant entre différentes organisations ou avec les personnes associées à cette organisation. La TLA peut également permettre d'extraire des opinions sur différents sujets comme le ressenti des gens face à un produit ou une expérience.

Pour bénéficier de l'analyse TLA, vous devez disposer de ressources contenant des règles de liens du texte (TLA). Lorsque vous sélectionnez un modèle, vous pouvez voir les modèles qui ont des règles TLA selon qu'ils ont ou non une icône dans la colonne TLA.

Figure 19-1
Boîtes de dialogue Modèle affichant la colonne TLA

Modèle	Propriét...	Version	Date	Annotati...	TLA (A...	Lan...
Bank CRM (English)	claired	1	sept.-30...			English
Insurance CRM (English)	claired	1	sept.-30...			English
Ads Opinions (English)	claired	1	sept.-17...			English
Bank Satisfaction Opinio...	claired	1	sept.-17...			English
Security Intelligence (En...	claired	1	sept.-30...			English

Les patrons d'analyse des liens du texte sont extraits des données textuelles pendant la phase de mise en correspondance des patrons du processus d'extraction. Pendant cette phase, les règles sont comparées aux données textuelles et lorsqu'une correspondance est trouvée, ces informations sont extraites en tant que patron. Il est possible que vous exigiez davantage de l'analyse des liens du texte ou que vous souhaitiez modifier les critères des correspondances. Dans ces cas, vous pouvez affiner les règles pour les adapter à vos besoins spécifiques. Pour cela, utilisez l'onglet Règles des liens du texte.

Remarque : La prise en charge des variables a été suspendue dans la version 13. Veuillez utiliser des macros à la place. Pour plus d'informations, reportez-vous à la section Utilisation des macros sur p. 348.

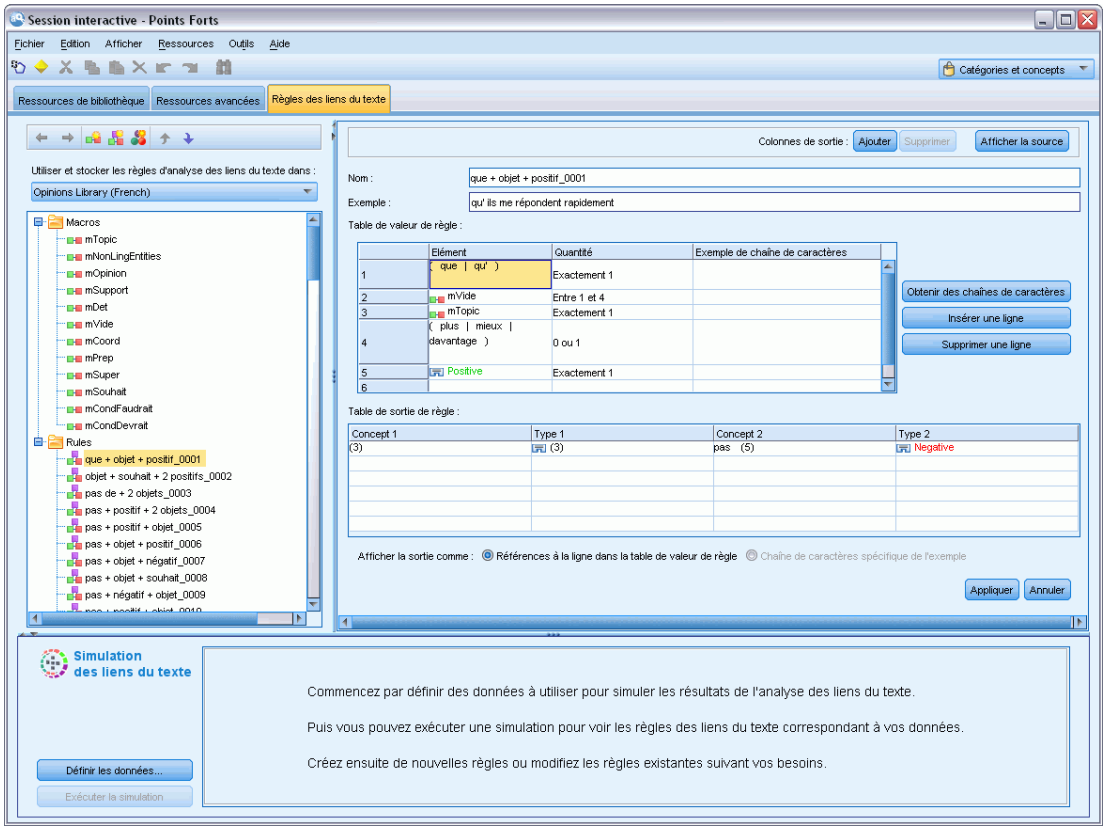
Où travailler sur les règles des liens du texte

Vous pouvez modifier et créer des règles directement dans l'onglet Règles des liens du texte de la vue Editeur de modèle ou Editeur de ressources. Pour obtenir un aperçu de la façon dont les règles peuvent correspondre au texte, vous pouvez effectuer une simulation dans cet onglet. Pendant la simulation, une extraction est effectuée sur l'échantillon de données de simulation uniquement et les règles des liens du texte sont appliquées pour savoir si des patrons correspondants existent. Toute règle qui correspond au texte apparaît ensuite dans le panneau de simulation. En fonction

de ces correspondances, vous pouvez choisir d'éditer les règles et des macros pour modifier les critères de correspondance du texte.

Contrairement à d'autres ressources avancées, les règles TLA sont spécifiques à la bibliothèque ; c'est pourquoi vous pouvez uniquement utiliser les règles TLA d'une seule bibliothèque à la fois. Depuis Editeur de modèle ou Editeur de ressources, allez dans l'onglet Règles des liens du texte. Dans cet onglet, vous pouvez spécifier la bibliothèque de votre modèle qui contient les règles TLA que vous souhaitez utiliser ou modifier. C'est pourquoi nous vous conseillons fortement de stocker toutes vos règles dans une seule bibliothèque, à moins que vous ne le souhaitiez pas pour une raison particulière.

Figure 19-2
Onglet Règles des liens du texte



Important ! Cet onglet n'est pas disponible pour les ressources en japonais.

Où commencer

Il existe plusieurs façons de commencer à travailler dans l'éditeur de l'onglet Règles des liens du texte :

- Commencez par simuler des résultats avec un échantillon de texte et éditez ou créez des règles de correspondance en fonction de la façon dont l'ensemble de règles actuel extrait les patrons des données de simulation.

- Créez une nouvelle règle à partir de notes ou modifiez une règle existante.
- Travaillez directement dans la vue des sources.

Quand éditer ou créer des règles

Bien que les règles des liens du texte fournies avec chaque modèle soient généralement adaptées à l'extraction de nombreuses relations de votre texte, qu'elles soient simples ou compliquées, il se peut que vous souhaitiez parfois modifier ces règles ou en créer de nouvelles. Par exemple :

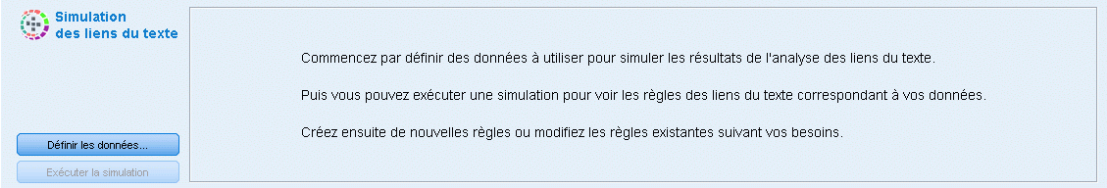
- pour capturer une idée ou une relation qui n'est pas extraite avec des règles existantes en créant une nouvelle règle ou une nouvelle macro.
- Pour modifier le comportement par défaut d'un type que vous avez ajouté aux ressources. Ceci nécessite généralement la modification d'une macro telle que `mTopic` ou `mNonLingEntities`. Pour plus d'informations, reportez-vous à la section Macros spécifiques : mTopic, mNonLingEntities, SEP sur p. 352.
- Pour ajouter de nouveaux types aux macros et règles d'analyse des liens du texte existantes. Par exemple, si vous trouvez que le type `<Organization>` est trop large, vous pouvez créer de nouveaux types pour les organisations dans différents secteurs du marché tels que `<Pharmaceuticals>`, `<Car Manufacturing>`, `<Finance>`, etc. Dans ce cas, vous devez modifier les règles d'analyse des liens du texte et/ou créer une macro pour prendre en compte ces nouveaux types et les traiter en conséquence.
- Pour ajouter des types à une règle d'analyse des liens du texte existante. Par exemple, supposons que vous ayez une règle qui capture le texte suivant `Pierre Martin appelle Martine Dupond`, et vous voulez que cette règle, qui capture les communications téléphoniques, capture également les échanges par email. Vous pourriez ajouter le type d'entité non linguistique pour les e-mails à la règle, afin que celle-ci capture également du texte comme par exemple : `pierremartin@aaa-email.com a envoyé un e-mail à martinedupond@aaa-email.com`
- Pour modifier légèrement une règle existante au lieu d'en créer une nouvelle. Par exemple, supposons que vous avez une règle correspondant au texte suivant `xyz est très bien`, et vous voulez que cette règle capture également `xyz est très très bien`.

Simulation des résultats d'analyse des liens du texte

Afin de définir plus facilement de nouvelles règles d'analyse des liens du texte ou de mieux comprendre la façon dont certaines phrases sont mises en correspondance pendant l'analyse des liens du texte, il est souvent utile de prendre un extrait de texte et d'effectuer une simulation. Pendant la simulation, une extraction est effectuée sur l'échantillon de données de simulation à l'aide de l'ensemble des ressources linguistiques actuelles et des paramètres d'extraction actuels. L'objectif est d'obtenir les résultats simulés et d'utiliser ces résultats pour améliorer vos règles, en créer de nouvelles ou mieux comprendre les mises en correspondance. Pour chaque extrait de texte (phrase, mot ou proposition selon le contexte), le résultat de la simulation affiche l'ensemble des chaînes de caractères et présente les règles TLA ayant trouvé un patron dans ce texte. Une **chaîne de caractères**est définie comme tout mot ou toute phrase identifié/e pendant l'extraction.

Contrairement à d'autres ressources avancées, les règles TLA sont spécifiques à la bibliothèque ; c'est pourquoi vous pouvez uniquement utiliser les règles TLA d'une seule bibliothèque à la fois. Depuis Editeur de modèle ou Editeur de ressources, allez dans l'onglet Règles des liens du texte. Dans cet onglet, vous pouvez spécifier la bibliothèque de votre modèle qui contient les règles TLA que vous souhaitez utiliser ou modifier. C'est pourquoi nous vous conseillons fortement de stocker toutes vos règles dans une seule bibliothèque, à moins que vous ne le souhaitiez pas pour une raison particulière.

Figure 19-3
Panneau de simulation des liens du texte avant la définition des données



Important ! Si vous utilisez un fichier de données, nous vous conseillons fortement de vous assurer que le texte qu'il contient est court afin de minimiser le temps de traitement. L'objectif d'une simulation est de voir comment une partie de texte est interprétée et de comprendre de quelle manière les règles correspondent à ce texte. Ces informations vont vous aider à rédiger et à modifier vos règles. Utilisez le noeud analyse des liens du texte ou exécutez un flux avec une session interactive, en ayant activé l'extraction TLA afin d'obtenir des résultats sur un ensemble plus complet de données. Cette simulation a pour seuls objectifs de tester et de créer des règles.

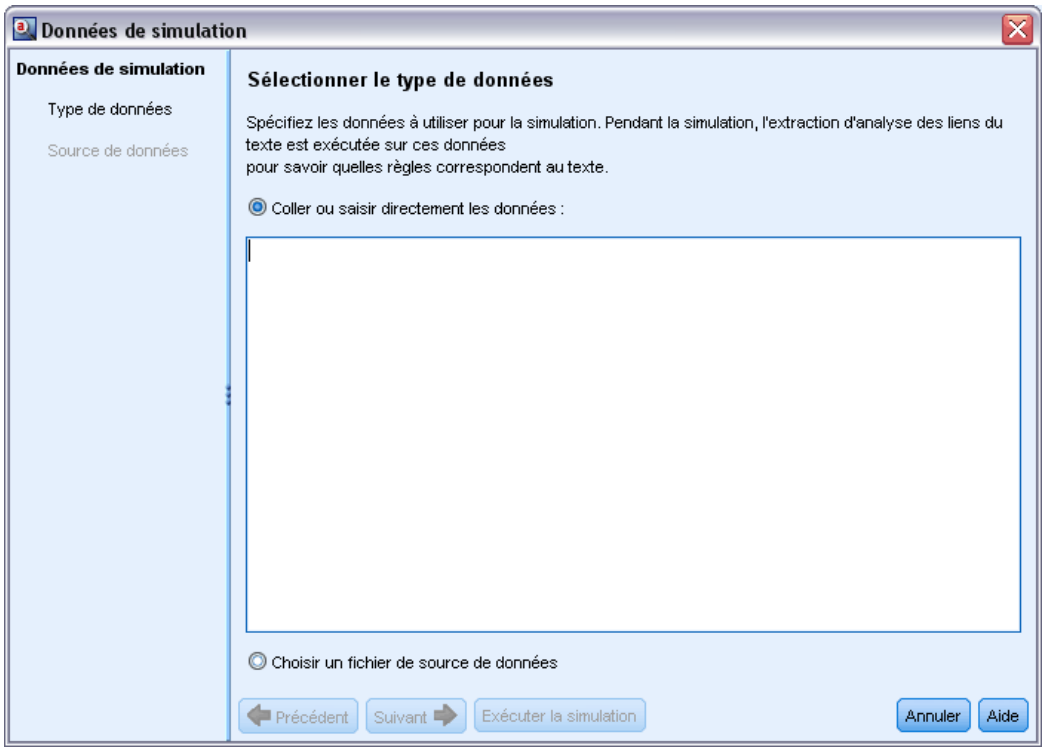
Définition des données pour la simulation

Pour obtenir un aperçu de la façon dont les règles peuvent correspondre au texte, vous pouvez effectuer une simulation à l'aide d'un échantillon de données. La première étape est de définir ces données.

Définition des données

- Cliquez sur Définir les données dans le panneau de simulation au bas de l'onglet Règles des liens du texte. Sinon, dans le cas où aucune donnée n'a été définie auparavant, choisissez Outils > Exécuter la simulation dans le menu. L'assistant Données de simulation s'ouvre.

Figure 19-4
Assistant de simulation

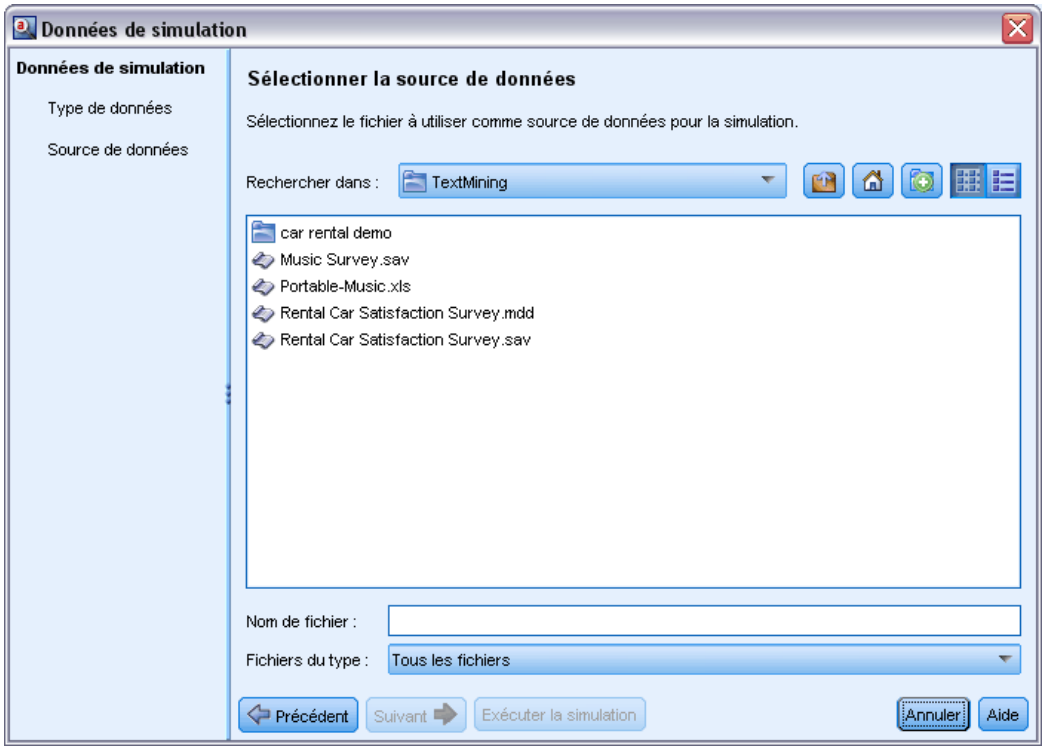


- Spécifiez le type de données en sélectionnant une des options suivantes :
  - Collez ou entrez directement le texte. Une zone de texte vous permet de coller du texte depuis le presse-papiers ou de saisir manuellement le texte à traiter. Vous pouvez entrer une phrase par ligne ou utiliser la ponctuation (virgules ou points) pour séparer les phrases. Une fois le texte saisi, vous pouvez commencer la simulation en cliquant sur Exécuter la simulation.
  - Spécifier une source de données de fichier. Cette option indique que vous souhaitez utiliser un fichier qui contient du texte. Cliquez sur Suivant pour accéder à l'étape de l'assistant dans laquelle vous pouvez définir le fichier à traiter. Une fois le fichier sélectionné, vous pouvez commencer la simulation en cliquant sur Exécuter la simulation. Les types de fichier suivants sont pris en charge : *.rtf*, *.doc*, *.docx*, *.docm*, *.xls*, *.xlsx*, *.xlsm*, *.htm*, *.html*, *.txt* et les fichiers sans extension. Le fichier de données choisi est lu 'tel quel' pendant la simulation. Par exemple, si vous sélectionnez un fichier Microsoft Excel, vous ne pouvez pas sélectionner une feuille de calcul ou une colonne en particulier. À la place, l'ensemble du classeur est lu comme si vous utilisiez un noeud source Microsoft Excel dans IBM® SPSS® Modeler. Le fichier entier est traité de la même façon que si vous aviez connecté un noeud Liste de fichiers à un noeud de Text Mining.

Important ! Si vous utilisez un fichier de données, nous vous conseillons fortement de vous assurer que le texte qu'il contient est court afin de minimiser le temps de traitement. L'objectif d'une simulation est de voir comment une partie de texte est interprétée et de comprendre de quelle manière les règles correspondent à ce texte. Ces informations vont vous aider à rédiger et à modifier vos règles. Utilisez le noeud analyse des liens du texte ou exécutez un flux avec une

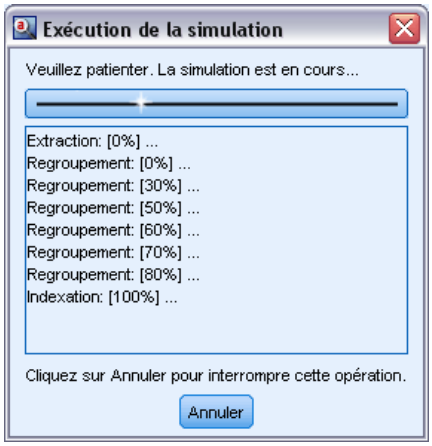
session interactive, en ayant activé l'extraction TLA afin d'obtenir des résultats sur un ensemble plus complet de données. Cette simulation a pour seuls objectifs de tester et de créer des règles.

Figure 19-5
Assistant de simulation - Sélectionner la source de données



- Cliquez sur Exécuter la simulation pour lancer le processus de simulation. Une boîte de dialogue de progression apparaît. Si vous vous trouvez dans une session interactive, les paramètres d'extraction utilisés pendant la simulation sont ceux qui sont actuellement sélectionnés dans la session interactive (voir Outils > Paramètres d'extraction dans la vue Concepts et Catégories) Si vous vous trouvez dans l'Editeur de modèle, les paramètres d'extraction utilisés durant la simulation sont les paramètres d'extraction par défaut ; ce sont les mêmes que ceux affichés dans l'onglet Expert du noeud Analyse des liens du texte. Pour plus d'informations, reportez-vous à la section Comprendre les résultats de la simulation sur p. 344.

Figure 19-6
Boîte de dialogue Progression de la simulation



Comprendre les résultats de la simulation

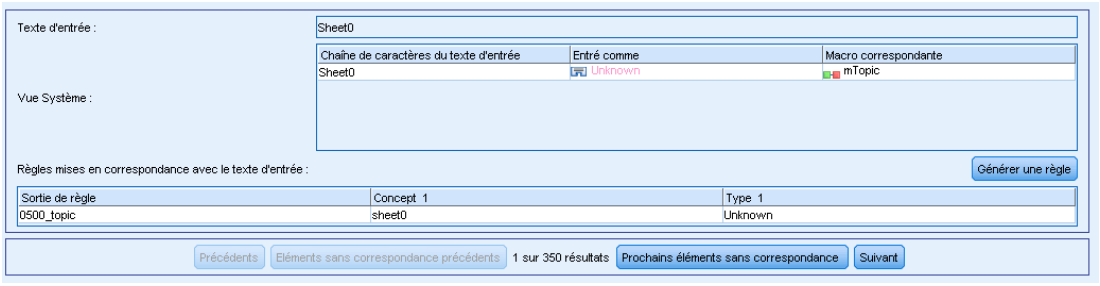
Pour obtenir un aperçu de la façon dont les règles peuvent correspondre au texte, vous pouvez effectuer une simulation à l'aide d'un échantillon de données et consulter les résultats. Ensuite, vous pouvez modifier votre ensemble de règles pour mieux l'adapter à vos données. Lorsque le processus d'extraction et de simulation est terminé, les résultats de la simulation s'afficheront.

Pour chaque "phrase" identifiée pendant l'extraction, plusieurs informations vous sont présentées, notamment la 'phrase' exacte, la répartition des chaînes de caractères trouvées dans cette phrase de texte d'entrée et enfin, toutes les règles qui correspondaient au texte de cette phrase. Par "**phrase**", nous entendons un mot, une phrase ou une proposition en fonction de la façon dont l'extracteur a découpé le texte en parties lisibles.

Une **chaîne de caractères**est définie comme tout mot ou toute phrase identifié/e pendant l'extraction. Par exemple, dans la phrase *Mon oncle vit à New York*, les chaînes de caractères suivantes peuvent être trouvées pendant l'extraction : *mon*, *oncle*, *vit*, *à*, et *new york*. De plus, *oncle* peut être extrait comme concept et entré comme `<Unknown>`, et *new york* peut également être extrait comme concept et entré comme `<Location>`. Tous les concepts sont des chaînes de caractères mais toutes les chaînes de caractères ne sont pas des concepts. Les chaînes de caractères peuvent également être des macros, des chaînes littérales et des intervalles de mots. Seuls les mots ou les phrases entrés peuvent être des concepts.

Lorsque vous travaillez dans la session interactive ou dans l'éditeur de ressources, vous travaillez au niveau du concept. Les règles TLA sont plus précises et les chaînes de caractères individuelles d'une phrase peuvent s'utiliser dans la définition d'une règle si elles n'ont jamais été extraites ou dotées d'un type. Pouvoir utiliser des chaînes de caractères qui ne sont pas des concepts donne encore plus de flexibilité aux règles pour capturer des relations complexes dans votre texte.

Figure 19-7
Exemple de résultats de simulation montrant une correspondance avec une règle



Si vous avez plusieurs phrases dans vos données de simulation, vous pouvez avancer et reculer dans les résultats en cliquant sur Suivant et sur Précédent.

Dans ce genre de cas où une phrase ne correspond à aucune règle TLA dans la bibliothèque sélectionnée (voir le nom de la bibliothèque au-dessus de l'arborescence dans cet onglet), les résultats sont considérés comme étant sans correspondance et les boutons Prochains éléments sans correspondance et Eléments sans correspondance précédents sont activés pour vous dire qu'il existe du texte pour lequel aucune règle n'a trouvé de correspondance et pour vous permettre d'accéder rapidement à ces instances.

Après avoir créé de nouvelles règles, avoir édité vos règles ou modifié vos ressources ou paramètres d'extraction, il est possible que vous souhaitiez effectuer une nouvelle simulation. Pour effectuer une nouvelle simulation, cliquez sur Exécuter une simulation dans le panneau de simulation et les même données d'entrée seront réutilisées.

Les champs et tableaux suivants apparaissent dans les résultats de simulation :

Texte d'entrée. La phrase proprement dite, identifiée par le processus d'extraction se trouvant dans les données de simulation que vous avez définies dans l'assistant. Par phrase, nous entendons un mot, une phrase ou une proposition en fonction de la façon dont l'extracteur a découpé le texte en parties lisibles.

Vue Système. Un ensemble de chaînes de caractères que le processus d'extraction a identifié.

- **Chaîne de caractères du texte d'entrée.** Chaque chaîne de caractères trouvée dans le texte d'entrée. Les chaînes de caractères ont été définies précédemment dans cette rubrique.
- **De type.** Si une chaîne de caractères a été identifiée comme concept et dotée d'un type, alors le nom du type associé (tel que `<Unknown>`, `<Person>`, `<Location>`) appparaît dans cette colonne.
- **Macro correspondante.** Si une chaîne de caractères correspond à une macro existante, le nom de la macro associée apparaît dans cette colonne.

Règles correspondant à du texte d'entrée. Ce tableau présente les règles TLA mises en correspondance avec le texte d'entrée. Pour chaque règle mise en correspondance, vous voyez apparaître le nom de la règle dans la colonne Sortie de règle et les valeurs de sortie associées pour cette règle (Paires Concept + Type). Vous pouvez double-cliquer sur le nom de la règle correspondante pour ouvrir la règle dans le panneau de l'éditeur au-dessus du panneau de simulation.

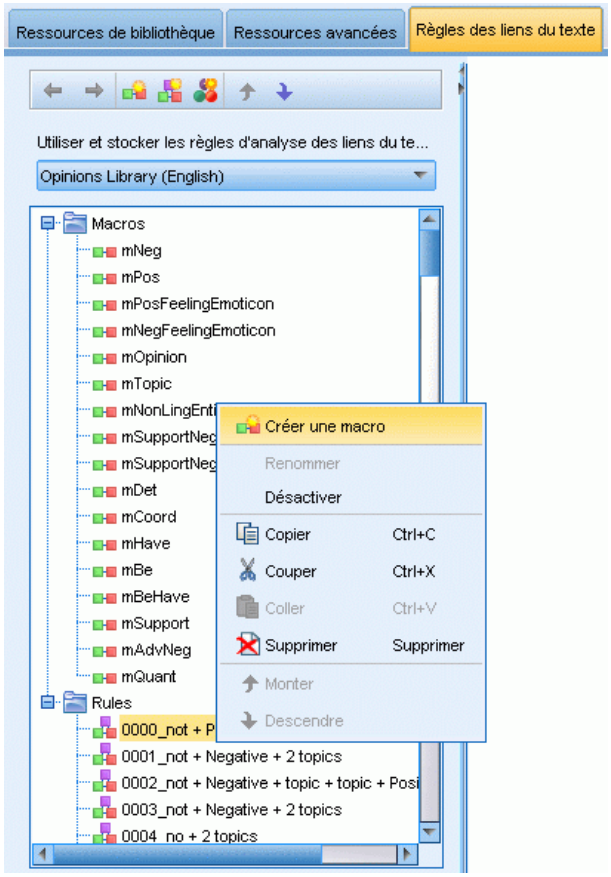
Bouton **Générer une règle**. Si vous cliquez sur ce bouton dans le panneau de simulation, une nouvelle règle s'ouvre dans le panneau de l'éditeur de règle au-dessus du panneau de simulation. Il utilise le texte d'entrée comme exemple. De même, toutes les chaînes de caractères dotées d'un type ou mises en correspondance avec une macro pendant la simulation sont automatiquement insérées en tant qu'éléments de colonne dans la table Valeurs de règle. Si une chaîne de caractères a été dotée d'un type *et* mise en correspondance avec une macro, la valeur de la macro est celle qui sera utilisée dans la règle afin de la simplifier. Par exemple, la phrase "*J'aime la pizza*" peut être dotée du type `<Unknown>` pendant la simulation et être mise en correspondance avec la macro `mTopic` si vous utilisiez les ressources Anglais de base. Dans ce cas, `mTopic` sera utilisée comme l'élément de la règle générée. Pour plus d'informations, reportez-vous à la section Utilisation des règles des liens du texte sur p. 353.

Navigation parmi les règles et les macros de l'arborescence

Lorsqu'une analyse des liens du texte est effectuée pendant l'extraction, les règles des liens du texte stockées dans la bibliothèque sélectionnée dans l'onglet Règles des liens du texte seront utilisées.

Contrairement à d'autres ressources avancées, les règles TLA sont spécifiques à la bibliothèque ; c'est pourquoi vous pouvez uniquement utiliser les règles TLA d'une seule bibliothèque à la fois. Depuis Editeur de modèle ou Editeur de ressources, allez dans l'onglet Règles des liens du texte. Dans cet onglet, vous pouvez spécifier la bibliothèque de votre modèle qui contient les règles TLA que vous souhaitez utiliser ou modifier. C'est pourquoi nous vous conseillons vivement de stocker toutes vos règles dans une seule bibliothèque, à moins que vous ne le souhaitiez pas pour une raison particulière.

Figure 19-8
Onglet Règles des liens du texte : Arborescence des règles et des macros



Vous pouvez spécifier dans quelle bibliothèque vous souhaitez travailler dans l'onglet Règles des liens du texte en sélectionnant cette bibliothèque dans la liste déroulante Utiliser et stocker les règles d'analyse des liens du texte dans : liste déroulante de cet onglet. Lorsqu'une analyse des liens du texte est effectuée pendant l'extraction, les règles des liens du texte stockées dans la bibliothèque sélectionnée dans l'onglet Règles des liens du texte seront utilisées. Ainsi, si vous définissez des règles des liens du texte (règles TLA) dans plusieurs bibliothèques, seule la première bibliothèque dans laquelle les règles TLA sont trouvées sera utilisée pour l'analyse des liens du texte. C'est pourquoi nous vous conseillons fortement de stocker toutes vos règles dans une seule bibliothèque, à moins que vous ne le souhaitiez pas pour une raison particulière.

Lorsque vous sélectionnez une macro ou une règle dans l'arborescence, son contenu apparaît dans le panneau de l'éditeur à droite. Si vous faites un clic droit sur un des éléments de l'arborescence, un menu contextuel vous présente les autres tâches possibles, comme :

- créer une nouvelle macro dans l'arborescence et l'ouvrir dans l'éditeur à droite.
- Créer une nouvelle règle dans l'arborescence et l'ouvrir dans l'éditeur à droite.
- Créer un nouvel ensemble de règles dans l'arborescence.
- Couper, copier et coller des éléments pour simplifier l'édition.
- Supprimer des macros, des règles et des ensembles de règles pour les effacer des ressources.

- Désactiver des macros, des règles et des ensembles de règles pour indiquer qu'ils doivent être ignorés pendant le traitement.
- Faire monter ou descendre les règles d'un niveau pour changer l'ordre de traitement.

Avertissements dans l'arbre

Les avertissements apparaissent avec un triangle jaune dans l'arbre et vous informent d'un problème potentiel. Placez le curseur de la souris sur la macro ou la règle ayant un problème pour afficher l'explication dans l'info-bulle. Dans la plupart des cas, vous verrez quelque chose comme : Avertissement : Aucun exemple fourni ; Entrez un exemple vous devez donc entrer un exemple.

S'il manque un exemple ou si l'exemple ne correspond à aucune règle, vous ne pourrez pas utiliser la fonctionnalité Obtenir des chaînes de caractères, par conséquent nous vous recommandons de n'entrer qu'un seul exemple par règle.

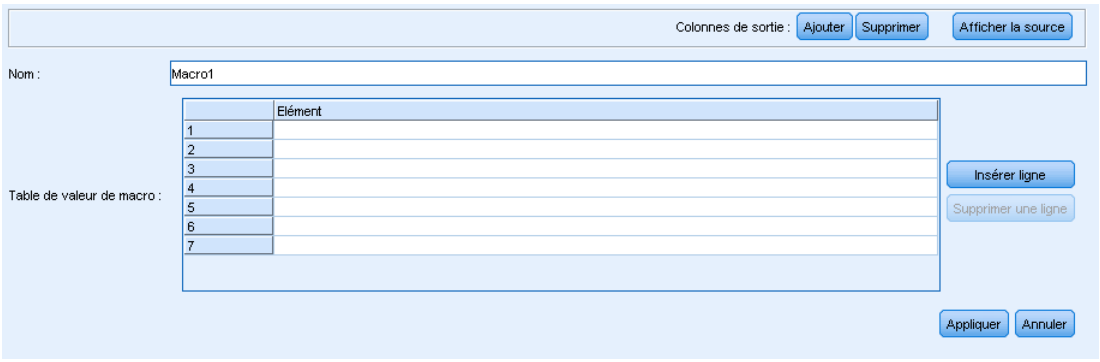
Lorsque la règle est surlignée en jaune, cela signifie qu'un type ou une macro est inconnu de l'éditeur TLA. Le message ressemblera à : Avertissement : Type ou macro inconnu. Ce message vous informe qu'un élément qui serait défini par $something dans la vue source, par exemple $myType, n'est pas un type standard dans votre bibliothèque et n'est pas non plus une macro.

Pour mettre à jour le vérificateur de syntaxe, vous devez passer à une autre règle ou macro ; il est inutile d'effectuer une recompilation. Ainsi, par exemple, si la règle A affiche un avertissement parce que l'exemple est manquant, vous devez ajouter un exemple, cliquer sur une règle plus haute ou plus basse et revenir à la règle A pour vérifier qu'elle est désormais correcte.

Utilisation des macros

Les macros peuvent simplifier l'apparence des règles d'analyse des liens du texte puisqu'elles vous permettent de regrouper des types, d'autres macros et des chaînes (de mots) littérales à l'aide d'un opérateur OR (|). L'avantage de l'utilisation des macros est que non seulement vous pouvez réutiliser les macros dans différentes règles d'analyse des liens du texte pour les simplifier, mais cela vous permet également d'effectuer des mises à jour dans une macro au lieu de devoir effectuer des mises à jour à travers toutes vos règles d'analyse des liens du texte. La plupart des règles TLA fournies contiennent des macros prédéfinies. Les macros apparaissent au-dessus de l'arborescence dans le panneau le plus à gauche de l'onglet Règles des liens du texte.

Figure 19-9
Onglet Règles des liens du texte : Editeur de macro



Les champs et tableaux suivants apparaissent dans les résultats de simulation :

Nom. Un nom unique identifiant cette macro. Nous vous conseillons d'ajouter le préfixe m minuscule au noms des macros pour les identifier rapidement dans vos règles. Lorsque vous faites manuellement référence aux macros dans les règles (par modification d'une ligne ou dans la vue des sources), vous devez utiliser le préfixe $ afin que le processus d'extraction puisse rechercher ce nom spécifique. Mais si vous déplacez et collez le nom de la macro ou l'ajoutez avec les menus contextuels, le produit le reconnaît immédiatement comme macro et aucun $ n'est ajouté.

Table **Valeur de macro**.

- Plusieurs lignes représentant toutes les valeurs possibles que cette macro peut représenter. Ces valeurs distinguent les majuscules des minuscules.
- Ces valeurs peuvent comprendre un type, une chaîne littérale, un intervalle de mots ou une macro ou une combinaison de ces formes. Pour plus d'informations, reportez-vous à la section Eléments pris en charge pour les règles et les macros sur p. 361.
- Pour entrer une valeur pour un élément d'une macro, double-cliquez sur la ligne dans laquelle vous souhaitez travailler. Une zone de texte modifiable apparaît dans laquelle vous pouvez saisir une référence de type, une référence de macro, une chaîne littérale ou un intervalle de mots. Sinon, cliquez avec le bouton droit sur la cellule pour afficher un menu contextuel proposant des listes de macros communes, des noms de types et des noms de types non linguistiques. Pour référencer un type ou une macro, le nom de la macro ou du type doit être précédé d'un '`$`' comme dans `$mTopic` pour la macro `mTopic`. Lors de la combinaison d'arguments, vous devez utiliser des parenthèses `( )` pour regrouper les arguments et le caractère `|` pour indiquer un booléen `OR`.
- Vous pouvez ajouter ou supprimer des lignes dans la table Valeur de macro à l'aide des boutons à sa droite.
- Saisissez chaque élément dans sa propre ligne. Par exemple, si vous souhaitez créer une macro qui représente une des 3 chaînes littérales comme `suis OR étais OR est`, vous devez saisir chaque chaîne littérale dans une ligne distincte de la vue et votre tableau Macro doit contenir 3 lignes.

Création et édition de macros

Vous pouvez créer de nouvelles macros ou éditer les macros existantes. Suivez les conseils et descriptions de l'éditeur de macro. Pour plus d'informations, reportez-vous à la section Utilisation des macros sur p. 348.

Création de nouvelles macros

- A partir des menus, sélectionnez Outils> Nouvelle macro. Sinon, cliquez sur l'icône Nouvelle macro dans la barre d'outils en arborescence pour ouvrir une nouvelle macro dans l'éditeur.
- Saisissez un nom unique et définissez les éléments de valeur de la macro.
- Cliquez sur Appliquer lorsque vous avez terminé de rechercher les erreurs.

Edition des macros

- Cliquez sur le nom de la macro dans l'arborescence. La macro s'ouvre dans le panneau de l'éditeur à droite.
- Effectuez vos modifications.
- Cliquez sur Appliquer lorsque vous avez terminé de rechercher les erreurs.

Désactiver et supprimer des macros

Désactiver des macros

Si vous souhaitez qu'une macro soit ignorée pendant le processus, vous pouvez la désactiver. Cette action peut provoquer des avertissements et des erreurs dans les règles qui référencent encore cette macro désactivée. Soyez prudent lors de la suppression et de la désactivation des macros.

- Cliquez sur le nom de la macro dans l'arborescence. La macro s'ouvre dans le panneau de l'éditeur à droite.
- Faites un clic droit sur le nom.
- Dans les menus contextuels, choisissez Désactiver. L'icône de la macro se grise et la macro ne peut plus être modifiée.

Suppression des macros

Si vous souhaitez effacer une macro, vous pouvez la supprimer. Cette action peut provoquer des erreurs dans les règles qui référencent encore cette macro. Soyez prudent lors de la suppression et de la désactivation des macros.

- Cliquez sur le nom de la macro dans l'arborescence. La macro s'ouvre dans le panneau de l'éditeur à droite.
- Faites un clic droit sur le nom.
- Dans les menus contextuels, choisissez Supprimer. La macro disparaît de la liste.

Vérification des erreurs, enregistrement et annulation

Application des modifications de macro

Si vous cliquez en-dehors de l'éditeur de macro ou si vous cliquez sur Appliquer, la macro est automatiquement analysée pour savoir s'il y a des erreurs. Si une erreur est trouvée, vous devez la résoudre avant de passer à une autre partie de l'application.

Mais si des erreurs moins graves sont détectées, un simple avertissement est donné. Par exemple, si votre macro contient des définitions incomplètes ou non référencées de types ou d'autres macros, un avertissement apparaît. Lorsque vous cliquez sur Appliquer, tout avertissement non corrigé fait apparaître une icône d'avertissement à gauche du nom de la macro dans l'arborescence des règles et des macros dans le panneau de gauche.

Appliquer une macro ne signifie pas que votre macro est enregistrée en permanence. Lorsque vous appliquez une macro, le processus de validation recherche les erreurs et les avertissements.

Enregistrement des ressources à l'intérieur d'une session interactive

- Pour enregistrer les modifications effectuées sur vos ressources pendant une session interactive et pouvoir les utiliser à la prochaine utilisation de votre flux, vous devez :

  mettre à jour votre noeud de modélisation pour vous assurer d'obtenir les mêmes ressources à la prochaine exécution de votre flux. Pour plus d'informations, reportez-vous à la section Mise à jour des nœuds de modélisation et enregistrement dans le chapitre 8 sur p. 136. Vous pouvez ensuite enregistrer votre flux. Pour enregistrer votre flux, rendez-vous dans la fenêtre principale de IBM® SPSS® Modeler après la mise à jour du noeud de modélisation.

- Pour enregistrer les modifications effectuées sur vos ressources pendant une session interactive et pouvoir les utiliser dans un autre flux, vous devez :
  - mettre à jour le modèle utilisé ou en créer un nouveau. Pour plus d'informations, reportez-vous à la section Création et mise à jour de modèles dans le chapitre 14 sur p. 269. Cela n'enregistrera pas les modifications du noeud actuel (voir étape précédente)
  - Ou mettre à jour le TAP utilisé. Pour plus d'informations, reportez-vous à la section Mise à jour des Packages d'analyse de texte dans le chapitre 10 sur p. 229.

Enregistrement des ressources dans l'Editeur de modèle

- D'abord, publiez la bibliothèque. Pour plus d'informations, reportez-vous à la section Publication de bibliothèques dans le chapitre 16 sur p. 300.

- Puis enregistrez le modèle avec Fichier > Enregistrer un modèle de ressources dans les menus.

Annuler les modifications des macros

- Si vous voulez ignorer les modifications, cliquez sur Annuler.

Macros spécifiques : mTopic, mNonLingEntities, SEP

Le modèle Opinions (et les modèles similaires) de même que les modèles Basic Resources sont fournis avec deux macros spéciales nommées `mTopic` et `mNonLingEntities`.

mTopic

Par défaut, la macro `mTopic` regroupe tous les types fournis dans le modèle susceptibles d'être liés à une opinion, comme les types de bibliothèques *Core* suivants : `<Person>`, `<Organization>`, `<Location>`, etc, tant que le type n'est pas un type d'opinion (par exemple, `<Negative>` ou `<Positive>`) ou un type défini comme une entité non linguistique dans les ressources avancées.

Lorsque vous créez un nouveau type dans un modèle Opinions (ou un modèle similaire), le produit suppose que sauf si ce type est spécifié dans une autre macro ou dans la section des entités non linguistiques de l'onglet Ressources avancées, il sera traité de la même façon que les autres types définis dans la macro `mTopic`.

Supposons que vous avez créé de nouveaux types dans les ressources à partir d'un modèle Opinions : `<Vegetables>` et `<Fruit>`. Sans avoir à effectuer aucune modification, vos nouveaux types sont traités comme des types `mTopic` pour que vous puissiez automatiquement connaître les opinions positives, négatives, neutres et contextuelles sur vos nouveaux types. Au cours de l'extraction, par exemple, la phrase "*J'aime les brocolis mais je déteste le pamplemousse*" produirait les 2 patrons de sortie suivants :

```
brocolis <Vegetables> + aime <Positive>
pamplemousse <Fruit> + déteste <Negative>
```

Mais, si vous souhaitez exécuter ces types autrement que pour les autres types dans mTopic, vous pouvez ajouter le nom du type dans une macro existante comme `mPos`, qui regroupe tous les types d'opinions positives, ou créer une nouvelle macro que vous pourrez ensuite référencer dans une ou plusieurs règles.

Important ! Si vous créez un nouveau type tel que `<Légumes>`, ce nouveau type sera inclus en tant que type dans mTopic. Cependant, ce nom de type ne sera pas explicitement visible dans la définition de la macro.

mNonLingEntities

De même, si vous ajoutez de nouvelles entités non linguistiques dans la section Entités non linguistiques de l'onglet Ressources avancées, elles seront automatiquement traitées comme `mNonLingEntities` sauf spécification contraire. Pour plus d'informations, reportez-vous à la section Entités non linguistiques dans le chapitre 18 sur p. 329.

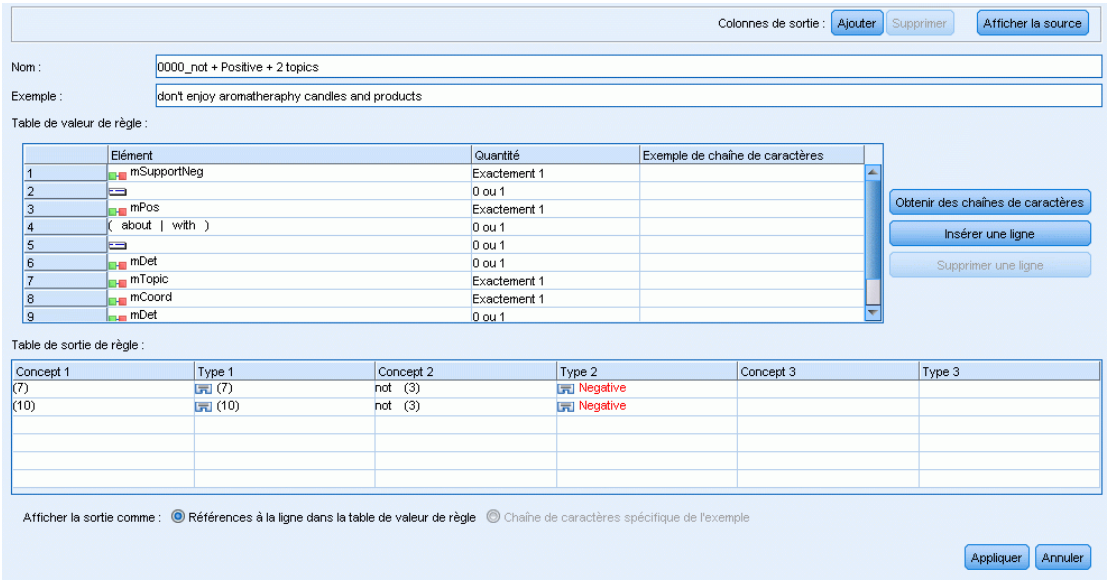
SEP

Vous pouvez également utiliser la macro prédéfinie `SEP`, qui correspond au séparateur global défini sur l'ordinateur local et qui est en général un virgule (,).

Utilisation des règles des liens du texte

Une règle d'analyse des liens du texte est une requête booléenne utilisée pour réaliser une mise en correspondance sur une phrase. Les règles d'analyse des liens du texte contiennent un ou plusieurs des arguments suivants : types, macros, chaînes littérales ou intervalles de mots. Vous devez avoir au moins une règle d'analyse des liens du texte pour extraire les résultats TLA.

Figure 19-10
Onglet Règles des liens du texte : Editeur de règle



Champ **Nom**. Le nom unique de la règle des liens du texte.

Champ **Exemple**. Vous pouvez éventuellement inclure un exemple de phrase ou de séquence de mots qui serait capturée par cette règle. Nous vous conseillons d'utiliser des exemples. Dans cet éditeur, vous pourrez générer des chaînes de caractères à partir de cet exemple de texte pour savoir quelle est sa correspondance avec la règle et quels en seront les résultats. Une **chaîne de caractères**est définie comme tout mot ou toute phrase identifié/e pendant l'extraction. Par exemple, dans la phrase *Mon oncle vit à New York*, les chaînes de caractères suivantes peuvent être trouvées pendant l'extraction : *mon*, *oncle*, *vit*, *à*, et *new york*. De plus, *oncle* peut être extrait comme concept et entré comme `<Unknown>`, et *new york* peut également être extrait comme concept et entré comme `<Location>`. Tous les concepts sont des chaînes de caractères mais toutes les chaînes de caractères ne sont pas des concepts. Les chaînes de caractères peuvent également être des macros, des chaînes littérales et des intervalles de mots. Seuls les mots ou les phrases entrés peuvent être des concepts.

Table de valeur de règle. Cette table contient les éléments de la règle qui sont utilisés pour faire correspondre une règle à une phrase. Vous pouvez ajouter ou supprimer des lignes de la table à l'aide des boutons à sa droite. La table est composée de 3 colonnes :

- La colonne **Elément**. Saisissez des valeurs sous la forme d'un type, d'une chaîne littérale, d'un intervalle de mots (`<Toutes les chaînes de caractères>`) ou d'une macro ou d'une combinaison de ces formes. Pour plus d'informations, reportez-vous à la section

Eléments pris en charge pour les règles et les macros sur p. 361. Double-cliquez sur la cellule d'élément pour y entrer directement les informations. Sinon, cliquez avec le bouton droit sur la cellule pour afficher un menu contextuel proposant des listes de macros communes, des noms de types et des noms de types non linguistiques. Conservez à l'esprit que si vous entrez les informations dans la cellule en les saisissant directement, vous devez faire précéder le nom ou le type de la macro par le caractère '$', par exemple `$mTopic` pour la macro `mTopic`. L'ordre dans lequel vous créez vos lignes d'élément joue un rôle critique pour la façon dont la règle sera mise en correspondance avec le texte. Lors de la combinaison d'arguments, vous devez utiliser des parenthèses `( )` pour regrouper les arguments et le caractère `|` pour indiquer un booléen `OR`. N'oubliez pas que les valeurs distinguent les majuscules des minuscules.

- La colonne **Quantité**. Elle indique le nombre d'occurrences minimal et maximal d'un élément présent dans un texte pour effectuer une correspondance. Par exemple, si vous souhaitez définir un intervalle ou une série de mots, entre deux autres éléments de 0 à 3 mots, vous pouvez choisir Entre 0 et 3 dans la liste ou entrer directement les nombres dans la boîte de dialogue. La valeur par défaut est 'Exactement 1'. Il est possible que dans certains cas vous souhaitiez rendre un élément optionnel. Si c'est le cas, il aura alors une quantité minimale de 0 et une quantité maximale supérieure à 0 (c'est-à-dire 0 ou 1, entre 0 et 2). Veuillez noter que le premier élément d'une règle ne peut pas être optionnel, c'est-à-dire qu'il ne peut pas avoir une quantité de 0.
- La colonne **Exemple de chaîne de caractères**. Si vous cliquez sur Obtenir des chaînes de caractères, le programme découpe l'Exemple de texte en chaînes de caractères et utilise ces dernières pour remplir cette colonne avec les chaînes de caractères qui correspondent aux éléments définis. Si vous le désirez, vous pouvez également voir ces chaînes de caractères dans la table de sortie.

Table de sortie de règle. Chaque ligne de cette table définit la façon dont la sortie des patrons TLA apparaîtra dans les résultats. La sortie de règle peut produire des patrons contenant jusqu'à 6 paires de colonnes Concept/Type, chacune représentant une *propriété*. Par exemple, le patron de type `<Location> + <Positive>` est un patron à deux propriétés c'est-à-dire qu'il est composé de deux paires de colonnes Concept/Type.

Alors que le langage nous donne la liberté d'exprimer les mêmes idées de différentes façons, vous pouvez définir plusieurs règles pour capturer la même idée. Par exemple, le texte *« Paris est un endroit que j'aime »* et le texte *« J'aime vraiment Paris et Florence »* représentent la même idée de base (qu'on aime Paris) mais exprimée de différentes façons. Il faudrait donc deux règles différentes pour les capturer tous les deux. Mais il est plus simple d'utiliser les résultats de patron si les mêmes idées sont regroupées. C'est pourquoi, tout en ayant deux règles différentes pour capturer ces deux phrases, vous pouvez définir la même sortie pour ces deux règles, par exemple, le patron de type `<Emplacement> + <Positif>`, de façon à ce qu'elle représente les deux textes. Vous pouvez alors voir que les résultats ne suivent pas toujours la structure ou l'ordre des mots trouvés dans le texte d'origine. De plus, un tel patron de type pourrait correspondre à d'autres expressions et pourrait produire des patrons de concept tels que : `paris + aime` et `tokyo + aime`.

Pour vous aider à définir plus rapidement les résultats en faisant moins d'erreurs, vous pouvez utiliser le menu contextuel pour choisir l'élément que vous voulez voir apparaître dans les résultats. Vous pouvez également faire glisser et déposer des éléments du tableau Valeur de règle dans les résultats. Par exemple, si vous avez une règle qui contient une référence à la macro `mTopic` dans la ligne 2 du tableau Valeur de règle et que vous souhaitez que cette valeur fasse

partie de vos résultats, vous pouvez simplement faire glisser l'élément de `mTopic` et le déposer dans la première paire de colonnes de la table Sortie de règle. Vous remplirez ainsi le concept et le type pour la paire sélectionnée. Ou si vous souhaitez que les résultats commencent par le type défini par le troisième élément (ligne 3) de la table de valeur de règle, faites glisser ce type de la table Valeur de règle jusqu'à la cellule Type 1 de la table de sortie. La table sera mise à jour et affichera la référence de la ligne entre parenthèses (3).

Vous pouvez également saisir ces références manuellement dans le tableau en double-cliquant sur la cellule de chaque colonne Concept devant faire partie des résultats et en entrant le symbole `$` suivi du numéro de la ligne, comme `$2` pour faire référence à l'élément défini dans la ligne 2 du tableau Valeur de règle. Lorsque vous entrez manuellement les informations, vous devez également définir la colonne Type, entrer le symbole `#` suivi du numéro de ligne, comme `#2` pour faire référence à l'élément défini dans la ligne 2 de la table Valeur de règle.

De plus, vous pouvez même combiner les méthodes. Imaginons le type `<Positive>` dans la ligne 4 de votre table Valeur de règle. Vous pouvez le faire glisser et le déposer dans la colonne `Type 2` puis double-cliquer sur la cellule dans la colonne `Concept 2` pour entrer ensuite manuellement le mot « *pas* » devant elle. Le titre de la colonne des résultats serait alors `pas (4)` dans le tableau, ou si vous étiez en mode édition ou en mode source `pas $4`. Vous pouvez ensuite faire un clic droit dans la colonne Type 1 et sélectionner, par exemple, la macro appelée `mTopic`. Ces résultats pourraient donner un patron de concept comme : `voiture + mauvaise`.

La majorité des règles n'ont qu'une seule ligne de résultats mais parfois, plusieurs résultats sont possibles et souhaités. Dans ce cas, définissez un résultat par ligne dans la table Sortie de règle.

Important ! Gardez à l'esprit que les autres opérations de traitement linguistique sont exécutées lors de l'extraction des patrons TLA. Par conséquent, lorsque la sortie lit `t$3\t#3`, cela signifie que le patron va afficher le concept final et le type final pour le troisième élément après l'application de tous les traitements linguistiques (synonymes et autres regroupements).

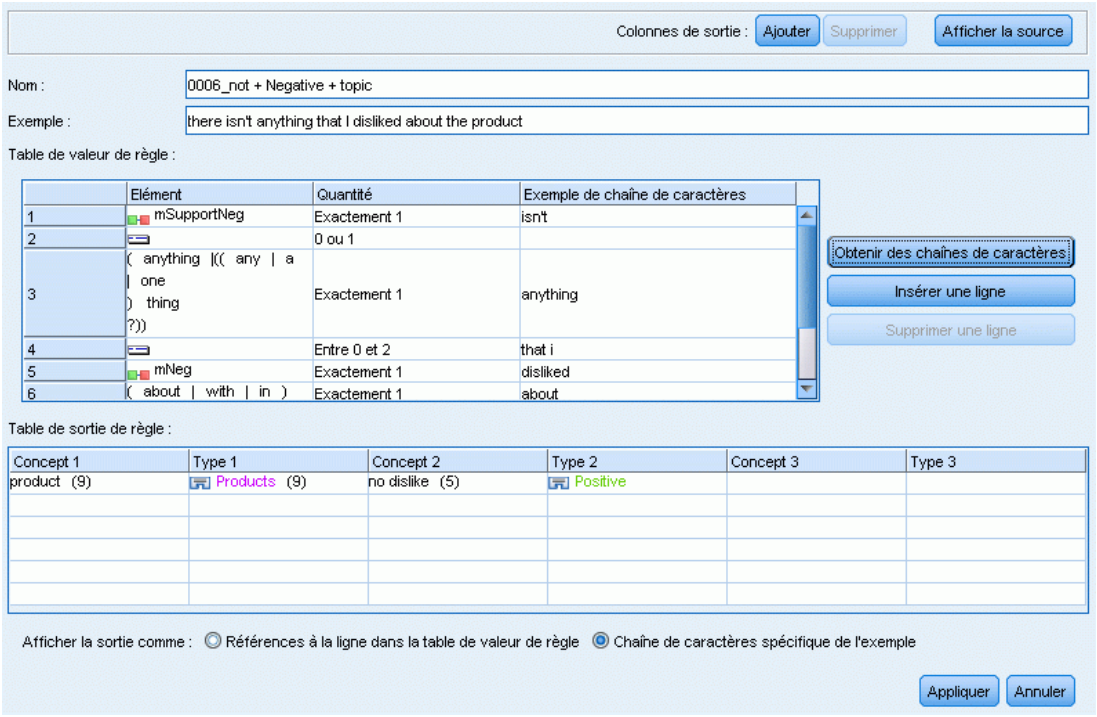
- **Afficher la sortie comme.** Par défaut, l'option Références à la ligne dans la table de valeur de règle est sélectionnée et les résultats utilisent les références numériques de la ligne comme défini dans l'onglet Valeur de règle. Si vous avez auparavant cliqué sur Obtenir des chaînes de caractères et que vous en avez dans la colonne Exemple de chaînes de caractères de la table Valeur de règle, vous pouvez choisir d'afficher les résultats de ces chaînes de caractères spécifiques en choisissant cette option.

Remarque : s'il n'y a pas assez de paires de résultats de concept/type qui apparaissent dans la table de sortie, vous pouvez ajouter une autre paire en cliquant sur le bouton Ajouter dans la barre d'outils de l'éditeur. Si 3 paires apparaissent et que vous cliquez sur Ajouter, 2 colonnes supplémentaires (Concept 4 et Type 4) sont ajoutées à la table. Cela signifie que vous verrez désormais 4 paires dans la table de sortie pour toutes les règles. Vous pouvez également supprimer les paires non utilisées tant qu'aucune autre règle dans l'ensemble de règles de cette bibliothèque n'utilise cette paire.

Exemple de règle

Supposons que vos ressources contiennent la règle d'analyse des liens du texte TLA suivante et que vous avez activé l'extraction des résultats TLA :

Figure 19-11
Onglet Règles des liens du texte : Editeur de règle



Lors de l'extraction, le moteur du programme d'extraction lit chaque phrase et essaie de faire correspondre la séquence suivante :

Elément (ligne)	Description des arguments
1	Le concept d'un des types représentés par les macros `mPos` ou `mNeg` ou du type `<Uncertain>`.
2	Un concept entré comme un des types représentés par la macro `mTopic`.
3	Un des mots représentés par la macro `mBe`.
4	Un élément optionnel, 0 ou 1 mot également, également appelé un intervalle de mots ou `<Toutes les chaînes de caractères>`
5	Un concept entré comme un des types représentés par la macro `mTopic`.

La table de sortie montre que tout ce qui est attendu de cette règle est un patron où tout concept ou type correspondant à la macro `mTopic` qui a été définie dans la ligne 5 de la table Valeur de règle + tout concept ou type correspondant à la macro `mPos`, `mNeg`, ou `<Uncertain>` comme défini dans la ligne 1 de la table Valeur de règle. Cela peut être `saucisse + aimer` ou `<Unknown> + <Positive>`.

Création et édition des règles

Vous pouvez créer de nouvelles règles ou éditer les règles existantes. Suivez les conseils et descriptions de l'éditeur de règle. Pour plus d'informations, reportez-vous à la section Utilisation des règles des liens du texte sur p. 353.

Création de nouvelles règles

- Dans les menus, sélectionnez Outils > Nouvelle règle. Vous pouvez également cliquer sur l'icône Nouvelle règle dans la barre d'outils en arborescence pour ouvrir une nouvelle règle dans l'éditeur.
- Saisissez un nom unique et définissez les éléments de valeur de la règle.
- Cliquez sur Appliquer lorsque vous avez terminé de rechercher les erreurs.

Modification de règles

- Cliquez sur le nom de la règle dans l'arborescence. La règle s'ouvre dans le panneau de l'éditeur à droite.
- Effectuez vos modifications.
- Cliquez sur Appliquer lorsque vous avez terminé de rechercher les erreurs.

Désactivation et suppression des règles

Désactivation des règles

Si vous souhaitez qu'une règle soit ignorée pendant le processus, vous pouvez la désactiver. Soyez prudent lors de la suppression et de la désactivation des règles.

- Cliquez sur le nom de la règle dans l'arborescence. La règle s'ouvre dans le panneau de l'éditeur à droite.
- Faites un clic droit sur le nom.
- Dans les menus contextuels, choisissez Désactiver. L'icône de la règle se grise et la règle ne peut plus être modifiée.

Suppression de règles

Si vous souhaitez effacer une règle, vous pouvez la supprimer. Soyez prudent lors de la suppression et de la désactivation des règles.

- Cliquez sur le nom de la règle dans l'arborescence. La règle s'ouvre dans le panneau de l'éditeur à droite.
- Faites un clic droit sur le nom.
- Dans les menus contextuels, choisissez Supprimer. La règle disparaît de la liste.

Vérification des erreurs, enregistrement et annulation

Application des modifications des règles

Si vous cliquez en-dehors de l'éditeur de règle ou si vous cliquez sur Appliquer, la règle est automatiquement analysée pour savoir s'il y a des erreurs. Si une erreur est trouvée, vous devez la résoudre avant de passer à une autre partie de l'application.

Mais si des erreurs moins graves sont détectées, un simple avertissement est donné. Par exemple, si votre règle contient des définitions incomplètes ou non référencées de types ou de macros, un avertissement apparaît. Lorsque vous cliquez sur Appliquer, tout avertissement non corrigé fait apparaître une icône d'avertissement à gauche du nom de la règle dans l'arborescence du panneau de gauche.

Appliquer une règle ne signifie pas que votre règle est enregistrée en permanence. Lorsque vous appliquez une macro, le processus de validation recherche les erreurs et les avertissements.

Enregistrement des ressources à l'intérieur d'une session interactive

- Pour enregistrer les modifications effectuées sur vos ressources pendant une session interactive et pouvoir les utiliser à la prochaine utilisation de votre flux, vous devez :

  mettre à jour votre noeud de modélisation pour vous assurer d'obtenir les mêmes ressources à la prochaine exécution de votre flux. Pour plus d'informations, reportez-vous à la section Mise à jour des nœuds de modélisation et enregistrement dans le chapitre 8 sur p. 136. Vous pouvez ensuite enregistrer votre flux. Pour enregistrer votre flux, rendez-vous dans la fenêtre principale de IBM® SPSS® Modeler après la mise à jour du noeud de modélisation.

- Pour enregistrer les modifications effectuées sur vos ressources pendant une session interactive et pouvoir les utiliser dans un autre flux, vous devez :
  - mettre à jour le modèle utilisé ou en créer un nouveau. Pour plus d'informations, reportez-vous à la section Création et mise à jour de modèles dans le chapitre 14 sur p. 269. Cela n'enregistrera pas les modifications du noeud actuel (voir étape précédente)
  - Ou mettre à jour le TAP utilisé. Pour plus d'informations, reportez-vous à la section Mise à jour des Packages d'analyse de texte dans le chapitre 10 sur p. 229.

Enregistrement des ressources dans l'Editeur de modèle

- D'abord, publiez la bibliothèque. Pour plus d'informations, reportez-vous à la section Publication de bibliothèques dans le chapitre 16 sur p. 300.

- Puis enregistrez le modèle avec Fichier > Enregistrer un modèle de ressources dans les menus.

Annulation des modifications de règles

- Si vous voulez ignorer les modifications, cliquez sur Annuler dans le panneau de l'éditeur.

Ordre de traitement des règles

Lorsque l'analyse des liens du texte est effectuée pendant l'extraction, une « phrase » (proposition, mot, expression), sera mise en correspondance avec chaque règle l'une après l'autre, jusqu'à ce qu'une correspondance soit trouvée, ou que toutes les règles aient été épuisées. La position au sein de l'arborescence indique l'ordre dans lequel les règles sont testées. La meilleure façon de faire est d'ordonner vos règles de la plus spécifique à la plus générique. Les règles les plus spécifiques doivent se trouver en haut de l'arborescence. Pour modifier l'ordre d'une règle ou d'un ensemble de règles spécifique, sélectionnez Monter d'un niveau ou Descendre d'un niveau dans le menu contextuel Arborescence des règles et des macros ou avec les flèches vers le haut et vers le bas de la barre d'outils.

Si vous vous trouvez *dans la vue des sources*, vous ne pouvez pas modifier l'ordre des règles en les déplaçant dans l'éditeur. Plus la règle est haute dans la vue des sources, plus vite elle sera traitée. Nous vous conseillons vivement de réordonner les règles dans l'arborescence uniquement pour éviter les problèmes de copier/coller.

Important ! Dans les versions précédentes de IBM® SPSS® Modeler Text Analytics , vous aviez un ID de règle numérique unique. Depuis la version 15, vous ne pouvez indiquer l'ordre de traitement qu'en déplaçant la règle vers le haut ou le bas dans l'arborescence, ou avec sa position dans la vue des sources.

Supposons que votre texte contienne les deux phrases suivantes :

J'aime les anchois
J'aime les anchois et les poivrons verts

Supposons également que deux règles d'analyse des liens du texte existent pour lesquelles les valeurs sont les suivantes :

Figure 19-12
2 exemples de règles

A

	Elément	Quantité	Exemple de chaîne de caractères
1	Positive	Exactement 1	
2	mDet	0 ou 1	
3	mTopic	Exactement 1	
4			
5			
6			
7			

B

	Elément	Quantité	Exemple de chaîne de caractères
1	Positive	Exactement 1	
2	mDet	0 ou 1	
3	mTopic	Exactement 1	
4	(SEP | and | or)	1 ou 2	
5	mDet	0 ou 1	
6	mTopic	Exactement 1	
7			

Dans la vue source, les valeurs de règles peuvent ressembler à ceci :

A: `value = $Positive $mDet?  $mTopic`

B: `value = $Positive $mDet?  $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

Si **A** est plus haute dans l'arborescence (plus près du haut) que la règle **B**, alors **A** sera d'abord traitée et la phrase *J'aime les anchois et les poivrons verts* sera d'abord mise en correspondance avec `$Positive $mDet?  $mTopic`, et produira une sortie de patron incomplète (`anchois + aimer`) car elle aura été mise en correspondance avec une règle qui ne recherchait pas 2 correspondances `$mTopic`.

Ainsi, pour capturer l'essence réelle du texte, la valeur la plus spécifique, dans ce cas **B** doit être placée plus haut dans l'arborescence que la valeur plus générique, dans ce cas la règle **A**.

Utilisation d'ensembles de règles (traitement en plusieurs étapes)

Un ensemble de règles permet de regrouper logiquement un ensemble de règles dans l'arborescence Règles et macros afin de pouvoir effectuer un traitement en plusieurs étapes. Un ensemble de règles n'a pas de définition en soi autre qu'un nom et permet d'organiser vos règles dans des groupes de manière logique. Dans certains contextes, le texte est trop riche et varié pour être traité en une seule fois. Par exemple, lorsque vous travaillez avec des données de renseignement, le texte peut contenir des liens entre les individus qui sont découverts à travers les méthodes de contact (*x nommé y*), des relations familiales (*x beau-frère de y*), l'échange d'argent (*x a viré 100$ à y*), etc. Dans ce cas, il est utile de créer des ensembles spécialisés de règles d'analyse des liens du texte, dont chacun est ciblé sur un certain type de relation comme celui destiné pour découvrir les contacts, un autre pour découvrir les membres de la famille, etc.

Pour créer un ensemble de règles, sélectionnez "Créer un ensemble de règles" dans le menu contextuel Arborescence des règles et des macros ou dans la barre d'outils. Vous pouvez ensuite créer de nouvelles règles directement dans un noeud Ensemble de règles de l'arborescence ou déplacer les règles existantes vers un ensemble de règles.

Lorsque vous effectuez une extraction à l'aide de ressources dans lesquelles des règles sont regroupées dans des ensembles de règles, le moteur d'extraction est forcé d'effectuer plusieurs analyses du texte afin de faire correspondre différents patrons dans chaque analyse. De cette façon, une « phrase » peut être mise en correspondance avec une règle de chaque ensemble de règles, alors que sans ensemble de règles, elle ne peut être mise en correspondance qu'avec une seule règle.

Remarque : Vous pouvez ajouter jusqu'à 512 règles par ensemble de règles.

Création de nouveaux ensembles de règles

- Dans les menus, sélectionnez Outils > Nouvel ensemble de règles. Vous pouvez également cliquer sur l'icône Nouvel ensemble de règles dans la barre d'outils de l'arborescence. Un ensemble de règles apparaît dans l'arborescence des règles.
- Ajoutez ces nouvelles règles à cet ensemble de règles ou déplacez les règles existantes dans l'ensemble.

Désactivation des ensembles de règles

- Faites un clic droit sur le nom de l'ensemble de règles dans l'arborescence.

- Dans les menus contextuels, choisissez Désactiver. L'icône de l'ensemble de règles se grise et toutes les règles de cet ensemble de règles sont également désactivées et ignorées pendant le traitement.

Supprimer des ensembles de règles

- Faites un clic droit sur le nom de l'ensemble de règles dans l'arborescence.
- Dans les menus contextuels, choisissez Supprimer. L'ensemble de règles et toutes les règles qu'il contient sont supprimées des ressources.

Eléments pris en charge pour les règles et les macros

Les arguments suivants sont acceptés pour les paramètres de valeur dans les règles d'analyse des liens du texte et les macros :

Macros

Vous pouvez utiliser une macro directement dans une règle d'analyse des liens du texte ou dans une autre macro. Si vous saisissez manuellement le nom d'une macro ou depuis une vue des sources (en opposition à la sélection du nom d'une macro dans un menu contextuel), vérifiez que le nom comporte un préfixe sous la forme du caractère dollar (`$`), comme `$mTopic`. Le nom de macro distingue les majuscules des minuscules. Lorsque vous sélectionnez des macros à partir des menus contextuels, vous pouvez choisir parmi toutes les macros définies dans l'onglet Règles des liens du texte en cours.

Types

Vous pouvez utiliser un type directement dans une règle d'analyse des liens du texte ou une macro. Si vous saisissez manuellement le nom d'un type ou depuis une vue des sources (en opposition à la sélection d'un type dans un menu contextuel), vérifiez que le nom du type comporte un préfixe sous la forme du caractère dollar (`$`), comme `$Person`. Le nom de type distingue les majuscules des minuscules. Si vous utilisez les menus contextuels, vous pouvez choisir parmi tous les types de l'ensemble de ressources utilisé.

Si vous référencez un type non reconnu, vous recevrez un message d'avertissement et la règle aura une icône d'avertissement dans l'arborescence Règles et macros jusqu'à ce que vous le corrigiez.

Chaînes littérales

Pour inclure des informations qui n'ont jamais été extraites, vous pouvez définir une chaîne littérale que le moteur du programme d'extraction va rechercher. Tous les mots ou expressions extraits ont été attribués à un type, c'est pourquoi ils ne peuvent pas être utilisés dans les chaînes littérales. Si vous utilisez un mot qui a été extrait, il sera ignoré, même si son type est `<Unknown>`.

Une chaîne littérale peut être constituée d'un ou de plusieurs mots. Les règles suivantes s'appliquent lors de la définition d'une liste de chaînes littérales :

- Incluez la liste de chaînes entre parenthèses, par exemple (son). Si vous avez le choix entre les chaînes littérales, alors chaque chaîne doit être séparée par l'opérateur OR, comme par exemple `(un|une|le)` ou `(son|sa|ses)`.
- Utilisez des mots uniques ou composés.
- Séparez chaque terme de la liste par une barre verticale (`|`), qui équivaut à l'opérateur booléen `OR`.
- Entrez les formes au singulier et au pluriel si votre recherche porte sur les deux formes. Les flexions ne sont pas générées automatiquement.
- Utilisez uniquement des minuscules.
- Pour réutiliser des chaînes littérales, définissez-les en tant que macro, puis utilisez cette macro dans vos autres macros et règles d'analyse des liens du texte.
- Si une chaîne contient des points ou des traits d'union, vous devez les inclure. Par exemple, pour trouver les correspondances de `a.k.a` dans le texte, entrez les points avec les lettres `a.k.a` comme chaîne littérale.

Opérateur d'exclusion

Utilisez `!` comme opérateur d'exclusion pour empêcher toute expression de la négation d'occuper une propriété particulière. Vous ne pouvez ajouter un opérateur d'exclusion que manuellement avec l'édition de cellule (double-cliquez sur la cellule dans la table Valeur de règle ou Valeur de macro) ou dans la vue des sources. Par exemple, si vous ajoutez `$mTopic @{0,2} !($Positive) $Budget` à votre règle d'analyse des liens du texte, vous recherchez du texte contenant (1) un terme attribué à l'un des types dans la macro `mTopic`, (2) un intervalle de zéro à deux mots, (3) aucune instance d'un terme attribué au type `<Positive>` et (4) un terme attribué au type `<Budget>`. Ceci peut capturer « *ces voitures ont un prix correct* » mais ignorer « *les magasins proposent d'excellentes réductions* ».

Pour utiliser cet opérateur, vous devez saisir le point d'exclamation et les parenthèses manuellement dans la cellule d'élément en double-cliquant sur la cellule.

Intervalles de mots (<Toutes les chaînes de caractères>)

Un intervalle de mots, également appelé `<Toutes les chaînes de caractères>`, définit une plage numérique de chaînes de caractères pouvant être présente entre deux éléments. Les intervalles de mots sont très utiles lors de la mise en correspondance d'expressions très semblables qui ne diffèrent que très légèrement en raison de la présence de déterminants supplémentaires, de prépositions, d'adjectifs ou autres.

Table 19-1
Exemple des éléments dans une table Valeur de règle sans intervalle de mots

#	Elément
1	 `Unknown`

2	mBeHave
3	Positive

Remarque : Dans la vue des sources, cette valeur est définie comme : `$Unknown $mBeHave $Positive`

Cette valeur correspondra à des phrases comme "*le personnel de l'hôtel était gentil*", où *personnel de l'hôtel* appartient au type `<Unknown>`, *était* est dans la macro `mBeHave` et *gentil* est `<Positive>`. Mais elle ne correspond pas à "*le personnel de l'hôtel était très sympathique*".

Table 19-2
Exemple des éléments dans une table Valeur de règle avec un intervalle de mots <Toutes les chaînes de caractères>

#	Elément
1	Unknown
2	mBeHave
3	
4	Positive

Remarque : Dans la vue des sources, cette valeur est définie comme : `$Unknown $mBeHave @{0,1} $Positive`

Si vous ajoutez un intervalle de mots à votre valeur de règle, il sera mis en correspondance avec "*le personnel de l'hôtel était gentil*" et "*le personnel de l'hôtel était très gentil*".

Dans la vue source ou avec l'édition de lignes, la syntaxe d'un intervalle de mots est `@{#,#}`, où `@` signifie un intervalle de mots et où `{#,#}` définit le nombre minimum et maximum de mots acceptés entre l'élément précédent et l'élément suivant. Par exemple, `@{1,3}` signifie qu'une correspondance peut être trouvée entre les deux éléments définis, si ces derniers sont séparés par au moins un mot et par pas plus de trois mots. `@{0,3}` signifie qu'une correspondance peut être trouvée entre les deux éléments définis si 0, 1, 2 ou 3 mots sont présents mais pas plus de trois mots.

Affichage et utilisation du mode Source

Pour chaque règle ou macro, l'éditeur TLA génère le code source sous-jacent qui est utilisé par l'extracteur pour mettre en correspondance et produire la sortie TLA. Si vous préférez travailler directement dans le code lui-même, vous pouvez voir et modifier ce code source directement en cliquant sur le bouton "Afficher la source" en haut de l'éditeur. La vue des sources vous amène directement à la règle ou macro sélectionnée et la met en surbrillance. Cependant, nous vous conseillons d'utiliser les panneaux de l'éditeur afin de réduire les risques d'erreurs.

Lorsque vous avez terminé de visualiser ou de modifier la source, cliquez sur Quitter la source. Si vous générez une syntaxe non valide pour une règle, vous devrez résoudre le problème avant de quitter la vue des sources.

Important ! Si vous effectuez une modification dans la vue des sources, nous vous recommandons vivement de modifier les règles et les macros une par une. Après avoir modifié une macro, validez les résultats en les extrayant. Si vous êtes satisfait du résultat, nous vous recommandons d'enregistrer le modèle avant d'effectuer une autre modification. Si vous n'êtes pas satisfait du résultat ou si une erreur se produit, rétablissez vos ressources enregistrées.

Macros dans la vue des sources

```
[macro]
name = macro_name
value = ([type_name|macro_name|literal_string|word_gap])
```

`[macro]`	Chaque macro doit commencer avec la ligne marquée [macro] pour désigner le début d'une macro.
`name`	Le nom de la définition de macro. Chaque nom doit être unique.
`value`	Une combinaison d'un ou de plusieurs types, chaînes littérales, intervalles de mots ou macros.Pour plus d'informations, reportez-vous à la section Eléments pris en charge pour les règles et les macros sur p. 361. Lors de la combinaison d'arguments, vous devez utiliser des parenthèses ( ) pour regrouper les arguments et le caractère | pour indiquer un booléen OR.

En plus des conseils et de la syntaxe que présente la rubrique sur les macros, la vue des sources contient quelques conseils supplémentaires qui ne sont pas obligatoires lors de l'utilisation de la vue de l'éditeur. Les macros doivent également respecter les règles suivantes lors de l'utilisation du mode Source :

- Chaque macro doit commencer avec la ligne marquée `[macro]` pour désigner le début d'une macro.
- Pour désactiver un élément, ajoutez un indicateur de commentaire (#) au début de chaque ligne concernée.

Exemple. Cet exemple définit une macro nommée `mTopic`. La valeur de `mTopic` est la présence d'un terme correspondant à l'*un* des types suivants : `<Product>`, `<Person>`, `<Location>`, `<Organization>`, `<Budget>`, ou `<Unknown>`.

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

Règles dans la vue des sources

```
[pattern(ID)]
name = pattern_name
value = [$type_name|macro_name|word_gaps|literal_strings]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]
```

`[pattern (<ID>)]`	Indique le début de la règle d'analyse des liens du texte et fournit un ID numérique unique pour déterminer l'ordre de traitement.
`name`	Génère un nom unique pour cette règle d'analyse des liens du texte.

`value` Fournit la syntaxe et les arguments à mettre en correspondance avec le texte. Pour plus d'informations, reportez-vous à la section Eléments pris en charge pour les règles et les macros sur p. 361.

`output` Le format de sortie des patrons mis en correspondance résultants trouvés dans le texte. La sortie ne ressemble pas toujours à la position originale exacte des éléments du texte source. De plus, il est possible d'avoir plusieurs lignes de sortie pour une règle d'analyse des liens du texte donnée en plaçant chaque sortie sur une ligne séparée.

Syntaxe de la sortie :

- sortie séparée avec le code de tabulation `\t`, tel que `$1\t#1\t$3\t#3`
- `$` et un numéro indique le terme trouvé correspondant à l'argument défini dans le paramètre de valeur à cette position. Ainsi, `$1` signifie le terme correspondant au premier argument défini pour la valeur.
- `#` et un numéro indique le nom du type de l'élément à cette position. Si un élément est une liste de chaînes littérales, le type `<Unknown>` est attribué.
- Une valeur `Null\tNull` ne créera aucun résultat.

En plus des conseils et de la syntaxe que présente la rubrique sur les règles, la vue des sources contient quelques conseils supplémentaires qui ne sont pas obligatoires lors de l'utilisation de la vue de l'éditeur. Les règles doivent également respecter les règles suivantes lors de l'utilisation du mode Source :

- Lorsque vous définissez deux éléments ou plus, incluez-les entre parenthèses, qu'ils soient optionnels ou non (par exemple, `($Negative|$Positive)` *ou* `($mCoord|$SEP)?`). `$SEP` représente une virgule.
- Le premier élément d'une règle d'analyse des liens du texte ne peut pas être un élément facultatif. Ainsi, il est impossible de faire commencer une définition de patrons par `value = $mTopic?` ou `value = @{0,1}`.
- Vous pouvez associer une quantité (ou un nombre d'instances) à une chaîne de caractères. Cela vous permet de créer une règle qui englobe tous les cas, plutôt que de rédiger une règle distincte par cas. Vous pouvez, par exemple, utiliser la chaîne littérale `($SEP|et)` si vous recherchez une virgule (`,`) ou la conjonction `et`. Si vous ajoutez à cela un nombre d'instances, la chaîne littérale se présente sous la forme `($SEP|et){1,2}` et vous obtenez comme correspondance l'une des trois instances ci-dessous : "`,`" "`et`" " `et`".
- Les espaces ne sont pas pris en charge entre le nom de la macro et les signes `$` et `?` dans la `valeur` de la règle d'analyse des liens du texte.
- Les espaces ne sont pas pris en charge dans la `sortie` de la règle d'analyse des liens du texte.
- Pour désactiver un élément, ajoutez un indicateur de commentaire (#) au début de chaque ligne concernée.

Exemple. Supposons que vos ressources contiennent la règle d'analyse des liens du texte TLA suivante et que vous avez activé l'extraction des résultats TLA :

```
##     Pierre Martin est l'ancien DRH d'IBM en France
[pattern(201)]
name= 1_201
value = $Person ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Function
 (of|with|for|in|to|at) @{0,1} $Organization @{0,2} $Location
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

Lors de l'extraction, le moteur du programme d'extraction lit chaque phrase et essaie de faire correspondre la séquence suivante :

Position	Description des arguments
1	Le nom d'une personne (`$Person`),
2	L'un des suivants : virgule (`$SEP`), déterminant (`$mDet`), verbe auxiliaire (`$mSupport`), les chaînes "`puis`" ou "`comme`",
3	0 ou 1 mot (`@{0,1}`)
4	Une fonction (`$Function`)
5	L'une des chaînes suivantes : "`de`", "`avec`", "`pour`", "`dans`", "`vers`" ou "`à`",
6	0 ou 1 mot (`@{0,1}`)
7	Le nom d'une organisation (`$Organization`),
8	0, 1 ou 2 mots (`@{0,2}`)
9	Le nom d'un emplacement (`$Location`),

Cet exemple de règle d'analyse des liens du texte correspondrait à des phrases ou des expressions comme :

Pierre Martin, le DRH d'IBM en France
Pierre Martin est l'ancien DRH d'IBM en France
IBM a nommé Pierre Martin comme DRH d'IBM en France

Cet exemple de règle d'analyse des liens du texte produirait la sortie suivante :

```
Pierre Martin  <Person> directeur ressources humaines <Function> ibm <Organization> france <Location>
```

Où :

- `pierre martin` est le terme correspondant à `$1` (le premier élément dans la règle d'analyse des liens du texte) et `<Person>` est le type de `pierre martin` (`#1`),
- `directeur ressources humaines` est le terme correspondant à `$4` (le 4ème élément dans la règle d'analyse des liens du texte) et `<Function>` est le type de `directeur ressources humaines` (`#4`),
- `ibm` est le terme correspondant à `$7` (le 7ème élément dans la règle d'analyse des liens du texte) et `<Organization>` est le type de `ibm`. (`#7`),
- `france` est le terme correspondant à `$9` (le 9ème élément dans la règle d'analyse des liens du texte) et `<Location>` est le type de `france` (`#9`)

Ensembles de règles dans la vue des sources

[ensemble(<ID>)]

`[ensemble (<ID>)]`	Indique le début d'un ensemble de règles et fournit un ID numérique unique pour déterminer l'ordre de traitement des ensembles.

Exemple. La phrase suivante contient des informations à propos des individus, leur fonction au sein d'une entreprise ainsi que les activités de fusion/acquisition de cette entreprise.

```
IBM a conclu un accord définitif de fusion avec SPSS, a déclaré Jack
Noonan, PDG de SPSS.
```

Vous pouvez écrire une règle avec plusieurs sorties pour traiter les différents résultats possibles :

```
## IBM a conclu un accord définitif de fusion avec SPSS, a déclaré
Jack Noonan, PDG de SPSS.

[pattern(020)]
name=020
value = $Organization @{0,4} $ActionNouns @{0,6}  $mOrg @{1,2}
 $Person @{0,2} $Function @{0,1} $Organization
output = $1\t#1\t$3\t#3\t$5\t#5
output = $7\t#7\t$9\t#9\t$11\t#11
```

qui produirait les 2 patrons de sortie suivants :

- `ibm <Organization> + fusionner <ActiveVerb> + spss <Organization>`
- `jack noonan <Person> + pdg <Function> + spss <Organization>`

Important ! Gardez à l'esprit que les autres opérations de traitement linguistique sont exécutées lors de l'extraction des patrons TLA. Dans ce cas, `fusion` est regroupé sous `fusionner` lors de la phase de regroupement des synonymes du processus d'extraction. Et puisque `fusionner` est du type `<ActiveVerb>`, ce nom de type apparaît dans la sortie du patron TLA finale. Par conséquent, lorsque la sortie lit `t$3\t#3`, cela signifie que le patron va afficher le concept final et le type final pour le troisième élément après l'application de tous les traitements linguistiques (synonymes et autres regroupements).

A la place de règles complexes comme la précédente, il est plus facile de gérer et d'utiliser deux règles. La première est spécialisée dans la recherche de fusions/acquisition entre les entreprises :

```
[set(1)]
## IBM a conclu un accord définitif de fusion avec SPSS
[pattern(44)]
name=firm + action + firm_0044
value=$mOrg @{0,20} $ActionNouns @{0,6} $mOrg
output(1)=$1\t#1\t$3\t#3\t$5\t#5
```

ce qui produirait `ibm <Organization> + fusionner <ActiveVerb> + spss <Organization>`

La seconde est spécialisée dans l'individu/fonction/entreprise :

```
[set(2)]
## a déclaré Jack Noonan, PDG de SPSS
[pattern(52)]
name=individual + role + firm_0007
value=$Person @{0,3} $mFunction (at|of)? ($mOrg|$Media|$Unknown)
output(1)=$1\t#1\t$3\tFunction\t$5\t#5
```

ce qui produirait `jack noonan <Person> + pdg <Function> + spss <Organization>`

Annexe

A

Exceptions pour le texte en japonais

Remarque : les fonctionnalités de ce chapitre sont uniquement disponibles dans IBM® SPSS® Modeler Premium.

Bien que le texte en japonais soit traité et exploité d'une manière similaire aux autres langues prises en charge dans IBM® SPSS® Modeler Text Analytics , il existe de nombreuses différences. Les différences minimes sont décrites avec les instructions concernant toutes les autres langues dans cette documentation. Toutefois, certaines des différences les plus importantes sont décrites dans ce chapitre de l'annexe.

Remarque : de nombreuses améliorations ont vu le jour depuis Text Mining for Clementine version 2.2 en japonais. Pour de plus amples informations à ce sujet, veuillez contacter votre représentant commercial.

Extraction et catégorisation du texte japonais

Lors de l'exploitation d'un texte en japonais, le processus est similaire aux autres langues prises en charge. Pour plus d'informations, reportez-vous à la section A propos de Text Mining dans le chapitre 1 sur p. 2. Cependant, pour le japonais, on peut noter les différences suivantes.

Fonctionnement de l'extraction

Lors de l'extraction des principaux concepts et idées de vos réponses, IBM® SPSS® Modeler Text Analytics s'appuie sur une analyse de texte linguistique. Cette approche a la même efficacité en temps et en argent que les systèmes statistiques. Mais elle offre un plus grand degré de précision tout en ne nécessitant que peu d'intervention humaine. L'analyse de texte linguistique repose sur un domaine d'étude appelé processus de langage naturel, également connu sous le nom de linguistique computationnelle.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Pour du texte en japonais, la différence entre les approches basées sur les statistiques et celles basées sur la linguistique pendant l'extraction peut être illustrée à l'aide du mot 沈む comme exemple. A l'aide de ce mot, nous pouvons rechercher des expressions telles que 日が沈む, traduites par *le soleil se couche*, ou 気分が沈む, traduite par *avoir le blues*. Si vous utilisez uniquement les techniques de statistiques, 日 (traduit par *soleil*), 気分 (traduit par *ressentir*), et 沈む (traduit par *en bas*) sont extraits séparément. Mais si vous utilisez l'analyseur de sentiment, qui utilise des techniques linguistiques, 日, 気分, et 沈む sont extraits, mais 気分が沈む (traduit comme *avoir le blues*) est également extrait et affecté au type <悪い － 悲しみ全般>. L'utilisation de techniques basées sur la linguistique à l'aide de l'analyseur de sentiment permet d'extraire plus d'expressions significatives. L'analyse et la capture des émotions permettent de lever toute ambiguïté dans le texte, et fait de l'exploration de texte linguistique la méthode la plus fiable, par définition.

Si vous comprenez le fonctionnement du processus d'extraction, vous êtes plus à même de prendre les décisions-clés lorsque vous affinez vos ressources linguistiques (bibliothèques, types, synonymes, etc.). Les principales étapes du processus d'extraction sont les suivantes :

- Conversion des données source en un format standard
- Identification des termes susceptibles d'être extraits
- Identification des classes d'équivalence et intégration des synonymes
- Affectation d'un type
- Indexation et, si nécessaire, mise en correspondance de patrons avec un deuxième analyseur

Etape 1. Conversion des données source en un format standard

Au cours de cette première étape, les données que vous importez sont converties dans un format uniforme pouvant être utilisé pour effectuer d'autres analyses. Cette conversion, qui s'effectue en interne, ne modifie pas les données d'origine.

Etape 2. Identification des termes susceptibles d'être extraits

Il est important de comprendre le rôle des ressources linguistiques dans l'identification des termes susceptibles d'être extraits lors de l'extraction linguistique. Les ressources linguistiques sont utilisées lors de chaque exécution d'une extraction. Elles se présentent sous la forme de ressources compilées, de bibliothèques et de modèles. Les bibliothèques comportent des listes de mots, des relations et des informations complémentaires qui permettent de spécifier ou d'affiner l'extraction. Vous ne pouvez pas afficher ni éditer les ressources compilées. Toutefois, les autres ressources peuvent être modifiées dans l'Editeur de modèle ou, si vous êtes dans une session interactive, dans l'Editeur de ressources.

Les ressources compilées sont des composants internes essentiels du moteur du programme d'extraction de SPSS Modeler Text Analytics . Ces ressources comportent un dictionnaire général qui répertorie les formes de base avec un code concernant la catégorie grammaticale (nom, verbe, adjectif, etc.). Les ressources comprennent également des types intégrés et réservés qui permettent d'affecter de nombreux termes extraits aux types suivants : <地名>, <組織>, ou <人名>. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais sur p. 378.*Remarque* : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Outre ces ressources compilées, plusieurs bibliothèques sont fournies avec le produit et peuvent être utilisées pour compléter les types et les définitions de concept figurant dans les ressources compilées, ainsi que pour proposer des synonymes. Ces bibliothèques —et toutes les bibliothèques personnalisées que vous créez— comprennent plusieurs dictionnaires : déclarations de types, dictionnaires de synonymes et dictionnaires d'exclusions. Pour plus d'informations, reportez-vous à la section Modification des ressources pour du texte en japonais sur p. 374.*Remarque* : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Une fois les données importées et converties, le moteur du programme d'extraction commence à identifier les termes susceptibles d'être extraits. Ces termes sont des mots ou des groupes de mots qui permettent d'identifier des concepts du texte. Pendant le traitement du texte, les mots uniques (**unitermes**) et les mots composés (**multitermes**) sont identifiés à l'aide d'extracteurs de patrons de catégorie grammaticale. Par exemple, l'expression multiterme 青森りんご, qui suit le patron de catégorie grammaticale <地名> + <名詞> possède deux composants. Par conséquent, les mots-clés de sentiment susceptibles d'être extraits sont identifiés à l'aide de l'analyse des liens du texte de sentiment.

Imaginons par exemple que vous ayez le texte suivant en japonais : 写真が新鮮で良かった. Dans ce cas, le moteur d'extraction affecte le type de sentiment 良い － 褒め・賞賛, après avoir mis en correspondance（品物） ＋ が ＋ 良い à l'aide des règles de liens du texte de sentiment.

Remarque : les termes qui figurent dans le dictionnaire général compilé susmentionné constituent la liste de tous les mots qui risquent de s'avérer inintéressants ou de présenter une ambiguïté linguistique en tant qu'unitermes. Ces mots sont exclus de l'extraction lorsque vous identifiez les unitermes. Ils font toutefois l'objet d'une réévaluation lorsque vous déterminez les catégories grammaticales ou que vous recherchez des mots composés (expressions multitermes) plus longs, susceptibles d'être extraits.

Etape 3. Identification des classes d'équivalence et intégration des synonymes

Une fois les expressions unitermes et multitermes susceptibles d'être extraites identifiées, le logiciel utilise un dictionnaire de normalisation afin d'identifier des classes d'équivalence. Une classe d'équivalence désigne la forme de base d'une phrase ou la forme unique de deux variantes d'une même phrase. L'affectation d'expression à des classes d'équivalence a pour objectif de veiller à ce que, par exemple, `effet secondaire` et 副作用 ne soient pas traités comme des concepts distincts. Pour déterminer quel concept utiliser pour la classe d'équivalence (c'est-à-dire si `effet secondaire` ou 副作用 est utilisé en tant que terme principal), le moteur du programme d'extraction applique les règles suivantes dans l'ordre indiqué ci-dessous :

- Forme définie par l'utilisateur dans une bibliothèque.
- La forme la plus fréquente, comme définie par les ressources précompilées.

Etape 4. Affectation d'un type

Des types sont ensuite affectés aux concepts extraits. Un type correspond à un regroupement sémantique de concepts. Les ressources compilées et les bibliothèques sont utilisées au cours de cette étape. Les types comprennent des éléments tels que des concepts de niveau supérieur, des mots positifs et négatifs, des prénoms, des lieux, des organisations, etc. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

Les ressources en japonais contiennent un ensemble de types distinct. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais sur p. 378.*Remarque* : l'extraction de texte japonais est disponible dans SPSS Modeler Premium.

Etape 5. Indexation et mise en correspondance des patrons avec extraction d'événements

L'indexation de l'ensemble des documents ou des enregistrements s'effectue en définissant un pointeur entre une position de texte et le terme représentatif de chaque classe d'équivalence. Cela suppose que toutes les instances de forme fléchies d'un concept susceptible d'être extrait sont indexées en tant que forme de base susceptible d'être extraite. La fréquence globale est calculée pour chaque forme de base.

SPSS Modeler Text Analytics peut non seulement détecter les types et les concepts, mais également les relations qui existent entre eux. Plusieurs algorithmes et bibliothèques sont fournis avec ce produit ; ils permettent d'extraire les patrons de relations d'analyse de liens du texte entre les types et les concepts. Ils s'avèrent particulièrement utiles lorsque vous tentez de découvrir des opinions spécifiques (par exemple, des réactions sur des produits).

Fonctionnement de l'extraction secondaire

Lorsque vous exécutez une extraction sur du texte japonais, vous obtenez automatiquement des concepts à partir des mots-clés de base et de 8 types de base, y compris 人名, 地名, 組織名, 名詞, 形容詞, 動詞, 形容動詞, et その他. Cependant, afin de profiter pleinement des ressources par défaut fournies pour le texte en japonais, vous devez sélectionner un des analyseurs secondaires suivants : Sentiment ou Dépendance.

Choisir un deuxième analyseur vous permet également d'extraire des patrons d'analyse des liens du texte et de connaître les relations entre les termes du texte. Lorsque vous définissez votre nœud ou lorsque vous choisissez une option d'extraction dans une session interactive, vous pouvez choisir d'ajouter un analyseur secondaire au processus d'extraction.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Analyse secondaire. Lorsqu'une extraction est lancée, l'extraction des mots-clés de base est effectuée à l'aide de l'ensemble de types par défaut. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais sur p. 378. Mais lorsque vous sélectionnez un analyseur secondaire, vous pouvez obtenir des concepts plus nombreux et plus riches car l'extracteur inclura désormais des verbes à particules et des auxiliaires comme faisant partie du concept. Par exemple, supposons que nous avons une phrase 肩の荷が下りた, traduite par "*Ca m'a enlevé un gros poids*". Dans cet exemple, l'extraction des mots-clés de base peut extraire chaque concept séparément comme suit : 肩 (*poids*), 荷 (*gros*), 下りる (*a enlevé*), mais la relation entre ces mots n'est pas extraite. Cependant, si vous avez appliqué l'analyse de sentiment, vous pouvez extraire des concepts plus riches relatifs à un type de sentiment comme le concept =肩の荷が下りた, qui est traduit par "*avoir enlevé un gros poids*", affecté au type <良い-安心>. Dans le cas d'une analyse de sentiment, un grand nombre de types supplémentaires est également inclus. De plus, choisir un analyseur secondaire vous permet également de générer des résultats d'analyse des liens du texte.

Remarque : Lorsqu'un analyseur secondaire est appelé, le processus d'extraction nécessite plus de temps. Pour plus d'informations, reportez-vous à la section Fonctionnement de l'extraction secondaire sur p. 371.

- **Analyse des dépendances** Choisir cette option génère des particules étendues pour les concepts d'extraction du type de base et de l'extraction des mots-clés. Vous pouvez également obtenir des résultats de patrons plus riches avec l'analyse de dépendance des liens du texte (TLA).
- **Analyse des sentiments**. Choisir cet analyseur génère l'extraction de concepts supplémentaires et, le cas échéant, l'extraction de résultats de patrons TLA. En plus des types de base, vous bénéficiez également de plus de 80 types de sentiments, notamment 嬉しい, 吉報, 幸運, 安心, 幸福, etc. Ces types permettent de découvrir des concepts et des patrons dans le texte grâce à l'expression des émotions, des sentiments et des opinions. Ce sont trois options qui dictent la cible de l'analyse des sentiments : Tous les sentiments, Sentiment représentatif uniquement et Conclusions uniquement.

Options de l'analyse des sentiments

Lorsque vous travaillez sur du texte en japonais et vous pouvez choisir d'extraire des concepts et des types supplémentaires à l'aide de l'analyseur de sentiments. Cet analyseur comprend plus de 80 types supplémentaires afin de vous aider à extraire les opinions, les sentiments et les

émotions de vos données textuelles. En outre, lorsque vous choisissez l'analyse de sentiments comme analyseur secondaire, vous devez également sélectionner l'une des options suivantes pour indiquer au moteur d'extraction quel sentiment extraire :

- **Tous les sentiments**
- **Uniquement le sentiment représentatif**
- **Uniquement les conclusions**

Au cours de l'extraction, l'analyse de sentiments commence par diviser un enregistrement ou un document en propositions, chacune contenant un prédicat. Par exemple, le texte, « 月になったがまだ寒い。 », qui est traduit « *On est en avril, mais il fait encore froid.* », est interprété comme 2 propositions par l'analyseur en dépit du fait qu'il ne contient qu'un seul caractère d'arrêt。. Chaque proposition est alors examinée par le moteur d'extraction afin de voir si elle correspond à l'option sélectionnée.

Examinons les trois options à l'aide de l'exemple de texte : "案内してくれた仲居さんは無愛想だったが、部屋は広くて申し分なかった。夕食も満足。". Ce texte est traduit : "*Une domestique n'était pas aimable, mais la pièce était grande et tout à fait satisfaisante. J'ai aussi apprécié le dîner* ». Au cours de l'extraction, le texte original est scindé en propositions qui sont :

- 案内してくれた仲居さんは無愛想だったが、, qui signifie « *Une des domestiques n'était pas aimable mais* »
- 部屋は広くて申し分なかった。, qui signifie « *La salle était grande et assez plaisante* »
- 夕食も満足。, qui signifie « *J'ai aussi apprécié le dîner* »

Tous les sentiments

Cette option extrait tous les sentiments, opinions et émotions qui correspondent aux ressources et aux règles des liens du texte de sentiment. Dans notre exemple, les concepts suivants peuvent être extraits l'échantillon de texte.

Table A-1
Résultats possibles pour l'échantillon en utilisant l'option « Tous les sentiments »

Concept	Type
仲居さんは無愛想だった	<悪い - 対応が不親切>
部屋は広くて	<良い - 満足>
申し分なかった	<良い - 満足>
満足	<良い - 満足>

Remarque : Dans le tableau précédent, les deuxième et troisième lignes indiquent la façon dont l'extracteur peut obtenir deux concepts à partir de la même proposition.

Uniquement le sentiment représentatif

Cette option n'extrait que les opinions et les émotions exprimées les plus représentatives dans chaque proposition. S'il existe plusieurs opinions et émotions dans le texte, un algorithme est appliqué. Cet algorithme tente de déterminer l'importance des sentiments trouvés ainsi que la position des mots dans une proposition. Dans certains cas où deux mots-clés de sentiments

ayant la même importance sont trouvés, le mot-clé de sentiments en dernière position dans la proposition est extrait plutôt que le premier.

部屋は広くて qui est traduit par *la salle était grande*, n'est pas extrait car le second 申し分なかった est considéré comme ayant plus d'importance que le premier 部屋は広くて dans cette proposition, lorsque l'algorithme interne et la position des mots sont appliqués.

Table A-2
Résultats possibles pour le texte en utilisant l'option « Uniquement les sentiments représentatifs »

Concept	Type
仲居さんは無愛想だった	<悪い - 対応が不親切>
申し分なかった	<満足>
満足	<満足>

Uniquement les conclusions

Cette option force l'extracteur à identifier et à extraire un mot clé de sentiment comme représentant la conclusion de la totalité de l'enregistrement ou du document. Tous les textes n'ont pas de conclusion, de sorte que dans certains cas, il est possible que rien ne soit extrait avec cette option pour une partie de texte donnée. En outre, plus l'enregistrement ou le document est long, plus l'analyse a du mal à identifier la conclusion principale. Bien que cela soit rare, il est possible que plusieurs conclusions soient extraites.

満足, traduit par *satisfait*, est considéré comme la conclusion principale des sentiments exprimés dans le texte.

Table A-3
Résultats possibles pour le texte en utilisant l'option « Uniquement les conclusions »

Concept	Type
満足	<満足>

Fonctionnement de la catégorisation

Lorsque vous créez des modèles de catégories dans IBM® SPSS® Modeler Text Analytics , vous disposez de plusieurs techniques de création de catégories. Etant donné que chaque ensemble de données est unique, le nombre de techniques et leur ordre d'application peuvent varier. Votre interprétation des résultats pouvant être différente de celle d'une autre personne, vous pouvez être amené à essayer plusieurs techniques de manière à déterminer celle qui donne les meilleurs résultats pour vos données textuelles. Dans SPSS Modeler Text Analytics , vous pouvez créer des modèles de catégories dans une session interactive qui vous permettra d'explorer et d'affiner vos catégories.

Dans ce manuel, la **création de catégories** fait référence à la génération de définitions et de classification de catégories à l'aide d'une ou de plusieurs techniques intégrées et la **catégorisation** fait référence au scoring, ou à l'étiquetage, processus par lequel des identificateurs uniques (nom/ID/valeur) sont affectés aux définitions de catégorie de chaque enregistrement ou de chaque document.

Pendant la création de catégories, les concepts et les types qui ont été extraits sont utilisés en tant que blocs de construction de vos catégories. Lorsque vous créez des catégories, les enregistrements ou les documents sont affectés aux catégories s'ils contiennent du texte qui correspond à un élément d'une définition de catégorie.

SPSS Modeler Text Analytics vous propose plusieurs techniques automatisées de classification supervisée qui vous permettent de catégoriser rapidement vos documents ou vos enregistrements. Chaque technique convient à certains types de données et de situation. Cependant, il est souvent judicieux de combiner des techniques dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. Aussi est-il possible de voir un concept figurer dans plusieurs catégories ou de rencontrer des catégories redondantes.

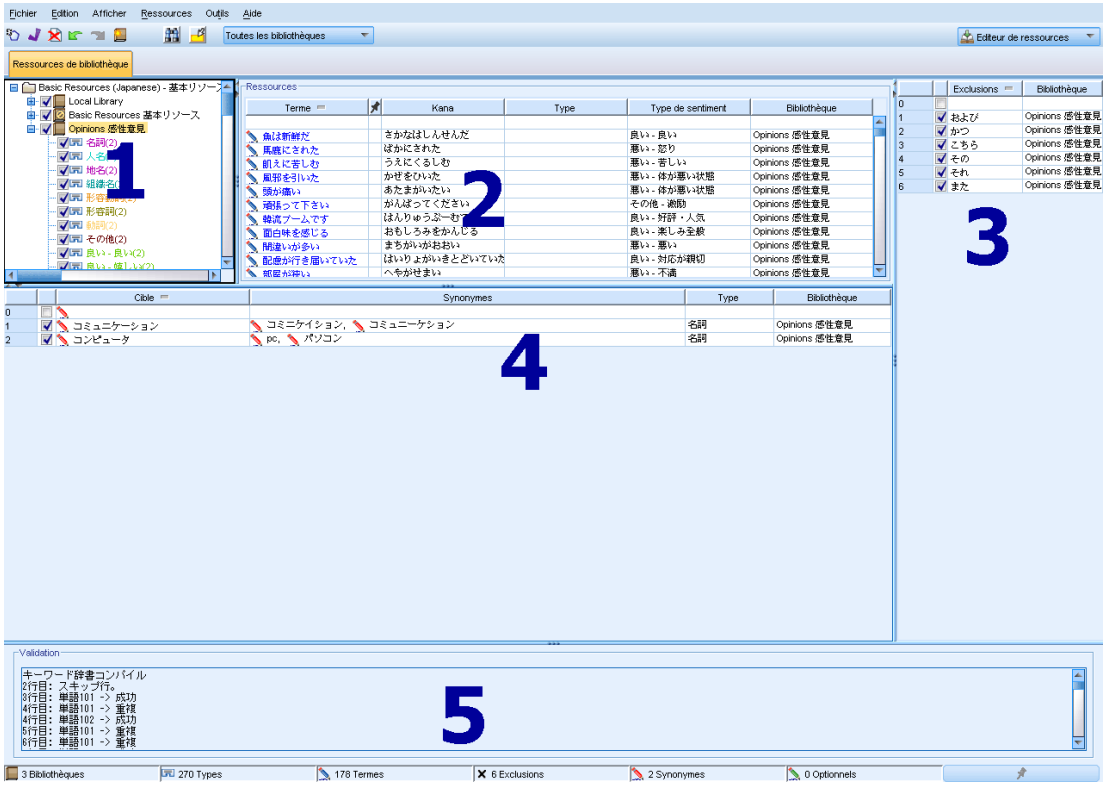
Modification des ressources pour du texte en japonais

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Depuis la version 14 de IBM® SPSS® Modeler Text Analytics , un nouveau modèle et un package d'analyse de texte (TAP) sont disponibles pour la langue japonaise. Vous pouvez effectuer les modifications des ressources en ajoutant et éditant des termes pour les personnaliser en fonction de vos données. Le package d'analyse de texte contient également une ensemble de catégories composé de catégories représentant des sentiments positifs, négatifs et des sentiments contextuels/génériques.

Vous pouvez utiliser vos ressources dans l'Editeur de ressources et l'Editeur de modèle. Les éditeurs fonctionnent de manière similaire pour toutes les langues ; il y a toutefois des différences significatives pour le texte en japonais, comme décrit ici.

Figure A-1
Vue Editeur de ressources pour du texte en japonais



Les points suivants mettent l'accent sur les différences majeures qui existent lors de l'utilisation de ressources en japonais. Pour une description générale des quatre panneaux principaux dans l'onglet Ressources de bibliothèque, consultez Interface de l'éditeur sur p. 275.

1. Panneau des bibliothèques. Cette zone, située dans le coin supérieur gauche, fonctionne globalement comme pour les autres langues. Il existe cependant quelques différences telles que l'impossibilité de créer de nouveaux types ou d'en renommer. Pour plus d'informations, reportez-vous à la section Utilisation des bibliothèques dans le chapitre 16 sur p. 290.

2. Panneaux des termes pour les déclarations de types. Ce panneau, situé à droite du panneau de l'arborescence des bibliothèques, est différent pour le texte japonais. En plus du nom du terme, vous pouvez également ajouter le nom en kana et sélectionner un ou deux types auxquels vous souhaitez associer ce terme. Cependant, vous ne pouvez pas générer de formes fléchies des termes ni affecter des options de mise en correspondance pour des termes japonais comme vous le pourriez dans les autres langues. Pour plus d'informations, reportez-vous à la section Panneau Arborescence des bibliothèques japonaises, panneau des types et panneau des termes sur p. 376.

3. Panneau Dictionnaire de substitutions/synonymes. Dans les ressources de texte japonais, vous trouverez un onglet Synonyme dans lequel vous pouvez définir tous les synonymes pour vos ressources. Dans l'onglet Synonyme, il existe une colonne supplémentaire Type dans laquelle vous devez désigner un type pour les synonymes entrés. Pour plus d'informations, reportez-vous à la section Utilisation du dictionnaire des synonymes pour du texte japonais sur p. 383.*Remarque* : L'onglet Eléments optionnels n'est pas disponible car il ne s'applique pas au texte japonais.

4. Panneau Dictionnaire d'exclusions. Il n'y a pas de différences dans ce panneau pour les ressources en japonais en-dehors du fait que le caractère générique * n'est pas pris en charge.

5. Panneau Validation. Pour le texte japonais, il existe un panneau de validation supplémentaire servant à vérifier vos ressources avant l'extraction. Lors de l'extraction du texte japonais, le moteur d'extraction recompile automatiquement les ressources si les modifications sont détectées avant le début du processus d'extraction. Pour éviter des erreurs potentielles pendant l'extraction, vous pouvez recompiler et valider les ressources avant l'extraction afin de pouvoir corriger les erreurs rencontrées. Pour plus d'informations, reportez-vous à la section Validation et compilation des ressources en japonais sur p. 385.

Remarque : ces onglets ne sont pas disponibles car il n'existe pas de ressources avancées ou de règles des liens du texte pouvant être modifiées pour du texte japonais.

Panneau Arborescence des bibliothèques japonaises, panneau des types et panneau des termes

La façon dont vous travaillez avec les bibliothèques et les types pour les ressources en japonais est très similaire à celle des autres langues. Pour plus d'informations, reportez-vous à la section Déclarations de types dans le chapitre 17 sur p. 304.

Mais il existe quelques différences majeures, notamment :

- Les ressources en japonais contiennent un ensemble de types distinct. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais sur p. 378.
- Les types ne peuvent pas être créés ni renommés, mais leurs propriétés peuvent être modifiées. Pour plus d'informations, reportez-vous à la section Modification des propriétés de types japonais sur p. 382.
- Vous pouvez ajouter et modifier des termes, y compris la spécification d'un nom en kana pour un terme, ainsi que l'attribution à un type et à un type de sentiment secondaire. Pour plus d'informations, reportez-vous à la section Panneau Arborescence des bibliothèques japonaises, panneau des types et panneau des termes sur p. 376.

Le panneau de l'arborescence de bibliothèques affiche les bibliothèques ainsi que les déclarations de types. Si vous sélectionnez une bibliothèque ou un type dans le panneau de gauche, un panneau des termes affiche à droite les termes des bibliothèques ou des déclarations de types sélectionnées. Vous pouvez ajouter des termes directement à une déclaration de types : soit directement dans le panneau de termes, soit par l'intermédiaire de la boîte de dialogue Ajouter des termes. Les termes que vous ajoutez peuvent être des mots simples ou composés. La ligne vide, située en haut de la liste, vous permet d'ajouter de nouveaux termes à tout moment.

Lorsque vous définissez un terme dans une déclaration de types, il est considéré comme un nom par défaut et est automatiquement affecté au type <名詞>. Mais vous pouvez modifier le type en un autre type de base comme <動詞>, <形容詞>, <地名> etc. Si le moteur d'extraction trouve ce terme dans la même catégorie grammaticale que le type auquel vous l'avez affecté dans la colonne Type, il sera alors affecté à ce type et extrait. Vous pouvez également affecter ce terme à un des types de sentiment dans la colonne Type de sentiment. Ensuite, lorsque vous utilisez l'analyseur secondaire de sentiments, le texte est traité une deuxième fois afin de rechercher des

termes et de les affecter aux types de sentiment. De plus, si vous définissez à la fois un type de sentiment et un type de base, et que le moteur d'extraction découvre que ce terme correspond aux deux types lorsque l'analyse secondaire de sentiment a été effectuée, alors le type de sentiment est prioritaire et est affiché dans le panneau des résultats d'extraction et dans les résultats de l'analyse des liens du texte. Par exemple, si un verbe a été extrait comme type de verbe <**動詞**> et est également un type positif comme « aimé » alors ce terme sera affiché comme appartenant au type positif dans l'interface car la capture de sentiments est souvent plus intéressante qu'une simple catégorie grammaticale.

Figure A-2
Panneaux Bibliothèque et Terme pour les ressources en japonais

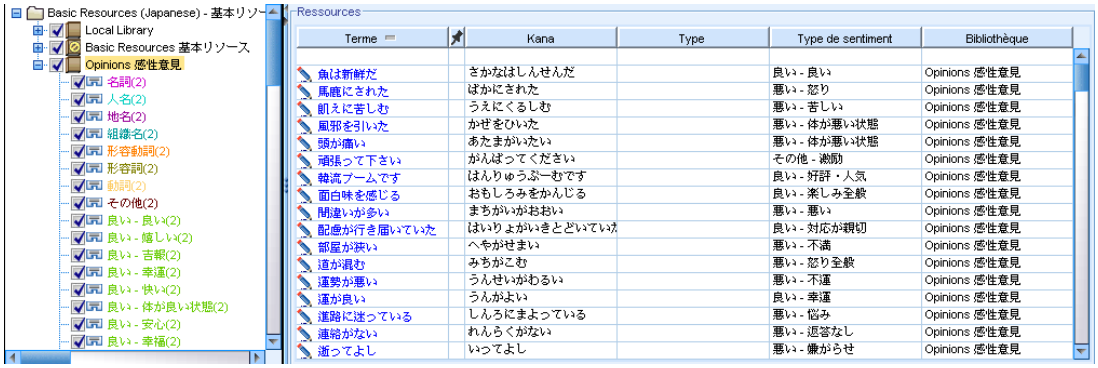


Table A-4
Descriptions des colonnes du panneau des termes

Nom de colonne	Description de colonne
Terme	Entrez des mots simples ou composés dans la cellule. La couleur d'affichage du terme dépend de la couleur du type dans lequel le terme est enregistré ou a été imposé. Vous pouvez changer les couleurs des types dans la boîte de dialogue Propriétés de type. Pour plus d'informations, reportez-vous à la section Modification des propriétés de types japonais sur p. 382. En général le terme est écrit en kanji, mais il peut aussi comprendre des kana. **Important !** La saisie de verbes utilisant des caractères Katakana n'est pas prise en charge.
Forcer	Cliquer sur une icône de punaise et la placer dans cette cellule indique au moteur d'extraction d'ignorer les autres occurrences de ce même terme dans les autres bibliothèques. Pour plus d'informations, reportez-vous à la section Ajout des termes forcés dans le chapitre 17 sur p. 312. Ceci fonctionne de la même manière pour toutes les langues.
Kana	Entrez l'orthographe kana du nom du terme Kanji.
Type	Sélectionnez le nom du type de base auquel ce terme doit être affecté. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais sur p. 378.
Type de sentiment	Si une analyse secondaire est réalisée, sélectionnez le nom du type de sentiment auquel le terme doit être attribué. Pour plus d'informations, reportez-vous à la section Types disponibles pour du texte en japonais sur p. 378.
Bibliothèque	Sélectionne la bibliothèque dans laquelle votre terme est stocké. Pour transférer un terme dans un autre type de l'arborescence, faites-le glisser et déposez-le sur l'autre bibliothèque.

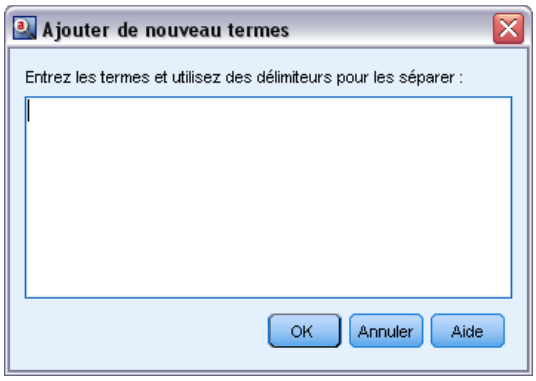
Pour ajouter un terme simple à une déclaration de types

- Dans l'arborescence de la bibliothèque, sélectionnez la déclaration de types à laquelle vous voulez ajouter le terme.
- Dans la liste de termes du panneau central, entrez votre terme dans la première cellule vide disponible et définissez les options souhaitées pour ce terme.

Pour ajouter plusieurs termes à une déclaration de types

- Dans l'arborescence de la bibliothèque, sélectionnez la déclaration de types à laquelle vous voulez ajouter des termes.
- A partir des menus, sélectionnez Outils> Nouveaux termes. La boîte de dialogue Ajouter des nouveaux termes apparaît.

Figure A-3
Boîte de dialogue Ajouter des nouveaux termes



- Entrez les termes que vous souhaitez ajouter à la déclaration de types sélectionnée en saisissant les termes ou en collant un ensemble de termes. Si vous entrez plusieurs termes, séparez-les au moyen du séparateur défini dans la boîte de dialogue Options ou ajoutez un terme par nouvelle ligne. Pour plus d'informations, reportez-vous à la section Définition des options dans le chapitre 8 sur p. 131.
- Cliquez sur OK pour ajouter les termes à la déclaration. La boîte de dialogue se ferme et les nouveaux termes apparaissent dans la déclaration.

Types disponibles pour du texte en japonais

Vous ne pouvez pas ajouter de nouveaux types pour les ressources en japonais, cependant vous pouvez ajouter ou supprimer des termes pour celles-ci. Les tableaux suivant incluent l'ensemble des types japonais actuellement disponibles.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Types pour l'extraction de base

À chaque fois qu'une extraction est lancée, les types suivants sont utilisés.

Table A-5
Types pour l'extraction de base

Types	Description
名詞	Mots faisant référence à des choses, comme « voiture » et « film ». Néanmoins, les noms de personnes, de lieux et d'organisations sont classés dans des catégories distinctes.
人名	Noms correspondant à des noms de personnes spécifiques comme « Tokugawa » et « leyasu ». Les combinaisons de prénoms et de noms de famille, comme « Tokugawa leyasu » sont également des noms propres.
地名	Noms comme « Tokyo » et « Londres » faisant référence à un lieu précis.
組織名	Noms faisant référence à des entreprises et des organisations spécifiques, comme « La Fédération des Organismes Economiques ».
形容動詞	Mots comme « calme (shizuka) » qui décrivent les caractéristiques ou l'état d'une chose et qui peuvent être utilisés dans des phrases comme « pas [adjectif] (~de nai) »et « [adjectif] chose (~ na koto) ».
形容詞	Mots comme « amusant (tanoshi) » qui décrivent les caractéristiques ou l'état d'une chose et qui peuvent être utilisés dans des phrases comme « devenir [adjectif] (~ku naru) » et « une chose [adjectif] (~ i koto) ».
動詞	Les mots qui décrivent le mouvement ou l'action, y compris les verbes de type I (base verbale consonne), de type II (base verbale voyelle) et les verbes irréguliers (sagyou henkaku et kagyou henkaku).
その他	Mots comme les adverbes, les adjectifs pronominaux, les conjonctions et les interjections : par exemple, « calme », "« quel que soit », « ensuite » et « merci ».

Types pour l'analyse des sentiments

À chaque fois que vous sélectionnez l'analyseur secondaire pour l'extraction des sentiments, vous obtenez un grand nombre de types en plus des 8 types de base.

Table A-6
Types pour l'analyse des sentiments

Types	Description
良い - 良い	Expressions de choses positives pouvant être classifiées comme « bonnes ».
良い - 嬉しい	Décrit un évènement désirable qui produit une stimulation agréable.
良い - 吉報	Décrit un évènement plaisant qui ne peut être possible qu'avec des efforts considérables.
良い - 幸運	Décrit un évènement heureux qui ne peut être possible que par hasard ou en raison d'une coïncidence incroyable.
良い - 快い	Suggère que quelque chose est un stimulus ou un environnement qui déclenche une sensation physiologique agréable.
良い - 体が良い状態	Décrit un état dans lequel le corps n'est ni malade, ni blessé, ni fatigué ou un état dans lequel l'état physique s'améliore.
良い - 安心	Suggère que quelqu'un est calme et ne risque ni blessure ni dérangement.
良い - 幸福	Indique que quelqu'un a obtenu des conditions favorables ou un effet positif grâce à sa propre action ou aux circonstances de sa naissance.
良い - 満足	Décrit un évènement désirable qui apaise l'esprit.
良い - 美味しい	Indique qu'une nourriture a bon goût.
良い - 効果が満足	Implique qu'une chose a produit l'effet désiré.
良い - 感動	Suggère que l'importance, la signification ou la valeur de quelque chose est extrêmement positive.

Types	Description
良い - 感謝	Suggère que quelqu'un considère les actions de quelqu'un autre de manière positive.
良い - 祝福	Exprime l'idée que la situation d'une autre personne est favorable (à un degré acceptable pour l'interlocuteur).
良い - 喜び全般	Autres évènements positifs avec peu de rapport avec l'interlocuteur.
良い - 楽しい	Indique ou anticipe des activités en lien avec les relations amicales, les loisirs et le divertissement.
良い - 可笑しい	Indique que quelque chose a une qualité humoristique qui génère une stimulation plaisante.
良い - 笑い	Exprime un sourire ou un rire provoqué par quelque chose de positif et/ou de drôle.
良い - 期待	Prédit qu'un évènement positif se produira dans le futur.
良い - 楽しみ全般	Autres évènements agréables et/ou activités/comportements positifs avec peu de rapport avec l'interlocuteur.
良い - 金額への賞賛	Suggère que, du point de vue de l'acheteur, quelque chose a une valeur monétaire intéressante.
良い - 対応が早い	Suggère qu'un service a été fourni ou terminé à temps.
良い - 対応が親切	Suggère que l'attitude ou le comportement du fournisseur de service était plein de sollicitude.
良い - 説明が良い	Exprime l'idée que le type et/ou la quantité d'informations, et/ou le moyen de les fournir, est approprié.
良い - 対応への賞賛	Points de vue, autres que ceux cités précédemment, qui font l'éloge du fournisseur de service.
良い - 褒め・ 賞賛	Points de vue, autres que ceux cités précédemment, qui font l'éloge des caractéristiques, capacités et/ou fonctionnement d'un objet.
良い - 好き	Exprime le désire de posséder ou de se rapprocher de quelque chose.
良い - 入会希望	Décrit le désir de faire partie ou de rester membre d'un groupe.
良い - 買いたい	Suggère que quelqu'un veut ou prévoit d'utiliser de l'argent pour obtenir quelque chose.
良い - 好評・ 人気	Indique que le nombre de personnes qui souhaitent ou apprécient quelque chose a dépassé un certain niveau.
良い - 売れた	Indique la présence de personnes qui achètent quelque chose ou que le nombre ou la valeur des achats dépassent un certain niveau.
悪い - 悪い	Expressions de choses négatives pouvant être classifiées comme « mauvaises ».
悪い - 怒り	Une sensation précise de colère ressentie lorsque quelque chose ne se passe pas comme prévu.
悪い - 批判	Exprime l'idée qu'une autre personne a échoué à faire le bon choix.
悪い - お叱り	Mots ou actions qui intimident une autre personne pour pouvoir faire selon son désir.
悪い - 誹謗・ 中傷	Mots utilisés pour démontrer une opinion excessivement négative de quelqu'un d'autre.
悪い - 軽蔑	Suggère qu'une autre personne manque de caractère, de capacités et/ou d'autres qualités.
悪い - 恨み	Exprime la revanche ou le ressentiment face à un préjudice provoqué par une autre personne.
悪い - 嫌がらせ	Mots utilisés dans le but d'empêcher une communication.
悪い - 不満	Un sentiment déplaisant provoqué par l'incapacité à obtenir une chose ou un état désiré.
悪い - 不味い	Indique qu'une nourriture a mauvais goût.

Types	Description
悪い - 効果が不満	Suggère que quelque chose n'a pas produit l'effet escompté.
悪い - 金額が不満	Suggère que, du point de vue de l'acheteur, la valeur monétaire d'une chose n'est pas acceptable.
悪い - 対応への不満	Suggère que le fournisseur d'un service est en faute.
悪い - 対応が遅い	Suggère qu'un service n'a pas été effectué / terminé dans les temps ou qu'un service doit encore être effectué.
悪い - 対応が不親切	Indique un sentiment désagréable provoqué par l'attitude ou le comportement d'un fournisseur de service.
悪い - 説明が悪い	Exprime l'idée que le type et/ou la quantité d'informations, et/ou le moyen de les fournir, n'est pas approprié.
悪い - 返答なし	Indique que le fournisseur de service ne parvient pas à proposer la réponse désirée, même si la situation en demande une.
悪い - 不快	Suggère que quelque chose est un stimulus ou un environnement qui déclenche une sensation physiologique négative.
悪い - 怒り全般	Sentiments de colère autres que ceux cités précédemment. Colère générale ressentie par l'organisation ou l'entreprise de l'interlocuteur, ou description des évènements provoqués par cette colère.
悪い - 悲しい	Un sentiment déplaisant spécifique ressenti lorsque quelqu'un perd ou ne peut pas obtenir quelque chose.
悪い - 凶報	Exprime l'idée qu'un certain objectif ne peut pas être atteint malgré de gros efforts.
悪い - 不運	Indique un résultat négatif provoqué par de la malchance, et qui n'est pas de sa faute.
悪い - ショック	Suggère que quelqu'un est contrarié par quelque chose d'imprévu ou de négatif et n'est pas capable de trouver une solution appropriée.
悪い - 残念	Un sentiment négatif ressenti lorsque quelque chose qui est prévu n'arrive pas.
悪い - 落胆	Un état dans lequel quelqu'un est envahi par un sentiment de malheur et de déception.
悪い - 諦め	Suggère qu'une chose négative, expérimentée par l'interlocuteur ou une autre personne ne peut pas s'améliorer.
悪い - 後悔	Exprime l'idée que dans le passé, quelqu'un n'est pas parvenu à faire le bon choix, même si ce choix était disponible.
悪い - 謝罪	Indique que l'interlocuteur sait qu'il a causé du tort à quelqu'un.
悪い - 淋しい	Exprime l'idée que le contact avec les autres est rare ou que ceux avec qui le contact est possible sont rares.
悪い - 哀れみ	Exprime l'idée que la situation de quelqu'un d'autre est bien pire que celle de l'interlocuteur.
悪い - 悩み	Indique que quelqu'un doit faire un choix mais est incapable de choisir parmi les options disponibles.
悪い - 困っている	Exprime l'idée qu'il n'existe pas de réponse adaptée à une situation demandant une action.
悪い - 苦しい	Exprime un état psychologique déplaisant dans lequel quelqu'un est incapable d'agir normalement en raison de causes externes ou en raison de ses propres erreurs.
悪い - 体が悪い状態	Décrit un état dans lequel le corps est malade, blessé, et/ou fatigué ou un état dans lequel l'état physique ne s'améliore pas.
悪い - 不安	Exprime l'idée que quelque chose de positif peut s'arrêter ou peut ne pas répondre aux attentes.
悪い - 恐怖	Suggère qu'une chose pourrait causer du tort ou blesser quelqu'un.

Types	Description
悪い - 悲しみ全般	Sentiments de tristesse autres que ceux mentionnés précédemment, comme la tristesse générale à propos de quelque chose qui n'a pas été précisé.
悪い - 嫌い	Indique que quelqu'un souhaite éloigner quelque chose ou s'en éloigner.
悪い - 退会希望	Décrit le désir de quitter un groupe ou de ne pas en devenir membre.
悪い - 買いたくない	Suggère que quelqu'un ne souhaite pas quelque chose ou n'a pas l'intention de payer pour cela.
悪い - 不評・ 不人気	Indique que le nombre de personnes aimant une certaine chose n'a pas atteint un certain niveau ou qu'il y a trop de monde ayant un sentiment négatif envers cette chose.
悪い - 売れていない	Indique l'absence de personnes qui achètent quelque chose ou que le nombre ou la valeur des achats n'a pas atteint un certain niveau.
その他 - 疑問	Expressions qui demandent des informations nécessitant un examen plus poussé ou des pensées supplémentaires de la part d'une autre personne.
その他 - 問い合わせ	Expressions qui demandent des informations qu'une autre personne possède déjà.
その他 - 要望	Expressions qui ordonnent à une autre personne (lorsque l'autre personne est directement en faute ou d'un rang inférieur à celui de l'interlocuteur) de résoudre un problème.
その他 - 提案・ 忠告	Expressions qui ordonnent à une autre personne (lorsque l'autre personne est directement en faute ou d'un rang inférieur à celui de l'interlocuteur) de mieux se comporter.
その他 - お願い	Expressions qui ordonnent à une autre personne, lorsque l'autre personne n'est pas en faute ou n'est pas d'un rang inférieur à celui de l'interlocuteur, de faire quelque chose.
その他 - 激励	Expressions qui encouragent une autre personne ou descriptions de comportement encourageant.
その他 - 勧誘	Expressions qui ordonnent à une autre personne de faire quelque chose avec l'interlocuteur.
その他 - 驚き	Exprime l'idée que la soudaineté ou la taille d'un évènement transcende le jugement/la compréhension rationnelle.
評価なし - 評価なし	Pas d'expression de jugement.

Modification des propriétés de types japonais

Bien que vous ne puissiez pas créer de type dans les ressources en japonais, vous pouvez afficher et modifier les propriétés des types. Notez que certaines options, telles que l'option de mise en correspondance ou les formes fléchies, ne s'appliquent pas au texte en japonais.

Figure A-4
Boîte de dialogue Propriétés du type pour les ressources en japonais

Nom. Le nom de la déclaration de types.

Ajouter à. Ce champ indique la bibliothèque dans laquelle vous allez créer votre nouvelle déclaration de types.

Couleur. Ce champ vous permet de distinguer les résultats de ce type dans l'interface. Si vous sélectionnez Par défaut, la couleur par défaut du type est également utilisée pour cette déclaration de types. La couleur par défaut se configure dans la boîte de dialogue des options. Pour plus d'informations, reportez-vous à la section Options : Onglet Affichage dans le chapitre 8 sur p. 133. Si vous sélectionnez Personnalisé, sélectionnez une couleur dans la liste déroulante.

Annotations. Ce champ est facultatif et sert à entrer des commentaires ou des descriptions.

Pour afficher ou modifier les propriétés de type

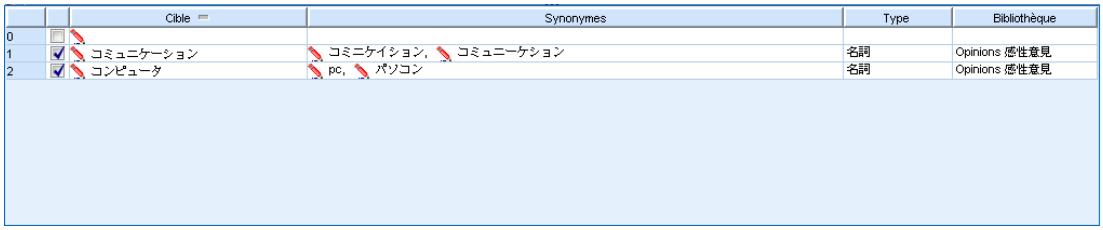
- Sélectionnez le type dont vous souhaitez afficher les propriétés.
- Cliquez avec le bouton droit de la souris, puis choisissez Propriétés de type dans le menu contextuel. La boîte de dialogue Propriétés de type apparaît.
- Procédez aux modifications nécessaires.
- Cliquez sur OK pour enregistrer les modifications dans la déclaration de types.

Utilisation du dictionnaire des synonymes pour du texte japonais

Pour le texte japonais, le dictionnaire des substitutions ne contient qu'un onglet permettant de gérer vos synonymes, l'onglet Synonyme. Les synonymes sont des mots ayant le même sens. Vous pouvez également utiliser des synonymes pour regrouper des termes et leurs abréviations, ou pour réunir les mots fréquemment mal orthographiés sous la version correcte du mot.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Figure A-5
Entrées de synonymes pour le texte en japonais



Une définition de synonyme est constituée de deux parties. Le terme cible est le terme sous lequel vous souhaitez que le moteur du programme d'extraction regroupe tous les termes synonymes. A moins que ce terme cible ne soit utilisé comme synonyme d'un autre terme cible ou ne soit exclu, il deviendra probablement le concept apparaissant dans le panneau Résultats d'extraction. La liste des synonymes est la liste des termes qui seront regroupés sous le terme cible.

Dans l'onglet Synonymes, vous pouvez entrer une définition de synonyme dans la ligne vide qui figure en haut du tableau. Commencez par définir le terme cible et ses synonymes. Vous pouvez également sélectionner la bibliothèque dans laquelle vous souhaitez stocker la définition. Toutes les occurrences des synonymes sont ensuite regroupées sous le terme cible lors de l'extraction finale. Pour plus d'informations, reportez-vous à la section Ajout de termes dans le chapitre 17 sur p. 308.

Quand vous créez vos déclarations de types, vous pouvez entrer un terme puis penser à trois ou quatre synonymes de ce terme. Dans ce cas, vous pouvez saisir tous les termes puis votre terme cible dans le dictionnaire de substitution avant de faire glisser les synonymes.

Important ! Les synonymes pour le texte en japonais ne prennent pas en charge les caractères génériques et les caractères spéciaux.

Pour ajouter une entrée de synonyme

- Dans la ligne vide en haut du tableau dans l'onglet Synonyme du panneau Substitution, entrez le terme cible dans la colonne Cible. Le terme cible que vous entrez apparaît en couleur. Cette couleur représente le type dans lequel le terme apparaît ou a été imposé, le cas échéant. Si le terme apparaît en noir, il n'est alors inclus dans aucune déclaration de types.

- Cliquez dans la deuxième cellule à droite de la cible et entrez l'ensemble de synonymes. Séparez chaque entrée à l'aide du séparateur global défini dans la boîte de dialogue Options. Tous les synonymes saisis doivent être du même type. Pour plus d'informations, reportez-vous à la section Définition des options dans le chapitre 8 sur p. 131. Les termes que vous entrez apparaissent en couleur. Cette couleur représente le type dans lequel le terme apparaît. Si le terme apparaît en noir, il n'est alors inclus dans aucune déclaration de types.

- Dans la troisième colonne, la colonne Type, choisissez un type pour ces synonymes. Cependant, la cible prend le type attribué pendant l'extraction. Mais si la cible n'a pas été extraite en tant que concept, alors le type répertorié dans cette colonne est affecté à la cible dans les résultats de l'extraction.

- Cliquez dans la dernière cellule pour sélectionner la bibliothèque dans laquelle vous voulez stocker cette définition de synonyme.

Remarque : Ces instructions indiquent la façon d'effectuer des modifications dans la vue Editeur de ressources ou dans la vue Editeur de modèle. N'oubliez pas que vous pouvez également réaliser ce type de réglage directement à partir du panneau Résultats d'extraction , du panneau Données, du panneau Catégorie ou de la boîte de dialogue Définitions de clusters dans les autres vues. Pour plus d'informations, reportez-vous à la section Affinage des résultats de l'extraction dans le chapitre 9 sur p. 154.

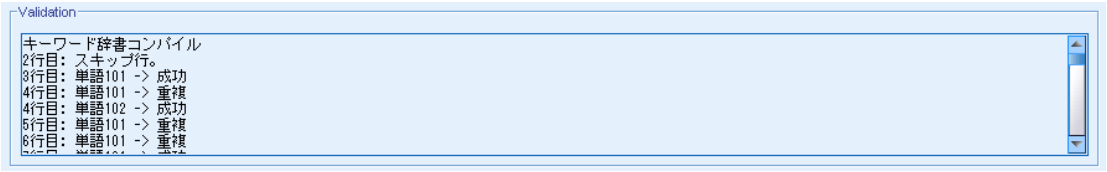
Validation et compilation des ressources en japonais

Pour le texte japonais, il existe un panneau de validation supplémentaire servant à vérifier vos ressources avant l'extraction. Avant le début de l'extraction d'un texte en japonais, le moteur d'extraction recompile automatiquement les ressources lorsque des modifications sont détectées. Si une erreur est trouvée pendant l'extraction, le processus risque de ne pas s'exécuter correctement.

Pour éviter les erreurs de compilation, nous vous recommandons de valider et de compiler les ressources après avoir effectué les modifications dans l'Editeur de ressources ou l'Editeur de modèle. Si un message d'erreur s'affiche, vous pouvez procéder aux corrections et tenter une nouvelle validation.

Remarque : l'extraction de texte japonais est disponible dans IBM® SPSS® Modeler Premium.

Figure A-6
Panneau de validation pour le texte japonais



Pour valider les ressources

- A partir des menus, sélectionnez Outils> Valider des ressources. Le panneau de validation s'ouvre et affiche les messages de compilation et d'erreur.

Autres exceptions pour le japonais

Ressources internes remplaçant les ressources définies par l'utilisateur

Pour le texte en japonais, les ressources par défaut contiennent des ressources de base internes précompilées. Ces ressources internes ne peuvent pas être modifiées. C'est la raison pour laquelle vous pouvez utiliser l'Editeur de ressources ou l'Editeur de modèle pour effectuer les modifications et les réglages. La majeure partie du temps, les termes, les synonymes et les entrées de listes d'exclusion que vous définissez dans vos ressources seront prioritaires aux ressources internes précompilées. Cependant il existe plusieurs exceptions comme les exemples suivants l'indiquent.

- Il existe des cas où l'ajout de ces termes à un type particulier n'a pas d'incidence sur les résultats d'extraction. Ceci a de fortes chances de se produire lorsque les données comportent des phrases longues qui comprennent plusieurs éléments morphologiques, ponctuations ou symboles. En outre, comme les ressources pour le texte japonais comprennent un grand nombre de termes communs précompilés, il existe quelques mots communs qui seront toujours forcés dans une définition linguistique spécifique.
- Il est possible que vous ne puissiez pas exclure de termes tels que ある, いる, ou なる car le moteur d'extraction forcera toujours l'extraction de ces termes.
- Bien qu'il soit possible de modifier le type du terme 東京 de <地名> en <名詞>, le moteur d'extraction ignorera ces changements si vous essayez de modifier le type d'un terme de <地名> en <動詞> ou en <形容詞> à l'aide du dictionnaire de mots-clés (types).
- Parfois les modifications effectuées dans l'Editeur de ressources ou l'Editeur de modèle ont une incidence sur les résultats d'extraction d'une phrase mais pas d'une autre car le processus d'extraction se termine en référençant les mots cooccurrents dans chaque phrase.

Problème d'affichage des katakanas de demi-largeur

Les caractères katakana de demi-largeur sont convertis en interne en caractères de pleine largeur pendant l'extraction mais apparaissent toujours dans la demi-largeur d'origine lors de l'affichage dans le panneau Données qui se trouve dans la session interactive (pour le nœud de Text Mining uniquement). Veuillez noter que les caractères katakana de demi-largeur ne peuvent pas être mis en surbrillance dans le panneau Données. Pour éviter ce problème, convertissez tous vos enregistrements en katakanas de pleine largeur avant l'exécution.

Utilisation des minuscules et des majuscules

Les caractères alphabétiques en majuscule sont temporairement convertis en minuscules lorsqu'ils sont lus dans l'application. Mais le panneau Données affichera le texte à l'aide de la même casse que le texte d'origine. Les minuscules et les majuscules sont traitées de la même façon dans ce produit.

Annexe
B

Avis

Ces informations sont développées pour des produits et des services proposés dans le monde entier.

Il est possible qu'IBM ne propose pas les produits, services ou fonctionnalités présentés dans ce document dans d'autres pays. Consultez votre représentant IBM local pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'a pas pour but d'affirmer ou de sous-entendre que seul le produit, programme ou service IBM peut être utilisé. Tout produit, programme ou service de fonctionnalité équivalente, conforme au droit sur la propriété intellectuelle d'IBM, peut être utilisé à la place. Cependant, il est de la responsabilité de l'utilisateur d'évaluer et de vérifier le fonctionnement de tout produit, programme ou service autre qu'IBM.

IBM peut posséder des brevets ou des applications de brevet en cours qui couvrent le sujet décrit dans ce document. En aucun cas, l'obtention de ce document ne vous accorde de licence pour ces brevets. Vous pouvez envoyer vos questions sur les licences par écrit à :

IBM Director of LicensingIBM CorporationNorth Castle DriveArmonk, NY 10504-1785U.S.A.

Pour des informations sur le jeu de caractères à deux-octets (DBCS) des licences, contactez le service sur la propriété intellectuelle d'IBM de votre pays ou envoyez vos questions par écrit à :

Intellectual Property LicensingLegal and Intellectual Property LawIBM Japan Ltd.1623-14, Shimotsuruma, Yamato-shiKanagawa 242-8502 Japon

Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun autre pays dans lequel ces dispositions sont contraires au droit local : INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ÉTAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, NOTAMMENT, LES GARANTIES IMPLICITES DE NON-VIOLATION, DE VALEUR MARCHANDE OU D'ADÉQUATION À UN BUT PRÉCIS. Certains pays n'autorisent pas les clauses de non responsabilité des garanties explicites ou implicites de certaines transactions et par conséquent, il est possible que cette déclaration ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou typographiques. Ces informations sont modifiées périodiquement ; les modifications seront intégrées aux nouvelles éditions de la publication. IBM peut effectuer des améliorations et/ou des modifications sur les produits et/ou les programmes décrits dans cette publication à tout moment et sans notification préalable.

Toutes les références, contenues dans ces informations, à des sites Web autres qu'IBM sont fournies dans un but pratique uniquement et ne constituent en aucun cas une approbation de ces sites Web. Le matériel de ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM a le droit d'utiliser ou de distribuer toutes les informations que vous avez fournies, comme l'entreprise IBM le juge approprié, sans aucune obligation envers vous.

Les détenteurs de licence pour ce programme qui souhaitent obtenir des informations dans le but d'autoriser : (i) l'échange d'informations entre les programmes créés de manière indépendante et d'autres programmes (y compris celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

IBM Software GroupÀ l'attention de : Obtention de licences 233 S. Wacker Dr.Chicago, IL 60606USA

Ces informations peuvent être disponibles conformément aux conditions générales appropriées et dans certains cas, être soumises à des frais.

Le programme sous licence décrit dans ce document et tout le matériel sous licence associé disponible sont fournis par IBM conformément à l'accord client IBM, au contrat de licence sur les programmes internationaux IBM et à tout accord équivalent entre nous.

Toutes les données sur les performances contenues ici ont été obtenues dans un environnement contrôlé. Par conséquent, les résultats obtenus dans d'autres environnements d'exploitation peuvent varier de manière significative. Il est possible que certaines mesures aient été obtenues sur des systèmes en cours de développement et il n'existe aucune garantie quant au fait que ces mesures seront les mêmes sur les systèmes publics. De plus, certaines mesures peuvent avoir été obtenues par extrapolation. Les résultats réels peuvent varier. Les utilisateurs de ce document doivent vérifier les données applicables à leur environnement spécifique.

Les informations concernant les produits autres qu'IBM ont été obtenues auprès des fournisseurs de ces produits, de leurs annonces publiées ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut pas confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre information associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Ces informations contiennent des exemples de données et de rapports utilisés dans les opérations quotidiennes d'entreprises. Pour les illustrer aussi précisément que possible, ces exemples contiennent des noms de personnes, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute similarité avec des noms et adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous utilisez la version électronique de ces informations, il est possible que les photographies et illustrations en couleurs n'apparaissent pas.

Marques

IBM, le logo IBM, ibm.com et SPSS sont des marques d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste actuelle des marques d'IBM est disponible sur le Web à l'adresse *http://www.ibm.com/legal/copytrade.hmtl*.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques commerciales ou déposées d'Adobe Systems Incorporated aux États-Unis et/ou dans d'autres pays.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium et Pentium sont des marques commerciales ou déposées d'Intel Corporation ou de ses filiales aux États-Unis ou dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux États-Unis, dans d'autres pays ou les deux.

Microsoft, Windows, Windows NT et le logo Windows sont des marques déposées de Microsoft Corporation aux États-Unis, dans d'autres pays ou les deux.

UNIX est une marque déposée de The Open Group aux États-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques de Sun Microsystems, Inc. aux États-Unis, dans d'autres pays, ou les deux.

Les autres noms de produits ou de services peuvent être des marques d'IBM ou d'autres entreprises.

Index