IBM SPSS Modeler Text Analytics 15 - Benutzerhandbuch

Hinweis: Lesen Sie vor der Verwendung dieser Informationen und des zugehörigen Produkts die allgemeinen Informationen unter Hinweise auf S. 396.

Diese Ausgabe gilt für IBM® SPSS® Modeler Text Analytics 15 und alle nachfolgenden Versionen und Abwandlungen, und zwar bis es in neuen Ausgaben anderweitig angegeben wird.

Vorwort

IBM® SPSS® Modeler Text Analytics bietet leistungsstarke Textanalysefunktionen, die hoch entwickelte linguistische Technologien und Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) benutzt, um eine schnelle Verarbeitung einer großen Vielfalt an unstrukturierten Textdaten zu ermöglichen und die Schlüsselkonzepte aus diesem Texte zu extrahieren und zu ordnen. Darüber hinaus können diese Konzepte mit SPSS Modeler Text Analytics in Kategorien zusammengefasst werden.

Bei ungefähr 80 % aller Daten, die in einem Unternehmen gespeichert sind, handelt es sich um Textdokumente— z. B. Berichte, Webseiten, E-Mails und Callcenter-Notizen. Text ist ein Schlüsselfaktor, der es einem Unternehmen ermöglicht, das Verhalten seiner Kunden besser zu verstehen. Ein System, das NLP verwendet, kann Konzepte, u. a. zusammengesetzte Ausdrücke, auf intelligente Art und Weise extrahieren. Außerdem ermöglicht die Kenntnis der zugrunde liegenden Sprache eine Klassifizierung von Begriffen in verwandte Gruppen, beispielsweise Produkte, Unternehmen oder Personen, wobei Bedeutung und Kontext verwendet werden. Folglich können Sie schnell ermitteln, ob die Informationen für Ihren Bedarf relevant sind. Diese extrahierten Konzepte und Kategorien können mit bestehenden strukturierten Daten, beispielsweise demografischen Informationen, kombiniert und in der vollständigen Suite der Data Mining-Tools von IBM® SPSS® Modeler auf die Modellierung angewendet werden, um bessere und fokussiertere Entscheidungen zu ermöglichen.

Linguistische Systeme sind wissensintensiv: Je mehr Informationen in den Wörterbüchern enthalten sind, desto höher ist die Qualität der Ergebnisse. SPSS Modeler Text Analytics enthält eine Reihe von linguistischen Ressourcen, z. B. Wörterbücher für Begriffe und Synonyme, Bibliotheken und Vorlagen. Darüber hinaus können Sie mit diesem Produkt die linguistischen Ressourcen Ihrem Umfeld entsprechend entwickeln und verfeinern. Bei der Feinabstimmung der linguistischen Ressourcen handelt es sich häufig um einen schrittweisen Prozess, der für einen genauen Abruf und die Kategorisierung der Konzepte erforderlich ist. Benutzerdefinierte Vorlagen, Bibliotheken und Wörterbücher für bestimmte Domänen, wie CRM und Genomforschung, sind ebenfalls eingeschlossen.

Informationen zu IBM Business Analytics

Die Software IBM Business Analytics liefert umfassende, einheitliche und korrekte Informationen, mit denen Entscheidungsträger die Unternehmensleistung verbessern können. Ein umfassendes Portfolio aus Business Intelligence, Vorhersageanalyse, Finanz- und Strategiemanagement sowie Analyseanwendungen bietet Ihnen sofort klare und umsetzbare Einblicke in die aktuelle Leistung und gibt Ihnen die Möglichkeit, zukünftige Ergebnisse vorherzusagen. Durch umfassende Branchenlösungen, bewährte Vorgehensweisen und professionellen Service können Unternehmen jeder Größe die Produktivität maximieren, Entscheidungen automatisieren und bessere Ergebnisse erzielen.

Als Teil dieses Portfolios unterstützt IBM SPSS Predictive Analytics-Software Unternehmen dabei, zukünftige Ereignisse vorherzusagen und proaktiv Maßnahmen zu ergreifen, um bessere Geschäftsergebnisse zu erzielen. Kunden aus Wirtschaft, öffentlichem Dienst und dem Bildungsbereich weltweit nutzen IBM SPSS-Technologie als Wettbewerbsvorteil für Kundengewinnung, Kundenbindung und Erhöhung der Kundenumsätze bei gleichzeitiger Eindämmung der Betrugsmöglichkeiten und Minderung von Risiken. Durch die Einbindung

von IBM SPSS-Software in ihre täglichen Operationen wandeln sich Organisationen zu "Predictive Enterprises" – die Entscheidungen auf Geschäftsziele ausrichten und automatisieren und einen messbaren Wettbewerbsvorteil erzielen können. Wenn Sie weitere Informationen wünschen oder Kontakt zu einem Mitarbeiter aufnehmen möchten, besuchen Sie die Seite *http://www.ibm.com/spss*.

Technischer Support

Kunden mit Wartungsvertrag können den technischen Support in Anspruch nehmen. Kunden können sich an den technischen Support wenden, wenn sie Hilfe bei der Arbeit mit IBM Corp.-Produkten oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Zur Kontaktaufnahme mit dem technischen Support besuchen Sie die Website von IBM Corp. unter *http://www.ibm.com/support*. Beachten Sie, dass Sie nach Ihrem Namen, dem Namen Ihrer Organisation und Ihrer Supportvereinbarung gefragt werden, wenn Sie Unterstützung anfordern.

Inhalt

Teil II: Interaktive Workbench

8 Modus “Interaktive Workbench” 119

9 Konzepte und Typen extrahieren 143

10 Kategorisieren von Textdaten 167

Anhänge

Teil I:
Text-Mining-Knoten

Kapitel 1

Info zu IBM SPSS Modeler Text Analytics

IBM® SPSS® Modeler Text Analytics bietet leistungsstarke Textanalysefunktionen, die hoch entwickelte linguistische Technologien und Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) benutzt, um eine schnelle Verarbeitung einer großen Vielfalt an unstrukturierten Textdaten zu ermöglichen und die Schlüsselkonzepte aus diesem Texte zu extrahieren und zu ordnen. Darüber hinaus können diese Konzepte mit SPSS Modeler Text Analytics in Kategorien zusammengefasst werden.

Bei ungefähr 80 % aller Daten, die in einem Unternehmen gespeichert sind, handelt es sich um Textdokumente— z. B. Berichte, Webseiten, E-Mails und Callcenter-Notizen. Text ist ein Schlüsselfaktor, der es einem Unternehmen ermöglicht, das Verhalten seiner Kunden besser zu verstehen. Ein System, das NLP verwendet, kann Konzepte, u. a. zusammengesetzte Ausdrücke, auf intelligente Art und Weise extrahieren. Außerdem ermöglicht die Kenntnis der zugrunde liegenden Sprache eine Klassifizierung von Begriffen in verwandte Gruppen, beispielsweise Produkte, Unternehmen oder Personen, wobei Bedeutung und Kontext verwendet werden. Folglich können Sie schnell ermitteln, ob die Informationen für Ihren Bedarf relevant sind. Diese extrahierten Konzepte und Kategorien können mit bestehenden strukturierten Daten, beispielsweise demografischen Informationen, kombiniert und in der vollständigen Suite der Data Mining-Tools von IBM® SPSS® Modeler auf die Modellierung angewendet werden, um bessere und fokussiertere Entscheidungen zu ermöglichen.

Linguistische Systeme sind wissensintensiv: Je mehr Informationen in den Wörterbüchern enthalten sind, desto höher ist die Qualität der Ergebnisse. SPSS Modeler Text Analytics enthält eine Reihe von linguistischen Ressourcen, z. B. Wörterbücher für Begriffe und Synonyme, Bibliotheken und Vorlagen. Darüber hinaus können Sie mit diesem Produkt die linguistischen Ressourcen Ihrem Umfeld entsprechend entwickeln und verfeinern. Bei der Feinabstimmung der linguistischen Ressourcen handelt es sich häufig um einen schrittweisen Prozess, der für einen genauen Abruf und die Kategorisierung der Konzepte erforderlich ist. Benutzerdefinierte Vorlagen, Bibliotheken und Wörterbücher für bestimmte Domänen, wie CRM und Genomforschung, sind ebenfalls eingeschlossen.

Bereitstellung. Mit IBM® SPSS® Modeler Solution Publisher können Sie Text-Mining-Streams für die Echtzeitbewertung unstrukturierter Daten verwenden. Die Möglichkeit, diese Streams zu verwenden, gewährleistet erfolgreiche Text-Mining-Implementierungen in einer geschlossenen Schleife. Ihre Organisation kann nun beispielsweise durch Anwendung von Vorhersagemodellen Notizen von eingehenden oder ausgehenden Anrufern analysieren, um die Güte Ihrer Marketingaussage in Echtzeit zu überprüfen.

Hinweis: Um SPSS Modeler Text Analytics mit SPSS Modeler Solution Publisher auszuführen, fügen Sie das Verzeichnis `<Installationsverzeichnis>/ext/bin/spss.TMWBServer` der Umgebungsvariablen `$LD_LIBRARY_PATH` hinzu.

Automatische Übersetzung unterstützter Sprachen. SPSS Modeler Text Analytics ermöglicht Ihnen in Verbindung mit Software as a Service (SaaS) von SDL die Übersetzung von Texten aus einer Reihe von unterstützten Sprachen – z. B. aus dem Arabischen, Chinesischen und Persischen – ins Englische. Sie können anschließend die Textanalyse auf den übersetzten Text anwenden und die Ergebnisse Personen bereitstellen, die den Textinhalt in der betreffenden Ausgangssprache nicht verstanden hätten. Da die Text-Mining-Ergebnisse automatisch wieder mit dem entsprechenden fremdsprachigen Text verknüpft werden, können sich die viel beschäftigten muttersprachlichen Mitarbeiter in Ihrem Unternehmen dann auf die wichtigsten Ergebnisse der Analyse konzentrieren. SDL bietet automatische Sprachübersetzung mithilfe statistischer Übersetzungsalgorithmen, die das Ergebnis von 20 Mannjahren hoch entwickelter Forschung auf dem Gebiet der maschinellen Übersetzung sind.

Upgrade auf IBM SPSS Modeler Text Analytics Version 15

Durchführen von Upgrades von früheren Versionen von PASW Text Analytics bzw. Text Mining for Clementine

Vor der Installation von IBM® SPSS® Modeler Text Analytics Version 15 sollten Sie alle TAPs, Vorlagen und Bibliotheken aus Ihrer aktuellen Version speichern und exportieren, die Sie in der neuen Version verwenden möchten. Diese Dateien sollten in einem Verzeichnis gespeichert werden, das bei der Installation der aktuellsten Version nicht gelöscht oder überschrieben wird.

Nach der Installation der aktuellsten Version von SPSS Modeler Text Analytics können Sie die gespeicherte TAP-Datei laden, gespeicherte Bibliotheken hinzufügen und gespeicherte Vorlagen importieren und laden, um sie in der aktuellsten Version zu verwenden.

Wichtig: Wenn Sie Ihre aktuelle Version deinstallieren, ohne zunächst die benötigten Dateien zu speichern und zu exportieren, gehen sämtliche TAPs, Vorlagen und Arbeiten an öffentlichen Bibliotheken in der vorherigen Version verloren und können nicht in SPSS Modeler Text Analytics Version 15 verwendet werden.

Informationen zum Text-Mining

Heutzutage wird eine Vielzahl von Informationen in unstrukturierter und halbstrukturierter Form gespeichert, beispielsweise Kunden-E-Mails, Callcenter-Notizen, offene Antworten bei Umfragen, News-Feeds, Webformulare usw. Diese Informationsflut ist ein Problem für Organisationen, die sich folgende Frage stellen: “Wie können wir diese Informationen erfassen, untersuchen und nutzen?”

Text-Mining besteht in der Analyse von gesammeltem Textmaterial mit dem Ziel, Schlüsselkonzepte und Themen zu erfassen und verborgene Beziehungen und Trends aufzudecken, ohne dass Sie die genauen Worte bzw. Begriffe kennen müssen, die die Autoren verwendet haben, um diese Konzepte auszudrücken. Obwohl es sehr große Unterschiede gibt, wird Text-Mining zuweilen mit Informationsrückgewinnung verwechselt. Das genaue Erfassen und Speichern von Informationen ist zwar eine große Herausforderung, doch Extrahierung und Verwaltung von qualitativ hochwertigen Inhalten, Terminologie und Beziehungen, die in den Informationen enthalten sind, stellen entscheidende und heikle Prozesse dar.

Text-Mining und Data-Mining

Für jeden Textartikel gibt linguistisch basiertes Text-Mining einen Index der Konzepte sowie Informationen zu diesen Konzepten aus. Diese destillierten und strukturierten Informationen können mit anderen Datenquellen kombiniert werden, um Fragen der folgenden Art zu beantworten:

- Welche Konzepte kommen zusammen vor?
- Womit sind sie außerdem verknüpft?
- Welche übergeordneten Kategorien können aus den extrahierten Daten gewonnen werden?
- Was sagen die Konzepte oder Kategorien vorher?
- Wie sagen die Konzepte oder Kategorien Verhalten vorher?

Eine Kombination von Text-Mining und Data-Mining bietet mehr Erkenntnisse als man allein aus strukturierten oder unstrukturierten Daten gewinnen kann. Dieser Prozess gliedert sich üblicherweise in folgende Schritte:

1. **Ermittlung des Texts, auf den das Mining angewendet werden soll.** Vorbereiten des Texts für das Mining. Wenn der Text aus mehreren Dateien besteht, müssen die Dateien im selben Verzeichnis gespeichert werden. Bei Datenbanken muss das Feld ermittelt werden, das den Text enthält.
2. **Anwenden des Minings auf den Text und Extrahierung strukturierter Daten.** Anwenden der Text-Mining-Algorithmen auf den Quellentext.
3. **Erstellen der Konzept- und Kategoriemodelle.** Ermittlung der Schlüsselkonzepte und/oder Erstellung von Kategorien. Die Zahl der aus unstrukturierten Daten erhaltenen Konzepte ist normalerweise sehr groß. Ermittlung der besten Konzepte und Kategorien für das Scoring.
4. **Analyse der strukturierten Daten.** Verwenden Sie traditionelle Data-Mining-Verfahren (wie Clustering, Klassifizierung und Erstellen von Vorhersagemodellen) zur Aufdeckung von Beziehungen zwischen den Konzepten. Führen Sie extrahierte Konzepte mit anderen strukturierten Daten zusammen, um auf der Grundlage der Konzepte zukünftiges Verhalten vorherzusagen.

Textanalyse und Kategorisierung

Bei der Textanalyse als Form einer qualitativen Analyse werden nützliche Daten aus Texten extrahiert , so dass die Schlüsselbegriffe und Konzepte, die im betreffenden Text enthalten sind, unter einer angemessenen Zahl von Kategorien zusammengefasst werden können. Textanalysen können auf Texte aller Arten und Längen angewendet werden. Allerdings unterscheiden sich die jeweiligen Analyseansätze dabei ein wenig.

Kürzere Datensätze oder Dokumente lassen sich am leichtesten kategorisieren, da sie eine geringere Komplexität aufweisen und für gewöhnlich weniger mehrdeutige Wörter und Antworten enthalten. Wenn Personen beispielsweise im Rahmen einer Umfrage mit offenen Antworten nach ihren drei Lieblingsaktivitäten im Urlaub gefragt werden, sind möglicherweise viele kurze Antworten zu erwarten, etwa: *an den Strand gehen*, *Nationalparks besuchen* oder *Nichtstun*. Längere offene Antworten können dagegen ziemlich komplex und weitschweifig ausfallen, besonders bei Befragten, die gebildet und motiviert sind und genug Zeit für das Ausfüllen eines Fragebogens zur Verfügung haben. Bei Umfragen zu den politischen Überzeugungen von

Personen oder bei einem langen Blog-Feed zum Thema Politik sind möglicherweise längere Kommentare zu allerlei Fragen und Positionen zu erwarten.

Einer der Hauptvorteile bei der Verwendung von IBM® SPSS® Modeler Text Analytics besteht darin, dass sehr schnell Schlüsselkonzepte extrahiert und aufschlussreiche Kategorien auf der Grundlage dieser längeren Textquellen erstellt werden können. Dieser Vorteil wird durch die Kombination von automatisierten linguistischen und statistischen Methoden erreicht. Damit werden bei jedem Schritt des Textanalyseprozesses die verlässlichsten Ergebnisse erzielt.

Linguistische Verarbeitung und NLP

Das Hauptproblem bei der Verwaltung dieser unstrukturierten Textdaten besteht darin, dass es keine Standardregeln dafür gibt, wie Texte so abgefasst werden können, dass der Computer sie versteht. Die Sprache, und damit die Bedeutung variiert zwischen den verschiedenen Dokumenten und Textstücken. Die einzige Möglichkeit, diese unstrukturierten Daten genau zu erfassen und zu organisieren, besteht darin, die Sprache zu analysieren und dadurch die Bedeutung aufzudecken. Es gibt mehrere verschiedene automatisierte Ansätze für die Extrahierung von Konzepten aus unstrukturierten Informationen. Diese Ansätze lassen sich in zwei Arten unterteilen: in linguistische und nicht linguistische Ansätze.

Einige Unternehmen haben versucht, automatisierte nicht linguistische Lösungen auf der Grundlage von Statistiken und neuronalen Netzen einzusetzen. Mithilfe von Computertechnologie können diese Lösungen Schlüsselkonzepte einfacher suchen und erfassen als menschliche Leser. Leider ist die Genauigkeit derartiger Lösungen ziemlich niedrig. Die meisten statistischen Systeme zählen einfach, wie oft bestimmte Wörter vorkommen und berechnen ihre statistische Nähe zu verwandten Konzepten. Sie produzieren viele irrelevante Ergebnisse, so genanntes "Rauschen" und finden manche gültigen Ergebnisse nicht ("Stille").

Um ihre begrenzte Genauigkeit auszugleichen, beinhalten einige Lösungen komplexe nicht linguistische Regeln, die die Unterscheidung zwischen relevanten und nicht relevanten Ergebnissen erleichtern sollen. Diese Vorgehensweise wird als *regelbasiertes Text-Mining* bezeichnet.

Beim *linguistisch basierten Text-Mining* dagegen werden die Prinzipien der Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) – der computergestützten Analyse menschlicher Sprachen – auf die Analyse der Wörter, Ausdrücke und der Syntax (Struktur) des Textes angewendet. Ein System, das NLP verwendet, kann Konzepte, u. a. zusammengesetzte Ausdrücke, auf intelligente Art und Weise extrahieren. Außerdem ermöglicht die Kenntnis der zugrunde liegenden Sprache eine Klassifizierung von Konzepten in verwandte Gruppen, beispielsweise Produkte, Organisationen oder Personen, wobei Bedeutung und Kontext verwendet werden.

Linguistisch basiertes Text-Mining sucht auf dieselbe Weise nach der Bedeutung im Text, wie Menschen es tun—indem sie erkennen, dass eine Reihe von Wortformen eine ähnliche Bedeutung haben und indem sie die Satzstruktur als Rahmen für das Textverständnis analysieren. Dieser Ansatz bietet dieselbe Geschwindigkeit und Kosteneffektivität wie statistikbasierte Systeme, er bietet jedoch einen wesentlich höheren Genauigkeitsgrad, während ein wesentlich geringerer Grad an Benutzereingriffen erforderlich ist.

Zur Veranschaulichung des Unterschieds zwischen statistikbasierten und linguistisch basierten Ansätzen beim Extrahierungsprozess mit Texten in allen Sprachen außer Japanisch dient die Überlegung, wie der jeweilige Ansatz auf eine Abfrage zum Begriff `Dokumentreproduktion`

reagieren würde. Sowohl bei den statistikbasierten als auch bei den linguistisch basierten Lösungen müsste eine Begriffserweiterung für das Wort `Reproduktion` erfolgen, damit auch Synonyme wie `Kopie` und `Vervielfältigung` berücksichtigt werden. Andernfalls werden relevante Informationen übersehen. Wenn jedoch bei einer statistikbasierten Lösung eine derartige Synonymik—die Suche nach anderen Termen mit derselben Bedeutung—angewendet werden soll, wird wahrscheinlich auch der Term `Geburt` berücksichtigt, was zur Erzeugung einer Reihe von irrelevanten Ergebnissen führt. Das Verstehen von Sprache beseitigt die Mehrdeutigkeit von Texten, weshalb das linguistisch basierte Text-Mining definitionsgemäß den verlässlicheren Ansatz darstellt.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Für Text in japanischer Sprache lässt sich der Unterschied zwischen statistikbasierten und linguistikbasierten Methoden beim Extrahierungsvorgang am Wort 沈む als ein Beispiel illustrieren. Mit diesem Wort können wir Ausdrücke wie 日が沈む finden, übersetzt als *Sonne geht unter*, oder 気分が沈む, übersetzt als *sich traurig fühlen*. Wenn Sie nur statistische Techniken einsetzen, werden 日 (übersetzt als *Sonne*), 気分 (übersetzt als *fühlen*) und 沈む (übersetzt als *traurig*) separat extrahiert. Bei Einsatz der Stimmungsanalyse, die linguistische Verfahren verwendet, werden jedoch nicht nur 日, 気分 und 沈む, sondern auch 気分が沈む (übersetzt als *sich traurig fühlen*) extrahiert und dem Typ <悪い － 悲しみ全般> zugewiesen. Die Verwendung von linguistischen Verfahren durch die Stimmungsanalyse ermöglicht die Extrahierung von bedeutungsvolleren Ausdrücken. Die Analyse und Erfassung von Stimmungen beseitigt die Mehrdeutigkeit von Texten, weshalb das linguistisch basierte Text-Mining definitionsgemäß den verlässlicheren Ansatz darstellt.

Wenn Sie verstehen, wie der Extrahierungsprozess funktioniert, fällt es Ihnen leichter, bei der Feinabstimmung Ihrer linguistischen Ressourcen (Bibliotheken, Typen, Synonyme und anderer) zentrale Entscheidungen zu treffen. Der Extrahierungsprozess umfasst folgende Schritte:

- Konvertieren von Quelldaten in ein Standardformat
- Ermittlung von Kandidaten
- Ermittlung von Äquivalenzklassen und Integration von Synonymen
- Zuweisung eines Typs
- Indexerstellung und, falls gewünscht, Musterabgleich mit einem Sekundäranalysator

Schritt 1. Konvertieren von Quelldaten in ein Standardformat

Im ersten Schritt werden die importierten Daten in ein einheitliches Format konvertiert, das für weitergehende Analysen genutzt werden kann. Diese Konvertierung erfolgt intern. Ihre Ausgangsdaten werden dabei nicht geändert.

Schritt 2. Ermittlung von Kandidaten

Es ist wichtig zu verstehen, welche Rolle die linguistischen Ressourcen während der linguistischen Extrahierung bei der Ermittlung von Kandidaten spielen. Linguistische Ressourcen kommen jedes Mal zum Einsatz, wenn ein Extrahierungsvorgang ausgeführt wird. Sie liegen in Form von Vorlagen, Bibliotheken und zusammengestellten Ressourcen vor. Bibliotheken bestehen aus Wortlisten, Beziehungen und weiteren Informationen, die eingesetzt werden, um die Extrahierung abzustimmen oder zu spezifizieren. Die zusammengestellten Ressourcen können nicht angezeigt

oder bearbeitet werden. Die übrigen Ressourcen können jedoch im Template Editor bzw., wenn eine interaktive Workbench-Sitzung gestartet wurde, im Resource Editor bearbeitet werden.

Zusammengestellte Ressourcen sind interne Kernkomponenten der Extraktor-Engine in SPSS Modeler Text Analytics . Diese Ressourcen umfassen ein allgemeines Wörterbuch, in dem eine Liste von Grundformen mit einem Code für die Wortart (Part of Speech) enthalten ist (Nomen, Verb, Adjektiv usw.). Die Ressourcen beinhalten auch belegte integrierte Typen, die verwendet werden, um den folgenden Typen eine Vielzahl von extrahierten Termen zuzuweisen: <地名>, <組織> oder <人名>. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text in Anhang A auf S. 387.*Anmerkung*: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Zusätzlich zu diesen zusammengestellten Ressourcen sind auch mehrere Bibliotheken im Lieferumfang enthalten. Diese können verwendet werden, um die Typen und Konzeptdefinitionen der zusammengestellten Ressourcen zu ergänzen und Synonyme zu liefern. Diese Bibliotheken—sowie sämtliche benutzerdefinierte Bibliotheken, die Sie erstellen—bestehen aus mehreren Wörterbüchern. Diese umfassen Typ-Wörterbücher, Synonymwörterbücher sowie Ausschlusswörterbücher. Für weitere Informationen siehe Thema Bearbeitungsressourcen für japanischen Text in Anhang A auf S. 383.*Anmerkung*: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Sobald die Daten importiert und konvertiert wurden, beginnt die Extrahierungsengine, Kandidaten für die Extrahierung zu identifizieren. Kandidaten sind Wörter oder Wortgruppen, die verwendet werden, um Konzepte im Text zu ermitteln. Bei der Verarbeitung des Texts werden einzelne Wörter (**Uniterms**) und zusammengesetzte Wörter (**Multiterms**) über Extraktoren auf der Grundlage von Wortklasse-Mustern (Part of Speech) ermittelt. Der Multiterm 青森りんご, der dem Wortklasse-Muster <地名> + <名詞> entspricht, besteht beispielsweise aus zwei Komponenten. Anschließend werden mithilfe der Stimmungstextlinkanalyse Kandidaten für Stimmungsstichwörter identifiziert.

Angenommen, Sie verfügen über den folgenden japanischen Text: 写真が新鮮で良かった. In diesem Fall würde die Extrahierungsengine den Stimmungstyp 良い － 褒め・賞賛 zuweisen, nachdem（品物） + が + 良い mithilfe einer der Stimmungstextlinkregeln abgeglichen wurde.

Anmerkung: Die Terme aus dem oben genannten zusammengestellten allgemeinen Wörterbuch stellen eine Liste aller Wörter dar, die als Uniterms wahrscheinlich uninteressant sind oder sprachliche Mehrdeutigkeiten aufweisen. Diese Wörter werden von der Extrahierung ausgeschlossen, wenn die Uniterms ermittelt werden. Sie werden jedoch erneut ausgewertet, wenn Wortarten bestimmt oder längere zusammengesetzte Wörter (Multiterms) als Kandidaten geprüft werden.

Schritt 3. Ermittlung von Äquivalenzklassen und Integration von Synonymen

Im Anschluss an die Ermittlung von Uniterms und Multiterms, die als Kandidaten infrage kommen, werden über ein Normalisierungswörterbuch der Software Äquivalenzklassen ermittelt. Bei einer Äquivalenzklasse handelt es sich um eine Grundform eines Ausdrucks oder einer einzelnen Form von zwei Varianten desselben Ausdrucks. Ausdrücke werden Äquivalenzklassen zugeordnet, damit beispielsweise die Begriffe `Nebenwirkung` und 副作用 nicht als unterschiedliche Konzepte betrachtet werden. Um festzustellen, welches Konzept für die betreffende Äquivalenzklasse als

Hauptterm verwendet wird, `Nebenwirkung` oder 副作用, werden die folgenden Regeln in der aufgeführten Reihenfolge durch die Extrahierungsengine angewendet:

- Die vom Benutzer festgelegte Form in einer Bibliothek.
- Die häufigste Form, wie von vorkompilierten Ressourcen definiert.

Schritt 4. Zuweisen eines Typs

Anschließend werden den extrahierten Konzepten Typen zugewiesen. Bei einem Typ handelt es sich um Konzepte, die nach semantischen Gesichtspunkten gruppiert werden. Für diesen Schritt werden sowohl zusammengestellte Ressourcen als auch die Bibliotheken verwendet. Zu den Typen gehören beispielsweise übergeordnete Konzepte, positive und negative Wörter, Vornamen, Orte, Organisationen und anderes. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

Japanische Textressourcen verfügen über ein anderes Set an Typen. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text in Anhang A auf S. 387.*Anmerkung*: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Linguistische Systeme sind wissensintensiv: Je mehr Informationen in den Wörterbüchern enthalten sind, desto höher ist die Qualität der Ergebnisse. Eine Änderung des Wörterbuchinhalts, z. B. Synonymdefinitionen, kann die resultierenden Informationen vereinfachen. Dabei handelt es sich häufig um einen schrittweisen Prozess, der für einen genauen Konzeptabruf erforderlich ist. NLP ist ein Kernelement von SPSS Modeler Text Analytics .

Wie Extrahierung funktioniert

Während der Extrahierung von Schlüsselkonzepten und -begriffen aus Ihren Antworten wird bei IBM® SPSS® Modeler Text Analytics die linguistisch basierte Textanalyse angewendet. Diese Methode bietet die Geschwindigkeit und Kosteneffektivität von statistisch basierten Systemen. Sie ermöglicht jedoch ein weitaus höheres Maß an Genauigkeit, während ein deutlich geringeres Maß an Eingriffen seitens des Benutzers erforderlich ist. Linguistisch basierte Textanalyse baut auf einem Forschungsgebiet namens Verarbeitung natürlicher Sprache auf, das auch als Computerlinguistik bekannt ist.

Wichtig: Für Text in japanischer Sprache führt der Extrahierungsvorgang eine andere Schrittfolge aus. Für weitere Informationen siehe Thema Wie Extrahierung funktioniert in Anhang A auf S. 377.*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Wenn Sie verstehen, wie der Extrahierungsprozess funktioniert, fällt es Ihnen leichter, bei der Feinabstimmung Ihrer linguistischen Ressourcen (Bibliotheken, Typen, Synonyme und anderer) zentrale Entscheidungen zu treffen. Der Extrahierungsprozess umfasst folgende Schritte:

- Konvertieren von Quelldaten in ein Standardformat
- Ermittlung von Kandidaten
- Ermittlung von Äquivalenzklassen und Integration von Synonymen
- Zuweisung eines Typs
- Indexerstellung
- Musterabgleich und Ereignisextrahierung

Schritt 1. Konvertieren von Quelldaten in ein Standardformat

Im ersten Schritt werden die importierten Daten in ein einheitliches Format konvertiert, das für weitergehende Analysen genutzt werden kann. Diese Konvertierung erfolgt intern. Ihre Ausgangsdaten werden dabei nicht geändert.

Schritt 2. Ermittlung von Kandidaten

Es ist wichtig zu verstehen, welche Rolle die linguistischen Ressourcen während der linguistischen Extrahierung bei der Ermittlung von Kandidaten spielen. Linguistische Ressourcen kommen jedes Mal zum Einsatz, wenn ein Extrahierungsvorgang ausgeführt wird. Sie liegen in Form von Vorlagen, Bibliotheken und zusammengestellten Ressourcen vor. Bibliotheken bestehen aus Wortlisten, Beziehungen und weiteren Informationen, die eingesetzt werden, um die Extrahierung abzustimmen oder zu spezifizieren. Die zusammengestellten Ressourcen können nicht angezeigt oder bearbeitet werden. Die übrigen Ressourcen (Vorlagen) können jedoch im Template Editor bzw., wenn eine interaktive Workbench-Sitzung gestartet wurde, im Resource Editor bearbeitet werden.

Zusammengestellte Ressourcen sind interne Kernkomponenten der Extraktor-Engine in IBM® SPSS® Modeler Text Analytics . Diese Ressourcen umfassen ein allgemeines Wörterbuch, in dem eine Liste von Grundformen mit einem Code für die Wortart (Part of Speech) enthalten ist (Nomen, Verb, Adjektiv, Adverb, Partizip, Koordinator, Determinator oder Präposition). Die Ressourcen beinhalten auch belegte integrierte Typen, die verwendet werden, um den folgenden Typen eine Vielzahl von extrahierten Termen zuzuweisen: `<Location>`, `<Organization>` oder `<Person>`. Für weitere Informationen siehe Thema Integrierte Typen in Kapitel 17 auf S. 314.

Zusätzlich zu diesen zusammengestellten Ressourcen sind auch mehrere Bibliotheken im Lieferumfang enthalten. Diese können verwendet werden, um die Typen und Konzeptdefinitionen der zusammengestellten Ressourcen zu ergänzen und weitere Typen und Synonyme zu liefern. Diese Bibliotheken—sowie sämtliche benutzerdefinierte Bibliotheken, die Sie erstellen—bestehen aus mehreren Wörterbüchern. Diese umfassen Typ-Wörterbücher, Substitutionswörterbücher (Synonyme und optionale Elemente) sowie Ausschlusswörterbücher. Für weitere Informationen siehe Thema Mit Bibliotheken arbeiten in Kapitel 16 auf S. 297.

Sobald die Daten importiert und konvertiert wurden, beginnt die Extrahierungsengine, Kandidaten für die Extrahierung zu identifizieren. Kandidaten sind Wörter oder Wortgruppen, die verwendet werden, um Konzepte im Text zu ermitteln. Während der Verarbeitung des Texts werden einzelne Wörter (**Uniterms**), die nicht in den zusammengestellten Ressourcen enthalten sind, als Kandidaten für die Extrahierung betrachtet. Kandidaten, die aus zusammengesetzten Wörtern bestehen (**Multiterms**), werden über Extraktoren auf der Grundlage von Wortklasse-Mustern (Part of Speech) ermittelt. Der Multiterm `Sportwagen`, der dem Wortklasse-Muster “Adjektiv-Nomen” entspricht, besteht beispielsweise aus zwei Komponenten. Der Multiterm `schneller Sportwagen`, der dem Wortklasse-Muster “Adjektiv-Adjektiv-Nomen” entspricht, besteht aus drei Komponenten.

Anmerkung: Die Terme aus dem oben genannten zusammengestellten allgemeinen Wörterbuch stellen eine Liste aller Wörter dar, die als Uniterms wahrscheinlich uninteressant sind oder sprachliche Mehrdeutigkeiten aufweisen. Diese Wörter werden von der Extrahierung ausgeschlossen, wenn die Uniterms ermittelt werden. Sie werden jedoch erneut ausgewertet, wenn Wortarten bestimmt oder längere zusammengesetzte Wörter (Multiterms) als Kandidaten geprüft werden.

Abschließend wird ein bestimmter Algorithmus für die Verarbeitung von Zeichenfolgen verwendet, die aus Großbuchstaben bestehen (z. B. bei Berufsbezeichnungen), so dass diese speziellen Muster extrahiert werden können.

Schritt 3. Ermittlung von Äquivalenzklassen und Integration von Synonymen

Im Anschluss an die Ermittlung von Uniterms und Multiterms, die als Kandidaten infrage kommen, werden diese über eine Reihe von Algorithmen der Software miteinander verglichen und Äquivalenzklassen ermittelt. Bei einer Äquivalenzklasse handelt es sich um eine Grundform eines Ausdrucks oder einer einzelnen Form von zwei Varianten desselben Ausdrucks. Ausdrücke werden Äquivalenzklassen zugeordnet, damit beispielsweise die Begriffe `Unternehmensleiter` und `Leiter des Unternehmens` nicht als unterschiedliche Konzepte betrachtet werden. Um festzustellen, welches Konzept für die betreffende Äquivalenzklasse als Hauptterm verwendet wird, —`Leiter des Unternehmens` oder `Unternehmensleiter`, werden die folgenden Regeln in der aufgeführten Reihenfolge durch die Extrahierungsengine angewendet:

- Die vom Benutzer festgelegte Form in einer Bibliothek.
- Die Form, die im gesamten Textkörper am häufigsten vorkommt.
- Die kürzeste Form im gesamten Textkörper (die normalerweise der Grundform entspricht).

Schritt 4. Zuweisen eines Typs

Anschließend werden den extrahierten Konzepten Typen zugewiesen. Bei einem Typ handelt es sich um Konzepte, die nach semantischen Gesichtspunkten gruppiert werden. Für diesen Schritt werden sowohl zusammengestellte Ressourcen als auch die Bibliotheken verwendet. Zu den Typen gehören beispielsweise übergeordnete Konzepte, positive und negative Wörter, Vornamen, Orte, Organisationen und anderes. Der Benutzer kann zusätzliche Typen definieren. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

Schritt 5. Indexerstellung

Die gesamte Reihe der Datensätze oder Dokumente wird indiziert. Dazu wird ein Zeiger zwischen einer Textstelle und dem bezeichnenden Term für jede Äquivalenzklasse erstellt. Das setzt voraus, dass sämtliche gebeugten Formen, die als Kandidaten für ein Konzept vorkommen, als Grundform für den Kandidaten indiziert werden. Die globale Häufigkeit wird für jede Grundform berechnet.

Schritt 6. Musterabgleich und Ereignisextrahierung.

Mit IBM SPSS Modeler Text Analytics können nicht nur Typen und Konzepte sondern auch Beziehungen ermittelt werden, die zwischen diesen bestehen. Mit diesem Produkt stehen mehrere Algorithmen und Bibliotheken zur Verfügung, über die Beziehungsmuster zwischen Typen und Konzepten extrahiert werden können. Besonders hilfreich ist dies für die Erfassung von bestimmten Meinungen (z. B. Reaktionen auf ein Produkt) oder von Beziehungsgefügen, die zwischen Personen oder Objekten (etwa zwischen politischen Gruppen oder Genomen) bestehen.

Wie die Kategorisierung funktioniert

Bei der Erstellung von Kategoriemodellen mit IBM® SPSS® Modeler Text Analytics haben Sie die Wahl zwischen verschiedenen Methoden, um Kategorien zu erstellen. Da jeder Datensatz seine besonderen Eigenheiten aufweist, kann die Zahl der Methoden und die Reihenfolge, in der

sie angewendet werden, gegebenenfalls variieren. Da sich Ihre Interpretation der Ergebnisse möglicherweise von der einer anderen Person unterscheidet, müssen Sie gegebenenfalls ein wenig mit den unterschiedlichen Methoden experimentieren, um zu erkennen, mit welcher Sie die besten Ergebnisse für Ihre Textdaten erzielen. In SPSS Modeler Text Analytics können Sie Kategoriemodelle in einer Workbench-Sitzung erstellen, in denen Sie Ihre Kategorien weiter untersuchen und abstimmen können.

In diesem Handbuch bezieht sich **Kategorieerstellung** auf die Erstellung von Kategoriedefinitionen und Klassifikation über eine oder mehrere integrierte Methoden und **Kategorisierung** auf das Scoring oder Labeling, einen Prozess, bei dem den Kategoriedefinitionen für jeden Datensatz oder jedes Dokument unverwechselbare IDs (Name/ID/Wert) zugewiesen werden.

Während der Kategorieerstellung werden die extrahierten Konzepte und Typen als Bausteine für Ihre Kategorien verwendet. Bei der Erstellung von Kategorien werden den Kategorien automatisch die Datensätze oder Dokumente zugewiesen, die Text enthalten, der einem Element der jeweiligen Kategoriedefinition entspricht.

SPSS Modeler Text Analytics bietet Ihnen mehrere automatisierte Kategorieerstellungsmethoden, mit denen Sie Ihre Dokumente oder Datensätze schnell kategorisieren können.

Gruppierverfahren

Die einzelnen verfügbaren Verfahren sind für bestimmte Datentypen und Situationen jeweils sehr gut geeignet, doch ist es oftmals nützlich, bei einer Analyse mehrere Verfahren miteinander zu verbinden, um die Dokumente oder Datensätze vollständig zu erfassen. Sie können ein Konzept in mehreren Kategorien erkennen oder redundante Kategorien vorfinden.

Konzeptwurzelableitung. Mit diesem Verfahren werden Kategorien erstellt, indem ausgehend von einem Konzept andere verwandte Konzepte ermittelt werden (durch Analyse, ob bestimmte Konzeptkomponenten morphologisch verwandt sind oder gemeinsame Wurzeln haben). Dieses Verfahren ist sehr nützlich bei der Identifizierung von bedeutungsgleichen Konzepten aus zusammengesetzten Wörtern, da die Konzepte in jeder generierten Kategorie die gleiche oder ähnliche Bedeutung haben. Das Verfahren funktioniert mit Daten unterschiedlicher Länge und erzeugt eine geringere Anzahl an kompakten Kategorien. So wird beispielsweise das Konzept `Möglichkeiten zum Aufstieg` mit den Konzepten`Möglichkeit des Aufstiegs` und `Aufstiegsmöglichkeit` zu einer Kategorie zusammengefasst. Für weitere Informationen siehe Thema Konzeptwurzelableitung in Kapitel 10 auf S. 193.Diese Option ist nicht für japanischen Text verfügbar.

Semantisches Netz. Bei diesem Verfahren wird zunächst auf der Grundlage eines umfassenden Index von Wortbeziehungen jedes Konzept auf seine möglichen Bedeutungen untersucht. Anschließend werden Kategorien durch Gruppieren zusammenhängender Konzepte erstellt. Diese Technik empfiehlt sich, wenn die Konzepte dem semantischen Netz bekannt und nicht zu zweideutig sind. Sie ist weniger hilfreich, wenn der Text spezielle Terminologie oder Sprache enthält, die dem Netz unbekannt ist. Das Konzept `Granny Smith Apfel` würde zum Beispiel mit `Gala Apfel` und `Winesap Apfel` gruppiert, da es sich um Geschwister von Granny Smith handelt. In einem anderen Beispiel würde das Konzept `Tier` mit `Katze` und `Känguru` gruppiert,

da dies Hyponyme von `Tier` sind. Dieses Verfahren ist in dieser Version nur für englischen Text verfügbar. Für weitere Informationen siehe Thema Semantische Netze in Kapitel 10 auf S. 195.

Konzepteinbeziehung. Diese Technik erstellt Kategorien durch die Gruppierung von Multiterm-Konzepten (zusammengesetzte Wörter) basierend darauf, ob sie Wörter enthalten, die Unter- oder Übermengen eines Worts in dem anderen sind. So wird beispielsweise `Sitz` mit `Kindersitz`, `Sitzheizung` und `Kindersitzgurt` zu einer Gruppe zusammengefasst. Für weitere Informationen siehe Thema Konzepteinbeziehung in Kapitel 10 auf S. 194.

Kookkurrenz. Diese Technik erstellt Kategorien aus Kookkurrenz im Text. Dahinter steht folgende Idee: Wenn Konzepte oder Konzeptmuster häufig in Dokumenten bzw. Datensätzen gefunden werden, ist diese Kookkurrenz Ausdruck einer zugrunde liegenden Beziehung, die wahrscheinlich in Ihren Kategoriedefinitionen von Nutzen ist. Wenn Wörter eine signifikante Kookkurrenz aufweisen, wird eine Kookkurrenzregel erstellt, die als Kategoriedeskriptor für eine neue Unterkategorie verwendet werden kann. Wenn beispielsweise viele Datensätze die Wörter `Preis` und `Verfügbarkeit` enthalten (wenige jedoch nur eines von beiden), könnten diese Konzepte in eine Kookkurrenzregel zusammengefasst (`Preis &verfügbar`) und beispielsweise einer Unterkategorie der Kategorie `Preis` zugewiesen werden.Für weitere Informationen siehe Thema Kookkurrenzregeln in Kapitel 10 auf S. 197.

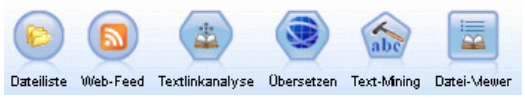
- **Minimale Anzahl anDokumenten.** Um festzustellen, wie interessant Kookkurrenzen sind, definieren Sie die minimale Anzahl an Dokumenten oder Datensätzen, die eine bestimmte Kookkurrenz enthalten muss, um als Deskriptor in einer Kategorie verwendet zu werden.

IBM SPSS Modeler Text Analytics-Knoten

Neben den vielen Standardknoten, die im Lieferumfang von IBM® SPSS® Modeler enthalten sind, können Sie außerdem mit Text-Mining-Knoten arbeiten, um die Möglichkeiten der Textanalyse in Ihre Streams aufzunehmen. In IBM® SPSS® Modeler Text Analytics stehen Ihnen hierzu mehrere Text-Mining-Knoten zur Verfügung. Diese Knoten sind auf der Registerkarte SPSS Modeler Text Analytics der Knotenpalette gespeichert.

Abbildung 1-1
Registerkarte IBM SPSS Modeler Text Analytics auf der Knotenpalette

Abbildung 1-2
Zoomen auf alle IBM SPSS Modeler Text Analytics-Knoten



Die folgenden Knoten sind enthalten:

- Der **Quellenknoten für die Dateiliste** erstellt eine Liste der Dokumentnamen als Eingabe in den Text-Mining-Prozess. Dies ist sinnvoll, wenn sich der Text in externen Dokumenten und nicht in einer Datenbank oder einer anderen strukturierten Datei befindet. Der Knoten gibt ein einzelnes Feld mit einem Datensatz für jedes aufgelistete Dokument bzw. jeden aufgelisteten Ordner aus. Dieses Feld kann dann als Eingabe in einen nachfolgenden Text-Mining-Knoten

verwendet werden. Für weitere Informationen siehe Thema Dateilistenknoten in Kapitel 2 auf S. 14.

- Der **Web-Feed-Quellenknoten** ermöglicht es, Text aus Web-Feeds einzulesen, beispielsweise aus Blogs oder News-Feeds in RSS- oder HTML-Formaten, und diese Daten im Text-Mining-Prozess zu verwenden. Der Knoten gibt ein einzelnes Feld oder mehrere Felder für jeden in den Feeds gefundenen Datensatz aus. Diese Felder können dann als Eingabe in einen nachfolgenden Text-Mining-Knoten verwendet werden. Für weitere Informationen siehe Thema Web-Feed-Knoten in Kapitel 2 auf S. 19.
- Über den **Text-Mining-Knoten** werden mit linguistischen Methoden Schlüsselkonzepte aus dem Text extrahiert. Der Text-Mining-Knoten ermöglicht es, mit diesen Konzepten und anderen Daten Kategorien zu erstellen. Daneben können Beziehungen und Verbindungen zwischen Konzepten ermittelt werden, die auf bekannten Mustern beruhen (Textlinkanalyse). Der Knoten kann genutzt werden, um Textdateninhalte zu untersuchen oder um ein Konzept- oder Kategoriemodell zu erstellen. Diese Konzepte und Kategorien können dann mit bestehenden strukturierten Daten, wie beispielsweise demografischen Informationen, kombiniert und auf die Modellierung angewendet werden. Für weitere Informationen siehe Thema Text-Mining-Modellierungsknoten in Kapitel 3 auf S. 32.
- Der **Textlinkanalyseknoten** extrahiert Konzepte und ermittelt auch Beziehungen zwischen Konzepten, die auf bekannten Mustern innerhalb des Texts beruhen. Mithilfe der Musterextrahierung können Beziehungen zwischen den Konzepten aufgedeckt werden, sowie alle möglicherweise diesen Konzepten beigefügten Meinungen und Vermerke. Der Textlinkanalyseknoten eröffnet einen direkteren Weg, Muster in Ihrem Text zu ermitteln und zu extrahieren und die Musterergebnisse anschließend dem Datensatz im Stream hinzuzufügen. Eine TLA kann auch über eine interaktive Workbench-Sitzung im Text-Mining-Modellierungsknoten durchgeführt werden. Für weitere Informationen siehe Thema Text Link Analysis-Knoten in Kapitel 4 auf S. 83.
- Mit dem **Übersetzungsknoten** können Texte aus unterstützten Sprachen wie dem Arabischen, Chinesischen oder Persischen zum Zwecke der Modellierung ins Englische oder in andere Sprachen übersetzt werden. Dadurch kann Text-Mining in Dokumenten durchgeführt werden, die in Double-Byte-Sprachen verfasst sind und andernfalls nicht unterstützt würden. Außerdem können Analysten Konzepte aus diesen Dokumenten extrahieren, selbst wenn sie die betreffende Sprache nicht beherrschen. Diese Funktion kann von jedem Textmodellierungsknoten aus aufgerufen werden, doch durch die Verwendung eines eigenen Übersetzungsknotens kann eine Übersetzung im Cache gespeichert und in mehreren Knoten wiederverwendet werden. Für weitere Informationen siehe Thema Übersetzungsknoten in Kapitel 5 auf S. 96.
- Bei der Durchführung von Text-Mining aus externen Dokumenten kann mithilfe des **Text-Mining-Ausgabeknotens** eine HTML-Seite erstellt werden, die Links zu den Dokumenten enthält, aus denen Konzepte extrahiert wurden. Für weitere Informationen siehe Thema Datei-Viewer-Knoten in Kapitel 6 auf S. 105.

Anwendungen

Im Allgemeinen können alle Personen, die routinemäßig große Mengen von Dokumenten sichten müssen, um Schlüsselelemente für eine eingehende Untersuchung zu ermitteln, von IBM® SPSS® Modeler Text Analytics profitieren.

Hier einige Anwendungsbereiche:

- **Naturwissenschaftliche und medizinische Forschung.** Untersuchen von sekundären Forschungsunterlagen wie Patentberichten, Artikel aus Fachzeitschriften und Protokollveröffentlichungen. Erkennen von Verbindungen, die zuvor nicht bekannt waren (beispielsweise zwischen einem Arzt und einem bestimmten Produkt), was Möglichkeiten zur weiteren Untersuchung bietet. Verringerung des Zeitaufwands für die Medikamentenentdeckung. Verwendung als Hilfsmittel bei der Genomforschung.
- **Investitionsforschung.** Überprüfung von täglichen Analystenberichten, Nachrichtenartikeln und Presseerklärungen von Unternehmen zur Ermittlung wichtiger Strategiepunkte oder Marktveränderungen. Durch die Trendanalyse solcher Informationen können über einen bestimmten Zeitraum aufkommende Probleme und Chancen für eine Firma oder Branche ermittelt werden.
- **Betrugserkennung.** Verwendung bei Betrug im Banken- und Gesundheitswesen, um markante Stellen in großen Textmengen zu erkennen.
- **Marktanalyse.** Verwendung in der Marktanalyse zur Ermittlung von zentralen Themen in Umfragen mit offenen Antworten.
- **Analyse von Blogs und Web-Feeds.** Untersuchung und Erstellung von Modellen unter Verwendung der zentralen Konzepte aus News-Feeds, Blogs usw.
- **CRM.** Erstellen von Modellen mithilfe aller Kundenberührungspunkte wie E-Mails, Transaktionen und Umfragen.

Kapitel
2

Einlesen von Quellentext

Daten für Text Mining können in allen von IBM® SPSS® Modeler verwendeten Standardformaten vorliegen, einschließlich Datenbanken und anderen "rechteckigen" Formaten, die Daten als Zeilen und Spalten darstellen, oder Dokumentenformaten, wie Microsoft Word, Adobe PDF oder HTML, die nicht dieser Struktur entsprechen.

- Zum Einlesen von Text aus Dokumenten, die nicht der Standarddatenstruktur entsprechen, wie beispielsweise Microsoft Word, Microsoft Excel und Microsoft PowerPoint sowie Adobe PDF, XML, HTML usw., kann der Dateilistenknoten verwendet werden, um eine Liste von Dokumenten oder Ordnern als Eingabe in den Text-Mining-Prozess zu erstellen. Für weitere Informationen siehe Thema Dateilistenknoten auf S. 14.
- Zum Einlesen von Web-Feeds, wie Blogs oder News Feeds im RSS- oder HTML-Format, kann der Web-Feed-Knoten verwendet werden, um Web-Feed-Daten für die Eingabe in den Text-Mining-Prozess zu formatieren. Für weitere Informationen siehe Thema Web-Feed-Knoten auf S. 19.
- Um Text in einem der von SPSS Modeler verwendeten Standarddatenformate einzulesen, wie beispielsweise einer Datenbank mit einem oder mehr Textfeldern für Kundenkommentare, kann jeder der Standard-Quellenknoten von SPSS Modeler verwendet werden. Weitere Informationen hierzu finden Sie in der SPSS Modeler-Knotendokumentation.

Dateilistenknoten

Zum Einlesen von Text aus unstrukturierten Dokumenten, die in Formaten wie beispielsweise Microsoft Word, Microsoft Excel und Microsoft PowerPoint sowie Adobe PDF, XML, HTML usw. gespeichert wurden, kann der Dateilistenknoten verwendet werden, um eine Liste von Dokumenten oder Ordnern als Eingabe in den Text-Mining-Prozess zu erstellen. Dies ist erforderlich, da unstrukturierte Textdokumente nicht als Felder und Datensätze – Zeilen und Spalten – dargestellt werden können, wie dies für andere von IBM® SPSS® Modeler verwendete Daten möglich ist. Dieser Knoten steht auf der Text-Mining-Palette zur Verfügung.

Der Dateilistenknoten funktioniert als Quellenknoten, mit der Ausnahme, dass der Knoten nicht die tatsächlichen Daten einliest, sondern die Namen der Dokumente bzw. Verzeichnisse unterhalb des angegebenen Stammknotens und diese als Liste ausgibt. Die Ausgabe ist ein einzelnes Feld mit einem Datensatz für jede aufgelistete Datei. Dieses Feld kann dann als Eingabe für einen nachfolgenden Text-Mining-Knoten verwendet werden.

Sie finden diesen Knoten auf der IBM® SPSS® Modeler Text Analytics -Registerkarte der Knotenpalette am unteren Rand des SPSS Modeler-Fensters. Für weitere Informationen siehe Thema IBM SPSS Modeler Text Analytics-Knoten in Kapitel 1 auf S. 11.

Wichtig: Jegliche Verzeichnis- und Dateinamen, die Zeichen enthalten, die nicht in der lokalen Rechnerkodierung enthalten sind, werden nicht unterstützt. Wenn Sie versuchen, einen Stream auszuführen, der einen Dateilistenknoten enthält, führen etwaige Verzeichnis- oder Dateinamen mit diesen Zeichen dazu, dass die Ausführung des Streams fehlschlägt. Dies könnte mit

Verzeichnis- oder Dateinamen in einer Fremdsprache geschehen, z. B. einem japanischen Dateinamen in einem französischen Gebietsschema.

RTF-Verarbeitung. Zur Verarbeitung von RTF-Dateien ist ein Filter erforderlich. Sie können einen RTF-Filter von der Microsoft-Website herunterladen und manuell registrieren.

Adobe PDF -Verarbeitung. Um Texte aus Adobe PDF-Dateien extrahieren zu können, muss Adobe Reader Version 9 auf dem Rechner installiert sein, auf dem sich SPSS Modeler Text Analytics und IBM® SPSS® Modeler Text Analytics Server befinden.

- **Anmerkung**: Führen Sie kein Upgrade auf Adobe Reader Version 10 oder höher aus, da diese Versionen den erforderlichen Filter nicht enthalten.
- Durch ein Upgrade auf Adobe Reader Version 9 können Sie ein beträchtliches Speicherleck im Filter vermeiden, das bei der Arbeit mit großen Mengen an Adobe PDF-Dokumenten (etwa 1000 und mehr) zu Verarbeitungsfehlern führte. Wenn Sie die Verarbeitung von Adobe PDF-Dokumenten auf 32-Bit- oder 64-Bit-Microsoft Windows-Betriebssystemen planen, sollten Sie eine Aktualisierung auf Adobe Reader Version 9.x für 32-Bit-Systeme oder Adobe PDF iFilter 9 für 64-Bit-Systeme durchführen, die beide auf der Adobe-Website zur Verfügung stehen.
- AdobeAdobe hat die verwendete Filtersoftware ab Adobe Reader 8.x geändert. Ältere Adobe PDF-Dateien sind eventuell nicht lesbar oder enthalten fremde Zeichen. Das ist ein Adobe-Problem, auf das SPSS Modeler Text Analytics keinen Einfluss hat.
- Wenn die Sicherheitseinstellung einer Adobe PDF im Dialogfeld “Dokumenteigenschaften” auf der Registerkarte “Sicherheit” für “Kopieren bzw. Entnehmen von Inhalt” auf “Nicht zulässig” gesetzt ist, kann das Dokument außerdem nicht gefiltert und in das Produkt eingelesen werden.
- Adobe PDF-Dateien können nur auf Microsoft Windows-Plattformen verarbeitet werden.
- Aufgrund von Beschränkungen bei Adobe ist es nicht möglich, Text aus bildbasierten Adobe PDF-Dateien zu extrahieren.

Microsoft Office -Verarbeitung.

- Zur Verarbeitung der neueren Formate von Microsoft Word-, Microsoft Excel- und Microsoft PowerPoint-Dokumenten, die mit Microsoft Office 2007 eingeführt wurden, muss entweder Microsoft Office 2007 auf dem Computer installiert sein, auf dem SPSS Modeler Text Analytics Server (lokal oder remote) ausgeführt wird, oder Sie müssen das neue Microsoft Microsoft Office2007-Filterpaket installieren (verfügbar auf der Microsoft-Website).
- Dateien aus Microsoft Office-Dateien können nur auf Microsoft Windows-Plattformen verarbeitet werden.

Lokale Datenunterstützung. Wenn Sie mit einem entfernten SPSS Modeler Text Analytics Server verbunden sind und einen Stream mit einem Dateilistenknoten haben, müssen sich die Daten auf demselben Computer wie der SPSS Modeler Text Analytics Server befinden oder der Servercomputer muss Zugriff auf den Ordner haben, auf dem die Quelldaten im Dateilistenknoten gespeichert sind.

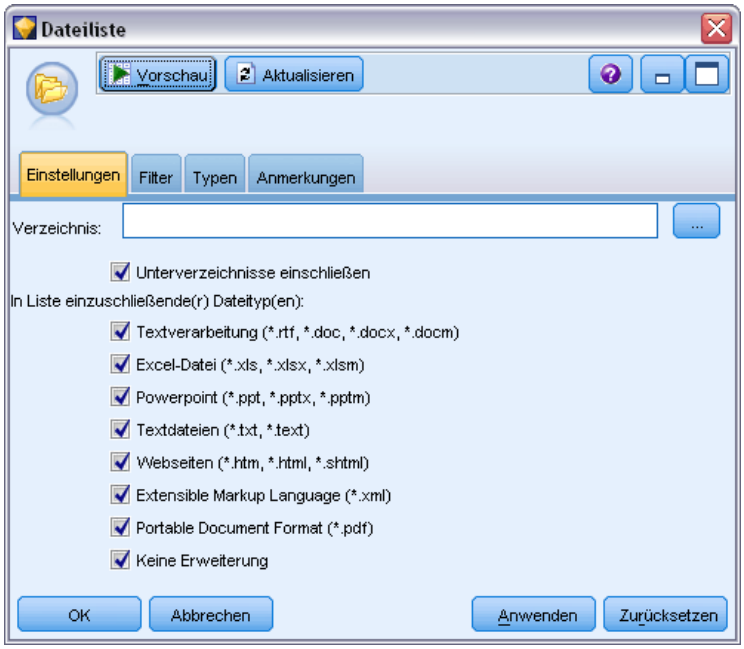
Dateilistenknoten: Registerkarte "Einstellungen"

In dieser Registerkarte können Sie die Verzeichnisse, Dateinamenerweiterungen und Ausgaben definieren, die von diesem Knoten erwünscht sind.

Anmerkung: Die Text-Mining-Extrahierung verarbeitet keine Microsoft Office- und Adobe PDF-Dateien auf Nicht-Microsoft Windows-Plattformen. XML-, HTML- oder Textdateien können jedoch immer verarbeitet werden.

Jegliche Verzeichnis- und Dateinamen, die Zeichen enthalten, die nicht in der lokalen Rechnerkodierung enthalten sind, werden nicht unterstützt. Wenn Sie versuchen, einen Stream auszuführen, der einen Dateilistenknoten enthält, führen etwaige Verzeichnis- oder Dateinamen mit diesen Zeichen dazu, dass die Ausführung des Streams fehlschlägt. Dies könnte mit Verzeichnis- oder Dateinamen in einer Fremdsprache geschehen, z. B. einem japanischen Dateinamen in einem französischen Gebietsschema.

Abbildung 2-1
Dialogfeld des Dateilistenknotens: Registerkarte "Einstellungen"



Verzeichnis. Gibt den Stammordner an, der die Dokumente enthält, die Sie auflisten möchten.

- **Unterverzeichnisse einschließen.** Gibt an, dass auch die Unterverzeichnisse durchsucht werden sollen.

In die Liste aufzunehmende Dateityp(en): Sie können die gewünschten Dateitypen bzw. Erweiterungen auswählen bzw. ihre Auswahl aufheben. Wenn Sie die Auswahl einer Dateierweiterung aufheben, werden Dateien mit dieser Erweiterung ignoriert. Eine Filterung ist nach folgenden Erweiterungen möglich:

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

Anmerkung: Für weitere Informationen siehe Thema Dateilistenknoten auf S. 14.

Wichtig: Ab Version 14 ist die Option “Verzeichnisliste” nicht mehr verfügbar; die einzige Ausgabe ist eine Dateiliste.

Dateilistenknoten: Andere Registerkarten

Die Registerkarte “Typen” ist eine Standardregisterkarte in IBM® SPSS® Modeler-Knoten, ebenso wie die Registerkarte “Anmerkungen”.

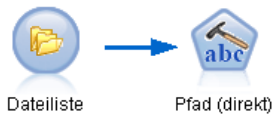
Verwenden des Dateilistenknotens in Text Mining

Der Dateilistenknoten wird verwendet, wenn sich Textdaten in externen unstrukturierten Dokumenten befinden, die in Formaten wie Microsoft Word, Microsoft Excel und Microsoft PowerPointsowie Adobe PDF, XML, HTML usw. vorliegen. Dieser Knoten wird verwendet, um eine Liste der Dokumente oder Ordner als Eingabe in den Text-Mining-Prozess zu generieren (einen nachfolgenden Text-Mining- oder Textlinkanalyseknoten).

Wenn Sie den Dateilistenknoten verwenden, dann stellen Sie sicher, dass im Text-Mining- oder Textlinkanalyseknoten angegeben ist, dass das Textfeld Pfadangaben zu den Dokumenten enthält, um anzuzeigen, dass das ausgewählte Feld anstelle des Texts, der ermittelt werden soll, Pfade zu den Dokumenten enthält, die den Text enthalten.

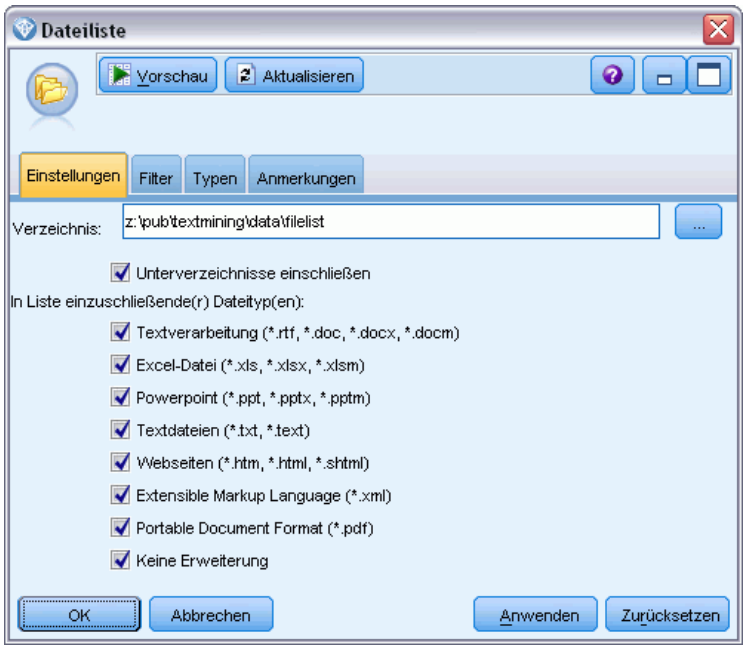
Im folgenden Beispiel wurde ein Dateilistenknoten mit einem Text-Mining-Knoten verbunden, um Text zu liefern, der sich in externen Dokumenten befindet.

Abbildung 2-2
Beispiel-Stream: Dateilistenknoten mit Text-Mining-Knoten



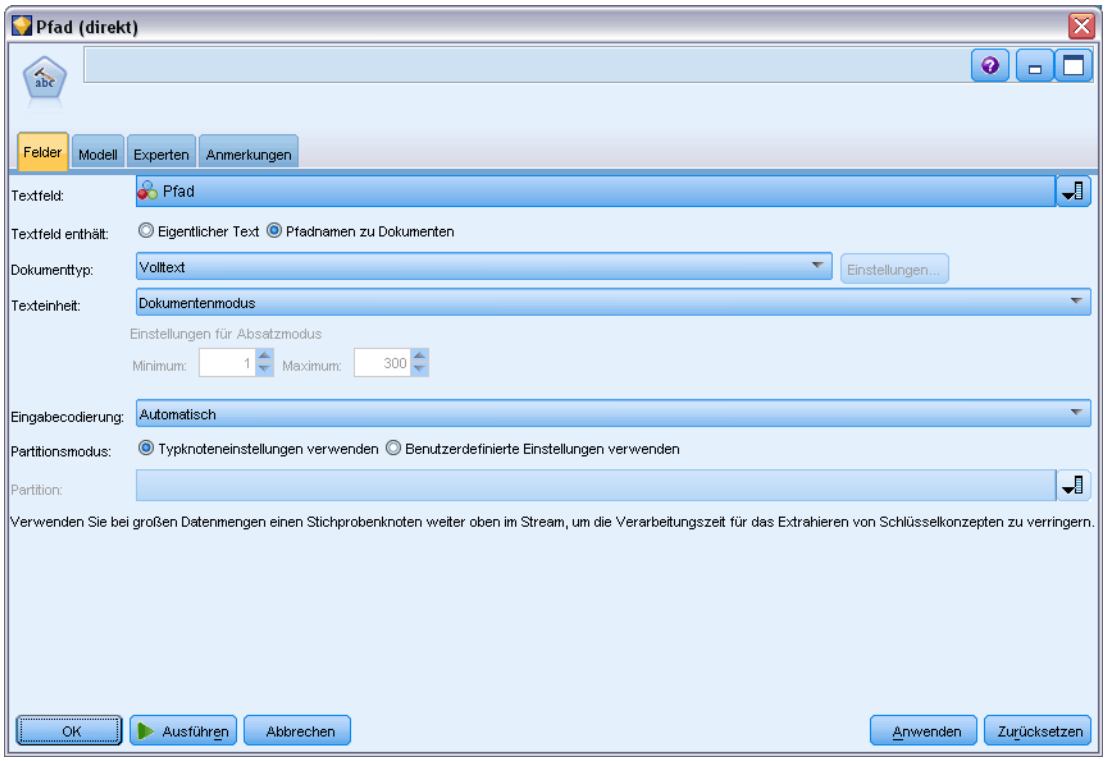
- **Dateilistenknoten (Registerkarte “Einstellungen”).** Zuerst fügten wir diesen Knoten dem Stream hinzu, um anzugeben, wo die Textdokumente gespeichert sind. Wir wählten das Verzeichnis aus mit allen Dokumenten, in denen wir eine Textdatensuche durchführen wollten.

Abbildung 2-3
Dialogfeld des Dateilistenknotens: Registerkarte "Einstellungen"



- **Text-Mining-Knoten (Registerkarte "Felder").** Anschließend fügten wir dem Dateilistenknoten einen Text-Mining-Knoten hinzu und verbanden ihn. In diesem Knoten definierten wir unser Eingabeformat und die Ressourcenvorlage sowie das Ausgabeformat. Wir wählten den aus dem Dateilistenknoten erstellten Feldnamen aus und wählten die Option aus, in der das Textfeld Pfadnamen zu Dokumenten repräsentiert, sowie andere Einstellungen. Für weitere Informationen siehe Thema Verwenden des Text-Mining-Knotens in einem Stream in Kapitel 3 auf S. 50.

Abbildung 2-4
Dialogfeld des Text-Mining-Knotens: Registerkarte "Felder"



Weitere Informationen zur Verwendung des Text-Mining-Knotens finden Sie unter Kapitel 3.

Web-Feed-Knoten

Mit dem Web-Feed-Knoten können Textdaten aus Web-Feeds für den Text-Mining-Prozess vorbereitet werden. Dieser Knoten akzeptiert Web-Feeds in zwei Formaten:

- **RSS-Format.** RSS ist ein einfaches XML-basiertes, standardisiertes Format für Webinhalte. Die URL verweist für dieses Format auf eine Seite, die einen Satz verlinkter Artikel enthält, z. B. Nachrichtenquellen und Blogs. Da es sich bei RSS um ein standardisiertes Format handelt, werden verlinkte Artikel automatisch identifiziert und im resultierenden Daten-Stream als einzelne Datensätze behandelt. Es ist keine weitere Eingabe erforderlich, um wichtige Textdaten und Datensätze aus dem Feed identifizieren zu können, außer Sie möchten eine Filtermethode auf den Text anwenden.
- **HTML-Format.** Auf der Registerkarte "Eingabe" können Sie eine oder mehrere URLs zu HTML-Seiten definieren. Anschließend können Sie auf der Registerkarte "Datensätze" den Datensatz-Anfangs-Tag definieren und diejenigen Tags festlegen, die den Zielinhalt begrenzen und diese Tags den Ausgabefeldern Ihrer Wahl zuweisen (Beschreibung, Titel, Änderungsdatum usw.). Für weitere Informationen siehe Thema Web-Feed-Knoten: Registerkarte "Datensätze" auf S. 22.

Wichtig: Wenn Sie versuchen, Informationen über das Internet durch einen Proxy-Server abzurufen, müssen Sie den Proxy-Server in der Datei `net.properties` für die Client- und Server-Version von IBM® SPSS® Modeler Text Analytics aktivieren. Befolgen Sie die in der Datei enthaltenen Anweisungen. Dies trifft zu, wenn Sie über den Web-Feed-Knoten auf das Internet zugreifen oder eine Lizenz für SDL Software as a Service (SaaS) abrufen, da diese Verbindungen über Java aufgebaut werden. Beim Client befindet sich diese Datei standardmäßig im Verzeichnis *C:\Program Files\IBM\SPSS\Modeler\15\jre\lib\net.properties*. Beim Server befindet sich diese Datei standardmäßig im Verzeichnis *C:\Program Files\IBM\SPSS\Modeler\15\ext\bin\spss.TMWBServer\jre\lib\net.properties*.

Die Ausgabe dieses Knotens besteht aus einem Satz Feldern, der verwendet wird, um die Datensätze zu beschreiben. Das Feld Beschreibung ist das am häufigsten verwendete Feld, da es den Großteil des Textinhalts enthält. Es kann jedoch sein, dass Sie sich auch für die Inhalte der anderen Felder interessieren, wie die Kurzbeschreibung eines Datensatzes (Feld Kurzbeschreibung) oder den Titel des Datensatzes (Feld Titel). Alle Eingabefelder können als Eingabe für einen nachfolgenden Text-Mining-Knoten ausgewählt werden.

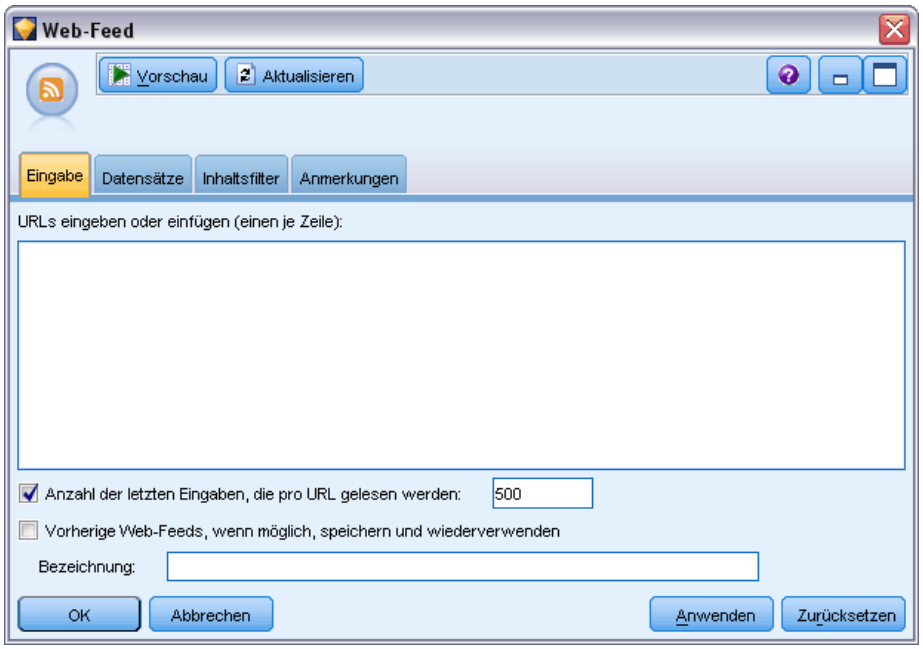
Sie finden diesen Knoten auf der SPSS Modeler Text Analytics -Registerkarte der Knotenpalette am unteren Rand des IBM® SPSS® Modeler-Fensters. Für weitere Informationen siehe Thema IBM SPSS Modeler Text Analytics-Knoten in Kapitel 1 auf S. 11.

Web-Feed-Knoten: Registerkarte “Eingabe”

Die Registerkarte “Eingabe” wird zur Angabe einer oder mehrerer Internetadressen bzw. URLs verwendet, um die Textdaten zu erfassen. Im Zusammenhang mit Text Mining können Sie URLs zu Feeds angeben, die Textdaten enthalten.

Wichtig: Beim Arbeiten mit Nicht-RSS-Daten ziehen Sie möglicherweise den Einsatz eines Scraping-Tools wie beispielsweise WebQL® vor, um das Sammeln von Inhalten und Verweisen auf die Ausgabe von diesem Tool mithilfe eines anderen Quellenknotens zu automatisieren.

Abbildung 2-5
Dialogfeld "Web-Feed-Knoten": Registerkarte "Eingabe"



Folgende Parameter können festgelegt werden:

Geben Sie URLs ein oder fügen Sie sie ein. In dieses Feld können sie URLs eingeben oder diese einfügen. Wenn Sie mehrere URLs eingeben, drücken Sie nach jeder URL die Eingabetaste, damit jeweils pro Zeile eine URL angegeben ist. Geben Sie den vollständigen URL-Pfad zu der Datei an. Diese URLs können bei Feeds in zwei Formaten vorliegen:

- *RSS-Format.* RSS ist ein einfaches XML-basiertes, standardisiertes Format für Webinhalte. Die URL verweist für dieses Format auf eine Seite, die einen Satz verlinkter Artikel enthält, z. B. Nachrichtenquellen und Blogs. Da es sich bei RSS um ein standardisiertes Format handelt, werden verlinkte Artikel automatisch identifiziert und im resultierenden Daten-Stream als einzelne Datensätze behandelt. Es ist keine weitere Eingabe erforderlich, um wichtige Textdaten und Datensätze aus dem Feed identifizieren zu können, außer Sie möchten eine Filtermethode auf den Text anwenden.
- *HTML-Format.* Auf der Registerkarte "Eingabe" können Sie eine oder mehrere URLs zu HTML-Seiten definieren. Anschließend können Sie auf der Registerkarte "Datensätze" den Datensatz-Anfangs-Tag definieren und diejenigen Tags festlegen, die den Zielinhalt begrenzen und diese Tags den Ausgabefeldern Ihrer Wahl zuweisen (Beschreibung, Titel, Änderungsdatum usw.). Beim Arbeiten mit Nicht-RSS-Daten ziehen Sie möglicherweise den Einsatz eines Scraping-Tools wie beispielsweise WebQL® vor, um das Sammeln von Inhalten und Verweisen auf die Ausgabe von diesem Tool mithilfe eines anderen Quellenknotens zu automatisieren. Für weitere Informationen siehe Thema Web-Feed-Knoten: Registerkarte "Datensätze" auf S. 22.

Anzahl der aktuellen, pro URL zu lesenden Einträge. Dieses Feld legt die maximale Anzahl der Datensätze fest, die pro URL im Feld aufgeführt werden, wobei mit dem ersten im Feed gefundenen Datensatz begonnen wird. Die Textmenge wirkt sich auf

die Verarbeitungsgeschwindigkeit während des Downstreams zur Extrahierung in einem Text-Mining-oder Text-Link-Analysis-Knoten aus.

Vorherige Web-Feeds, wenn möglich, speichern and wiederverwenden. Mit dieser Option werden Web-Feeds durchsucht und die verarbeiteten Ergebnisse im Cache zwischengespeichert. Wenn sich die Inhalte eines Feeds dann bei nachfolgenden Stream-Verarbeitungen nicht verändert haben oder wenn kein Zugriff auf den Feed möglich ist (z. B.: Verbindung unterbrochen), wird die Version im Cache verwendet, um die Verarbeitungszeit zu beschleunigen. Neue Inhalte, die in solchen Feeds gefunden werden, werden ebenfalls im Cache gespeichert und beim nächsten Ausführen des Knotens verwendet.

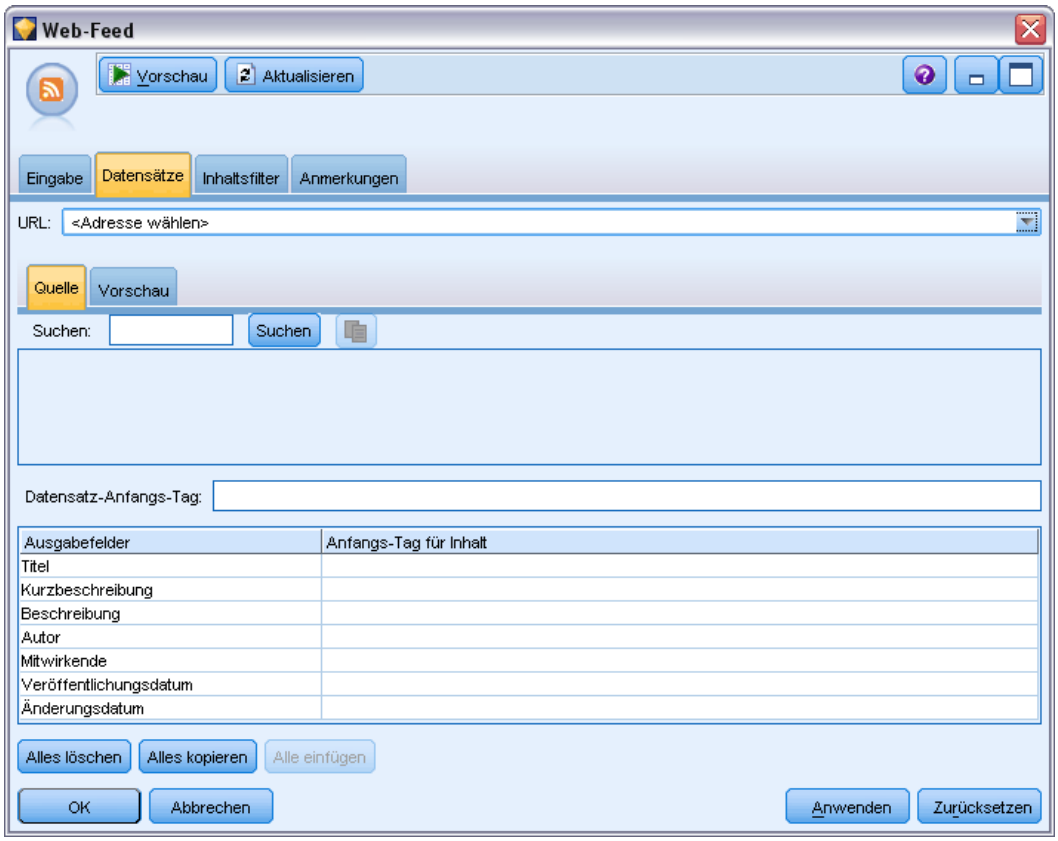
- **Beschriftung**. Wenn Sie die Option Nach Möglichkeit vorherige Web-Feeds speichern und wiederverwenden aktivieren, müssen Sie einen Beschriftungsnamen für die Ergebnisse angeben. Diese Beschriftung wird verwendet, um die auf dem Server zwischengespeicherten Feeds zu beschreiben. Wenn keine Beschriftung festgelegt wurde oder die Beschriftung nicht erkannt wurde, ist keine Wiederbenutzung möglich. Sie können diese Web-Feed-Zwischenspeicherungen in der Sitzungstabelle von IBM® SPSS® Text Analytics Administration Console verwalten. Im SPSS Text Analytics Administration Console -Benutzerhandbuch finden Sie weitere Informationen.

Web-Feed-Knoten: Registerkarte "Datensätze"

Auf der Registerkarte "Datensätze" wird der Textinhalt von nicht RSS-Feeds angegeben, indem festgelegt wird, wo ein neuer Datensatz beginnt, sowie andere Informationen zum Datensatz. Wenn Sie wissen, dass ein Nicht-RSS-Feed (HTML) Text enthält, der auf mehrere Datensätze verteilt ist, müssen Sie hier das Datensatz-Anfangs-Tag angeben, damit der Text nicht als ein Datensatz behandelt wird. Obwohl RSS-Feeds standardisiert sind und daher auf dieser Registerkarte keine Tag-Angaben erforderlich sind, können Sie den Inhalt auf der Registerkarte "Vorschau" vorab betrachten.

Wichtig: Beim Arbeiten mit Nicht-RSS-Daten ziehen Sie möglicherweise den Einsatz eines Scraping-Tools wie beispielsweise WebQL® vor, um das Sammeln von Inhalten und Verweisen auf die Ausgabe von diesem Tool mithilfe eines anderen Quellenknotens zu automatisieren.

Abbildung 2-6
Dialogfeld "Web-Feed-Knoten": Registerkarte "Datensätze"



URL. Diese Dropdown-Liste enthält eine Liste der URLs, die Sie auf der Registerkarte "Eingabe" erfasst haben. Es werden sowohl Feeds im RSS- als auch im HTML-Format angezeigt. Wenn die URL-Adresse für die Dropdown-Liste zu lang ist, wird automatisch der mittlere Teil durch Auslassungspunkte ersetzt, wie z. B. bei *http://www.ibm.com/beispiel/anfang-der-adresse...ende-der-adresse/path.htm*.

- Bei **HTML-Feeds** können Sie, wenn der Feed mehr als einen Datensatz (oder Eintrag) enthält, definieren, welche HTML-Tags die Daten enthalten, die dem in der Tabelle angezeigten Feld entsprechen. Sie können zum Beispiel das Anfangs-Tag definieren, das den Beginn eines neuen Datensatzes kennzeichnet, ein Tag für das Änderungsdatum oder eines für den Namen des Autors.
- Bei **RSS-Feeds** werden Sie nicht aufgefordert, Tags anzugeben, da es sich bei RSS um ein standardisiertes Format handelt. Falls gewünscht können Sie jedoch eine Vorschau der Ergebnisse auf der Registerkarte "Vorschau" ansehen. Allen erkannten RSS-Feeds wird das RSS-Logo vorangestellt.

Registerkarte "Quelle". Auf dieser Registerkarte können Sie den Quellcode aller HTML-Feeds anzeigen. Dieser Code ist nicht editierbar. Mit der Funktion zur Feldsuche können Sie auf dieser Seite bestimmte Tags oder Informationen suchen, die sie kopieren und unten in die Tabelle einfügen können. Bei der Feldsuche wird nicht zwischen Groß- und Kleinschreibung unterschieden. Außerdem werden Teilzeichenfolgen erkannt.

Registerkarte "Vorschau". Auf dieser Registerkarte können Sie eine Vorschau anzeigen und prüfen, wie der Web-Feed-Knoten den Datensatz einlesen wird. Dies ist besonders für HTML-Feeds sehr praktisch, da Sie die Art, wie ein Datensatz eingelesen wird, ändern können, indem Sie in der Tabelle unterhalb der Registerkarte HTML-Tags definieren.

Datensatz-Anfangs-Tag Diese Option gilt nur für Nicht-RSS-Feeds. Falls Ihr HTML-Feed Text enthält, den Sie auf mehrere Datensätze aufteilen möchten, geben Sie hier den HTML-Tag an, der den Anfang eines Datensatzes kennzeichnet (z. B. eines Artikels oder Blog-Eintrags). Wenn Sie für einen Nicht-RSS-Feed kein Anfangs-Tag definieren, wird die gesamte Seite als ein einziger Datensatz behandelt und der gesamte Inhalt wird im Feld Beschreibung ausgegeben. Als Datum der Änderung sowie als Datum der Veröffentlichung wird das Datum der Ausführung des Knotens eingesetzt.

Feldtabelle. Diese Option gilt nur für Nicht-RSS-Feeds. In dieser Tabelle können Sie den Textinhalt auf spezielle Ausgabefelder verteilen, indem Sie für jedes der vordefinierten Ausgabefelder einen Anfangs-Tag angeben. Geben Sie lediglich das Anfangs-Tag ein. Der Abgleich wird durchgeführt, indem der HTML-Code durchsucht und die Namen der Tags und Attribute mit dem Inhalt der Tabelle verglichen werden. Mithilfe der Schaltfläche am unteren Rand können Sie die von Ihnen definierten Tags kopieren und für andere Feeds erneut verwenden.

Tabelle 2-1
Mögliche Ausgabefelder für Nicht-RSS-Feeds (HTML-Formate)

Name des Ausgabefelds	Erwarteter Tag-Inhalt
Titel	Der Tag, der den Datensatztitel begrenzt. (optional)
Kurzbeschreibung	Der Tag, der die Kurzbeschreibung oder die Beschriftung begrenzt. (optional)
Beschreibung	Der Tag, der den Haupttext begrenzt. Wenn in diesem Feld keine Angabe erfolgt, enthält es den gesamten Inhalt, der sich innerhalb der `<body>`-Tags befindet (sofern es sich um einen einzigen Datensatz handelt) oder den Inhalt, der innerhalb des aktuellen Datensatzes gefunden wird (sofern ein Datensatztrennzeichen festgelegt wurde).
Author	Der Tag, der den Autor des Texts begrenzt. (optional)
Mitwirkende	Der Tag, der die Namen der beitragenden Personen begrenzt. (optional)
Datum der Veröffentlichung	Der Tag, der das Veröffentlichungsdatum des Texts begrenzt. Wenn keine Angabe erfolgt, enthält dieses Feld die Daten, wenn der Knoten die Daten liest.
Änderungsdatum	Der Tag, der das Änderungsdatum des Texts begrenzt. Wenn keine Angabe erfolgt, enthält dieses Feld die Daten, wenn der Knoten die Daten liest.

Wenn Sie ein Tag in die Tabelle eingeben, wird dieses Tag beim Durchsuchen des Feeds als ein minimales Tag für den Abgleich und nicht als exakte Übereinstimmung verwendet. Wenn Sie beispielsweise `<div>` für das Feld "Titel" eingegeben haben, werden alle im Feed gefundenen `<div>`-Tags als Treffer gewertet, auch solche, für die Attribute angegeben sind (wie `<div class="post three">`), solche, bei denen `<div>` dem Root-Tag (`<div>`) entspricht, sowie alle abgeleiteten Tags, die ein Attribut enthalten und diesen Inhalt für das Ausgabefeld "Titel" verwenden. Wenn Sie einen Root-Tag eingeben, sind alle weiteren Attribute ebenfalls enthalten.

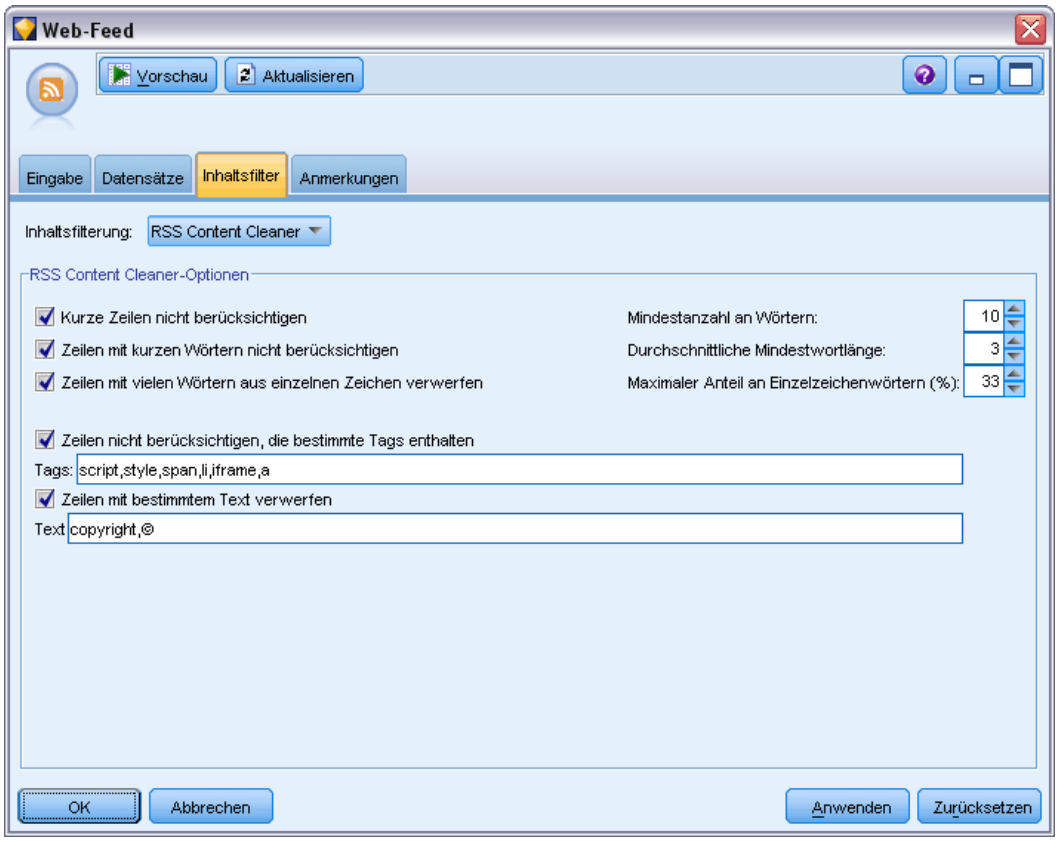
Tabelle 2-2
Beispiele für HTML-Tags, die zur Angabe des Texts für die Ausgabefelder verwendet werden

Eingabe:	Entspricht:	Entspricht außerdem:	Entspricht nicht:
`<div>`	`<div>`	`<div class="post">`	jeder andere Tag
`<p class="auth">`	`<p class="auth">`	`<p color="black" class="auth" id="85643">`	`<p color="black">`

Web-Feed-Knoten: Registerkarte "Inhaltsfilter"

Die Registerkarte "Inhaltsfilter" wird verwendet, um eine Filtermethode auf den Inhalt von RSS-Feeds anzuwenden. Diese Registerkarte gilt nicht für HTML-Feeds. Eventuell möchten Sie eine Filterung durchführen, wenn der Feed viel Text in Form von Überschriften, Fußzeilen, Menüs, Werbung und anderen unerwünschten Text enthält. Sie können diese Registerkarte verwenden, um unerwünschte HTML-Tags, JavaScript und kurze Wörter oder Textzeilen aus dem Inhalt herauszufiltern.

Abbildung 2-7
Dialogfeld "Web-Feed-Knoten": Registerkarte "Inhaltsfilter"



Inhaltsfilterung. Wenn Sie keine Filtermethode verwenden möchten, wählen Sie Keine. Ansonsten wählen Sie RSS Content Cleaner.

RSS Content Cleaner-Optionen Wenn Sie RSS Content Cleaner wählen, können Sie Zeilen basierend auf bestimmten Kriterien verwerfen. Eine Zeile wird durch einen HTML-Tag wie `<p>` und `<li>` begrenzt, nicht aber durch Inline-Tags wie `<span>`, `<b>` und `<font>`. Bitte beachten Sie, dass `<br>`-Tags als Zeilenumbrüche verarbeitet werden.

- **Kurze Zeilen verwerfen**. Diese Option ignoriert Zeilen, die nicht die hier definierte Mindestanzahl an Wörtern enthalten.
- **Zeilen mit kurzen Wörtern verwerfen**. Diese Option ignoriert Zeilen, die länger sind als die hier definierte durchschnittliche Mindestwortlänge.

- **Zeilen mit vielen Wörtern aus einzelnen Zeichen verwerfen** Diese Option ignoriert Zeilen, die mehr als einen bestimmten Anteil an aus einzelnen Zeichen bestehenden Wörtern enthalten.
- **Zeilen mit bestimmten Tags verwerfen**. Diese Option ignoriert Text in Zeilen, der in diesem Feld angegebene Tags enthält.
- **Zeilen mit bestimmtem Text verwerfen**. Diese Option ignoriert Zeilen, die in diesem Feld angegebenen Text enthalten.

Verwenden des Web-Feed-Knotens in Text Mining

Mit dem Web-Feed-Knoten können Textdaten aus Internet-Feeds für den Text-Mining-Prozess vorbereitet werden. Dieser Knoten akzeptiert Web-Feeds sowohl im HTML- als auch im RSS-Format. Diese Feeds dienen als Eingabe in den Text-Mining-Prozess (für nachfolgende Text-Mining- oder Textlinkanalyseknoten).

Wenn Sie den Web-Feed-Knoten verwenden, müssen Sie im Text-Mining- oder Textlinkanalyseknoten sicherstellen, dass angegeben ist, dass das Textfeld den Tatsächlichen Text enthält, um anzugeben, dass diese Feeds direkt auf die Artikel oder Blog-Einträge verweisen.

Wichtig: Wenn Sie versuchen, Informationen über das Internet durch einen Proxy-Server abzurufen, müssen Sie den Proxy-Server in der Datei `net.properties` für die Client- und Server-Version von IBM® SPSS® Modeler Text Analytics aktivieren. Befolgen Sie die in der Datei enthaltenen Anweisungen. Dies trifft zu, wenn Sie über den Web-Feed-Knoten auf das Internet zugreifen oder eine Lizenz für SDL Software as a Service (SaaS) abrufen, da diese Verbindungen über Java aufgebaut werden. Beim Client befindet sich diese Datei standardmäßig im Verzeichnis *C:\Program Files\IBM\SPSS\Modeler\15\jre\lib\net.properties*. Beim Server befindet sich diese Datei standardmäßig im Verzeichnis *C:\Program Files\IBM\SPSS\Modeler\15\ext\bin\spss.TMWBServer\jre\lib\net.properties*.

Beispiel: Web-Feed-Knoten (RSS-Feed) mit dem Text-Mining-Modellierungsknoten

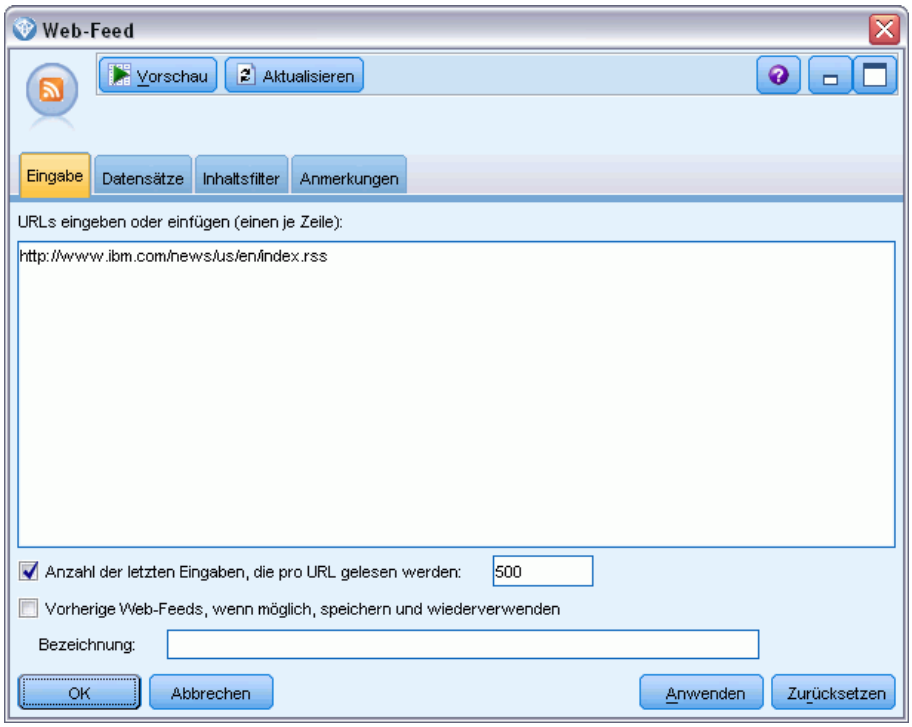
Im folgenden Beispiel stellen wir eine Verbindung zwischen einem Web-Feed-Knoten und einem Text-Mining-Knoten her, um Text aus einem RSS-Feed direkt in den Text-Mining-Prozess einzugeben.

Abbildung 2-8
Beispiel-Stream: Web-Feed-Knoten mit dem Text-Mining-Knoten

- **Web-Feed-Knoten (Registerkarte "Eingabe").** Zuerst haben wir diesen Knoten demm Stream hinzugefügt, um festzulegen, wo sich die Feed-Inhalte befinden und um die Inhaltsstruktur zu prüfen. Auf der ersten Registerkarte haben wir die URL zu einem RSS-Feed angegeben. Da es sich bei unseren Beispiel um einen RSS-Feed handelt, ist die Formatierung bereits definiert. Wir müssen daher auf der Registerkarte "Datensätze" keine Änderungen vornehmen. Für RSS-Feeds

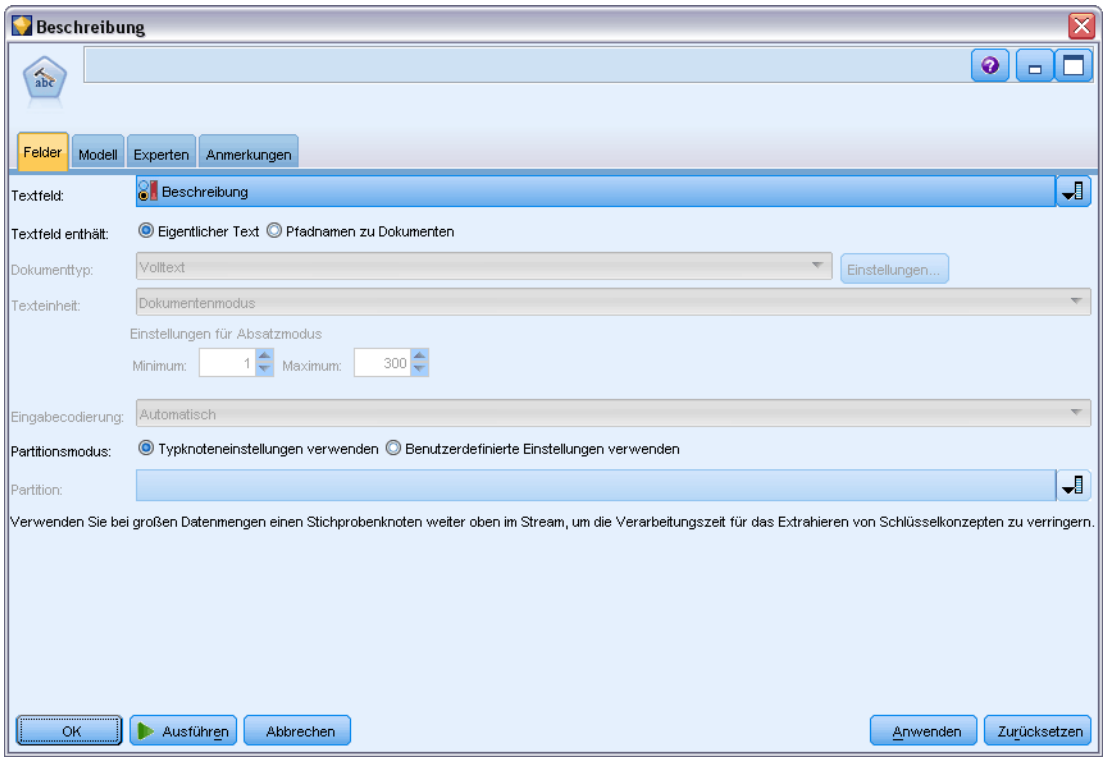
ist ein optionaler Algorithmus zur Inhaltsfilterung verfügbar, der in diesem Fall jedoch nicht angewendet wurde.

Abbildung 2-9
Dialogfeld "Web-Feed-Knoten": Registerkarte "Eingabe"



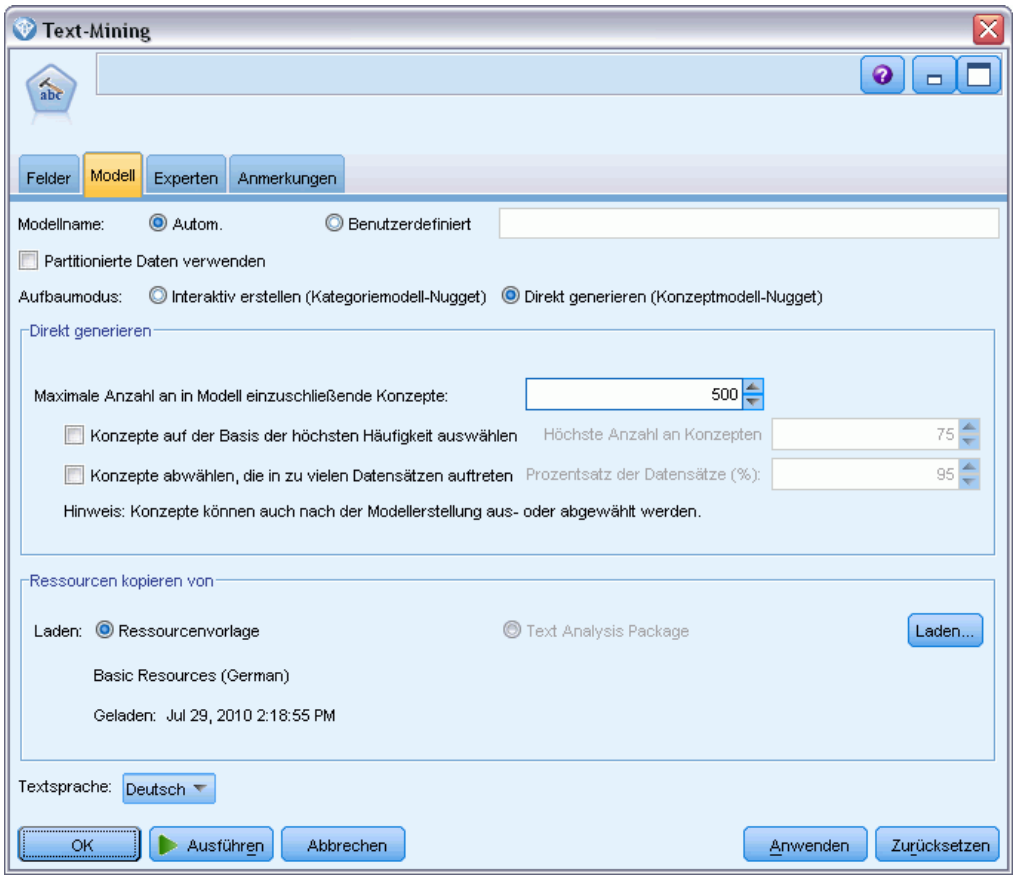
- **Text-Mining-Knoten (Registerkarte "Felder").** Im nächsten Schritt haben wir einen Text-Mining-Knoten mit dem Web-Feed-Knoten verbunden. Auf dieser Registerkarte haben wir die Textfeldausgabe des Web-Feed-Knotens definiert. In diesem Fall wollten wir das Feld Beschreibung verwenden. Wir haben für das Textfeld außerdem die Option aktiviert, dass es den Tatsächlichen Text enthält, und einige andere Einstellungen vorgenommen.

Abbildung 2-10
Dialogfeld des Text-Mining-Knotens: Registerkarte "Felder"



- **Text-Mining-Knoten (Registerkarte "Modell").** Als Nächstes haben wir auf der Registerkarte "Modell" den Aufbaumodus und die Ressourcen ausgewählt. In diesem Beispiel haben wir festgelegt, dass mit Hilfe der Standard-Ressourcenvorlage direkt aus diesem Knoten ein Konzeptmodell erzeugt werden soll.

Abbildung 2-11
Text-Mining-Knoten: Registerkarte "Modell"



Weitere Informationen zur Verwendung des Text-Mining-Knotens finden Sie unter Kapitel 3.

Kapitel
3

Mining nach Konzepten und Kategorien

Der Text-Mining-Modelllierungsknoten wird verwendet, um eines der beiden Text-Mining-Modell-Nuggets zu erzeugen:

- *Konzeptmodell-Nuggets* finden und extrahieren hervorstechende Konzepte in strukturierten oder unstrukturierten Textdaten.
- *Kategoriemodell-Nuggets* scoren Dokumente und Datensätze und ordnen sie Kategorien zu, die aus den extrahierten Konzepten (und Mustern) bestehen.

Die extrahierten Konzepte und Muster sowie die Kategorien aus Modell-Nuggets können mit bestehenden strukturierten Daten, beispielsweise demografischen Informationen, kombiniert und mithilfe der vollständigen Suite der Tools von IBM® SPSS® Modeler angewendet werden, um bessere und fokussiertere Entscheidungen zu ermöglichen. Wenn die Kunden beispielsweise häufig Anmeldeprobleme als Haupthindernis für die Online-Kontoverwaltung anführen, sollten Sie vermutlich "Anmeldeprobleme" in Ihre Modelle aufnehmen.

Außerdem ist der Text-Mining-Modellierungsknoten vollständig in SPSS Modeler integriert, sodass die Bereitstellung von Text-Mining-Streams über IBM® SPSS® Modeler Solution Publisher für ein Echtzeit-Scoring unstrukturierter Daten in Anwendungen wie PredictiveCallCenter möglich ist. Die Möglichkeit, diese Streams zu verwenden, gewährleistet erfolgreiche Text-Mining-Implementierungen in einer geschlossenen Schleife. Ihre Organisation kann nun beispielsweise durch Anwendung von Vorhersagemodellen Notizen von eingehenden oder ausgehenden Anrufern analysieren, um die Güte Ihrer Marketingaussage in Echtzeit zu überprüfen. Die Verwendung von Text-Mining-Modellergebnissen in Streams steigert erwiesenermaßen die Genauigkeit von Vorhersagedatenmodellen.

Hinweis: Um IBM® SPSS® Modeler Text Analytics mit SPSS Modeler Solution Publisher auszuführen, fügen Sie das Verzeichnis `<Installationsverzeichnis>/ext/bin/spss.TMWBServer` der Umgebungsvariablen `$LD_LIBRARY_PATH` hinzu.

In SPSS Modeler Text Analytics beziehen wir uns häufig auf extrahierte Konzepte und Kategorien. Es ist wichtig, die Bedeutung von Konzepten und Kategorien zu verstehen, da diese es ermöglichen, während der Untersuchung und Modellerstellung fundiertere Entscheidungen zu treffen.

Konzepte und Konzeptmodell-Nuggets

Während des Extrahierungsprozesses werden die Textdaten überprüft und analysiert, um interessante oder relevante Einzelwörter, beispielsweise `Wahl` oder `Frieden`, sowie Wortzusammensetzungen, beispielsweise `Präsidentschaftswahl`, `Wahl des Präsidenten` oder `Friedensverträge`, zu ermitteln. Diese Wörter und Zusammensetzungen werden

auch als *Fachausdrücke* bezeichnet. Unter Verwendung der linguistischen Ressourcen werden die relevanten Fachausdrücke extrahiert. Ähnliche Fachausdrücke werden dabei unter einem übergeordneten Fachausdruck zusammengefasst, der als **Konzept** bezeichnet wird.

So kann ein Konzept gegebenenfalls aus mehreren zugrundeliegenden Fachausdrücken bestehen. Dies hängt von dem betreffenden Text ab sowie von den verwendeten linguistischen Ressourcen. Nehmen wir zum Beispiel an, wir hätten eine Umfrage zur Angestelltenzufriedenheit und das Konzept `Gehalt` wurde extrahiert. Nehmen wir zudem an, dass Sie bei der Suche nach Datensätzen in Verbindung mit `Gehalt` festgestellt haben, dass `Gehalt` nicht immer im Text vorkommt, aber stattdessen bestimmte Datensätze etwas Ähnliches enthielten wie die Begriffe `Lohn`, `Einkommen` und `Verdienst`. Diese Begriffe werden unter `Gehalt` gruppiert, da die Extrahierungsengine sie als ähnlich eingestuft hat oder festgestellt hat, dass sie basierend auf Verarbeitungsregeln oder linguistischen Ressourcen Synonyme sind. In diesem Fall würden Dokumente oder Datensätze, in denen diese Fachausdrücke enthalten sind, so behandelt, als würden sie das Wort `Gehalt` beinhalten.

Um herauszufinden, welche Fachausdrücke unter einem Konzept zusammengefasst sind, können Sie das Konzept im Rahmen einer interaktiven Workbench-Sitzung untersuchen oder prüfen, welche Synonyme im Konzeptmodell angezeigt werden. Für weitere Informationen siehe Thema Zugrundeliegende Ausdrücke in Konzeptmodellen auf S. 62.

Ein **Konzeptmodell-Nugget** enthält eine Reihe von Konzepten, die eingesetzt werden können, um Datensätze oder Dokumente zu identifizieren, in denen das Konzept ebenfalls enthalten ist (zusammen mit sämtlichen Synonymen oder zugeordneten Fachausdrücken des Konzepts). Ein Konzeptmodell kann auf zwei Arten verwendet werden. Erstens kann es verwendet werden, um die Konzepte zu untersuchen und zu analysieren, die in dem ursprünglichen Quellentext ermittelt wurden, oder um schnell Dokumente zu identifizieren, die interessant erscheinen. Die zweite Verwendungsmöglichkeit besteht darin, dieses Modell auf weitere Textdatensätze oder Dokumente anzuwenden, um so rasch übereinstimmende Schlüsselkonzepte in den neuen Dokumenten/Datensätzen zu ermitteln, beispielsweise bei der Echtzeitermittlung von Schlüsselkonzepten in Notizen aus einem Callcenter.

Für weitere Informationen siehe Thema Text-Mining-Nugget: Konzeptmodell auf S. 57.

Kategorien und Kategoriemodell-Nuggets

Sie können **Kategorien** erstellen, die im Wesentlichen Konzepte oder Themen auf höherer Ebene darstellen, mit denen sich Schlüsselbegriffe, Wissensinhalte und Einstellungen erfassen lassen, die in dem jeweiligen Text zum Ausdruck kommen. Kategorien bestehen aus einer Reihe von Deskriptoren wie *Konzepten*, *Typen* und *Regeln*. Diese Deskriptoren werden zusammen verwendet, um zu bestimmen, ob ein Datensatz oder Dokument zu einer gegebenen Kategorie gehört oder nicht. Texte aus einem Dokument oder Datensatz können dahingehend überprüft werden, ob sie mit einem Deskriptor übereinstimmen. Liegt eine Übereinstimmung vor, wird das Dokument/der Datensatz dieser Kategorie zugeordnet. Dieser Prozess wird als **Kategorisierung** bezeichnet.

Kategorien können mit den leistungsfähigen automatisierten Methoden des Produkts automatisch erstellt werden. Daneben können Sie Kategorien auch manuell erstellen und dabei zusätzliche Erkenntnisse mit einbeziehen, die Sie hinsichtlich der Datengrundlage möglicherweise gewonnen haben. Es ist außerdem möglich, eine Kombination aus automatisierten und manuellen Methoden zu nutzen. Zudem können Sie über die Registerkarte “Modell” dieses Knotens

ein vordefiniertes Set von Kategorien aus einem Text Analysis Package laden. Die manuelle Erstellung und Verfeinerung von Kategorien kann ausschließlich über die interaktive Workbench erfolgen. Für weitere Informationen siehe Thema Text-Mining-Knoten: Registerkarte “Modell” auf S. 37.

Ein **Kategoriemodell-Nugget** enthält eine Reihe von Kategorien mit den zugehörigen Deskriptoren. Das Modell kann genutzt werden, um eine Reihe von Dokumenten oder Datensätzen auf der Grundlage des darin enthaltenen Textes zu kategorisieren. Jedes Dokument bzw. jeder Datensatz wird eingelesen und anschließend den einzelnen Kategorien zugeordnet, für die eine Übereinstimmung mit einem Deskriptor ermittelt wurde. So kann ein Dokument bzw. Datensatz mehr als einer Kategorie zugeordnet werden. Kategoriemodell-Nuggets können beispielsweise verwendet werden, um die wesentlichen Anschauungen in Umfragen mit offenen Antworten oder in einer Reihe von Blogeinträgen zu ermitteln.

Für weitere Informationen siehe Thema Text-Mining-Nugget: Kategoriemodell auf S. 72.

Text-Mining-Modellierungsknoten

Über den Text-Mining-Knoten werden mit linguistischen und häufigkeitsbasierten Verfahren Schlüsselkonzepte aus dem Text extrahiert und Kategorien mit diesen Konzepten und anderen Daten erstellt. Der Knoten kann genutzt werden, um Textdateninhalte zu untersuchen oder um ein Konzept- oder Kategoriemodell-Nugget zu erstellen. Bei der Ausführung dieses Modellierungsknotens führt eine interne linguistische Extrahierungsengine die Extrahierung und Anordnung der Konzepte, Muster und/oder Kategorien mithilfe von Verarbeitungsmethoden für natürliche Sprache durch.

Sie können den Text-Mining-Knoten ausführen und über die Option Direkt generieren automatisch ein Konzept- oder Kategoriemodell-Nugget erzeugen. Alternativ können Sie eine praktischere Untersuchungsmethode, den Modus Interaktiv erstellen verwenden, in dem Sie nicht nur Konzepte extrahieren, Kategorien erstellen und Ihre linguistischen Ressourcen verfeinern können, sondern auch eine Text Link Analysis durchführen und Cluster untersuchen können. Für weitere Informationen siehe Thema Text-Mining-Knoten: Registerkarte “Modell” auf S. 37.

Sie finden diesen Knoten auf der IBM® SPSS® Modeler Text Analytics -Registerkarte der Knotenpalette am unteren Rand des IBM® SPSS® Modeler-Fensters. Für weitere Informationen siehe Thema IBM SPSS Modeler Text Analytics-Knoten in Kapitel 1 auf S. 11.

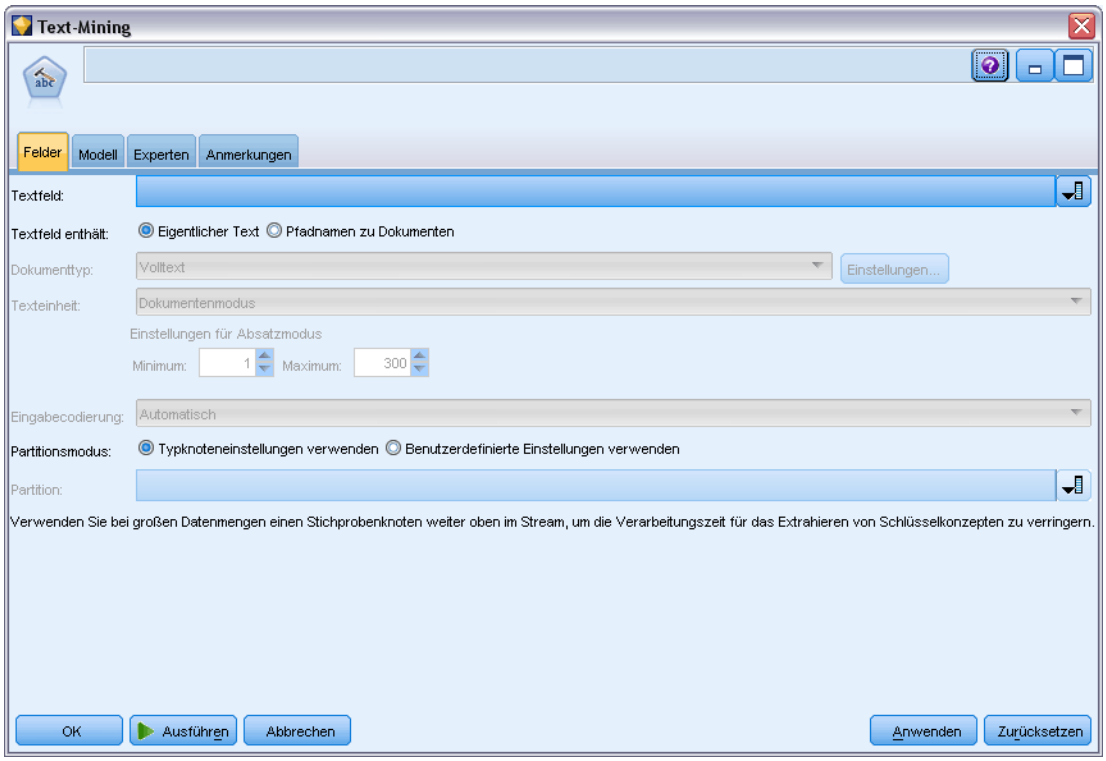
Anforderungen. Text-Mining-Modellierungsknoten nehmen Textdaten aus Web-Feed-Knoten, Dateilistenknoten sowie aus sämtlichen standardmäßigen Quellenknoten auf. Dieser Knoten wird mit SPSS Modeler Text Analytics installiert und ist über die SPSS Modeler Text Analytics -Palette verfügbar.

Anmerkung: Dieser Knoten ersetzt den Textextrahierungsknoten für alle Benutzer sowie den alten Text-Mining-Knoten für japanische Benutzer, der in früheren Versionen von Text-Mining für Clementine angeboten wurde. Wenn Sie über ältere Streams verfügen, in denen diese Knoten oder Modell-Nuggets genutzt werden, ist es erforderlich, die betreffenden Streams mithilfe des neuen Text-Mining-Knotens erneut zu erstellen. *Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Text-Mining-Knoten: Registerkarte "Felder"

Die Registerkarte "Felder" dient speziell zur Angabe der Feldeinstellungen für die Daten, aus denen Konzepte extrahiert werden sollen. Ziehen Sie bei der Arbeit mit größeren Daten-Sets die Verwendung eines Stichprobenknotens weiter oben im Stream in Erwägung; auf diese Weise lässt sich die Verarbeitungsdauer verkürzen. Für weitere Informationen siehe Thema Stichprobenziehung weiter oben im Stream zur Zeitersparnis auf S. 50.

Abbildung 3-1
Dialogfeld des Text-Mining-Modellierungsknotens: Registerkarte "Felder"



Folgende Parameter können festgelegt werden:

Textfeld. Wählen Sie das Feld, das den Text für das Mining enthält, den Dokumentenpfadnamen bzw. den Verzeichnispfadnamen, der zu den Dokumenten führt. Dieses Feld hängt von der Datenquelle ab.

Textfeld enthält. Geben Sie an, was das Textfeld enthalten soll, das in der vorausgehenden Einstellung festgelegt wurde. Folgende Optionen stehen zur Auswahl:

- **Eigentlicher Text.** Wählen Sie diese Option aus, wenn das Feld genau den Text enthält, aus dem Konzepte extrahiert werden sollen.
- **Pfadnamen zu Dokumenten.** Wählen Sie diese Option aus, wenn das Feld mindestens einen Pfadnamen zu dem oder den Speicherort(en) der Textdokumente enthält.

Dokumenttyp. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält. Der Dokumenttyp gibt die Struktur des Texts an. Wählen Sie einen der folgenden Typen aus:

- **Volltext.** Verwenden Sie diese Option für die meisten Dokumente bzw. Textquellen. Die gesamte Textmenge wird für die Extrahierung gescannt. Im Gegensatz zu den anderen Optionen gibt es keine weiteren Einstellungen für diese Option.
- **Gegliederter Text.** Verwenden Sie diese Option für bibliografische Formulare, Patente und alle Dateien, die reguläre Strukturen enthalten, die identifiziert und analysiert werden können. Dieser Dokumenttyp wird verwendet, um den gesamten Extrahierungsvorgang oder Teile des Extrahierungsvorgangs zu überspringen. Er ermöglicht das Definieren von Trennzeichen zu Fachausdrücken, das Zuweisen von Typen und das Festlegen eines minimalen Häufigkeitswerts. Wenn Sie diese Option aktivieren, müssen Sie auf die Schaltfläche Einstellungen klicken und im Bereich Formatierung als gegliederter Text des Dialogfelds "Dokumenteinstellungen" Texttrennzeichen eingeben. Für weitere Informationen siehe Thema Dokumenteinstellungen der Registerkarte "Felder" auf S. 35.
- **XML** Verwenden Sie diese Option, um die XML-Tags anzugeben, die den zu extrahierenden Text enthalten. Alle anderen Tags werden ignoriert. Wenn Sie diese Option aktivieren, müssen Sie auf die Schaltfläche Einstellungen klicken und im Bereich Formatierung als XML des Dialogfelds "Dokumenteinstellungen" explizit die XML-Elemente angeben, die den Text enthalten, der während des Extrahierungsprozesses gelesen werden soll. Für weitere Informationen siehe Thema Dokumenteinstellungen der Registerkarte "Felder" auf S. 35.

Texteinheit. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält und Sie Volltext als Dokumenttyp ausgewählt haben. Wählen Sie den Extrahierungsmodus aus folgenden Elementen aus:

- **Dokumentenmodus.** Wird für kurze, semantisch homogene Dokumente verwendet, beispielsweise Artikel von Nachrichtenagenturen.
- **Absatzmodus.** Verwenden Sie diese Option für Webseiten und Dokumente ohne Tags. Der Extrahierungsprozess teilt die Dokumente semantisch. Dabei nutzt er Merkmale wie interne Tags und Syntax. Bei Auswahl dieses Modus wird das Scoring absatzweise durchgeführt. Folglich ist die Regel `Apfel & Orange` nur erfüllt, wenn `Apfel` und `Orange` im gleichen Absatz gefunden werden.

Einstellungen für Absatzmodus. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält und als Texteinheitoption Absatzmodus angegeben haben. Geben Sie die für Extrahierungen zu verwendenden Zeichenschwellenwerte an. Die tatsächliche Größe wird auf den nächsten Punkt (Satzende) auf- bzw. abgerundet. Um sicherzustellen, dass die aus dem Text der Dokumentensammlung erstellten Wortzuordnungen repräsentativ sind, sollten Sie eine zu kleine Extrahierungsgröße vermeiden.

- **Minimum.** Geben Sie die Mindestzahl der bei Extrahierungen zu verwendenden Zeichen an.
- **Maximum.** Geben Sie die Höchstzahl der bei Extrahierungen zu verwendenden Zeichen an.

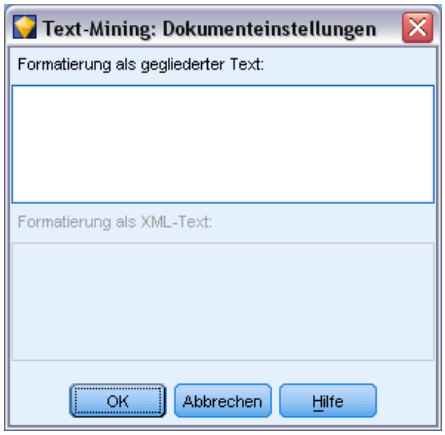
Eingabekodierung. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält. Es bestimmt die Standardtextkodierung. Für alle Sprachen außer Japanisch erfolgt eine Konvertierung von der angegebenen bzw. erkannten Kodierung in ISO-8859-1. Selbst wenn Sie also eine andere Kodierung angeben, wird diese vor der Verarbeitung von der Extrahierungsengine in iso-8859-1 konvertiert. Alle Zeichen, die nicht in die

ISO-8859-1-Kodierungsdefinition passen, werden in Leerzeichen umgewandelt. Für japanischen Text können Sie eine von mehreren Kodierungsoptionen wählen: SHIFT_JIS, EUC_JP, UTF-8 oder ISO-2022-JP.

Partitionsmodus. Mit dem Partitionsmodus können Sie auswählen, ob die Partitionierung auf der Grundlage der Typknoteneinstellungen erfolgen soll, oder eine andere Partition auswählen. Bei der Partitionierung werden die Daten in Trainings- und Teststichproben unterteilt.

Dokumenteinstellungen der Registerkarte "Felder"

Abbildung 3-2
Dialogfeld "Dokumenteinstellungen"



Formatierung als gegliederter Text

Wenn Sie den gesamten Extrahierungsvorgang oder Teile des Extrahierungsvorgangs überspringen möchten, da strukturierte Daten vorliegen oder Sie Regeln zur Behandlung des Texts festlegen möchten, verwenden Sie die Dokumenttypoption Gegliederter Text und deklarieren Sie die Felder bzw. Tags mit dem Text im Abschnitt Formatierung als gegliederter Text des Dialogfelds "Dokumenteinstellungen". Extrahierte Ausdrücke werden nur von dem Text abgeleitet, der sich in den deklarierten Feldern bzw. Tags (und untergeordneten Tags) befindet. Nicht deklarierte Felder oder Tags werden ignoriert.

In bestimmten Kontexten ist die linguistische Verarbeitung nicht erforderlich und die linguistische Extrahierungsengine kann durch explizite Deklarationen ersetzt werden. In einer Bibliografiedatei, in der die Stichwortfelder durch Trennzeichen getrennt sind, z. B. mit einem Semikolon (;) oder einem Komma (,), genügt es, die Zeichenkette zwischen zwei Trennzeichen zu extrahieren. Daher können Sie den gesamten Extrahierungsvorgang überspringen und stattdessen spezielle Regeln zum Umgang definieren, um Trennzeichen zu Fachausdrücken zu deklarieren, dem extrahierten Text Typen zuzuweisen oder einen minimalen Häufigkeitswert für die Extrahierung festzulegen.

Verwenden Sie beim Deklarieren von Elementen die folgenden Regeln:

- Pro Zeile kann nur ein Feld, Tag oder Element deklariert werden. Sie müssen nicht in den Daten vorhanden sein.
- Bei Deklarationen wird zwischen Groß- und Kleinschreibung unterschieden.

- Wenn beim Deklarieren eines Tags Attribute vorliegen, z. B. `<title id="1234">`, und Sie alle Variationen bzw. in diesem Fall alle IDs einbeziehen möchten, fügen Sie das Tag ohne das Attribut oder die spitze Klammer (>) wie folgt hinzu: `<Titel`.
- Fügen Sie nach dem Feld bzw. Tag-Namen einen Doppelpunkt hinzu, um anzugeben, dass es sich um gegliederten Text handelt. Fügen Sie diesen Doppelpunkt direkt nach dem Feld bzw. Tag und vor allen Trennzeichen, Typen oder Häufigkeitswerten (z. B. `author:` oder `<place>:`) hinzu.
- Um anzugeben, dass das Feld oder Tag mehrere Fachausdrücke enthält und zum Definieren der einzelnen Fachausdrücke ein Trennzeichen verwendet wird, deklarieren Sie das Trennzeichen nach dem Doppelpunkt (z. B. `author:,` oder `<section>:;`).
- Um dem Inhalt im Tag einen Typ zuzuweisen, deklarieren Sie den Typcode nach dem Doppelpunkt und einem Trennzeichen (z. B. `author:,Person` oder `<place>:;Location`. Deklarieren Sie den Typ mithilfe der Namen, die im Ressourceneditor angezeigt werden.
- Um einen minimalen Häufigkeitswert für ein Feld oder Tag anzugeben, deklarieren Sie am Ende der Zeile eine Zahl (z. B. `author:,Person1` oder `<place>:;Location5`). Dabei steht `n` für den Häufigkeitswert, den Sie definiert haben. Im Feld oder Tag gefundene Fachausdrücke müssen in der Gesamtmenge der Dokumente bzw. Datensätze, die extrahiert werden soll, mindestens `n`-mal vorkommen. Des Weiteren muss ein Trennzeichen definiert werden.
- Wenn ein Tag mit einem Doppelpunkt vorliegt, muss dem Doppelpunkt ein umgekehrter Schrägstrich vorangestellt werden, damit die Deklaration nicht ignoriert wird. Geben Sie also beispielsweise das Feld `<topic:source>` folgendermaßen ein: `<topic\:source>`.

Nehmen wir zur Veranschaulichung an, dass folgende wiederkehrende bibliografischen Felder vorliegen:

```
author:Morel, Kawashima
abstract:Dieser Artikel beschreibt die Deklarierung von Feldern.
publication:Text Mining-Dokumentation
datepub:März 2010
```

Falls sich der Extrahierungsvorgang in diesem Beispiel auf den Autor sowie die Zusammenfassung konzentrieren und den restlichen Inhalt ignorieren soll, werden nur die folgenden Felder deklariert:

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

In diesem Beispiel gibt die Felddeklaration `author:,Person1` an, dass die linguistische Verarbeitung beim Inhalt des Felds ausgesetzt wurde. Es wird angegeben, dass das Feld zum Autor mehrere Namen enthält, die mit einem Komma als Trennzeichen voneinander getrennt sind und die dem Typ “Person” zugewiesen werden sollen. Des weiteren wird angegeben, dass der Name extrahiert werden soll, wenn er mindestens einmal in der Gesamtmenge der Dokumente bzw. Datensätze vorkommt. Da das Feld `abstract:` ohne jegliche Deklarationen aufgeführt ist, wird es bei der Extrahierung gescannt; hierbei werden allerdings die standardmäßige linguistische Verarbeitung und die Typzuweisung herangezogen.

Formatierung als XML

Wenn Sie den Extrahierungsvorgang auf Text innerhalb bestimmter XML-Tags beschränken möchten, verwenden Sie die Dokumenttypoption XML und deklarieren Sie die Tags, die den Text enthalten, im Abschnitt Formatierung als XML des Dialogfelds “Dokumenteinstellungen”. Extrahierte Ausdrücke werden nur von dem Text abgeleitet, der sich in diesen Tags bzw. ihren untergeordneten Tags befindet.

Wichtig: Wenn Sie den Extrahierungsvorgang überspringen und Regeln für Trennzeichen zu Fachausdrücken festlegen, dem extrahierten Text Typen zuweisen oder einen Häufigkeitswert für extrahierte Fachausdrücke festlegen möchten, verwenden Sie die im Folgenden beschriebene Option Gegliederter Text.

Verwenden Sie beim Deklarieren von Tags für Formatierung als XML die folgenden Regeln:

- Pro Zeile kann nur ein XML-Tag deklariert werden.
- Bei Tag-Elementen wird zwischen Groß- und Kleinschreibung unterschieden.
- Wenn ein Tag Attribute hat, z. B. `<title id="1234">`, und Sie alle Variationen bzw. in diesem Fall alle IDs einbeziehen möchten, fügen Sie das Tag ohne das Attribut oder die spitze Klammer (>) wie folgt hinzu: `<Titel`.

Nehmen wir zur Veranschaulichung an, dass folgendes XML-Dokument vorliegt:

```
<section>Straßenverkehrsvorschriften
     <title id="01234">Verkehrssignale</title>
     <p>Straßenschilder sind hilfreich.</p>
</section>
<p>Das Erlernen der Vorschriften ist hilfreich.</p>
```

Für dieses Beispiel deklarieren wir die folgenden Tags:

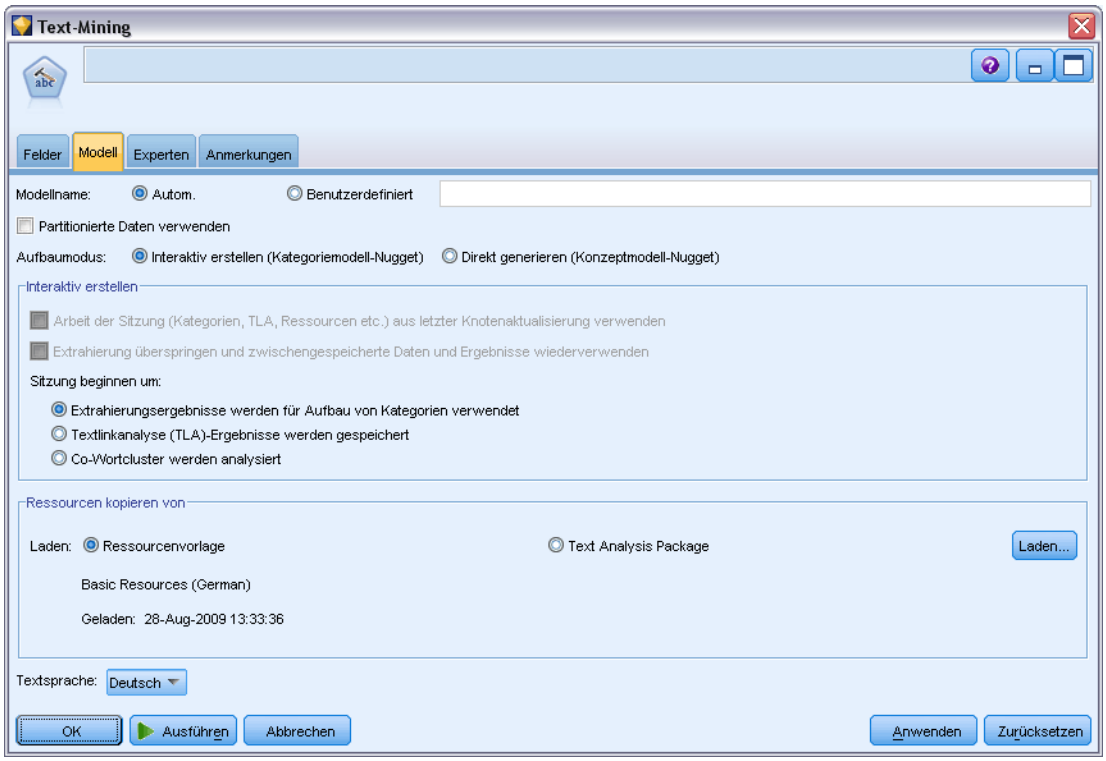
```
<section>
<title
```

In diesem Beispiel wird wegen der Deklarierung des Tags `<section>` der Text in diesem Tag und in seinen geschachtelten Tags (`Verkehrssignale` und `Straßenschilder sind hilfreich`) während des Extrahierungsvorgangs gescannt. Der Text `Das Erlernen der Vorschriften ist hilfreich` wird jedoch ignoriert, da das Tag `<p>` weder ausdrücklich deklariert wurde noch das Tag in einem deklarierten Tag verschachtelt ist.

Text-Mining-Knoten: Registerkarte “Modell”

Über die Registerkarte “Modell” können Sie die Erstellungsmethode sowie die allgemeinen Modelleinstellungen für die Knotenausgabe festlegen.

Abbildung 3-3
Dialogfeld des Text-Mining-Knotens: Registerkarte "Modell"



Folgende Parameter können festgelegt werden:

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufbaumodus. Legt fest, wie die Modell-Nuggets erstellt werden, wenn ein Stream mit diesem Text-Mining-Knoten ausgeführt wird. Alternativ können Sie eine praktischere Untersuchungsmethode, den Modus Interaktiv erstellen verwenden, in dem Sie nicht nur Konzepte extrahieren, Kategorien erstellen und Ihre linguistischen Ressourcen verfeinern können, sondern auch eine Text Link Analysis durchführen und Cluster untersuchen können.

- **Interaktiv erstellen** Bei der Ausführung eines Streams startet diese Option eine interaktive Schnittstelle, über die Sie Konzepte und Muster extrahieren, die extrahierten Ergebnisse untersuchen und optimieren, die linguistischen Ressourcen (Vorlagen, Synonyme, Typen, Bibliotheken usw.) optimieren und Kategoriemodell-Nuggets erstellen können. Für weitere Informationen siehe Thema Interaktiv erstellen auf S. 39.
- **Direkt generieren** Diese Option legt fest, dass bei der Ausführung des Streams automatisch ein Modell erstellt und der Modellpalette hinzugefügt werden soll. Im Gegensatz zur interaktiven Workbench ist bei der Ausführung neben den Einstellungen, die im Knoten definiert sind, keine weitere Manipulation erforderlich. Wenn Sie diese Option auswählen, werden

modellspezifische Optionen angezeigt, über die Sie festlegen können, welche Art von Modell Sie erstellen möchten. Für weitere Informationen siehe Thema Direkt generieren auf S. 41.

Ressourcen kopieren von: Beim Text Mining basierte die Extrahierung nicht nur auf den Einstellungen auf der Registerkarte “Experten”, sondern auch auf den linguistischen Ressourcen. Diese Ressourcen dienen als Grundlage für die Art der Be- und Verarbeitung von Daten während der Extrahierung, um die Konzepte, Typen und manchmal Muster zu erhalten. Sie können Ressourcen aus einer Ressourcenvorlage oder einem Textanalysepaket in diesen Knoten kopieren. Wählen Sie eine und klicken Sie dann auf Laden, um das Paket oder die Vorlage zu definieren, aus dem bzw. der die Ressourcen kopiert werden sollen. Wenn Sie den Ladevorgang starten, wird eine Kopie der Ressourcen im Knoten gespeichert. Wenn Sie also eine aktualisierte Vorlage oder TAP verwenden möchten, dann müssen Sie sie hier oder in einer interaktiven Workbench-Sitzung laden. Um Ihnen die Arbeit zu erleichtern, werden die Uhrzeit und das Datum im Knoten angezeigt, zu denen die Ressourcen kopiert und geladen wurden. Für weitere Informationen siehe Thema Kopieren von Ressourcen aus TAPs und Vorlagen auf S. 42.

Textsprache. Bestimmt die Sprache des für das Mining verwendeten Texts. Die in den Knoten kopierten Ressourcen steuern die angezeigten Sprachoptionen. Sie können die Sprache wählen, auf die die Ressourcen abgestimmt wurden, oder die Option ALLE wählen. Wir empfehlen, für die Textdaten die exakte Sprache anzugeben. Wenn Sie sich jedoch nicht sicher sind, können Sie die Option ALLE wählen. ALL ist für japanischen Text nicht verfügbar. Diese Option ALL verlängert die Ausführungsdauer, da mithilfe der automatischen Spracherkennung zunächst alle Dokumente und Datensätze gescannt werden, um die Textsprache zu ermitteln. Wenn Sie diese Option wählen, werden alle Datensätze oder Dokumente, die in einer unterstützten und lizenzierten Sprache vorliegen, durch die Extrahierungsengine mit den internen Wörterbüchern in der jeweiligen Sprache gelesen. Für weitere Informationen siehe Thema Language Identifier in Kapitel 18 auf S. 345. Wenden Sie sich an Ihren Kundendienstmitarbeiter, wenn Sie eine Lizenz für eine unterstützte Sprache erwerben möchten, auf die Sie zurzeit keinen Zugriff haben.

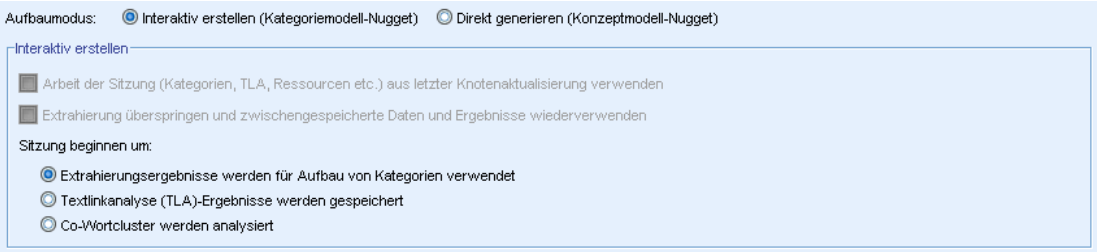
Interaktiv erstellen

Auf der Registerkarte “Modell” des Text-Mining-Modellierungsknotens können Sie einen Aufbaumodus für Ihre Modell-Nuggets auswählen. Wenn Sie Interaktiv erstellen auswählen, wird eine interaktive Schnittstelle geöffnet, wenn Sie den Stream ausführen. In dieser interaktiven Workbench können Sie:

- Die Extrahierungsergebnisse einschließlich Konzepten und Typen extrahieren und untersuchen, um hervorstechende Ideen in Ihren Textdaten zu erkennen.
- Nutzen Sie verschiedene Methoden zur Erstellung von Kategorien aus Konzepten, Typen, TLA-Mustern und Regeln sowie zu ihrer Erweiterung, um Ihre Dokumente und Datensätze in diesen Kategorien zu scoren.
- Verfeinern Sie Ihre linguistischen Ressourcen (Ressourcenvorlagen, Bibliotheken, Wörterbücher, Synonyme usw.), damit Sie Ihre Ergebnisse mit Hilfe eines iterativen Prozesses, in dem Konzepte extrahiert, untersucht und verfeinert werden, optimieren können.
- Eine Text Link Analysis (TLA) durchführen und die erkannten TLA-Muster verwenden, um bessere Kategoriemodell-Nuggets zu erstellen. Der Text Link Analysis-Knoten bietet nicht dieselben Untersuchungs- und Modellierungsfunktionen.

- Erstellen Sie Cluster zur Ermittlung neuer Beziehungen und untersuchen Sie Beziehungen zwischen Konzepten, Typen, Mustern und Kategorien im Visualisierungsbereich.
- Erstellen Sie optimierte Kategoriemodell-Nuggets für die Modellpalette in IBM® SPSS® Modeler und verwenden Sie sie in anderen Streams.

Abbildung 3-4
Optionen auf der Registerkarte "Modell" für die interaktive Erstellung



Arbeit der Sitzung (Kategorien, TLA, Ressourcen usw.) aus letzter Knotenaktualisierung verwenden. Bei einer interaktiven Workbench-Sitzung können Sie den Knoten mit Sitzungsdaten (Extrahierungsparameter, Ressourcen, Kategoriedefinitionen usw.) aktualisieren. Mit der Option Arbeit der Sitzung verwenden können Sie die interaktive Workbench mit den gespeicherten Sitzungsdaten erneut starten. Diese Option ist bei der erstmaligen Verwendung dieses Knotens deaktiviert, da keine Sitzungsdaten gespeichert werden konnten. Informationen dazu, wie der Knoten mit Sitzungsdaten aktualisiert und so die Verwendung dieser Option ermöglicht wird, finden Sie unter Aktualisieren von Modellierungsknoten und Speichern auf S. 140.

Wenn Sie eine Sitzung *mit* dieser Option starten, stehen die Extrahierungseinstellungen, Kategorien, Ressourcen und sämtliche anderen Arbeiten von der letzten Knotenaktualisierung im Rahmen einer interaktiven Workbench-Sitzung zur Verfügung, wenn Sie das nächste Mal eine Sitzung starten. Da bei dieser Option gespeicherte Sitzungsdaten verwendet werden, sind bestimmte Inhalte, beispielsweise aus der nachfolgenden Vorlage kopierte Ressourcen, sowie weitere Registerkarten deaktiviert und werden nicht berücksichtigt. Wenn Sie jedoch eine Sitzung *ohne* diese Option starten, werden nur die Inhalte des Knotens gemäß der aktuellen Definition verwendet, d. h., Arbeiten, die Sie zuvor in der Workbench durchgeführt haben, stehen nicht zur Verfügung.

Anmerkung: Wenn Sie den Quellenknoten für Ihren Stream ändern, nachdem Extrahierungsergebnisse mit der Option Arbeit der Sitzung verwenden... im Cache abgelegt wurden, müssen Sie eine neue Extrahierung ausführen, sobald die interaktive Workbench-Sitzung gestartet wird, damit Sie aktualisierte Extrahierungsergebnisse erhalten.

Extrahierung überspringen und zwischengespeicherte Daten und Ergebnisse wiederverwenden. In der interaktiven Workbench-Sitzung können sämtliche Extrahierungsergebnisse und Daten wiederverwendet werden. Diese Option ist besonders nützlich, wenn Sie Zeit sparen und Extrahierungsergebnisse wiederverwenden möchten, anstatt nach dem Start der Sitzung zu warten, bis eine ganz neue Extrahierung durchgeführt wurde. Um diese Option verwenden zu können, muss dieser Knoten im Rahmen einer interaktiven Workbench-Sitzung aktualisiert und die Option Sitzungsarbeit behalten und Textdaten mit Extrahierungsergebnissen für Wiederverwendung im Cache zwischenspeichern aktiviert worden sein. Informationen dazu, wie der Knoten mit Sitzungsdaten aktualisiert und so die Verwendung dieser Option ermöglicht wird, finden Sie unter Aktualisieren von Modellierungsknoten und Speichern auf S. 140.

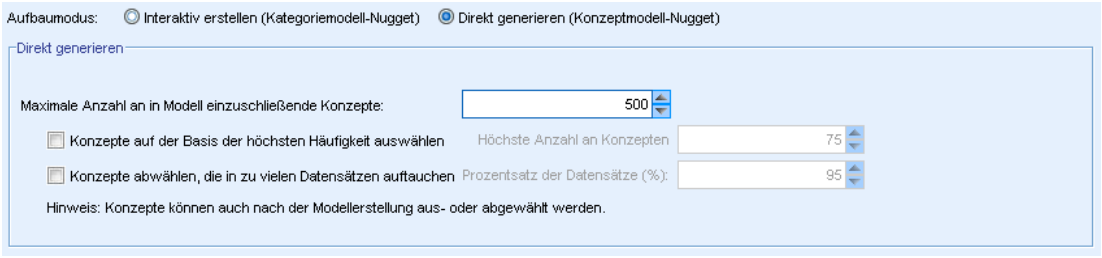
Workbench beginnen mit. Aktivieren Sie diese Option und geben Sie an, welche Ansicht angezeigt und welche Aktion durchgeführt werden soll, wenn die interaktive Workbench-Sitzung zum ersten Mal gestartet wird. Unabhängig von der Startansicht können Sie auf jede Ansicht umschalten, sobald die Sitzung gestartet wurde.

- **Verwendung von Extrahierungsergebnissen zum Aufbauen von Kategorien.** Mit dieser Option wird die interaktive Workbench in der Ansicht "Kategorien und Konzepte" gestartet und gegebenenfalls eine Extrahierung durchgeführt. In dieser Ansicht können Sie Kategorien erstellen und ein Kategoriemodell generieren. Außerdem können Sie eine andere Ansicht wählen. Für weitere Informationen siehe Thema Modus "Interaktive Workbench" in Kapitel 8 auf S. 119.
- **Untersuchen von Textlinkanalyse-(TLA-)Ergebnissen.** Mit dieser Option wird die Ansicht "Textlinkanalyse" aufgerufen und zunächst die Extrahierung durchgeführt. Dann werden Beziehungen zwischen Konzepten im Text identifiziert, beispielsweise Meinungen und andere Links. Es ist erforderlich, dass eine Vorlage oder ein Textanalysepaket ausgewählt wird, die bzw. das TLA-Musterregeln enthält, damit diese Option angezeigt wird und Ergebnisse ausgegeben werden. Wenn Sie mit größeren Datenbanken arbeiten, kann die TLA-Extrahierung einige Zeit dauern. In diesem Fall könnten Sie erwägen, upstream einen Stichprobenknoten zu verwenden. Für weitere Informationen siehe Thema Untersuchen von Textlinkanalyse in Kapitel 12 auf S. 252.
- **Analyse von Co-Wortclustern.** Mit dieser Option wird die Clusteransicht aufgerufen und veraltete Extrahierungsergebnisse werden aktualisiert. In dieser Ansicht können Sie eine Co-Wortcluster-Analyse durchführen, die zu einer Reihe von Clustern führt. Unter Co-Wortclustering versteht man einen Prozess, bei dem zunächst die Höhe des Zusammenhangwerts zwischen zwei Konzepten auf der Grundlage ihres gemeinsamen Auftretens in einem gegebenen Datensatz oder Dokument bewertet wird. Abschließend werden dann stark miteinander verbundene Konzepte in Cluster zusammengefasst. Für weitere Informationen siehe Thema Modus "Interaktive Workbench" in Kapitel 8 auf S. 119.

Direkt generieren

Auf der Registerkarte "Modell" des Text-Mining-Modellierungsknotens können Sie einen Aufbaumodus für Ihre Modell-Nuggets auswählen. Wenn Sie die Option Direkt generieren wählen, können Sie die Optionen im Knoten festlegen und anschließend einfach den Stream ausführen. Bei der Ausgabe handelt es sich um ein Konzeptmodell-Nugget, das direkt in der Modellpalette platziert wurde. Im Gegensatz zur interaktiven Workbench ist bei der Ausführung neben den Häufigkeitseinstellungen, die für diese Option im Knoten definiert sind, keine weitere Manipulation erforderlich.

Abbildung 3-5
Optionen auf der Registerkarte "Modell" für die Option "Modell aufbauen"



Maximal in das Modell aufzunehmende Konzepte. Diese Option ist nur für den automatischen (nicht interaktiven) Aufbau von Modellen verfügbar. Sie zeigt an, dass ein Konzeptmodell erstellt wird. Außerdem legt die Option fest, dass dieses Modell nicht mehr als die angegebene Zahl der Konzepte enthalten darf.

- **Konzepte nach größter Häufigkeit markieren. Größte Anzahl an Konzepten.** Gibt – ausgehend von dem Konzept mit der größten Häufigkeit – die Anzahl der zu markierenden Konzepte an. Häufigkeit bezieht sich in diesem Fall darauf, wie häufig die betreffenden Konzepte (und alle zugrundeliegenden Ausdrücke) in sämtlichen Dokumenten/Datensätzen insgesamt auftauchen. Dieser Wert kann höher sein, als der Datensatzwert, da ein Konzept mehrmals in einem Datensatz vorkommen kann.
- **Markierung von Konzepten aufheben, die in zu vielen Datensätzen auftreten. Prozentsatz der Datensätze.** Hebt die Markierung von Konzepten auf, deren Datensatzwert in Prozent höher ist als die von Ihnen angegebene Zahl. Diese Option ist nützlich, um Konzepte auszuschließen, die häufig im Text oder in jedem Datensatz vorkommen, aber keine Bedeutung für die Analyse haben.

Kopieren von Ressourcen aus TAPs und Vorlagen

Beim Text Mining basierte die Extrahierung nicht nur auf den Einstellungen auf der Registerkarte "Experten", sondern auch auf den linguistischen Ressourcen. Diese Ressourcen dienen als Grundlage für die Art der Be- und Verarbeitung von Daten während der Extrahierung, um die Konzepte, Typen und manchmal Muster zu erhalten. Sie können Ressourcen entweder aus einer *Ressourcenvorlage* in diesen Knoten kopieren, oder, falls Sie sich im Text-Mining-Knoten befinden, können Sie auch ein *Textanalysepaket* (TAP) auswählen.

Standardmäßig werden Ressourcen von der Basisvorlage für die lizenzierte Sprache für Ihr Produkt in den Knoten kopiert, wenn Sie den Knoten der Zeichenfläche hinzufügen. Falls Sie über Lizenzen für mehrere Sprachen verfügen, wird anhand der zuerst ausgewählten Sprache die Vorlage bestimmt, die automatisch geladen werden soll.

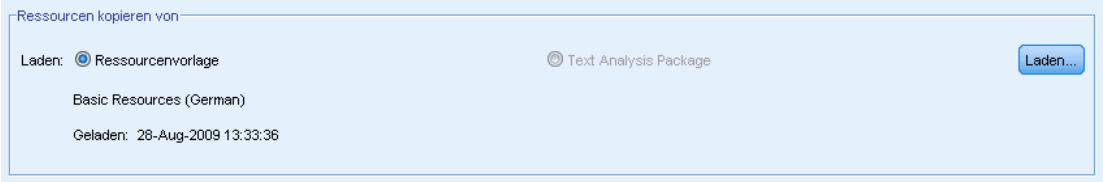
Wenn Sie den Ladevorgang starten, wird eine Kopie der ausgewählten Ressourcen im Knoten gespeichert. Dabei werden lediglich die Inhalte der Vorlage oder des TAP kopiert, während die Vorlage oder das TAP selbst nicht mit dem Knoten verknüpft wird. Das bedeutet, dass Aktualisierungen, die für diese Vorlage oder dieses TAP zu einem späteren Zeitpunkt vorgenommen werden, nicht automatisch in dem Knoten verfügbar sind. Kurz gesagt: Die Ressourcen, die in den Knoten geladen werden, werden immer verwendet, außer wenn Sie eine Kopie einer Vorlage oder eines TAP neu laden oder einen Text-Mining-Knoten aktualisieren

und die Option Arbeit der Sitzung verwenden... auswählen. Weitere Informationen zu Arbeit der Sitzung verwenden... finden Sie in diesem Thema.

Wählen Sie eine Vorlage oder ein TAP in derselben Sprache aus wie Ihre Textdaten. Sie können ausschließlich Vorlagen oder TAPs in Sprachen verwenden, für die Sie über eine Lizenz verfügen. Wenn Sie eine Textlinkanalyse ausführen möchten, wählen Sie eine Vorlage, die TLA-Muster enthält. Wenn eine Vorlage TLA-Muster enthält, wird in der TLA-Spalte des Dialogfelds "Ressourcenvorlage laden" ein Symbol angezeigt.

Anmerkung: Sie können keine TAPs in den Textlinkanalyseknoten laden.

Abbildung 3-6
Text-Mining-Knoten, Registerkarte "Modell": Optionen für das Kopieren von Ressourcen in den Knoten

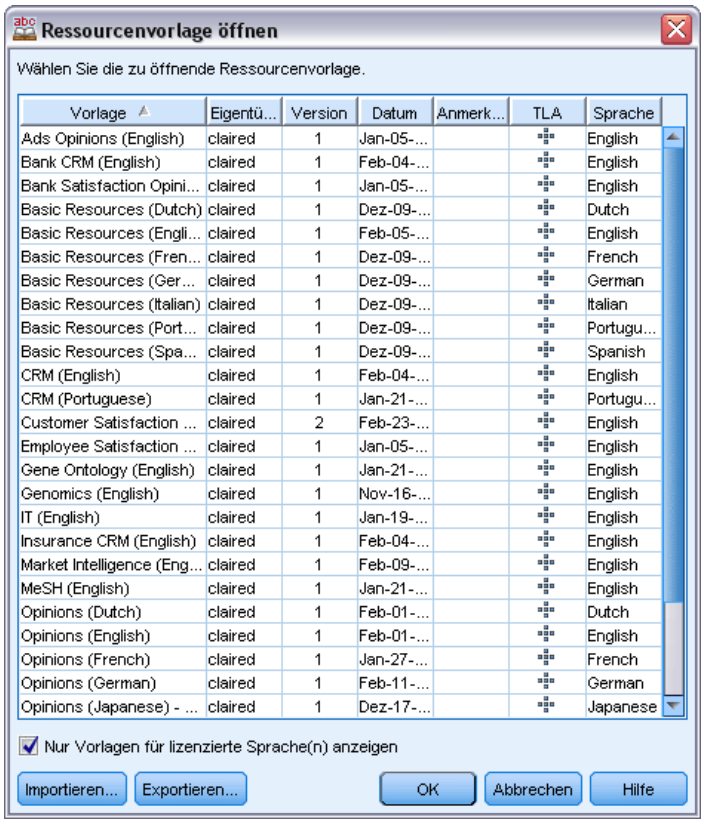


Resource Templates

Bei einer Ressourcenvorlage handelt es sich um eine vordefinierte Reihe von Bibliotheken und erweiterten linguistischen und nicht linguistischen Ressourcen, die auf eine bestimmte Domäne oder Nutzung feinabgestimmt worden sind. Im Text-Mining-Modellierungsknoten wird eine Kopie der Ressourcen von einer grundlegenden Vorlage bereits beim Hinzufügen des Knotens zum Stream in den Knoten geladen. Sie können jedoch die Vorlage ändern oder ein Textanalysepaket laden, indem Sie entweder Ressourcenvorlage oder Textanalysepaket auswählen und anschließend auf Laden klicken. Für Vorlagen können Sie die Vorlage dann im Dialogfeld "Ressourcenvorlage laden" auswählen.

Anmerkung: Wird die gewünschte Vorlage in der Liste nicht angezeigt, Sie jedoch über eine exportierte Kopie auf Ihrem Computer verfügen, können Sie sie jetzt importieren. Über dieses Dialogfeld können Sie auch exportieren, für die gemeinsame Nutzung mit anderen Benutzern. Für weitere Informationen siehe Thema Importieren und Exportieren von Vorlagen in Kapitel 15 auf S. 290.

Abbildung 3-7
Dialogfeld "Ressourcenvorlage laden"



Textanalysepakete (TAPs)

Bei einem Text Analysis Package (TAP) handelt es sich um eine vordefinierte Reihe von Bibliotheken und erweiterten linguistischen und nicht linguistischen Ressourcen sowie eines oder mehrere Sets vordefinierter Kategorien. IBM® SPSS® Modeler Text Analytics bietet nun mehrere vordefinierte TAPs für Text in englischer und auch in japanischer Sprache. Jedes TAP ist dabei auf eine bestimmte Domäne abgestimmt. Sie können diese TAPs nicht bearbeiten, aber sie dienen als Schnellstart zur Erstellung Ihres Kategoriemodells. Sie können auch eigene TAPs in der interaktiven Sitzung erstellen. Für weitere Informationen siehe Thema Laden von Text Analysis Packages in Kapitel 10 auf S. 233.*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Anmerkung: Sie können keine TAPs in den Textlinkanalyseknoten laden.

Verwenden der Option "Arbeit der Sitzung verwenden" (Registerkarte "Modell")

Obwohl Ressourcen in den Knoten auf der Registerkarte "Modell" kopiert werden, können Sie auch später noch in einer interaktiven Sitzung Änderungen an den Ressourcen vornehmen und den Text-Mining-Modellierungsknoten mit diesen aktuellen Änderungen aktualisieren. In diesem Fall würden Sie die Option Arbeit der Sitzung verwenden auf der Registerkarte "Modell" des Text-Mining-Modellierungsknotens auswählen.

Wenn Sie Arbeit der Sitzung verwenden auswählen, ist die Schaltfläche Laden im Knoten deaktiviert, um darauf hinzuweisen, dass diese Ressourcen aus der interaktiven Workbench-Sitzung an Stelle der zuvor geladenen Ressourcen verwendet werden.

Um nach Auswahl der Option Arbeit der Sitzung verwenden Änderungen an Ressourcen vorzunehmen, können Sie Ihre Ressourcen direkt in der interaktiven Workbench-Sitzung in der Resource Editor-Ansicht bearbeiten oder zwischen ihnen wechseln. Für weitere Informationen siehe Thema Aktualisieren von Knotenressourcen nach dem Laden in Kapitel 15 auf S. 288.

Text-Mining-Knoten: Registerkarte "Experten"

Die Registerkarte "Experten" enthält bestimmte erweiterte Parameter, die beeinflussen, wie der Text extrahiert und gehandhabt wird. Die Parameter in diesem Dialogfeld legen das Grundverhalten sowie einige erweiterte Verhaltensweisen des Extrahierungsprozesses fest. Sie stellen jedoch nur einen Teil der Ihnen zur Verfügung stehenden Optionen dar. Zudem werden die Extrahierungsergebnisse auch von einer Reihe linguistischer Ressourcen und Optionen beeinflusst. Diese werden über die Ressourcenvorlage gesteuert, die Sie auf der Registerkarte "Modell" auswählen. Für weitere Informationen siehe Thema Text-Mining-Knoten: Registerkarte "Modell" auf S. 37.

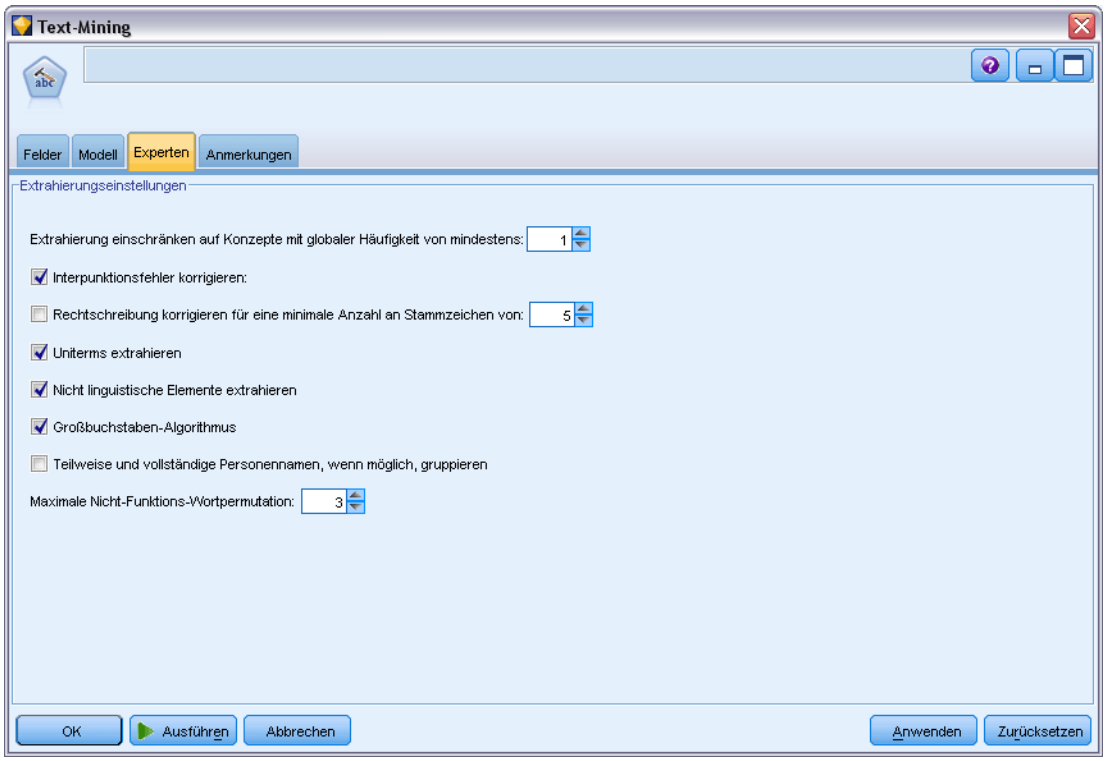
Anmerkung: Wenn Sie auf der Registerkarte "Modell" den Modus Interaktiv erstellen mit der Option "Gespeicherte Informationen aus interaktiver Sitzung verwenden" gewählt haben, ist die gesamte Registerkarte deaktiviert. In diesem Fall werden die Extrahierungseinstellungen von der letzten gespeicherten Workbench-Sitzung übernommen.

Für niederländischen, englischen, französischen, deutschen, italienischen, portugiesischen und spanischen Text

Die folgenden Parameter können Sie immer dann festlegen, wenn Sie von Japanisch abweichende Sprachen extrahieren, z. B. Englisch, Spanisch, Französisch, Deutsch usw.:

Anmerkung: In diesem Thema finden Sie auch Informationen zu den Experteneinstellungen für japanischen Text. Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Abbildung 3-8
Dialogfeld des Text-Mining-Knotens: Registerkarte "Experten"



Extrahierung einschränken auf Konzepte mit globaler Häufigkeit von mindestens [n] Gibt an, wie oft ein Wort oder Ausdruck mindestens im Text vorkommen muss, damit es bzw. er extrahiert wird. So begrenzt der Wert 5 die Extrahierung auf diejenigen Wörter oder Ausdrücke, die mindestens fünfmal in der Gesamtmenge der Datensätze bzw. Dokumente vorkommen.

In manchen Fällen kann eine Änderung dieser Grenze zu einem großen Unterschied bei den Extrahierungsergebnissen (und folglich auch Ihrer Kategorien) führen. Nehmen wir an, Sie arbeiten mit Restaurantdaten und belassen die Grenze für diese Option auf 1. In diesem Fall würden Sie unter Umständen *Pizza (1), dünne Pizza (2), Pizza Spinat (2)* und *beliebteste Pizza (2)* in Ihren Extrahierungsergebnissen finden. Würden Sie jedoch die Extrahierung auf eine globale Häufigkeit vom 5 oder mehr beschränken und erneut extrahieren, würden Sie drei dieser Konzepte nicht mehr erhalten. Stattdessen wäre das Ergebnis *Pizza (7)*, da *Pizza* die einfachste Form ist und dieses Wort außerdem bereits als ein möglicher Kandidat vorhanden war. Und abhängig vom Rest Ihres Textes hätten Sie unter Umständen auch eine Häufigkeit von mehr als sieben, nämlich dann, wenn im Text noch andere Wortfolgen mit "Pizza" vorkämen. Außerdem müssten Sie, falls *Pizza Spinat* bereits ein Kategoriedeskriptor wäre, stattdessen *Pizza* als Deskriptor hinzufügen, um alle Datensätze zu erfassen. Aus diesem Grund sollten Sie bei Änderungen dieser Grenze vorsichtig verfahren, wenn bereits Kategorien erstellt worden sind.

Beachten Sie, dass es sich hierbei um eine reine Extraktionsfunktion handelt. Wenn Ihre Vorlage Begriffe enthält (was normalerweise der Fall ist) und ein Begriff für die Vorlage im Text gefunden wird, wird der Begriff unabhängig von seiner Häufigkeit indiziert.

Nehmen wir beispielsweise an, Sie verwenden eine Vorlage vom Typ “Basic Resources”, die unter dem Typ <Location> in der Kernbibliothek “los angeles” enthält. Wenn in Ihrem Dokument nur ein einziges Mal “Los Angeles” vorkommt, wird Los Angeles in die Liste der Konzepte aufgenommen. Um dies zu verhindern, müssen Sie einen Filter festlegen, um nur Konzepte anzuzeigen, die mindestens so oft vorkommen wie im Feld Extrahierung einschränken auf Konzepte mit globaler Häufigkeit von mindestens: [n] angegeben.

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Punktuationsfehlern (zum Beispiel ungeeignete Verwendung) während der Extrahierung, um die Extrahierbarkeit von Konzepten zu verbessern. Diese Option ist besonders nützlich bei kurzem Text und niedriger Textqualität (wie dies beispielsweise bei offenen Antworten bei Umfragen, E-Mails und CRM-Daten der Fall ist) oder wenn der Text viele Abkürzungen enthält.

Rechtschreibung korrigieren für eine minimale Anzahl an Stammzeichen von [n]. Diese Option wendet ein unscharfes Gruppierungsverfahren an, das hilft, häufig falsch geschriebene Wörter oder genau geschriebene Wörter unter einem Konzept zu gruppieren. Der Algorithmus für die unscharfe Gruppierung entfernt alle Vokale (außer den ersten) und doppelte/dreifache Konsonanten temporär aus extrahierten Wörtern und vergleicht sie, um festzustellen, ob sie gleich sind, so dass `Modellierung` und `Modelierung` zusammen gruppiert werden würden. Wenn jedoch jeder Fachausdruck einem anderen Typ (ausschließlich des Typs `<Unknown>`) zugewiesen ist, wird das unscharfe Gruppierungsverfahren nicht angewendet.

Sie können auch die minimal erforderliche Zahl von *Stamm*zeichen definieren, bevor unscharfe Gruppierung eingesetzt wird. Die Anzahl der Stammzeichen in einem Ausdruck berechnet sich aus der Summe aller Zeichen abzüglich aller Zeichen, die Beugungsendungen und – bei zusammengesetzten Ausdrücken – Determinatoren und Präpositionen bilden. So würde beispielsweise der Ausdruck `Aufgaben` durch die Form “Aufgabe” mit 7 Stammzeichen gezählt werden, da der Buchstabe *n* am Ende des Worts eine Beugung darstellt (Pluralform). Gleichermaßen werden für `Apfelmus` 8 Stammzeichen (“Apfelmus”) gezählt und `Hersteller von Autos` zählt als 14 Stammzeichen (“Hersteller Auto”). Diese Zählmethode dient nur zur Überprüfung, ob die Fuzzy-Gruppierung angewendet werden soll, hat jedoch keinen Einfluss auf den Abgleich der Wörter.

Hinweis: Wenn Sie feststellen, dass bestimmte Wörter später falsch gruppiert werden, können Sie Wortpaare aus diesem Verfahren ausschließen, indem Sie sie explizit im Abschnitt Unscharfe Gruppierung: Ausnahmen im erweiterten Ressourceneditor deklarieren. Für weitere Informationen siehe Thema Unscharfe Gruppierung in Kapitel 18 auf S. 336.

uniterms extrahieren. Diese Option extrahiert einzelne Wörter (Uniterms), solange das Wort nicht bereits Teil eines zusammengesetzten Worts ist und es entweder ein Nomen oder eine nicht erkannte Wortart ist.

Nicht linguistische Elemente extrahieren. Diese Option extrahiert nicht linguistische Elemente wie beispielsweise Telefonnummern, Personalausweisnummern, Uhrzeiten, Datumsangaben, Währungen, Ziffern, Prozentsätze, E-Mail-Adressen und HTTP-Adressen. Sie können bestimmte Typen von nicht linguistischen Elementen im Abschnitt Nicht linguistische Elemente: Konfiguration des erweiterten Ressourceneditors ein- bzw. ausschließen. Durch Deaktivierung unnötiger Elemente vergeudet die Extrahierungsengine keine Verarbeitungszeit. Für weitere Informationen siehe Thema Konfiguration in Kapitel 18 auf S. 341.

Großbuchstaben-Algorithmus. Diese Option extrahiert einfache und zusammengesetzte Ausdrücke, die sich nicht in den integrierten Wörterbüchern befinden, solange der erste Buchstabe des Begriffs in Großbuchstaben geschrieben ist. Diese Option ist eine gute Möglichkeit, die geeignetsten Substantive zu extrahieren.

Teilweise und vollständige Personennamen, wenn möglich, gruppieren. Diese Option gruppiert Namen, die zusammen im Text unterschiedlich erscheinen. Diese Funktion ist nützlich, da Namen zu Beginn des Textes oft in voller Länge angegeben werden und später nur noch mit einer Kurzform auf sie verwiesen wird. Diese Option versucht, jeden Uniterm mit dem Typ `<Unknown>` mit dem letzten Wort aller zusammengesetzten Ausdrücke abzugleichen, die dem Typ `<Person>` zugeordnet sind. Wird beispielsweise *doe* gefunden und anfänglich dem Typ `<Unknown>` zugeordnet, überprüft die Extrahierungsengine, ob ein zusammengesetzter Ausdruck vom Typ `<Person>` als letztes Wort *doe* enthält, z. B. *john doe*. Diese Option wird nicht auf Vornamen angewendet, da sie in den meisten Fällen nicht als Uniterms extrahiert werden.

Maximale Nicht-Funktions-Wortpermutation. Diese Option gibt die maximale Anzahl von Füllwörtern an, die für die Anwendung des Permutationsverfahrens vorhanden sein müssen. Dieses Permutationsverfahren gruppiert ähnliche Phrasen, die sich nur durch die enthaltenen Füllwörter (zum Beispiel von und der) unabhängig von der Beugung unterscheiden. Nehmen wir zum Beispiel an, dass Sie diesen Wert auf höchstens zwei Wörter eingestellt haben und sowohl `Unternehmen des Vertreters` und `Vertreter des Unternehmens` extrahiert wurden. In diesem Fall würden beide extrahierte Ausdrücke in der endgültigen Konzeptliste zusammen gruppiert, da beide Ausdrücke als gleich betrachtet werden, wenn `des` ignoriert wird.

Anmerkung: Um die Extrahierung von Text-Link-Analysis-Ergebnissen zu ermöglichen, müssen Sie die Sitzung mit der Option Text-Link-Analysis-Ergebnisse untersuchen beginnen und außerdem Ressourcen mit TLA-Definitionen auswählen. Sie können TLA-Ergebnisse auch später während einer interaktiven Workbench-Sitzung über das Dialogfeld “Extrahierungseinstellungen” extrahieren. Für weitere Informationen siehe Thema Daten extrahieren in Kapitel 9 auf S. 147.

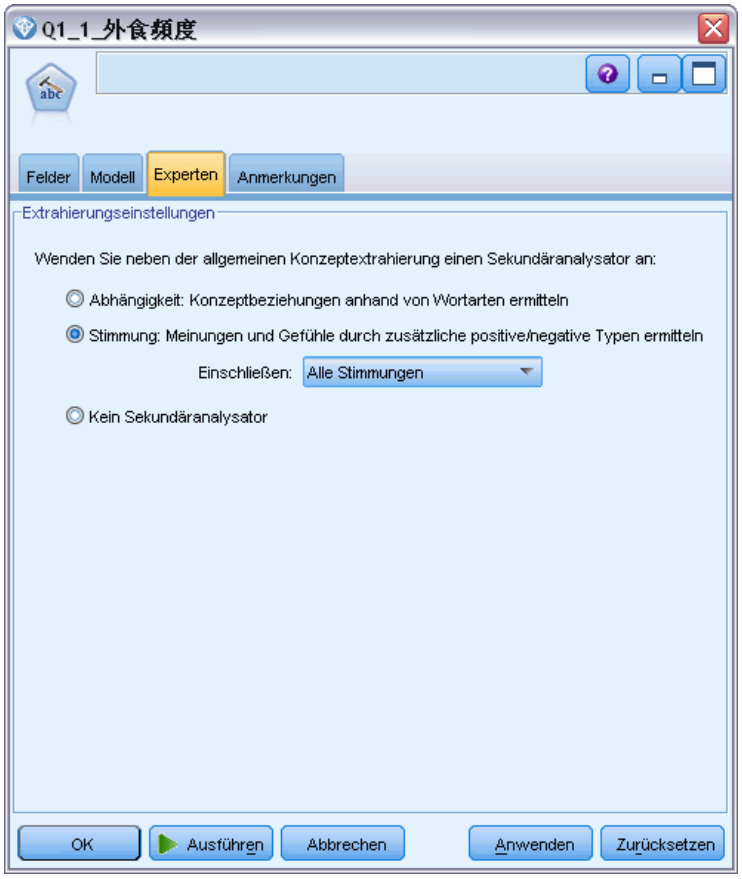
Für japanischen Text

Anmerkung: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Dieses Dialogfeld enthält abweichende Optionen für japanischen Text, da der Extrahierungsprozess einige Unterschiede aufweist. Für die Arbeit mit japanischem Text müssen Sie auch eine Vorlage oder ein Textanalysepaket für japanische Sprache in der Registerkarte “Modell” dieses Knotens wählen. Für weitere Informationen siehe Thema Kopieren von Ressourcen aus TAPs und Vorlagen auf S. 42.

Folgende Parameter können festgelegt werden:

Abbildung 3-9
Dialogfeld des Text-Mining-Knotens: Registerkarte "Experten" (japanischer Text)



Anmerkung: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Sekundäre Analyse. Beim Start einer Extrahierung werden anhand des Standardsatzes an Typen grundlegende Stichwörter extrahiert. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text in Anhang A auf S. 387. Wenn Sie jedoch eine sekundäre Analyse wählen, erhalten Sie mehr und vielfältigere Konzepte, da der Extraktor nun Partikel und Hilfsverben als Teil des Konzepts berücksichtigt. Nehmen Sie beispielsweise an, der Satz 肩の荷が下りた wurde in "*Mir wurde eine große Last von den Schultern genommen*". In diesem Beispiel kann die grundlegende Stichwortextrahierung jedes Konzept wie folgt separat extrahieren: 肩 (*Schultern*), 荷 (*Last*), 下りる (*wurde genommen*), jedoch wird die Beziehung zwischen diesen Wörtern nicht extrahiert. Wenn Sie jedoch die Stimmungsanalyse anwenden, können Sie vielfältigere Konzepte hinsichtlich eines Stimmungstyps extrahieren, z. B. das Konzept =肩の荷が下りた, das als "*eine große Last von den Schultern nehmen*", was dem Typ <良い-安心> zugeordnet ist. Im Fall der Stimmungsanalyse wird auch eine große Anzahl an zusätzlichen Typen eingeschlossen. Des Weiteren ermöglicht Ihnen die Wahl einer sekundären Analyse, auch Textlinkanalysis-Ergebnisse zu generieren.

Anmerkung: Wenn ein Sekundär-Analysator genannt wird, dauert der Extrahierungsprozess länger. Für weitere Informationen siehe Thema Wie sekundäre Extrahierung funktioniert in Anhang A auf S. 380.

- **Abhängigkeitsanalyse**. Die Wahl dieser Option führt zu erweiterten Partikeln für die Extrahierungskonzepte aus der grundlegenden Typ- und Stichwortextrahierung. Sie können auch vielfältigere Musterergebnisse aus einer Abhängigkeits-Textlinkanalyse (TLA) gewinnen.
- **Stimmungsanalyse**. Bei dieser Analyse werden zusätzliche Konzepte und – wann immer möglich – TLA-Musterergebnisse extrahiert. Neben den grundlegenden Typen erhalten Sie auch den Vorteil von über 80 Stimmungstypen, z. B. 嬉しい, 吉報, 幸運, 安心, 幸福 usw. Mithilfe dieser Typen werden Konzepte und Muster im Text durch den Ausdruck von Emotionen, Stimmungen und Meinungen aufgedeckt. Es gibt drei Optionen, die den Fokus für die Stimmungsanalyse festlegen: Alle Stimmungen, Nur repräsentative Stimmung und Nur Schlussfolgerungen.
- **Keine sekundäre Analyse**. Diese Option schaltet sämtliche sekundären Analysen aus. Diese Option ist ausgeblendet, wenn die Option Untersuchen von Textlinkanalyse- (TLA-) Ergebnissen in der Registerkarte "Modell" ausgewählt wurde, da eine sekundäre Analyse erforderlich ist, um TLA-Ergebnisse zu erhalten. Wenn Sie diese Option auswählen, später aber die Option Untersuchen von Textlinkanalyse- (TLA-) Ergebnissen aktivieren, tritt während der Stream-Ausführung ein Fehler auf.

Stichprobenziehung weiter oben im Stream zur Zeitersparnis

Bei einer großen Datenmenge kann die Verarbeitung Minuten oder Stunden in Anspruch nehmen, insbesondere bei einer interaktiven Workbench-Sitzung. Je größer der Umfang der Daten, desto mehr Zeit nehmen Extrahierung und Kategorisierung in Anspruch. Für effizienteres Arbeiten können Sie einen der Stichprobenknoten von IBM® SPSS® Modeler weiter oben im Stream hinzufügen. Ziehen Sie mithilfe dieses Stichprobenknotens eine Zufallsstichprobe mit einer kleineren Untergruppe von Dokumenten oder Datensätzen für die ersten paar Durchläufe.

Eine kleinere Stichprobe ist häufig absolut ausreichend, um die Bearbeitung der Ressourcen festzulegen und die meisten, wenn nicht sogar alle, Kategorien zu erstellen. Und wenn Sie das kleinere Daten-Set ausgeführt haben und die Ergebnisse Ihren Vorstellungen entsprechen, können Sie dieselbe Technik anwenden, um Kategorien für das gesamte Daten-Set zu erstellen. Im Anschluss können Sie nach Dokumenten oder Datensätzen suchen, die nicht in die von Ihnen definierten Kategorien fallen, und nach Bedarf Änderungen vornehmen.

Anmerkung: Bei dem Stichprobenknoten handelt es sich um einen standardmäßigen SPSS Modeler-Knoten.

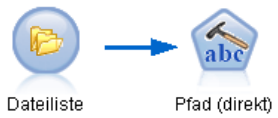
Verwenden des Text-Mining-Knotens in einem Stream

Der Text-Mining-Modellierungsknoten wird für den Zugriff auf Daten und zum Extrahieren von Konzepten in einem Stream verwendet. Sie können für den Zugriff auf die Daten jeden beliebigen Quellenknoten verwenden, beispielsweise den Datenbankknoten, den Knoten für variable Dateien, den Web-Feed-Knoten oder den Knoten für feste Dateien. Für Text in externen Dokumenten kann ein Dateilistenknoten verwendet werden.

Beispiel 1: Dateilistenknoten und Text-Mining-Knoten zum direkten Erzeugen eines Konzeptmodell-Nuggets

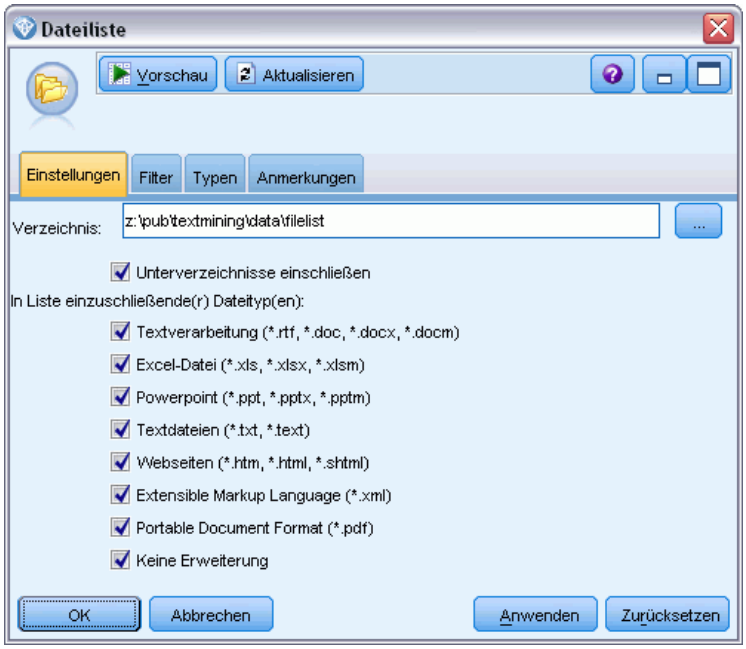
Das folgende Beispiel zeigt die Verwendung des Dateilistenknotens mit dem Text-Mining-Modellierungsknoten zum Erzeugen des Konzeptmodell-Nuggets. Weitere Informationen zur Verwendung des Dateilistenknotens finden Sie unter Kapitel 2.

Abbildung 3-10
Beispiel-Stream: Dateilistenknoten mit Text-Mining-Knoten



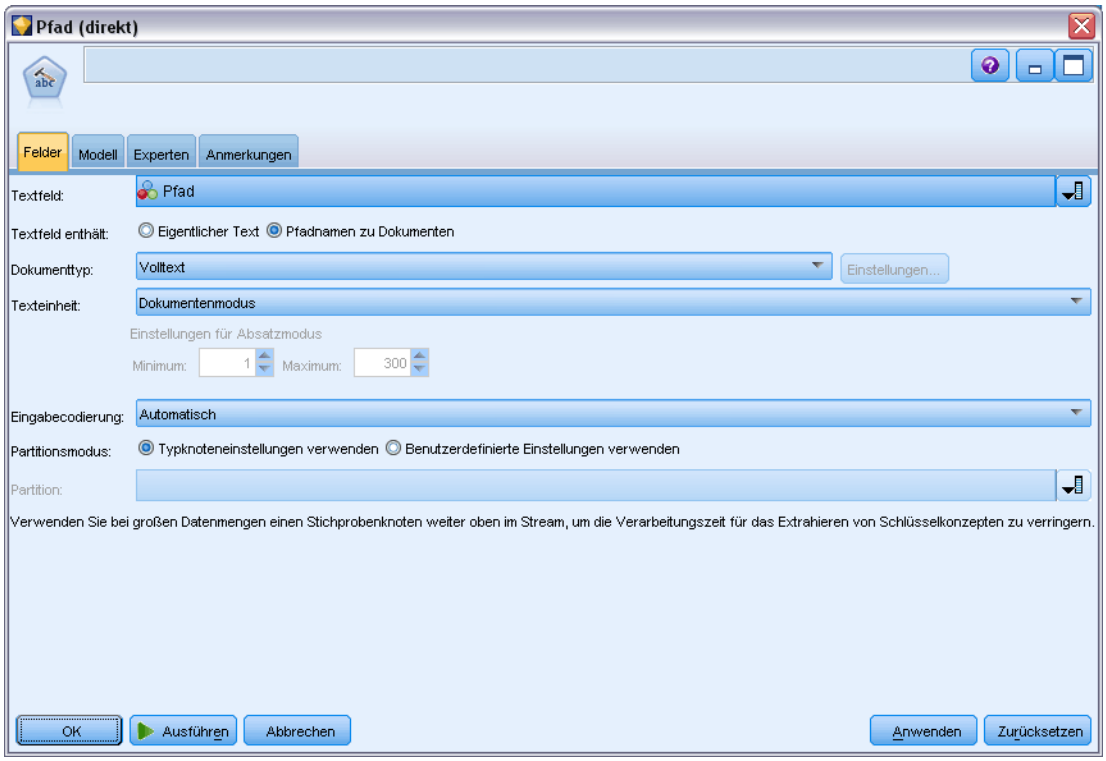
- **Dateilistenknoten (Registerkarte "Einstellungen").** Zuerst fügten wir diesen Knoten dem Stream hinzu, um anzugeben, wo die Textdokumente gespeichert sind. Wir wählten das Verzeichnis aus mit allen Dokumenten, in denen wir eine Textdatensuche durchführen wollten.

Abbildung 3-11
Dialogfeld des Dateilistenknotens: Registerkarte "Einstellungen"



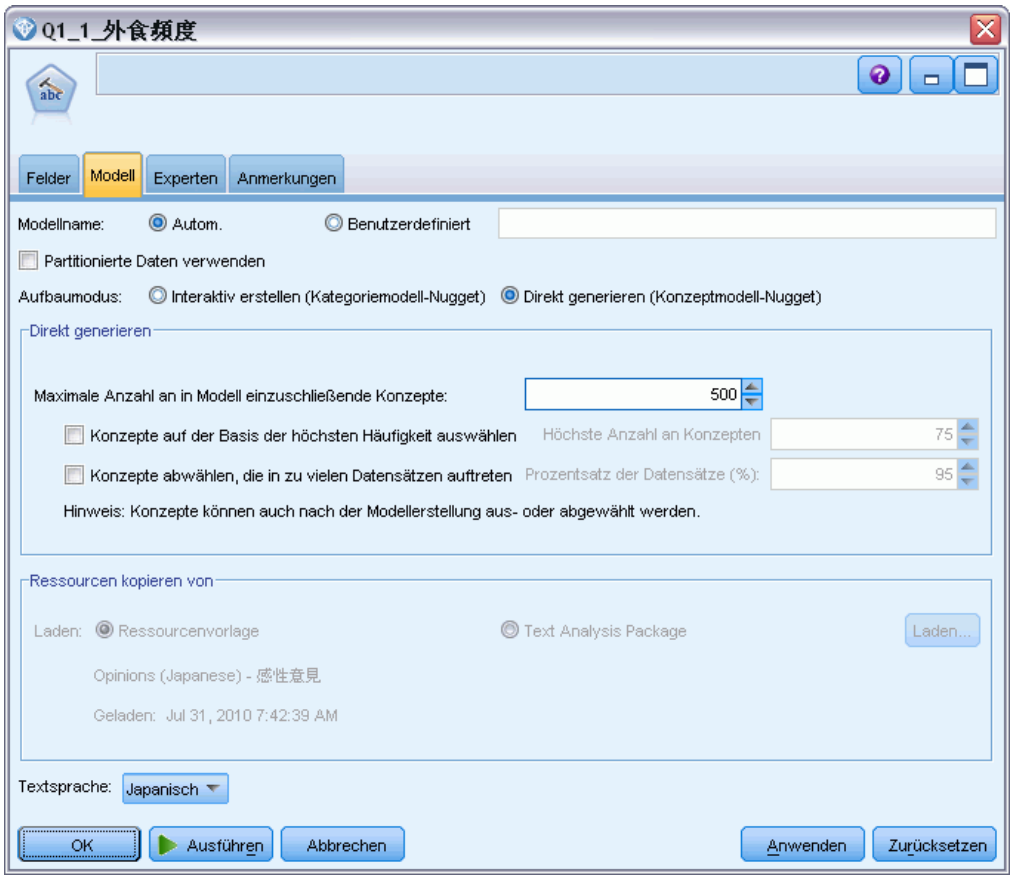
- **Text-Mining-Knoten (Registerkarte "Felder").** Anschließend fügten wir dem Dateilistenknoten einen Text-Mining-Knoten hinzu und verbanden ihn. In diesem Knoten definierten wir unser Eingabeformat und die Ressourcenvorlage sowie das Ausgabeformat. Wir wählten den aus dem Dateilistenknoten erstellten Feldnamen aus und wählten die Option aus, in der das Textfeld Pfadnamen zu Dokumenten repräsentiert, sowie andere Einstellungen. Für weitere Informationen siehe Thema Verwenden des Text-Mining-Knotens in einem Stream auf S. 50.

Abbildung 3-12
Dialogfeld des Text-Mining-Knotens: Registerkarte "Felder"



- **Text-Mining-Knoten (Registerkarte "Modell").** Als Nächstes haben wir auf der Registerkarte "Modell" den Modus zum Generieren eines Konzeptmodells direkt über diesen Knoten gewählt. Sie können eine andere Ressourcenvorlage wählen, doch für dieses Beispiel haben wir die grundlegenden Ressourcen beibehalten.

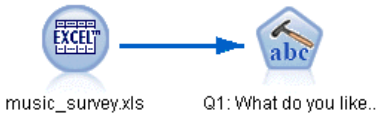
Abbildung 3-13
Dialogfeld des Text-Mining-Modellierungsknotens: Registerkarte "Modell"



Beispiel 2: Excel-Datei- und Text-Mining-Knoten zum interaktiven Aufbau eines Kategoriemodells

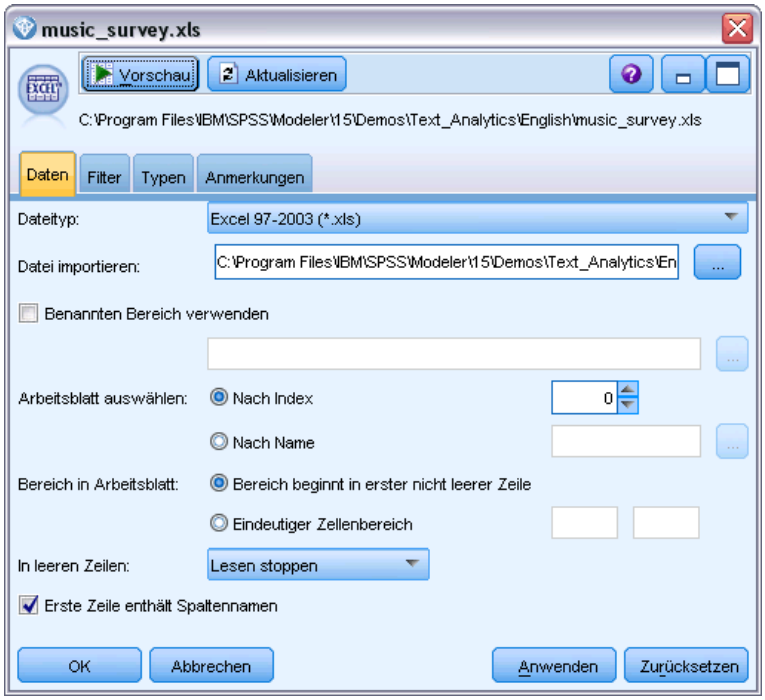
Dieses Beispiel zeigt, wie eine interaktive Workbench-Sitzung auch über den Text-Mining-Knoten gestartet werden kann. Weitere Informationen zur interaktiven Workbench finden Sie hier: Kapitel 8.

Abbildung 3-14
Beispiel-Stream: Excel-Quellenknoten mit Text-Mining-Knoten (interaktiv erstellen)



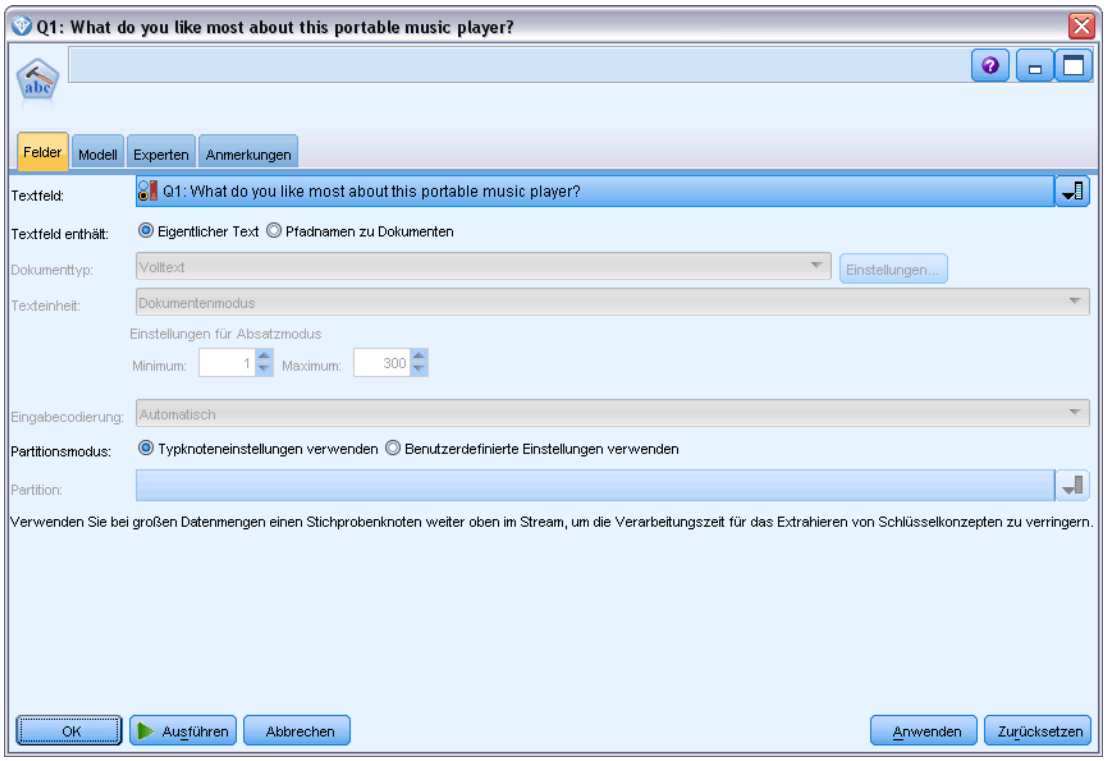
- **Excel-Quellenknoten (Registerkarte "Daten").** Zuerst fügten wir diesen Knoten dem Stream hinzu, um anzugeben, wo der Text gespeichert ist.

Abbildung 3-15
Dialogfeld für Excel-Quellenknoten: Registerkarte "Daten"



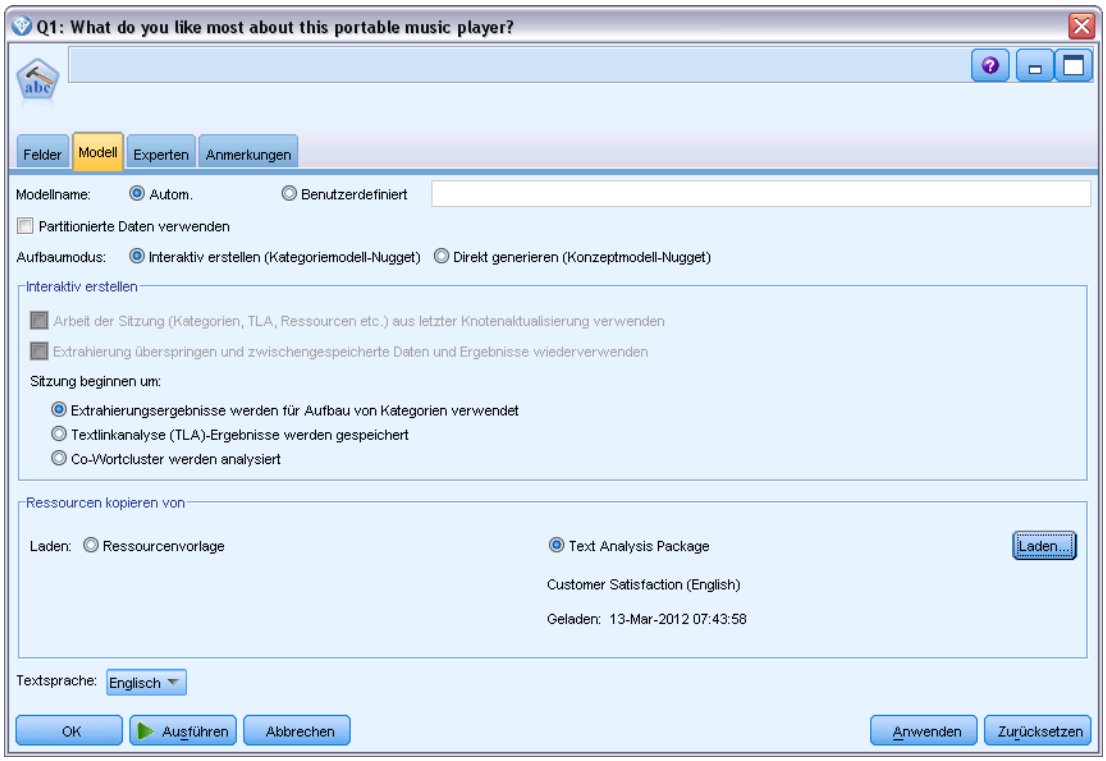
- **Text-Mining-Knoten (Registerkarte "Felder").** Als Nächstes haben wir einen Text-Mining-Knoten hinzugefügt und angeschlossen. Auf dieser ersten Registerkarte haben wir unser Eingabeformat festgelegt. Wir haben einen Feldnamen des Quellenknotens und die Option "Textfeld enthält" tatsächlichen Text ausgewählt, da die Daten direkt aus dem Excel-Quellenknoten stammen.

Abbildung 3-16
Dialogfeld des Text-Mining-Modellierungsknotens: Registerkarte "Felder"



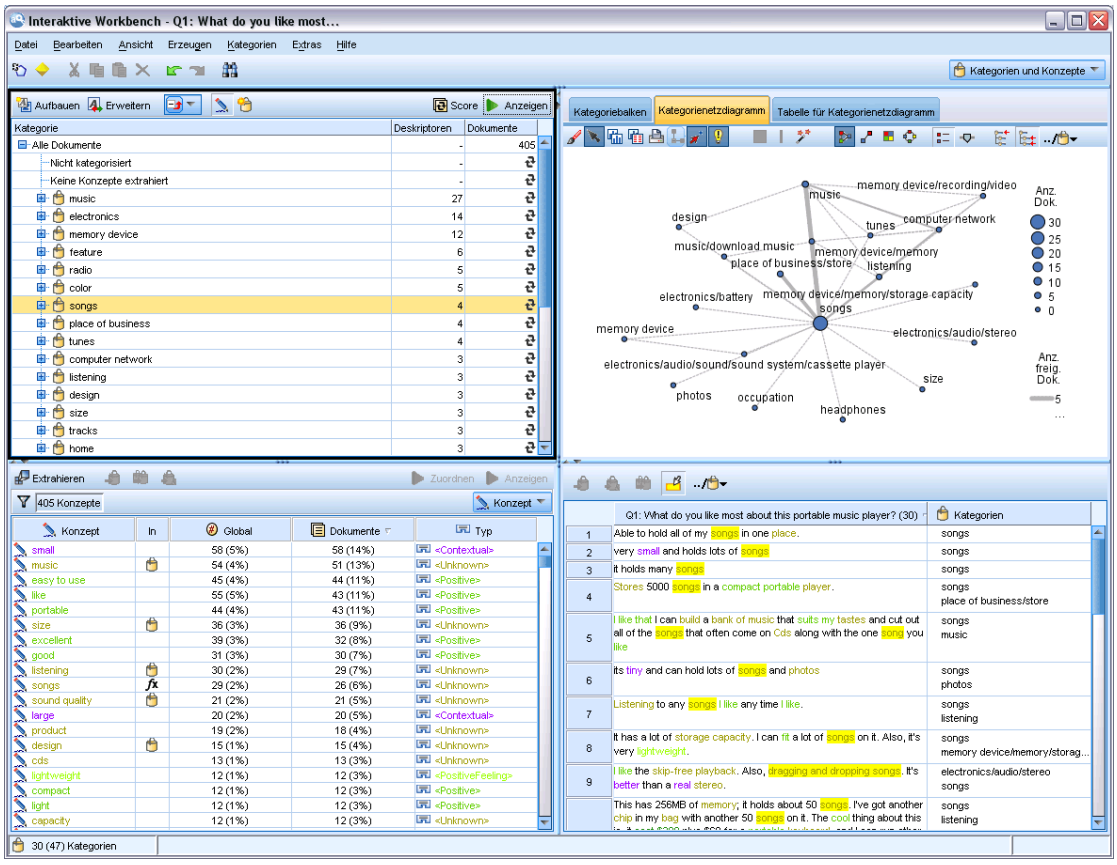
- **Text-Mining-Knoten (Registerkarte "Modell").** Als Nächstes haben wir auf der Registerkarte "Modell" die interaktive Erstellung eines Kategoriemodell-Nuggets und die Verwendung von Extrahierungsergebnissen zum Aufbauen von Kategorien gewählt. In diesem Beispiel haben wir eine Kopie von Ressourcen und ein Set von Kategorien aus einem Textanalysepaket geladen.

Abbildung 3-17
Dialogfeld des Text-Mining-Modellierungsknotens: Registerkarte "Modell"



- **Interaktive Workbench-Sitzung**. Als Nächstes haben wir den Stream ausgeführt. Die Benutzeroberfläche der interaktiven Workbench wurde angezeigt. Nach einer Extrahierung begannen wir, unsere Daten zu untersuchen und die Kategorien zu verbessern.

Abbildung 3-18
Interaktive Workbench-Sitzung



Text-Mining-Nugget: Konzeptmodell

Ein Text-Mining-Konzeptmodell-Nugget wird jedes Mal erstellt, wenn Sie einen Text-Mining-Modellknoten erfolgreich ausführen und dafür in der Registerkarte “Modell” die Option Modell direkt generieren ausgewählt haben. Ein Text-Mining-Konzeptmodell wird für die Echtzeitermittlung von Schlüsselkonzepten in anderen Textdaten, wie beispielsweise Notizen aus einem Callcenter, verwendet.

Das Konzeptmodell-Nugget selbst umfasst eine Liste von Konzepten, die Typen zugewiesen wurden. Sie können einige oder alle Konzepte in diesem Modell zum Scoring mit anderen Daten auswählen. Wenn Sie einen Stream ausführen, der ein Text-Mining-Modell-Nugget enthält, werden den Daten gemäß dem Aufbaumodus, der in der Registerkarte “Modell” des Text-Mining-Modellierungsknotens ausgewählt wurde, vor dem Aufbau des Modells neue Felder hinzugefügt. Für weitere Informationen siehe Thema Konzeptmodell: Registerkarte “Modell” auf S. 58.

Wenn das Modell-Nugget unter Verwendung von übersetzten Dokumenten generiert wurde, erfolgt das Scoring in der übersetzten Sprache. Umgekehrt können Sie, wenn das Modell-Nugget mit Englisch als Sprache generiert wurde, eine Übersetzungssprache im Modell-Nugget angeben, da die Dokumente anschließend ins Englische übersetzt werden.

Die Text-Mining-Modell-Nuggets werden nach der Erstellung in der Palette der Modell-Nuggets gespeichert (diese befindet sich rechts oben im IBM® SPSS® Modeler-Fenster auf der Registerkarte “Modelle”).

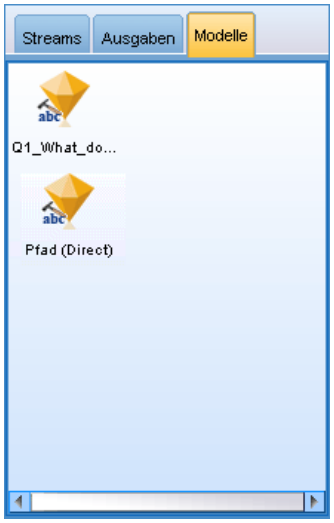
Anzeigen der Ergebnisse

Um Informationen zum Modell-Nugget anzuzeigen, klicken Sie mit der rechten Maustaste auf die Palette der Modell-Nuggets und wählen Sie im Kontextmenü die Option Durchsuchen (bzw. Bearbeiten bei Knoten in einem Stream) aus.

Hinzufügen von Modellen zu Streams

Um das Modell-Nugget zu Ihrem Stream hinzuzufügen, klicken Sie auf das Symbol in der Palette der Modell-Nuggets und dann auf den Stream-Zeichenbereich, in dem der Knoten platziert werden soll. Alternativ können Sie mit der rechten Maustaste auf das Symbol klicken. Wählen Sie im Kontextmenü Zu Stream hinzufügen. Verbinden Sie anschließend den Stream mit dem Knoten und Sie können die Daten weitergeben, um Prognosen zu erstellen.

Abbildung 3-19
Palette der Modell-Nuggets mit einem Text-Mining-Modell-Nugget

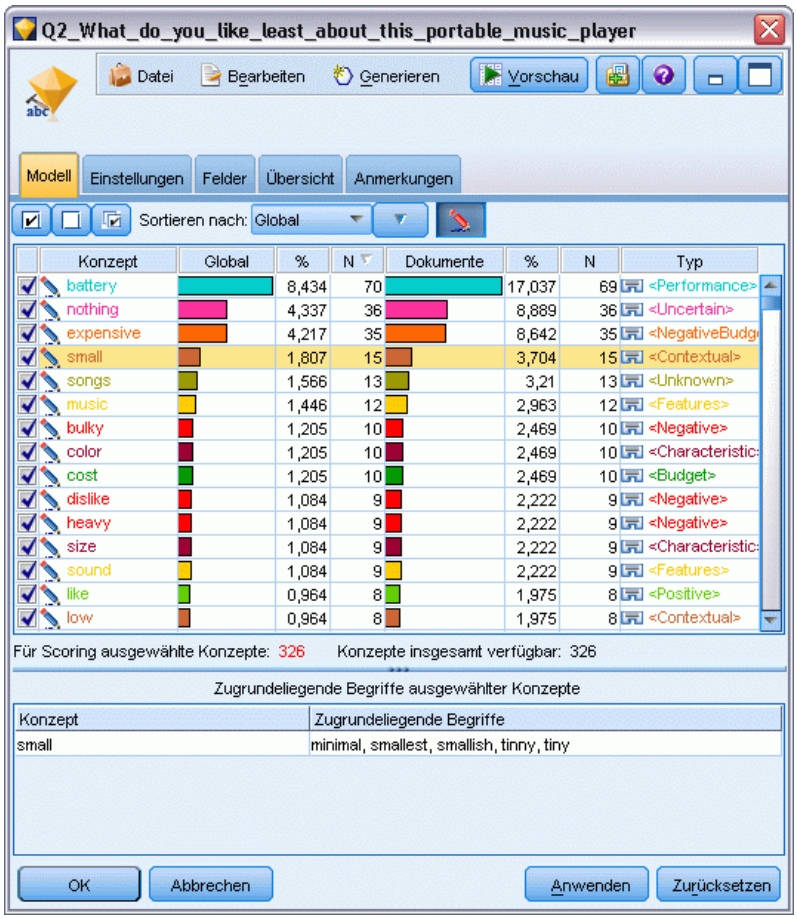


Konzeptmodell: Registerkarte “Modell”

Bei Konzeptmodellen werden auf der Registerkarte “Modell” die extrahierten Konzepte angezeigt. Die Konzepte werden als Tabelle dargestellt, mit einer Zeile für jedes Konzept. Diese Registerkarte ist zur Auswahl der Konzepte gedacht, die für das Scoring verwendet werden sollen.

Anmerkung: Wenn Sie stattdessen ein Kategoriemodell-Nugget generiert haben, enthält diese Registerkarte andere Informationen. Für weitere Informationen siehe Thema Kategoriemodell-Nugget: Registerkarte “Modell” auf S. 73.

Abbildung 3-20
Dialogfeld für Konzeptmodell-Nugget: Registerkarte "Modell"



Standardmäßig werden alle Konzepte für das Scoring ausgewählt, wie in den Kontrollkästchen in der äußersten linken Spalte gezeigt. Wenn das Kontrollkästchen aktiviert ist, wird das Konzept für das Scoring verwendet. Wenn das Kontrollkästchen nicht aktiviert ist, wird das Konzept vom Scoring ausgenommen. Sie können mehrere Zeilen aktivieren, indem Sie sie auswählen und auf eines der Kontrollkästchen in Ihrer Auswahl klicken.

Mehr über die einzelnen Konzepte können Sie aus den zusätzlichen Informationen erfahren, die in jeder der folgenden Spalten angegeben werden:

Konzept. Dies ist das extrahierte übergeordnete Wort bzw. der extrahierte übergeordnete Ausdruck. In einigen Fällen werden als Konzept der Konzeptname sowie einige weitere zugrundeliegende Ausdrücke, die mit diesem Konzept verbunden sind, angegeben. Um zu sehen, welche zugrundeliegende Ausdrücke zu einem Konzept gehören, zeigen Sie den Bereich "Zugrundeliegende Konzepte" in dieser Registerkarte an und wählen Sie das betreffende Konzept aus, um die entsprechenden Ausdrücke unten im Dialogfeld anzuzeigen. Für weitere Informationen siehe Thema Zugrundeliegende Ausdrücke in Konzeptmodellen auf S. 62.

Globalwert. Globalwert (Häufigkeit) bezieht sich in diesem Fall darauf, wie häufig die betreffenden Konzepte (und alle zugrundeliegenden Fachausdrücke) in sämtlichen Dokumenten/Datensätzen insgesamt vorkommen.

- **Balkendiagramm.** Die globale Häufigkeit des Konzepts in den Textdaten (als Balkendiagramm). Der Balken wird in der Farbe des Typs dargestellt, dem das Konzept zugeordnet ist, damit die Typen optisch voneinander unterschieden werden können.
- **%.** Die globale Häufigkeit dieses Konzepts in den Textdaten (als Prozentsatz).
- **N.** Die tatsächliche Anzahl der Vorkommnisse dieses Konzepts in den Textdaten.

Dokumente. In diesem Fall bezieht sich "Dokumente" auf die Dokumentanzahl, also die Zahl der Dokumente bzw. Datensätze, in denen das Konzept vorkommt (inklusive aller zugrundeliegenden Fachausdrücke).

- **Balkendiagramm.** Die Dokumentanzahl für dieses Konzept (als Balkendiagramm). Der Balken wird in der Farbe des Typs dargestellt, dem das Konzept zugeordnet ist, damit die Typen optisch voneinander unterschieden werden können.
- **%.** Die Dokumentanzahl für dieses Konzept (als Prozentsatz).
- **N.** Die tatsächliche Anzahl der Dokumente bzw. Datensätze, die dieses Konzept enthalten.

Typ. Der Typ, dem das Konzept zugewiesen ist. Die Spalten Globalwert und Dokumente werden in Farbe angezeigt, um den Typ anzugeben, dem das jeweilige Konzept zugewiesen ist. Bei einem **Typ** handelt es sich um Konzepte, die nach semantischen Gesichtspunkten gruppiert werden. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

Arbeiten mit Konzepten

Wenn Sie mit der rechten Maustaste auf eine Zelle in der Tabelle klicken, wird ein Kontextmenü angezeigt, in dem folgende Optionen zur Auswahl stehen:

- **Alles auswählen.** Alle Zeilen in der Tabelle werden ausgewählt.
- **Kopieren.** Die ausgewählten Konzepte werden in die Zwischenablage kopiert.
- **Mit Feldern kopieren** Die ausgewählten Konzepte werden zusammen mit der Spaltenüberschrift in die Zwischenablage kopiert.
- **Ausgewählte Elemente überprüfen.** Aktiviert alle Kontrollkästchen für die ausgewählten Zeilen in der Tabelle und berücksichtigt dabei die Konzepte für das Scoring.
- **Markierung der ausgewählten Elemente aufheben.** Deaktiviert alle Kontrollkästchen für die ausgewählten Zeilen in der Tabelle.
- **Alles markieren.** Aktiviert alle Kontrollkästchen in der Tabelle. Dadurch werden alle Konzepte in der Endausgabe verwendet.
- **Alle Markierungen aufheben.** Deaktiviert alle Kontrollkästchen in der Tabelle. Wenn die Markierung für ein Konzept aufgehoben wird, wird dieses nicht in der Endausgabe verwendet.
- **Konzepte einschließen.** Öffnet das Dialogfeld "Konzepte einschließen". Für weitere Informationen siehe Thema Optionen zum Einschluss von Konzepten für das Scoring auf S. 61.

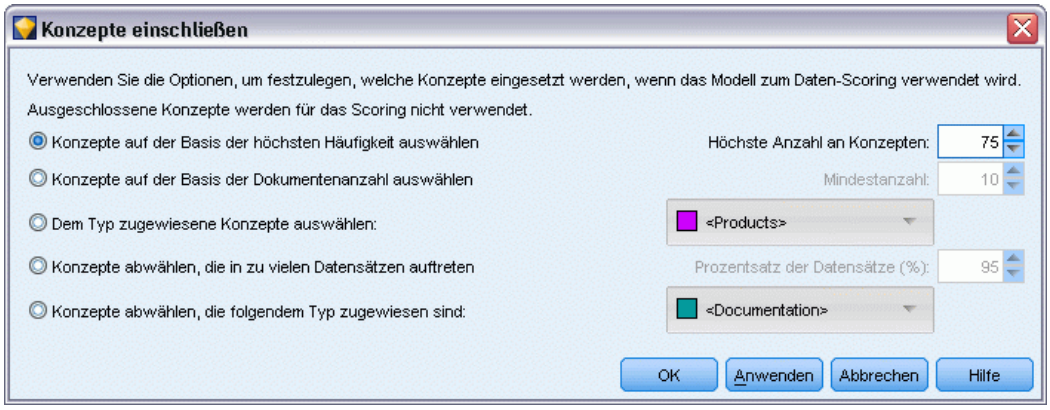
Optionen zum Einschluss von Konzepten für das Scoring

Um rasch die Konzepte zu aktivieren oder zu deaktivieren, die zum Scoring verwendet werden, klicken Sie auf das Symbolleistenfeld für **Konzepte einschließen.**.

Abbildung 3-21
Symbolleistenfeld "Konzepte einschließen"

Wenn Sie auf dieses Symbolleistenfeld klicken, wird das Dialogfeld "Konzepte einschließen" geöffnet, in dem Sie Konzepte auf der Grundlage von Regeln auswählen können. Alle Konzepte, die auf der Registerkarte "Modell" aktiviert sind, werden in das Scoring mit einbezogen. Wenden Sie in diesem untergeordneten Dialogfeld eine Regel an, um zu ändern, welche Konzepte für das Scoring verwendet werden.

Abbildung 3-22
Dialogfeld "Konzepte einschließen"



Sie können folgende Optionen auswählen:

Konzepte nach größter Häufigkeit markieren. Größte Anzahl an Konzepten. Gibt – ausgehend von dem Konzept mit der größten Häufigkeit – die Anzahl der zu markierenden Konzepte an. Häufigkeit bezieht sich in diesem Fall darauf, wie häufig die betreffenden Konzepte (und alle zugrundeliegenden Ausdrücke) in sämtlichen Dokumenten/Datensätzen insgesamt auftauchen. Dieser Wert kann höher sein, als der Datensatzwert, da ein Konzept mehrmals in einem Datensatz vorkommen kann.

Konzepte basierend auf Dokumentanzahl markieren. Mindestzahl. Dies ist die geringste Dokumentanzahl, die für die zu markierenden Konzepte erforderlich ist. In diesem Fall bezieht sich Dokumentanzahl auf die Zahl der Dokumente/Datensätze, in denen das Konzept vorkommt (inklusive aller zugrundeliegenden Fachausdrücke).

Aktivieren Sie die Konzepte, die dem Typ zugewiesen werden. Wählen Sie in der Dropdown-Liste einen Typ aus, um alle Konzepte zu markieren, die diesem Typ zugewiesen sind. Konzepte werden den Typen automatisch während des Extrahierungsprozesses zugewiesen. Bei einem **Typ** handelt es sich um Konzepte, die nach semantischen Gesichtspunkten gruppiert werden. Zu den Typen gehören beispielsweise übergeordnete Konzepte, positive und negative Wörter und

Bezeichnungen des Grades sowie des Kontexts, Vornamen, Orte, Organisationen und anderes mehr. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

Markierung von Konzepten aufheben, die in zu vielen Datensätzen auftreten. Prozentsatz der Datensätze. Hebt die Markierung von Konzepten auf, deren Datensatzwert in Prozent höher ist als die von Ihnen angegebene Zahl. Diese Option ist nützlich, um Konzepte auszuschließen, die häufig im Text oder in jedem Datensatz vorkommen, aber keine Bedeutung für die Analyse haben.

Deaktivieren Sie die Konzepte, die dem Typ zugewiesen werden. Hebt die Markierung von Konzepten auf, die mit dem in der Dropdown-Liste ausgewählten Typ übereinstimmen.

Zugrundeliegende Ausdrücke in Konzeptmodellen

Sie können die zugrundeliegenden Ausdrücke anzeigen, die für die in der Tabelle ausgewählten Konzepte definiert sind. Sie können die Tabelle für zugrundeliegende Ausdrücke in einem geteilten Fensterbereich unten im Bildschirm anzeigen, indem Sie auf die Umschaltfläche für zugrundeliegende Ausdrücke auf der Symbolleiste klicken.

Diese zugrundeliegenden Ausdrücke umfassen die Synonyme, die in den linguistischen Ressourcen definiert sind (unabhängig davo, ob sie im Text gefunden wurden) sowie extrahierte Plural-/Singularformen, die im Text zur Gernierung des Modell-Nuggets gefunden wurden, permutierte Ausdrücke, Ausdrücke aus unscharfen Gruppierungen usw.

Abbildung 3-23
Symbolleistenfeld "Zugrundeliegende Ausdrücke anzeigen"

Anmerkung: Die Liste der zugrundeliegenden Ausdrücke kann nicht geändert werden. Diese Liste wird durch Substitutionen, Synonymdefinitionen (im Substitutionswörterbuch), Fuzzy-Gruppierung u. a. erzeugt—Verfahren, die sämtlich in den linguistischen Ressourcen definiert sind. Nehmen Sie Änderungen an der Gruppierung von Ausdrücken unter einem Konzept oder deren Behandlung direkt in den Ressourcen vor (Bearbeitung ist im Resource Editor in der interaktiven Workbench bzw. im Template Editor möglich, anschließend erneutes Laden in den Knoten). Führen Sie dann den Stream erneut aus, um ein neues Modell-Nugget mit den aktualisierten Ergebnissen zu erhalten.

Wenn Sie mit der rechten Maustaste auf die Zelle eines zugrundeliegenden Ausdrucks oder Konzepts klicken, wird ein Kontextmenü angezeigt, in dem folgende Optionen zur Auswahl stehen:

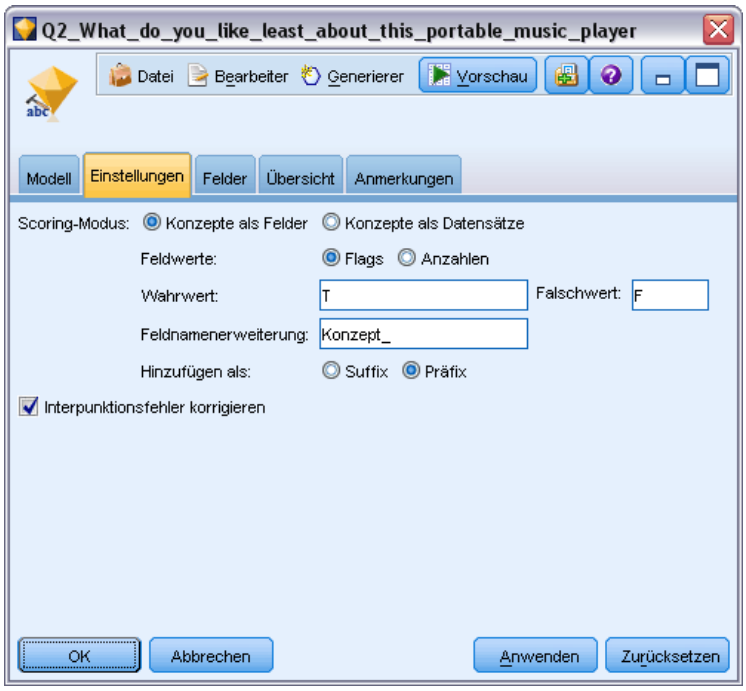
- **Kopieren.** Die ausgewählte Zelle wird in die Zwischenablage kopiert.
- **Mit Feldern kopieren.** Die ausgewählte Zelle wird zusammen mit den Spaltenüberschriften in die Zwischenablage kopiert.
- **Alles auswählen.** Alle Zellen in der Tabelle werden ausgewählt.

Konzeptmodell: Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" wird verwendet, um den Wert des Textfelds für die neuen Eingangsdaten zu definieren (sofern erforderlich). Außerdem können Sie dort das Datenmodell für die Ausgabe definieren (Scoring-Modus).

Anmerkung: Diese Registerkarte wird nur angezeigt, wenn das Modell-Nugget auf der Zeichenfläche platziert wird. Sie ist nicht vorhanden, wenn Sie direkt in der Modellpalette auf dieses Dialogfeld zugreifen.

Abbildung 3-24
Dialogfeld des Text-Mining-Konzeptmodell-Nuggets: Registerkarte "Einstellungen"



Scoring-Modus: Konzepte als Datensätze

Mit diesem Scoring-Modus wird für jedes concept/document-Paar ein neuer Datensatz erstellt. Normalerweise gibt es mehr Datensätze in der Ausgabe, als in der Eingabe vorhanden waren.

Neben den Eingabefeldern werden den Daten die folgenden neuen Felder hinzugefügt:

Tabelle 3-1
Ausgabefelder für "Konzepte als Datensätze"

Feld	Beschreibung
Concept	Enthält den im Textdatenfeld gefundenen Namen des extrahierten Konzepts.
Type	Speichert den Konzepttyp als vollständigen Typnamen, beispielsweise *Location* oder *Person*. Bei einem **Typ** handelt es sich um Konzepte, die nach semantischen Gesichtspunkten gruppiert werden. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.
Count	Zeigt an, wie häufig dieses Konzept (inklusive seiner zugrundeliegenden Fachausdrücke) im Textkörper (Datensatz/Dokument) vorkommt.

Wenn Sie diese Option aktivieren, werden alle anderen Optionen mit Ausnahme von Interpunktionsfehler korrigieren deaktiviert.

Scoring-Modus: Konzepte als Felder

Bei Konzeptmodellen wird für jeden Eingabedatensatz ein neuer Datensatz für jedes Konzept erstellt, das in einem gegebenen Dokument ermittelt wird. Daher gibt es in der Ausgabe genauso viele Datensätze wie in der Eingabe. Allerdings enthält jetzt jeder Datensatz (Zeile) ein weiteres Feld (Spalte) für jedes auf der Registerkarte "Modell" ausgewählte Konzept bzw. für jede ausgewählte Kategorie (Auswahl über Kontrollkästchen). Der Wert für jedes Konzeptfeld hängt davon ab, ob Sie auf dieser Registerkarte Flags oder Anzahl als Feldwert auswählen.

Feldwerte. Wählen Sie, ob das neue Feld für jedes Konzept eine Anzahl oder einen Flag-Wert enthalten soll.

- **Flags.** Diese Option wird verwendet, um Flags mit zwei verschiedenen Werten in der Ausgabe zu erhalten, beispielsweise *Ja*/*Nein*, *Wahr*/*Falsch*, *W*/*F* oder *1* und *2*. Die Speichertypen werden automatisch so eingestellt, dass sie die gewählten Werte reflektieren. Wenn Sie beispielsweise numerische Werte für die Flags eingeben, werden sie automatisch als Ganzzahl behandelt. Als Speichertypen für Flags sind "Zeichenkette", "Ganze Zahl", "Reelle Zahl" und "Datum/Uhrzeit" möglich. Geben Sie einen Flag-Wert für Wahr und für Falsch ein.
- **Anzahl.** Wird verwendet, um die Häufigkeit eines Konzepts in einem bestimmten Datensatz zu erhalten.

Feldnamenerweiterung. Dient zur Angabe einer Erweiterung für den Feldnamen. Feldnamen werden unter Verwendung des Konzeptnamens und dieser Erweiterung generiert.

- **Hinzufügen als.** Gibt an, an welcher Stelle die Erweiterung zum Feldnamen hinzugefügt werden soll. Wählen Sie Präfix, um die Erweiterung am Anfang der Zeichenkette einzufügen. Wählen Sie Suffix, um die Erweiterung am Ende der Zeichenkette einzufügen.

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Punktuationsfehlern (zum Beispiel ungeeignete Verwendung) während der Extrahierung, um die Extrahierbarkeit von Konzepten zu verbessern. Diese Option ist besonders nützlich bei kurzem Text und niedriger Textqualität (wie dies beispielsweise bei offenen Antworten bei Umfragen, E-Mails und CRM-Daten der Fall ist) oder wenn der Text viele Abkürzungen enthält.

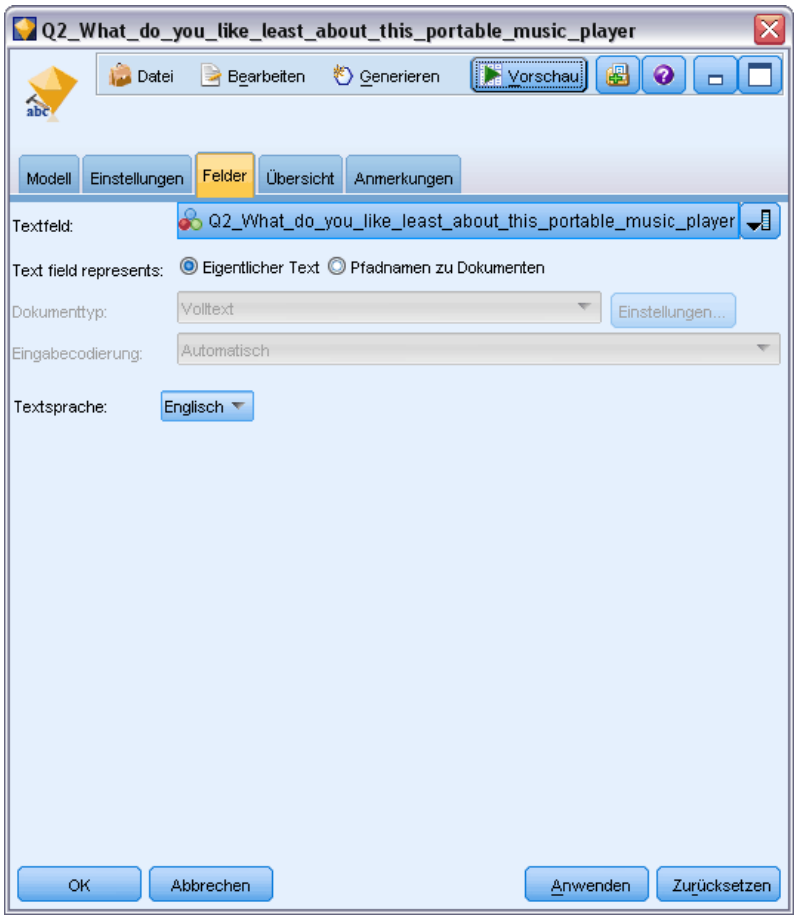
Anmerkung: Die Option Interpunktionsfehler korrigieren gilt nicht beim Arbeiten mit japanischem Text.

Konzeptmodell: Registerkarte "Felder"

Die Registerkarte "Felder" wird verwendet, um den Wert des Textfelds für die neuen Eingangsdaten zu definieren (sofern erforderlich).

Anmerkung: Diese Registerkarte wird nur angezeigt, wenn das Modell-Nugget im Stream platziert wird. Sie ist nicht vorhanden, wenn Sie direkt in der Modellpalette auf diese Ausgabe zugreifen.

Abbildung 3-25
Dialogfeld des Text-Mining-Konzeptmodell-Nuggets: Registerkarte "Felder"



Textfeld. Wählen Sie das Feld, das den Text für das Mining enthält, den Dokumentenpfadnamen bzw. den Verzeichnispfadnamen, der zu den Dokumenten führt. Dieses Feld hängt von der Datenquelle ab.

Textfeld enthält. Geben Sie an, was das Textfeld enthalten soll, das in der vorausgehenden Einstellung festgelegt wurde. Folgende Optionen stehen zur Auswahl:

- **Eigentlicher Text.** Wählen Sie diese Option aus, wenn das Feld genau den Text enthält, aus dem Konzepte extrahiert werden sollen.
- **Pfadnamen zu Dokumenten.** Wählen Sie diese Option aus, wenn das Feld mindestens einen Pfadnamen zu dem oder den Speicherort(en) der Textdokumente enthält.

Dokumenttyp. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält. Der Dokumenttyp gibt die Struktur des Texts an. Wählen Sie einen der folgenden Typen aus:

- **Volltext.** Verwenden Sie diese Option für die meisten Dokumente bzw. Textquellen. Die gesamte Textmenge wird für die Extrahierung gescannt. Im Gegensatz zu den anderen Optionen gibt es keine weiteren Einstellungen für diese Option.

- **Gegliederter Text.** Verwenden Sie diese Option für bibliografische Formulare, Patente und alle Dateien, die reguläre Strukturen enthalten, die identifiziert und analysiert werden können. Dieser Dokumenttyp wird verwendet, um den gesamten Extrahierungsvorgang oder Teile des Extrahierungsvorgangs zu überspringen. Er ermöglicht das Definieren von Trennzeichen zu Fachausdrücken, das Zuweisen von Typen und das Festlegen eines minimalen Häufigkeitswerts. Wenn Sie diese Option aktivieren, müssen Sie auf die Schaltfläche Einstellungen klicken und im Bereich Formatierung als gegliederter Text des Dialogfelds "Dokumenteinstellungen" Texttrennzeichen eingeben. Für weitere Informationen siehe Thema Dokumenteinstellungen der Registerkarte "Felder" auf S. 35.
- **XML** Verwenden Sie diese Option, um die XML-Tags anzugeben, die den zu extrahierenden Text enthalten. Alle anderen Tags werden ignoriert. Wenn Sie diese Option aktivieren, müssen Sie auf die Schaltfläche Einstellungen klicken und im Bereich Formatierung als XML des Dialogfelds "Dokumenteinstellungen" explizit die XML-Elemente angeben, die den Text enthalten, der während des Extrahierungsprozesses gelesen werden soll. Für weitere Informationen siehe Thema Dokumenteinstellungen der Registerkarte "Felder" auf S. 35.

Eingabekodierung. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält. Es bestimmt die Standardtextkodierung. Für alle Sprachen außer Japanisch erfolgt eine Konvertierung von der angegebenen bzw. erkannten Kodierung in ISO-8859-1. Selbst wenn Sie also eine andere Kodierung angeben, wird diese vor der Verarbeitung von der Extrahierungsengine in iso-8859-1 konvertiert. Alle Zeichen, die nicht in die ISO-8859-1-Kodierungsdefinition passen, werden in Leerzeichen umgewandelt. Für japanischen Text können Sie eine von mehreren Kodierungsoptionen wählen: SHIFT_JIS, EUC_JP, UTF-8 oder ISO-2022-JP.

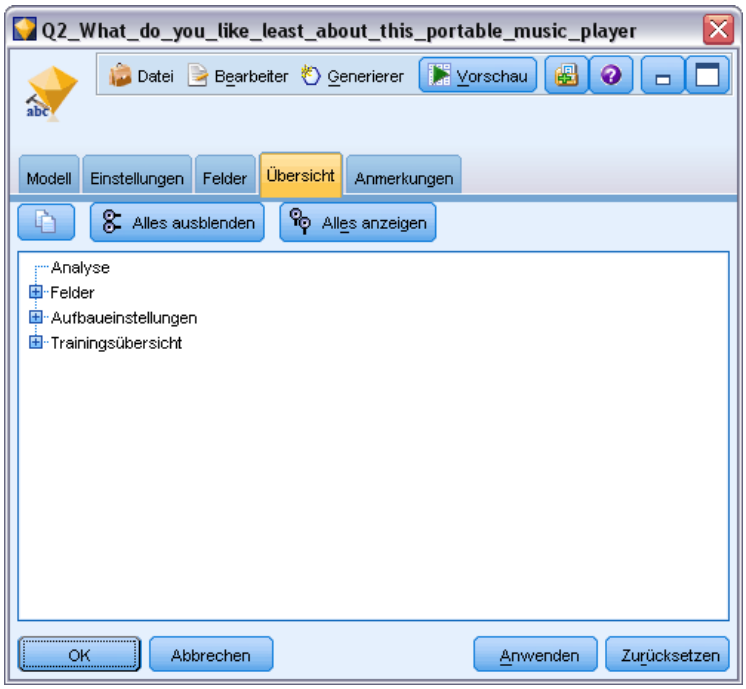
Textsprache. Gibt die Sprache des Texts an, für das das Mining erfolgt. Hierbei handelt es sich um die Hauptsprache, die während der Extraktion erkannt wird. Wenden Sie sich an Ihren Kundendienstmitarbeiter, wenn Sie eine Lizenz für eine unterstützte Sprache erwerben möchten, auf die Sie zurzeit keinen Zugriff haben.

Konzeptmodell: Registerkarte "Übersicht"

Die Registerkarte "Übersicht" bietet Informationen zum Modell selbst (Ordner *Analyse* folder), zu den im Modell verwendeten Feldern (Ordner *Felder*), zu den beim Erstellen des Modells verwendeten Einstellungen (Ordner *Aufbaueinstellungen*) und zum Modelltraining (Ordner *Trainingsübersicht*).

Beim ersten Durchsuchen eines Modellierungsknotens sind die Ordner auf der Registerkarte "Übersicht" minimiert. Um die für Sie relevanten Ergebnisse anzuzeigen, vergrößern Sie die gewünschten Ergebnisse mithilfe des Erweiterungssteuerelements links neben dem Ordner oder klicken Sie auf die Schaltfläche Alles anzeigen, um alle Ergebnisse anzuzeigen. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement den gewünschten Ordner reduzieren. Alternativ können Sie mit der Schaltfläche Alles ausblenden alle Ordner ausblenden.

Abbildung 3-26
Dialogfeld des Text-Mining-Modell-Nuggets: Registerkarte "Übersicht"



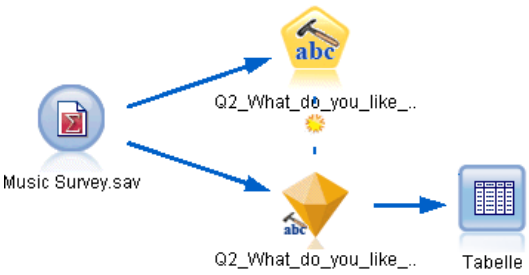
Verwenden von Konzeptmodell-Nuggets in einem Stream

Beim Verwenden eines Text-Mining-Modellierungsknotens können Sie ein Konzeptmodell-Nugget oder ein Kategoriemodell-Nugget (durch eine interaktive Workbench-Sitzung) generieren. Das folgende Beispiel zeigt die Verwendung eines Konzeptmodells in einem einfachen Stream.

Beispiel: Statistikdatei-Knoten mit Konzeptmodell-Nugget

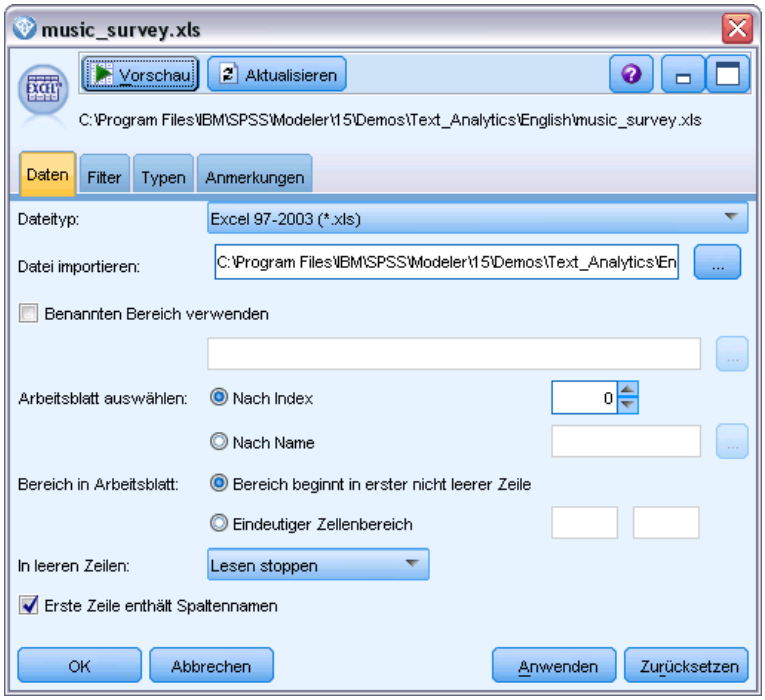
Das folgende Beispiel zeigt die Verwendung des Text-Mining-Konzeptmodell-Nuggets.

Abbildung 3-27
Beispiel-Stream: Statistikdateiknoten mit einem Text-Mining-Konzeptmodell-Nugget



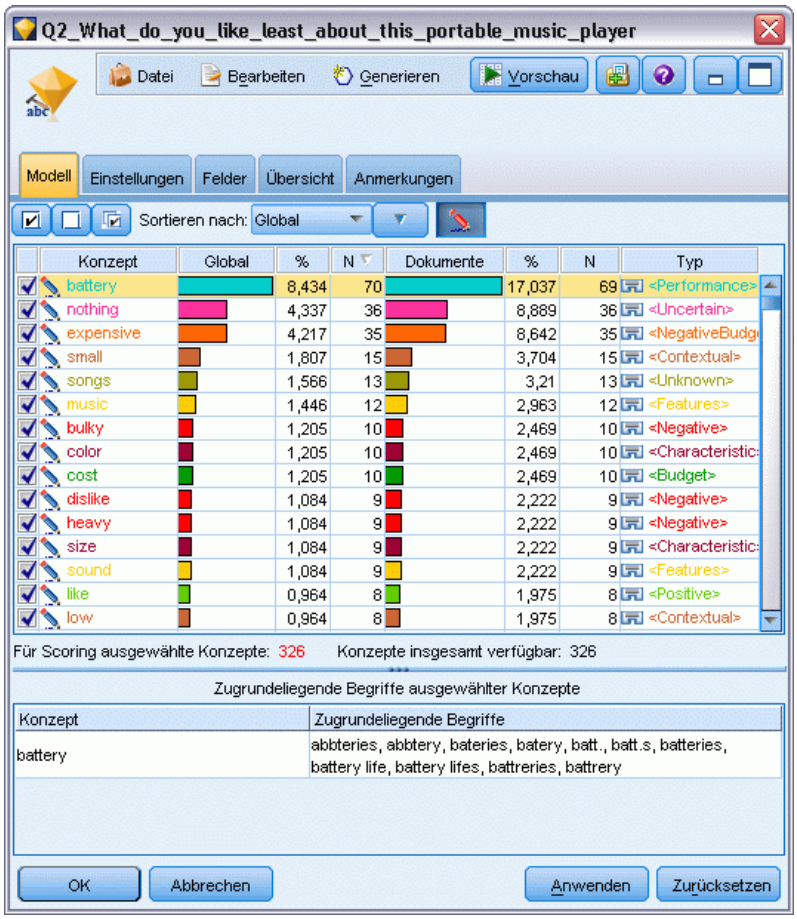
- **Statistikdateiknoten (Registerkarte "Einstellungen")** Zuerst fügten wir diesen Knoten dem Stream hinzu, um anzugeben, wo die Textdokumente gespeichert sind.

Abbildung 3-28
Dialogfeld des Knotens für Statistikdateien: Registerkarte "Daten"



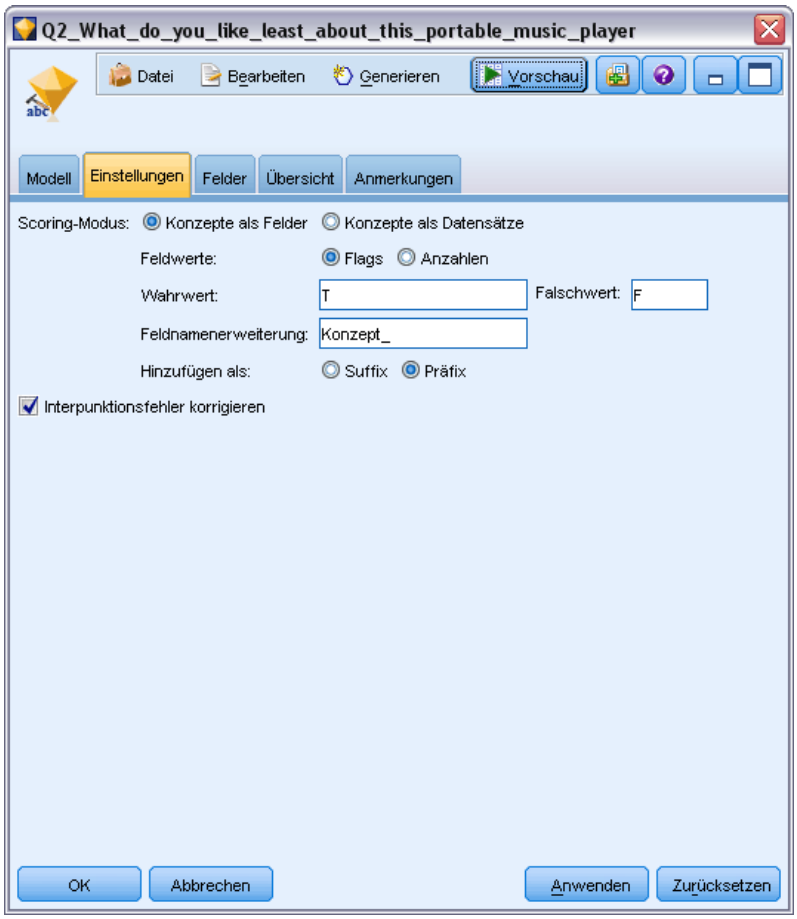
► **Text-Mining-Konzeptmodell-Nugget (Registerkarte "Modell").** Als Nächstes haben wir dem Dateilistenknoten ein Konzeptmodell-Nugget hinzugefügt und dieses mit dem Statistikdateiknoten verbunden. Wir haben die Konzepte ausgewählt, die wir für das Scoring unserer Daten verwenden wollten.

Abbildung 3-29
Dialogfeld des Text-Mining-Modell-Nuggets: Registerkarte "Modell"



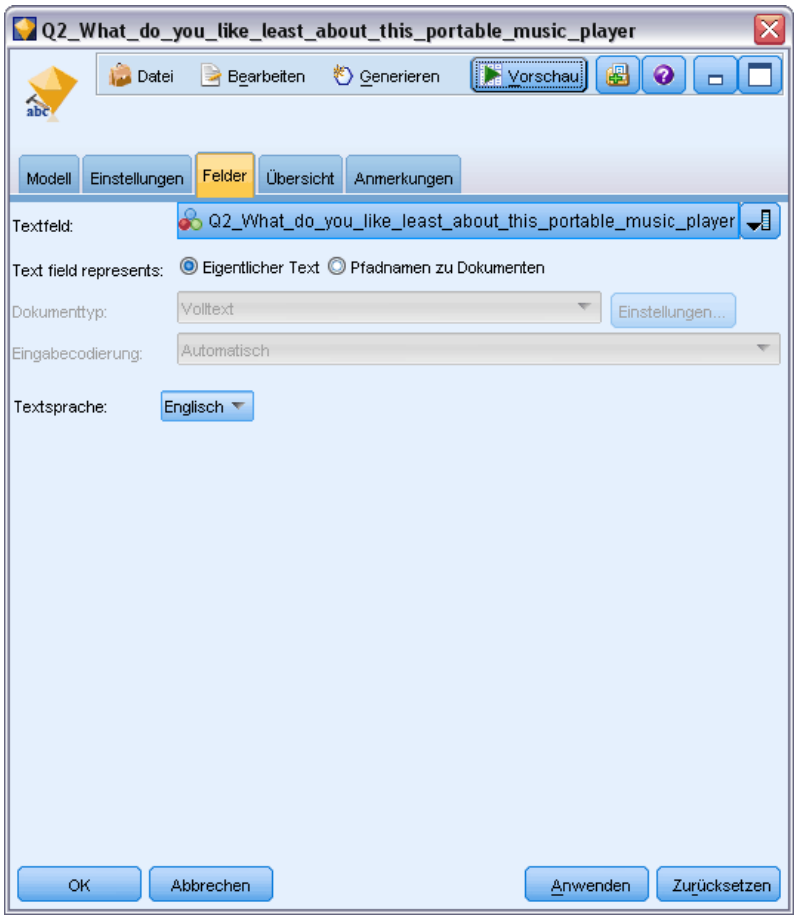
- **Text-Mining-Konzeptmodell-Nugget (Registerkarte "Einstellungen").** Als Nächstes haben wir das Ausgabeformat definiert und *Konzepte als Felder* ausgewählt. In der Ausgabe wird ein neues Feld für jedes Konzept erstellt, das in der Registerkarte "Modell" ausgewählt wurde. Jeder Feldname wird aus dem Konzeptnamen und dem Präfix "*Concept_*" gebildet.

Abbildung 3-30
Dialogfeld des Text-Mining-Konzeptmodell-Nuggets: Registerkarte "Einstellungen"



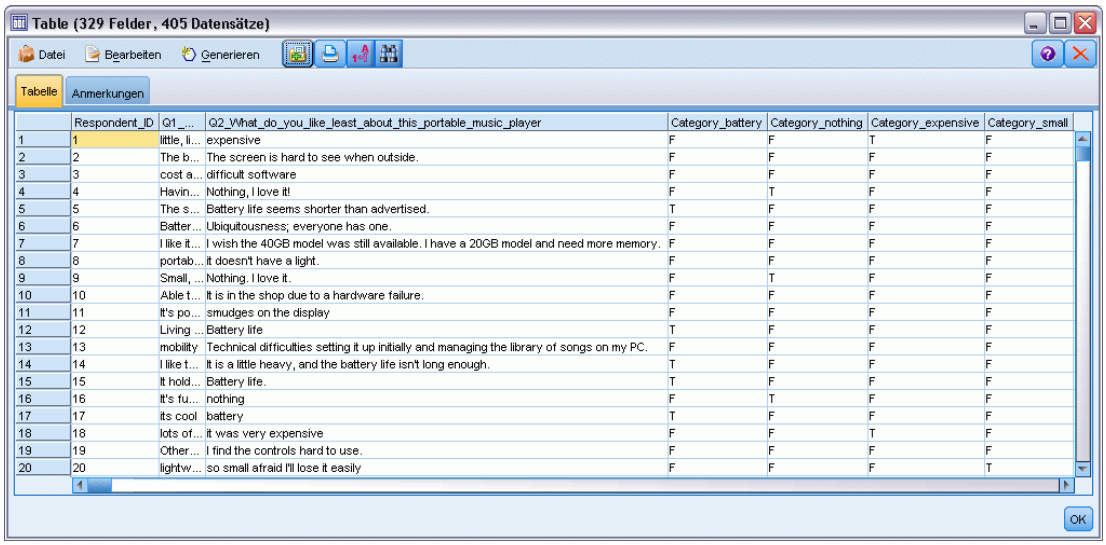
- **Text-Mining-Konzeptmodell-Nugget (Registerkarte "Felder").** Als Nächstes haben wir das Textfeld Q2_What_do_you_like_least_about_this_portable_music_player ausgewählt, das der Feldname aus dem Statistikdateiknoten ist. Wir haben auch die Option Textfeld enthält: tatsächlichen Text ausgewählt.

Abbildung 3-31
Dialogfeld des Text-Mining-Konzeptmodell-Nuggets: Registerkarte "Felder"



- **Tabellenknoten.** Als Nächstes fügten wir einen Tabellenknoten hinzu, um die Ergebnisse zu prüfen. Anschließend führten wir den Stream aus. Die Tabellenausgabe wird am Bildschirm geöffnet.

Abbildung 3-32
Nach unten abgerollte Tabellenausgabe, um die Konzept-Flags zu zeigen



	Respondent_ID	Q1_...	Q2_What_do_you_like_least_about_this_portable_music_player	Category_battery	Category_nothing	Category_expensive	Category_small
1	1	little, li...	expensive	F	F	T	F
2	2	The b...	The screen is hard to see when outside.	F	F	F	F
3	3	cost a...	difficult software	F	F	F	F
4	4	Havin...	Nothing, I love it!	F	T	F	F
5	5	The s...	Battery life seems shorter than advertised.	T	F	F	F
6	6	Batter...	Ubiquitousness; everyone has one.	F	F	F	F
7	7	I like it...	I wish the 40GB model was still available. I have a 20GB model and need more memory.	F	F	F	F
8	8	portab...	it doesn't have a light.	F	F	F	F
9	9	Small, ...	Nothing. I love it.	F	T	F	F
10	10	Able t...	It is in the shop due to a hardware failure.	F	F	F	F
11	11	It's po...	smudges on the display	F	F	F	F
12	12	Living ...	Battery life	T	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
14	14	I like t...	It is a little heavy, and the battery life isn't long enough.	T	F	F	F
15	15	It hold...	Battery life.	T	F	F	F
16	16	It's fu...	nothing	F	T	F	F
17	17	its cool	battery	T	F	F	F
18	18	lots of...	it was very expensive	F	F	T	F
19	19	Other...	I find the controls hard to use.	F	F	F	F
20	20	lightw...	so small afraid I'll lose it easily	F	F	F	T

Text-Mining-Nugget: Kategoriemodell

Ein Text-Mining-Kategoriemodell-Nugget wird jedes Mal erstellt, wenn Sie ein Kategoriemodell in der interaktiven Workbench generieren. Dieses Modellierungsnugget enthält ein Set an Kategorien, deren Definition aus Konzepten, Typen, TLA-Mustern und/oder Kategorieregeln besteht. Das Nugget wird verwendet, um Umfrageergebnisse, Blog-Einträge, andere Web-Feeds und sonstige Textdaten zu kategorisieren.

Wenn Sie im Modellierungsknoten eine interaktive Workbench-Sitzung starten, können Sie die Extrahierungsergebnisse untersuchen, die Ressourcen verfeinern und eine Feinabstimmung Ihrer Kategorien vornehmen, bevor Sie Kategoriemodelle erstellen. Wenn Sie einen Stream ausführen, der ein Text-Mining-Modell-Nugget enthält, werden den Daten gemäß dem Aufbaumodus, der in der Registerkarte “Modell” des Text-Mining-Modellierungsknotens ausgewählt wurde, vor dem Aufbau des Modells neue Felder hinzugefügt. Für weitere Informationen siehe Thema Kategoriemodell-Nugget: Registerkarte “Modell” auf S. 73.

Wenn das Modell-Nugget unter Verwendung von übersetzten Dokumenten generiert wurde, erfolgt das Scoring in der übersetzten Sprache. Umgekehrt können Sie, wenn das Modell-Nugget mit Englisch als Sprache generiert wurde, eine Übersetzungssprache im Modell-Nugget angeben, da die Dokumente anschließend ins Englische übersetzt werden.

Die Text-Mining-Modell-Nuggets werden nach der Erstellung in der Palette der Modell-Nuggets gespeichert (diese befindet sich rechts oben im IBM® SPSS® Modeler-Fenster auf der Registerkarte “Modelle”).

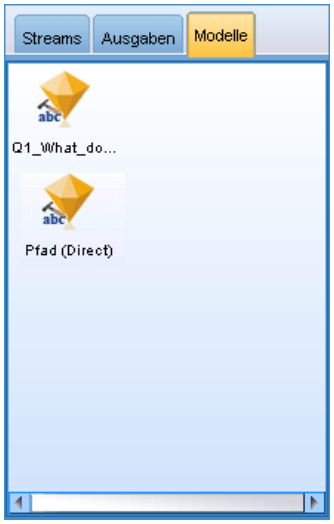
Anzeigen der Ergebnisse

Um Informationen zum Modell-Nugget anzuzeigen, klicken Sie mit der rechten Maustaste auf die Palette der Modell-Nuggets und wählen Sie im Kontextmenü die Option Durchsuchen (bzw. Bearbeiten bei Knoten in einem Stream) aus.

Hinzufügen von Modellen zu Streams

Um das Modell-Nugget zu Ihrem Stream hinzuzufügen, klicken Sie auf das Symbol in der Palette der Modell-Nuggets und dann auf den Stream-Zeichenbereich, in dem der Knoten platziert werden soll. Alternativ können Sie mit der rechten Maustaste auf das Symbol klicken. Wählen Sie im Kontextmenü Zu Stream hinzufügen. Verbinden Sie anschließend den Stream mit dem Knoten und Sie können die Daten weitergeben, um Prognosen zu erstellen.

Abbildung 3-33
Palette der Modell-Nuggets mit einem Text-Mining-Modell-Nugget



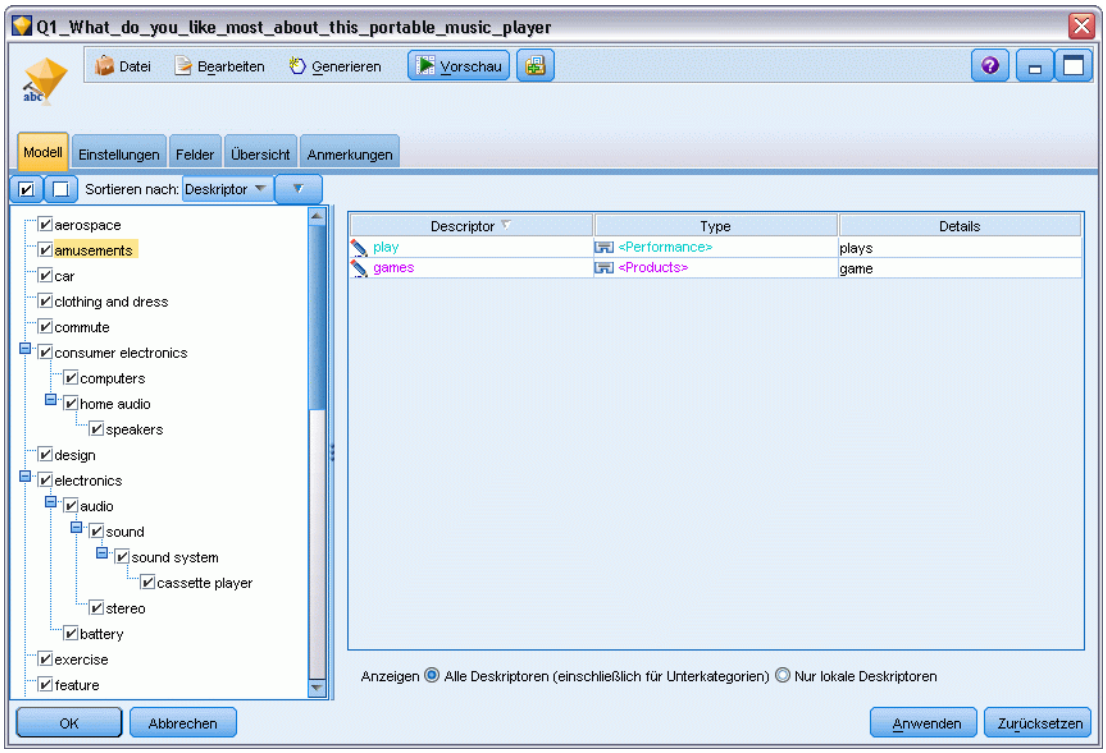
Kategoriemodell-Nugget: Registerkarte "Modell"

Bei Kategoriemodellen wird die Liste der Kategorien im Kategoriemodell auf der linken Seite der Registerkarte "Modell" angezeigt. Die Deskriptoren für eine ausgewählte Kategorie sind rechts zu sehen. Jede Kategorie besteht aus einer Anzahl von Deskriptoren. Für jede ausgewählte Kategorie werden die der Kategorie zugewiesenen Deskriptoren in der Tabelle angezeigt. Zu diesen Deskriptoren gehören u. a. Konzepte, Kategorieregeln, Typen und TLA-Muster. Ebenso wird die Art jedes Deskriptors angezeigt sowie einige Beispiele, die zeigen, wofür der betreffende Deskriptor steht.

Diese Registerkarte ist zur Auswahl der Kategorien gedacht, die für das Scoring verwendet werden sollen. Für ein Kategoriemodell werden Dokumente und Datensätze über das Scoring bestimmten Kategorien zugewiesen. Wenn der Text eines Dokuments bzw. Datensatzes einen oder mehrere der Deskriptoren oder zugrundeliegende Ausdrücke enthält, wird das Dokument bzw. der Datensatz der Kategorie zugewiesen, welcher der Deskriptor angehört. Diese zugrundeliegenden Ausdrücke umfassen die Synonyme, die in den linguistischen Ressourcen definiert sind (unabhängig davon, ob sie im Text gefunden wurden), sowie extrahierte Plural-/Singularausdrücke, die im Text zur Generierung des Modell-Nuggets gefunden wurden, permutierte Ausdrücke, Ausdrücke aus unscharfen Gruppierungen usw.

Anmerkung: Wenn Sie stattdessen ein Konzeptmodell-Nugget generiert haben, enthält diese Registerkarte andere Ergebnisse. Für weitere Informationen siehe Thema Konzeptmodell: Registerkarte "Modell" auf S. 58.

Abbildung 3-34
Dialogfeld für Kategoriemodell-Nugget: Registerkarte "Modell"



Kategoriebaum

Weitere Informationen zu den einzelnen Kategorien erhalten Sie, wenn Sie eine Kategorie auswählen und die Informationen prüfen, die für die Deskriptoren in der betreffenden Kategorie angezeigt werden. Zu jedem Deskriptor finden Sie die folgenden Informationen:

- **Deskriptor**name. Dieses Feld enthält ein Symbol, das anzeigt, um welche Art von Deskriptor es sich handelt, sowie den Namen des Deskriptors.

Konzepte	TLA-Muster
Typen	Kategorienregeln

- **Typ.** Dieses Feld enthält den Typennamen des Deskriptors. Bei Typen handelt es sich um Sammlungen von ähnlichen Konzepten (Gruppierungen nach semantischen Gesichtspunkten), z. B. Namen von Organisationen, Produkte oder positive Meinungen. Den Typen werden keine Regeln zugewiesen.
- **Details.** In diesem Feld ist aufgelistet, was zu dem betreffenden Deskriptor gehört. In Abhängigkeit von der Anzahl der Übereinstimmungen wird gegebenenfalls nicht für jeden Deskriptor die gesamte Liste angezeigt, da die Größe des Dialogfelds beschränkt ist.

Auswählen und Kopieren von Kategorien

Standardmäßig werden alle Kategorien der ersten Ebene für das Scoring ausgewählt, wie in den Kontrollkästchen im linken Fensterbereich angezeigt. Wenn das Kontrollkästchen aktiviert ist, wird die Kategorie für das Scoring verwendet. Wenn das Kontrollkästchen nicht aktiviert ist, wird die Kategorie vom Scoring ausgenommen. Sie können mehrere Zeilen aktivieren, indem Sie sie auswählen und auf eines der Kontrollkästchen in Ihrer Auswahl klicken. Wenn eine Kategorie oder Unterkategorie ausgewählt ist, aber eine ihrer Unterkategorien nicht ausgewählt ist, zeigt das Kontrollkästchen einen blauen Hintergrund, um darauf hinzuweisen, dass in den Unterkategorien der ausgewählten Kategorie nur eine Teilauswahl getroffen wurde.

Wenn Sie mit der rechten Maustaste auf eine Kategorie im Kategoriebaum klicken, wird ein Kontextmenü angezeigt, in dem folgende Optionen zur Auswahl stehen:

- **Ausgewählte Elemente überprüfen.** Aktiviert alle Kontrollkästchen für die ausgewählten Zeilen in der Tabelle.
- **Markierung der ausgewählten Elemente aufheben.** Deaktiviert alle Kontrollkästchen für die ausgewählten Zeilen in der Tabelle.
- **Alles markieren.** Aktiviert alle Kontrollkästchen in der Tabelle. Dadurch werden alle Kategorien in der Endausgabe verwendet. Sie können auch das entsprechende Kontrollkästchen auf der Symbolleiste verwenden.
- **Alle Markierungen aufheben.** Deaktiviert alle Kontrollkästchen in der Tabelle. Wenn Sie eine Kategorie deaktivieren, wird diese in der Endausgabe nicht verwendet. Sie können auch das entsprechende leere Kontrollkästchenfeld in der Symbolleiste verwenden.

Wenn Sie mit der rechten Maustaste auf eine Zelle in der Deskriptortabelle klicken, wird ein Kontextmenü angezeigt, in dem folgende Optionen zur Auswahl stehen:

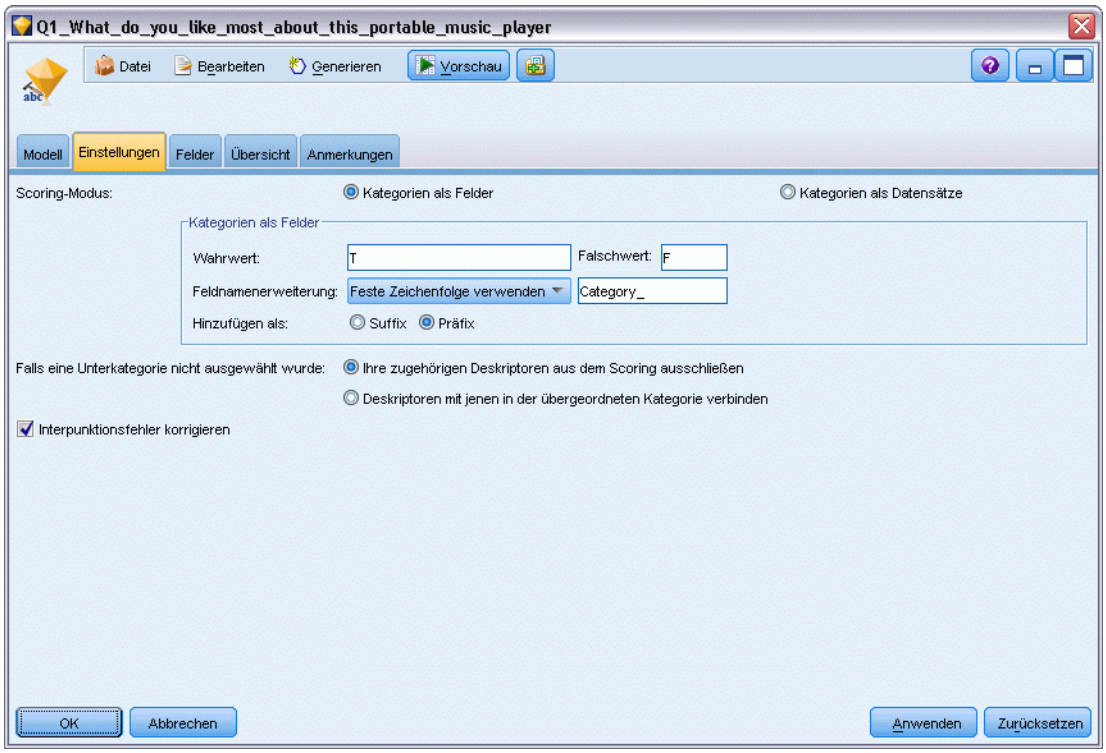
- **Kopieren.** Die ausgewählten Konzepte werden in die Zwischenablage kopiert.
- **Mit Feldern kopieren.** Der ausgewählte Deskriptor wird zusammen mit den Spaltenüberschriften in die Zwischenablage kopiert.
- **Alles auswählen.** Alle Zeilen in der Tabelle werden ausgewählt.

Kategoriemodell-Nugget: Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" wird verwendet, um den Wert des Textfelds für die neuen Eingangsdaten zu definieren (sofern erforderlich). Außerdem können Sie dort das Datenmodell für die Ausgabe definieren (Scoring-Modus).

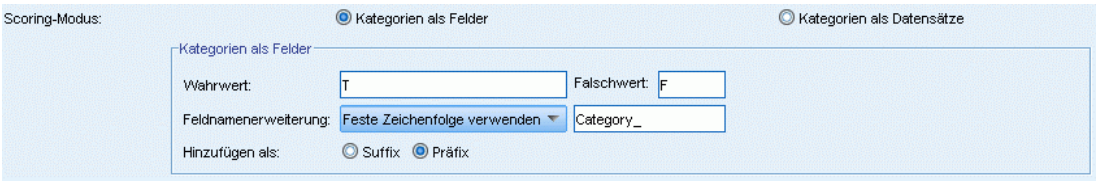
Anmerkung: Diese Registerkarte wird nur im Knotendialogfeld angezeigt, wenn das Modell-Nugget auf der Zeichenfläche oder im Stream platziert wird. Sie ist nicht vorhanden, wenn Sie direkt in der Modellpalette auf dieses Nugget zugreifen.

Abbildung 3-35
Dialogfeld des Text-Mining-Kategoriemodell-Nuggets: Registerkarte "Einstellungen"



Scoring-Modus: Kategorien als Felder

Abbildung 3-36
Registerkarte "Einstellungen" für "Kategorien als Felder"



Wenn diese Option ausgewählt ist, gibt es in der Ausgabe genauso viele Datensätze wie in der Eingabe. Allerdings enthält jetzt jeder Datensatz ein weiteres Feld für jede auf der Registerkarte "Modell" ausgewählte Kategorie (Auswahl über Kontrollkästchen). Geben Sie für jedes Feld einen Flag-Wert für Wahr und für Falsch ein, z. B. *Ja*/*Nein*, *Wahr*/*Falsch*, *W*/*F* oder *1* und *2*. Die Speichertypen werden automatisch so eingestellt, dass sie die gewählten Werte reflektieren. Wenn Sie beispielsweise numerische Werte für die Flags eingeben, werden sie automatisch als Ganzzahl behandelt. Als Speichertypen für Flags sind "Zeichenkette", "Ganze Zahl", "Reelle Zahl" und "Datum/Uhrzeit" möglich.

Feldnamenerweiterung. Sie können ein Erweiterungspräfix/-suffix für den Feldnamen angeben oder die Kategoriecodes verwenden. Feldnamen werden unter Verwendung des Kategorienamens und dieser Erweiterung generiert.

- **Hinzufügen als.** Gibt an, an welcher Stelle die Erweiterung zum Feldnamen hinzugefügt werden soll. Wählen Sie Präfix, um die Erweiterung am Anfang der Zeichenkette einzufügen. Wählen Sie Suffix, um die Erweiterung am Ende der Zeichenkette einzufügen.

Wenn eine Unterkategorie nicht ausgewählt ist. Mit dieser Option können Sie angeben, wie die Deskriptoren, die zu nicht für das Scoring ausgewählten Unterkategorien gehören, behandelt werden. Es gibt zwei Optionen.

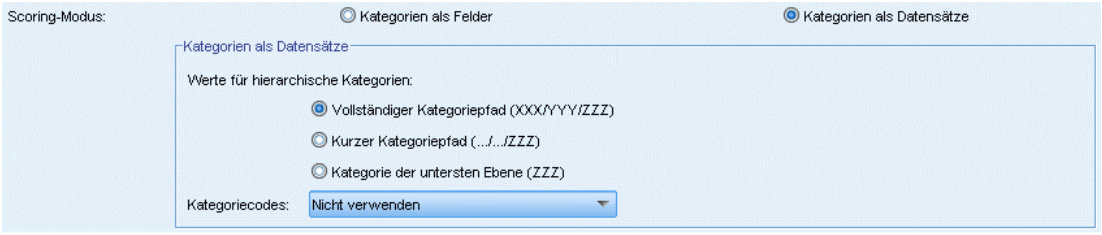
- Die Option **Deskriptoren vollständig vom Scoring ausschließen** verursacht, dass Deskriptoren von Unterkategorien ohne Häkchen (nicht ausgewählt) beim Scoring ignoriert und nicht verwendet werden.
- Die Option **Deskriptoren mit denen in der übergeordneten Kategorie aggregieren** verursacht, dass Deskriptoren von Unterkategorien ohne Häkchen (nicht ausgewählt) als Deskriptoren für die übergeordnete Kategorie (die Kategorie über dieser Unterkategorie) verwendet werden. Wenn mehrere Ebenen von Unterkategorien nicht ausgewählt sind, werden die Deskriptoren unter der ersten verfügbaren übergeordneten Kategorie zusammengefasst.

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Punktuationsfehlern (zum Beispiel ungeeignete Verwendung) während der Extrahierung, um die Extrahierbarkeit von Konzepten zu verbessern. Diese Option ist besonders nützlich bei kurzem Text und niedriger Textqualität (wie dies beispielsweise bei offenen Antworten bei Umfragen, E-Mails und CRM-Daten der Fall ist) oder wenn der Text viele Abkürzungen enthält.

Anmerkung: Die Option Interpunktionsfehler korrigieren gilt nicht beim Arbeiten mit japanischem Text.

Scoring-Modus: Kategorien als Datensätze

Abbildung 3-37
Registerkarte "Einstellungen" für "Kategorien als Datensätze"



Mit dieser Option wird für jedes category, document-Paar ein neuer Datensatz erstellt. Normalerweise gibt es mehr Datensätze in der Ausgabe, als in der Eingabe vorhanden waren. Zusätzlich zu den Eingabefeldern werden zu den Daten auch weitere Felder hinzugefügt. Dies hängt davon ab, um welche Art von Modell es sich handelt.

Tabelle 3-2
Ausgabefelder für "Kategorien als Datensätze"

Feld "Neue Ausgabe"	Beschreibung
Category	Enthält den Namen der Kategorie, der das Textdokument zugewiesen wurde. Wenn die Kategorie die Unterkategorie einer anderen Kategorie is, wird der vollständige Pfad zum Kategoriennamen durch den Wert gesteuert, den Sie in diesem Dialogfeld auswählen.

Werte für hierarchische Kategorien. Diese Option steuert, wie die Namen von Unterkategorien in der Ausgabe angezeigt werden.

- **Vollständiger Kategoriepfad.** Diese Option gibt den Namen der Kategorie und den vollständigen Pfad von übergeordneten Kategorien (falls zutreffend) mit Schrägstrichen zwischen den Namen von Kategorien und Unterkategorien an.
- **Kurzer Kategoriepfad.** Diese Option gibt nur den Namen der Kategorie aus, verwendet aber Auslassungszeichen, um die Anzahl der übergeordneten Kategorien für die betreffende Kategorie anzuzeigen.
- **Kategorie der untersten Ebene.** Diese Option gibt nur den Namen der Kategorie aus, ohne dass der vollständige Pfad oder übergeordnete Kategorien angezeigt werden.

Wenn eine Unterkategorie nicht ausgewählt ist. Mit dieser Option können Sie angeben, wie die Deskriptoren, die zu nicht für das Scoring ausgewählten Unterkategorien gehören, behandelt werden. Es gibt zwei Optionen.

- Die Option **Deskriptoren vollständig vom Scoring ausschließen** verursacht, dass Deskriptoren von Unterkategorien ohne Häkchen (nicht ausgewählt) beim Scoring ignoriert und nicht verwendet werden.
- Die Option **Deskriptoren mit denen in der übergeordneten Kategorie aggregieren** verursacht, dass Deskriptoren von Unterkategorien ohne Häkchen (nicht ausgewählt) als Deskriptoren für die übergeordnete Kategorie (die Kategorie über dieser Unterkategorie) verwendet werden. Wenn mehrere Ebenen von Unterkategorien nicht ausgewählt sind, werden die Deskriptoren unter der ersten verfügbaren übergeordneten Kategorie zusammengefasst.

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Punktuationsfehlern (zum Beispiel ungeeignete Verwendung) während der Extrahierung, um die Extrahierbarkeit von Konzepten zu verbessern. Diese Option ist besonders nützlich bei kurzem Text und niedriger Textqualität (wie dies beispielsweise bei offenen Antworten bei Umfragen, E-Mails und CRM-Daten der Fall ist) oder wenn der Text viele Abkürzungen enthält.

Anmerkung: Die Option Interpunktionsfehler korrigieren gilt nicht beim Arbeiten mit japanischem Text.

Kategoriemodell-Nugget: Andere Registerkarten

Die Registerkarten "Felder" und "Einstellungen" für das Kategoriemodell-Nugget sind dieselben wie für das Konzeptmodell-Nugget.

- Registerkarte “Felder”. Für weitere Informationen siehe Thema Konzeptmodell: Registerkarte “Felder” auf S. 64.
- Registerkarte “Übersicht”. Für weitere Informationen siehe Thema Konzeptmodell: Registerkarte “Übersicht” auf S. 66.

Verwenden von Kategoriemodell-Nuggets in einem Stream

Das Text-Mining-Kategoriemodell-Nugget wird aus einer interaktiven Workbench-Sitzung generiert. Sie können dieses Modell-Nugget in einem Stream verwenden.

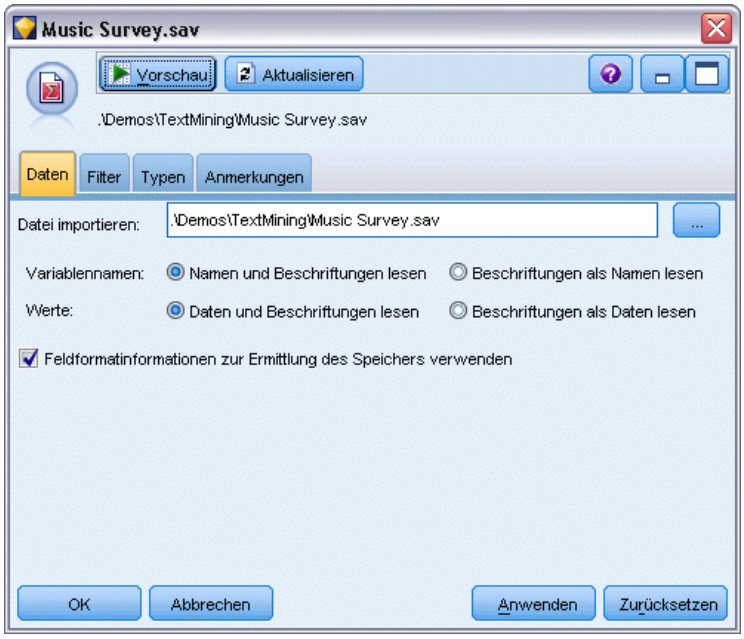
Beispiel: Statistikdatei-Knoten mit Kategoriemodell-Nugget

Das folgende Beispiel zeigt die Verwendung des Text-Mining-Modell-Nuggets.

Abbildung 3-38
Beispiel-Stream: Statistikdateiknoten mit einem Text-Mining-Kategoriemodell-Nugget

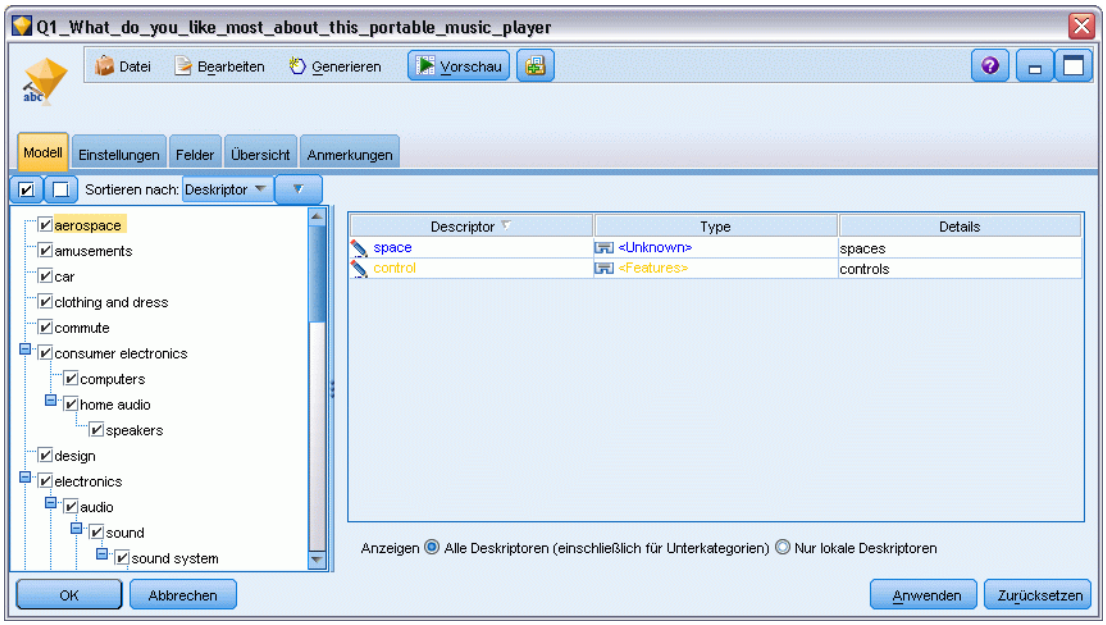
▶ **Statistikdateiknoten (Registerkarte “Einstellungen”)** Zuerst fügten wir diesen Knoten dem Stream hinzu, um anzugeben, wo die Textdokumente gespeichert sind.

Abbildung 3-39
Dialogfeld des Knotens für Statistikdateien: Registerkarte “Daten”



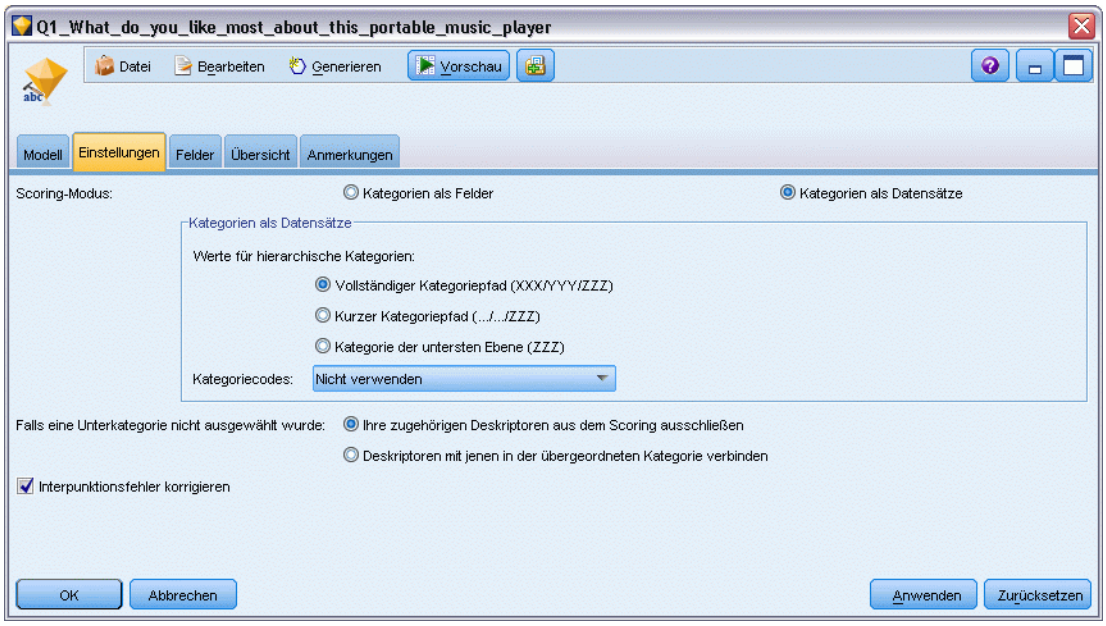
- **Text-Mining-Kategoriemodell-Nugget (Registerkarte "Modell").** Als Nächstes haben wir dem Dateilistenknoten ein Kategoriemodell-Nugget hinzugefügt und dieses mit dem Statistikdateiknoten verbunden. Wir haben die Kategorien ausgewählt, die wir für das Scoring unserer Daten verwenden wollten.

Abbildung 3-40
Dialogfeld des Text-Mining-Modell-Nuggets: Registerkarte "Modell"



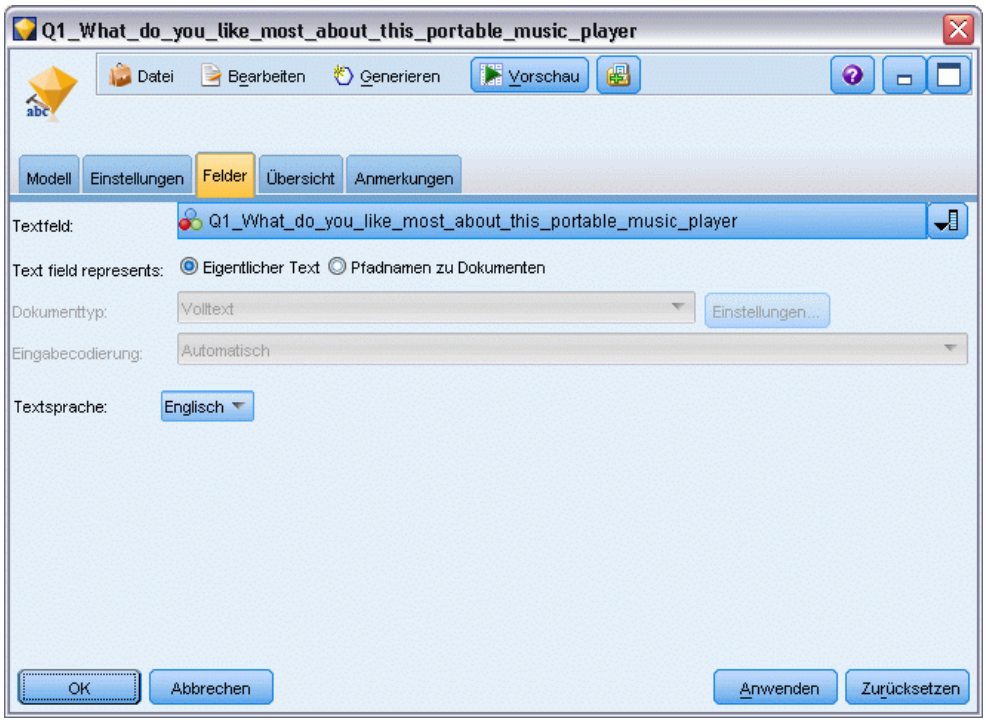
- **Text-Mining-Modell-Nugget (Registerkarte "Einstellungen").** Als Nächstes haben wir das Ausgabeformat Kategorien als Felder definiert.

Abbildung 3-41
Dialogfeld für Kategoriemodell-Nugget: Registerkarte "Einstellungen"



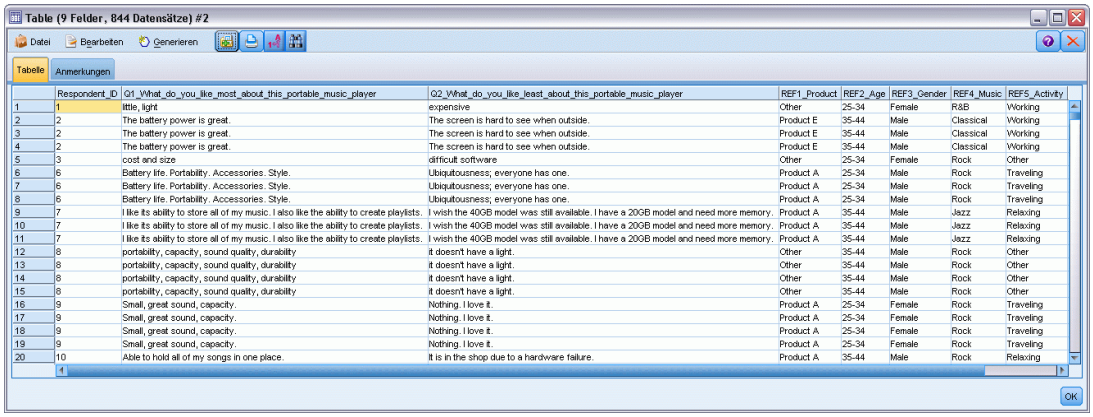
▶ **Text-Mining-Kategoriemodell-Nugget (Registerkarte "Felder").** Anschließend haben wir die Textfeldvariable ausgewählt. Dies ist der Feldname aus dem Statistikdateiknoten. Dann wählten wir die Option „Textfeld enthält tatsächlichen Text" sowie andere Einstellungen aus.

Abbildung 3-42
Dialogfeld des Text-Mining-Modell-Nuggets: Registerkarte "Felder"



- **Tabellenknoten.** Als Nächstes fügten wir einen Tabellenknoten hinzu, um die Ergebnisse zu prüfen. Anschließend führten wir den Stream aus.

Abbildung 3-43
Tabellenausgabe



	Respondent_ID	Q1_What_do_you_like_most_about_this_portable_music_player	Q2_What_do_you_like_least_about_this_portable_music_player	REF1_Product	REF2_Age	REF3_Gender	REF4_Music	REF5_Activity
1	1	little, light	expensive	Other	25-34	Female	R&B	Working
2	2	The battery power is great.	The screen is hard to see when outside.	Product E	35-44	Male	Classical	Working
3	2	The battery power is great.	The screen is hard to see when outside.	Product E	35-44	Male	Classical	Working
4	2	The battery power is great.	The screen is hard to see when outside.	Product E	35-44	Male	Classical	Working
5	3	cost and size	difficult software	Other	25-34	Female	Rock	Other
6	6	Battery life. Portability. Accessories. Style.	Ubiquitousness; everyone has one.	Product A	25-34	Male	Rock	Traveling
7	6	Battery life. Portability. Accessories. Style.	Ubiquitousness; everyone has one.	Product A	25-34	Male	Rock	Traveling
8	6	Battery life. Portability. Accessories. Style.	Ubiquitousness; everyone has one.	Product A	25-34	Male	Rock	Traveling
9	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.	Product A	35-44	Male	Jazz	Relaxing
10	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.	Product A	35-44	Male	Jazz	Relaxing
11	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.	Product A	35-44	Male	Jazz	Relaxing
12	8	portability, capacity, sound quality, durability	it doesn't have a light.	Other	35-44	Male	Rock	Other
13	8	portability, capacity, sound quality, durability	it doesn't have a light.	Other	35-44	Male	Rock	Other
14	8	portability, capacity, sound quality, durability	it doesn't have a light.	Other	35-44	Male	Rock	Other
15	8	portability, capacity, sound quality, durability	it doesn't have a light.	Other	35-44	Male	Rock	Other
16	9	Small, great sound, capacity.	Nothing. I love it.	Product A	25-34	Female	Rock	Traveling
17	9	Small, great sound, capacity.	Nothing. I love it.	Product A	25-34	Female	Rock	Traveling
18	9	Small, great sound, capacity.	Nothing. I love it.	Product A	25-34	Female	Rock	Traveling
19	9	Small, great sound, capacity.	Nothing. I love it.	Product A	25-34	Female	Rock	Traveling
20	10	Able to hold all of my songs in one place.	It is in the shop due to a hardware failure.	Product A	35-44	Male	Rock	Relaxing

Kapitel 4

Mining für Textlinks

Text Link Analysis-Knoten

Über den Textlinkanalyseknoten (TLA) wird die Konzeptextrahierung beim Text-Mining um eine Technologie zum Musterabgleich erweitert. Damit können Beziehungen zwischen den Konzepten, die in den Textdaten enthalten sind, über bekannte Muster ermittelt werden. Diese Beziehungen können Aufschluss darüber geben, wie ein Kunde über ein Produkt denkt, welche Unternehmen Geschäftsbeziehungen miteinander unterhalten und sogar darüber, welche Beziehungen zwischen verschiedenen Genen oder Arzneimittelwirkstoffen vorliegen.

So könnte es beispielsweise sein, dass es Ihnen nicht ausreicht, den Produktnamen Ihres Mitbewerbers zu extrahieren. Mit diesem Knoten können Sie außerdem erfahren, was die Kunden von diesem Produkt halten, wenn derartige Meinungen in den Daten vorliegen. Die Beziehungen und Zuordnungen (Assoziationen) werden ermittelt und extrahiert, indem bekannte Muster mit Ihren Textdaten abgeglichen werden.

Sie können die TLA-Musterregeln aus bestimmten Ressourcenvorlagen verwenden, die im Lieferumfang von IBM® SPSS® Modeler Text Analytics enthalten sind, oder Ihre eigenen erstellen bzw. bearbeiten. Musterregeln bestehen aus Makros, Wortlisten sowie Wortlücken und bilden eine Boole'sche Abfrage (Regel), die mit dem Eingangstext abgeglichen wird. Wenn eine TLA-Musterregel mit einem Text übereinstimmt, kann dieser Text als TLA-Ergebnis extrahiert und für die Datenausgabe neu strukturiert werden. Für weitere Informationen siehe Thema Textlinkregeln in Kapitel 19 auf S. 346.

Der Textlinkanalyseknoten eröffnet einen direkteren Weg, TLA-Musterergebnisse in Ihrem Text zu ermitteln und die Ergebnisse anschließend dem Datenset im Stream hinzuzufügen. Doch der Textlinkanalyseknoten stellt nicht die einzige Methode zur Textlinkanalyse dar. Sie können dazu auch eine interaktive Workbench-Sitzung im Text-Mining-Modellierungsknoten nutzen.

In der interaktiven Workbench können Sie die TLA-Musterergebnisse untersuchen und als Kategoriedeskriptoren nutzen und/oder mithilfe von Drilldown-Verfahren und Diagrammen mehr über die betreffenden Ergebnisse herausfinden. Für weitere Informationen siehe Thema Untersuchen von Textlinkanalyse in Kapitel 12 auf S. 252. Tatsächlich ist die Nutzung des Text-Mining-Knotens zur Extrahierung von TLA-Ergebnissen eine hervorragende Methode, um Vorlagen im Hinblick auf Ihre Daten zu untersuchen und eine entsprechende Feinabstimmung vorzunehmen. Anschließend können die Vorlagen direkt im TLA-Knoten verwendet werden.

Die Ausgabe kann in bis zu sechs Slots, oder Teilen, repräsentiert werden. Japanische Muster werden nur als einer oder zwei Slots ausgegeben. Für weitere Informationen siehe Thema Ausgabe des TLA-Knotens auf S. 90.

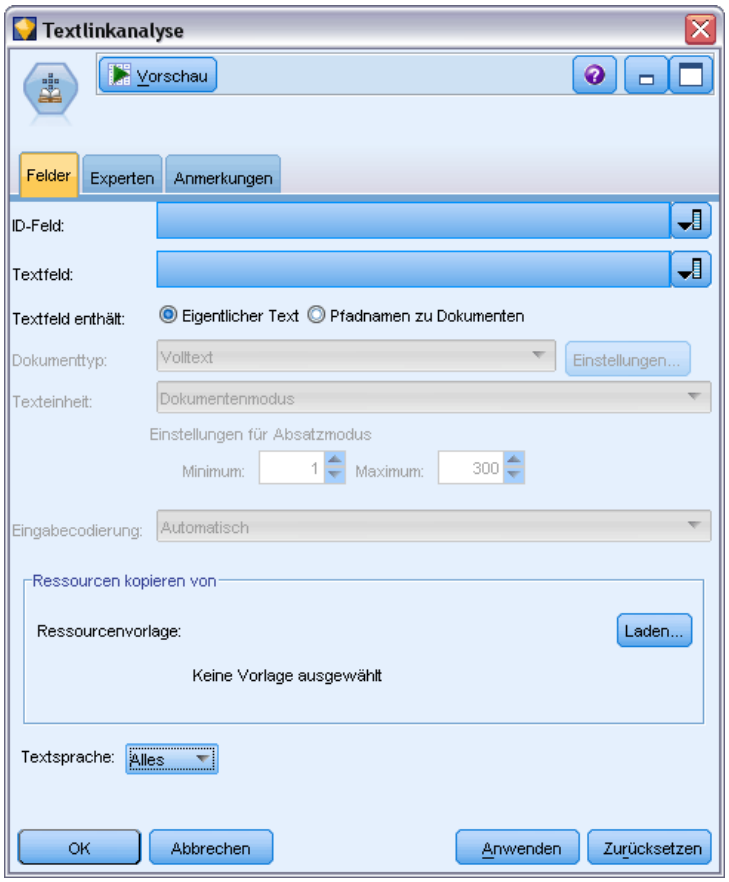
Sie finden diesen Knoten auf der SPSS Modeler Text Analytics -Registerkarte der Knotenpalette am unteren Rand des IBM® SPSS® Modeler-Fensters. Für weitere Informationen siehe Thema IBM SPSS Modeler Text Analytics-Knoten in Kapitel 1 auf S. 11.

Anforderungen. Der Textlinkanalyseknoten akzeptiert Textdaten, die unter Verwendung eines der standardmäßigen -Quellenknoten (Datenbankknoten, Textdateiknoten usw.) oder unter Auflistung der Pfade zu externen Dokumenten, die von einem Dateilisten- oder Web-Feed-Knoten erstellt wurden, in ein Feld eingelesen wurden.

Stärken. Der Textlinkanalyseknoten geht über die einfache Konzeptextrahierung hinaus und bietet Informationen über die Beziehungen *zwischen* Konzepten sowie verwandte Meinungen oder Vermerke, die in den Daten aufgedeckt werden können.

Textlinkanalyseknoten: Registerkarte "Felder"

Abbildung 4-1
Dialogfeld des Textlinkanalyseknotens: Registerkarte "Felder"



Die Registerkarte "Felder" dient speziell zur Angabe der Feldeinstellungen für die Daten, aus denen Konzepte extrahiert werden sollen. Folgende Parameter können festgelegt werden:

ID-Feld. Wählen Sie das Feld aus, das die ID für die Textdatensätze enthält. Bei den IDs muss es sich um Ganzzahlen handeln. Das ID-Feld dient als Index für die einzelnen Textdatensätze. Verwenden Sie ein ID-Feld, wenn das Textfeld den für das Mining verwendeten Text darstellt. Verwenden Sie kein ID-Feld, wenn das Textfeld Pfadnamen zu Dokumenten enthält.

Textfeld. Wählen Sie das Feld, das den Text für das Mining enthält, den Dokumentenpfadnamen bzw. den Verzeichnispfadnamen, der zu den Dokumenten führt. Dieses Feld hängt von der Datenquelle ab.

Textfeld enthält. Geben Sie an, was das Textfeld enthalten soll, das in der vorausgehenden Einstellung festgelegt wurde. Folgende Optionen stehen zur Auswahl:

- **Eigentlicher Text.** Wählen Sie diese Option aus, wenn das Feld genau den Text enthält, aus dem Konzepte extrahiert werden sollen.
- **Pfadnamen zu Dokumenten.** Wählen Sie diese Option aus, wenn das Feld mindestens einen Pfadnamen zu dem oder den Speicherort(en) der Textdokumente enthält.

Dokumenttyp. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält. Der Dokumenttyp gibt die Struktur des Texts an. Wählen Sie einen der folgenden Typen aus:

- **Volltext.** Verwenden Sie diese Option für die meisten Dokumente bzw. Textquellen. Die gesamte Textmenge wird für die Extrahierung gescannt. Im Gegensatz zu den anderen Optionen gibt es keine weiteren Einstellungen für diese Option.
- **Gegliederter Text.** Verwenden Sie diese Option für bibliografische Formulare, Patente und alle Dateien, die reguläre Strukturen enthalten, die identifiziert und analysiert werden können. Dieser Dokumenttyp wird verwendet, um den gesamten Extrahierungsvorgang oder Teile des Extrahierungsvorgangs zu überspringen. Er ermöglicht das Definieren von Trennzeichen zu Fachausdrücken, das Zuweisen von Typen und das Festlegen eines minimalen Häufigkeitswerts. Wenn Sie diese Option aktivieren, müssen Sie auf die Schaltfläche Einstellungen klicken und im Bereich Formatierung als gegliederter Text des Dialogfelds “Dokumenteinstellungen” Texttrennzeichen eingeben. Für weitere Informationen siehe Thema Dokumenteinstellungen der Registerkarte “Felder” in Kapitel 3 auf S. 35.
- **XML** Verwenden Sie diese Option, um die XML-Tags anzugeben, die den zu extrahierenden Text enthalten. Alle anderen Tags werden ignoriert. Wenn Sie diese Option aktivieren, müssen Sie auf die Schaltfläche Einstellungen klicken und im Bereich Formatierung als XML des Dialogfelds “Dokumenteinstellungen” explizit die XML-Elemente angeben, die den Text enthalten, der während des Extrahierungsprozesses gelesen werden soll. Für weitere Informationen siehe Thema Dokumenteinstellungen der Registerkarte “Felder” in Kapitel 3 auf S. 35.

Texteinheit. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält und Sie Volltext als Dokumenttyp ausgewählt haben. Wählen Sie den Extrahierungsmodus aus folgenden Elementen aus:

- **Dokumentenmodus.** Wird für kurze, semantisch homogene Dokumente verwendet, beispielsweise Artikel von Nachrichtenagenturen.
- **Absatzmodus.** Verwenden Sie diese Option für Webseiten und Dokumente ohne Tags. Der Extrahierungsprozess teilt die Dokumente semantisch. Dabei nutzt er Merkmale wie interne Tags und Syntax. Bei Auswahl dieses Modus wird das Scoring absatzweise durchgeführt. Folglich ist die Regel `Apfel & Orange` nur erfüllt, wenn `Apfel` und `Orange` im gleichen Absatz gefunden werden.

Einstellungen für Absatzmodus. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält und als Texteinheitoption Absatzmodus angegeben haben. Geben Sie die für Extrahierungen zu verwendenden Zeichenschwellenwerte

an. Die tatsächliche Größe wird auf den nächsten Punkt (Satzende) auf- bzw. abgerundet. Um sicherzustellen, dass die aus dem Text der Dokumentensammlung erstellten Wortzuordnungen repräsentativ sind, sollten Sie eine zu kleine Extrahierungsgröße vermeiden.

- **Minimum.** Geben Sie die Mindestzahl der bei Extrahierungen zu verwendenden Zeichen an.
- **Maximum.** Geben Sie die Höchstzahl der bei Extrahierungen zu verwendenden Zeichen an.

Eingabekodierung. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld Pfadnamen zu Dokumenten enthält. Es bestimmt die Standardtextkodierung. Für alle Sprachen außer Japanisch erfolgt eine Konvertierung von der angegebenen bzw. erkannten Kodierung in ISO-8859-1. Selbst wenn Sie also eine andere Kodierung angeben, wird diese vor der Verarbeitung von der Extrahierungsengine in iso-8859-1 konvertiert. Alle Zeichen, die nicht in die ISO-8859-1-Kodierungsdefinition passen, werden in Leerzeichen umgewandelt. Für japanischen Text können Sie eine von mehreren Kodierungsoptionen wählen: SHIFT_JIS, EUC_JP, UTF-8 oder ISO-2022-JP.

Ressourcen kopieren von: Beim Text Mining basierte die Extrahierung nicht nur auf den Einstellungen auf der Registerkarte “Experten”, sondern auch auf den linguistischen Ressourcen. Diese Ressourcen dienen als Grundlage für die Art der Be- und Verarbeitung von Daten während der Extrahierung, um die Konzepte, Typen und TLA-Muster zu erhalten. Sie können Ressourcen aus einer Ressourcenvorlage in diesen Knoten kopieren.

Bei einer Ressourcenvorlage handelt es sich um eine vordefinierte Reihe von Bibliotheken und erweiterten linguistischen und nicht linguistischen Ressourcen, die auf eine bestimmte Domäne oder Nutzung feinabgestimmt worden sind. Diese Ressourcen dienen als Grundlage für die Art der Be- und Verarbeitung von Daten während der Extrahierung. Klicken Sie auf Laden und wählen Sie die Vorlage, aus der Ihre Ressourcen kopiert werden sollen.

Vorlagen werden nicht bei der Ausführung des Streams geladen, sondern wenn sie ausgewählt werden. Wenn Sie den Ladevorgang starten, wird eine Kopie der Ressourcen im Knoten gespeichert. Wenn Sie also eine aktualisierte Vorlage verwenden möchten, dann müssen Sie sie hier neu laden. Für weitere Informationen siehe Thema Kopieren von Ressourcen aus TAPs und Vorlagen in Kapitel 3 auf S. 42.

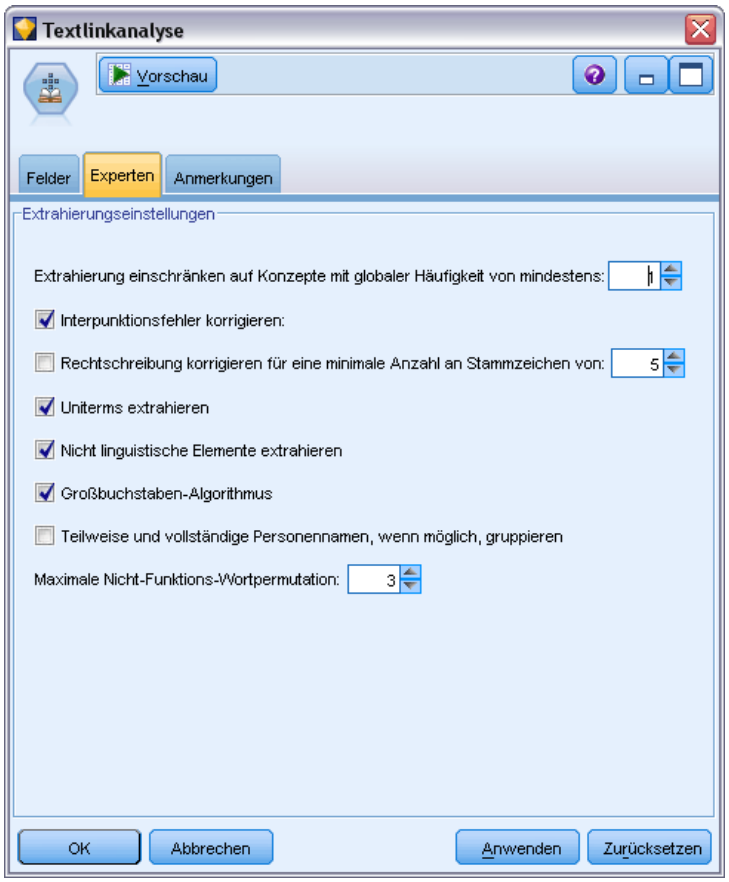
Textsprache. Bestimmt die Sprache des für das Mining verwendeten Texts. Die in den Knoten kopierten Ressourcen steuern die angezeigten Sprachoptionen. Sie können die Sprache wählen, auf die die Ressourcen abgestimmt wurden, oder die Option ALLE wählen. Wir empfehlen, für die Textdaten die exakte Sprache anzugeben. Wenn Sie sich jedoch nicht sicher sind, können Sie die Option ALLE wählen. ALL ist für japanischen Text nicht verfügbar. Diese Option ALL verlängert die Ausführungsdauer, da mithilfe der automatischen Spracherkennung zunächst alle Dokumente und Datensätze gescannt werden, um die Textsprache zu ermitteln. Wenn Sie diese Option wählen, werden alle Datensätze oder Dokumente, die in einer unterstützten und lizenzierten Sprache vorliegen, durch die Extrahierungsengine mit den internen Wörterbüchern in der jeweiligen Sprache gelesen. Für weitere Informationen siehe Thema Language Identifier in Kapitel 18 auf S. 345. Wenden Sie sich an Ihren Kundendienstmitarbeiter, wenn Sie eine Lizenz für eine unterstützte Sprache erwerben möchten, auf die Sie zurzeit keinen Zugriff haben.

Textlinkanalyseknoten: Registerkarte "Experten"

In diesem Knoten ist die Extrahierung von TLA-Musterergebnissen automatisch aktiviert. Die Registerkarte "Experten" enthält bestimmte zusätzliche Parameter, die beeinflussen, wie der Text extrahiert und gehandhabt wird. Die Parameter in diesem Dialogfeld legen das Grundverhalten sowie einige erweiterte Verhaltensweisen des Extrahierungsprozesses fest. Zudem werden die Extrahierungsergebnisse von einer Reihe von linguistischen Ressourcen und Optionen beeinflusst. Diese werden über die ausgewählte Ressourcenvorlage gesteuert.

Für niederländischen, englischen, französischen, deutschen, italienischen, portugiesischen und spanischen Text

Abbildung 4-2
Dialogfeld des Textlinkanalyseknotens: Registerkarte "Experten"



Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Punktuationsfehlern (zum Beispiel ungeeignete Verwendung) während der Extrahierung, um die Extrahierbarkeit von Konzepten zu verbessern. Diese Option ist besonders nützlich bei kurzem Text und niedriger Textqualität (wie dies beispielsweise bei offenen Antworten bei Umfragen, E-Mails und CRM-Daten der Fall ist) oder wenn der Text viele Abkürzungen enthält.

Rechtschreibung korrigieren für eine minimale Anzahl an Stammzeichen von [n]. Diese Option wendet ein unscharfes Gruppierungsverfahren an, das hilft, häufig falsch geschriebene Wörter oder genau geschriebene Wörter unter einem Konzept zu gruppieren. Der Algorithmus für die unscharfe Gruppierung entfernt alle Vokale (außer den ersten) und doppelte/dreifache Konsonanten temporär aus extrahierten Wörtern und vergleicht sie, um festzustellen, ob sie gleich sind, so dass `Modellierung` und `Modelierung` zusammen gruppiert werden würden. Wenn jedoch jeder Fachausdruck einem anderen Typ (ausschließlich des Typs `<Unknown>`) zugewiesen ist, wird das unscharfe Gruppierungsverfahren nicht angewendet.

Sie können auch die minimal erforderliche Zahl von *Stamm*zeichen definieren, bevor unscharfe Gruppierung eingesetzt wird. Die Anzahl der Stammzeichen in einem Ausdruck berechnet sich aus der Summe aller Zeichen abzüglich aller Zeichen, die Beugungsendungen und – bei zusammengesetzten Ausdrücken – Determinatoren und Präpositionen bilden. So würde beispielsweise der Ausdruck `Aufgaben` durch die Form "Aufgabe" mit 7 Stammzeichen gezählt werden, da der Buchstabe *n* am Ende des Worts eine Beugung darstellt (Pluralform). Gleichermaßen werden für `Apfelmus` 8 Stammzeichen ("Apfelmus") gezählt und `Hersteller von Autos` zählt als 14 Stammzeichen ("Hersteller Auto"). Diese Zählmethode dient nur zur Überprüfung, ob die Fuzzy-Gruppierung angewendet werden soll, hat jedoch keinen Einfluss auf den Abgleich der Wörter.

Hinweis: Wenn Sie feststellen, dass bestimmte Wörter später falsch gruppiert werden, können Sie Wortpaare aus diesem Verfahren ausschließen, indem Sie sie explizit im Abschnitt Unscharfe Gruppierung: Ausnahmen im erweiterten Ressourceneditor deklarieren. Für weitere Informationen siehe Thema Unscharfe Gruppierung in Kapitel 18 auf S. 336.

uniterms extrahieren. Diese Option extrahiert einzelne Wörter (Uniterms), solange das Wort nicht bereits Teil eines zusammengesetzten Worts ist und es entweder ein Nomen oder eine nicht erkannte Wortart ist.

Nicht linguistische Elemente extrahieren. Diese Option extrahiert nicht linguistische Elemente wie beispielsweise Telefonnummern, Personalausweisnummern, Uhrzeiten, Datumsangaben, Währungen, Ziffern, Prozentsätze, E-Mail-Adressen und HTTP-Adressen. Sie können bestimmte Typen von nicht linguistischen Elementen im Abschnitt Nicht linguistische Elemente: Konfiguration des erweiterten Ressourceneditors ein- bzw. ausschließen. Durch Deaktivierung unnötiger Elemente vergeudet die Extrahierungsengine keine Verarbeitungszeit. Für weitere Informationen siehe Thema Konfiguration in Kapitel 18 auf S. 341.

Großbuchstaben-Algorithmus. Diese Option extrahiert einfache und zusammengesetzte Ausdrücke, die sich nicht in den integrierten Wörterbüchern befinden, solange der erste Buchstabe des Begriffs in Großbuchstaben geschrieben ist. Diese Option ist eine gute Möglichkeit, die geeignetsten Substantive zu extrahieren.

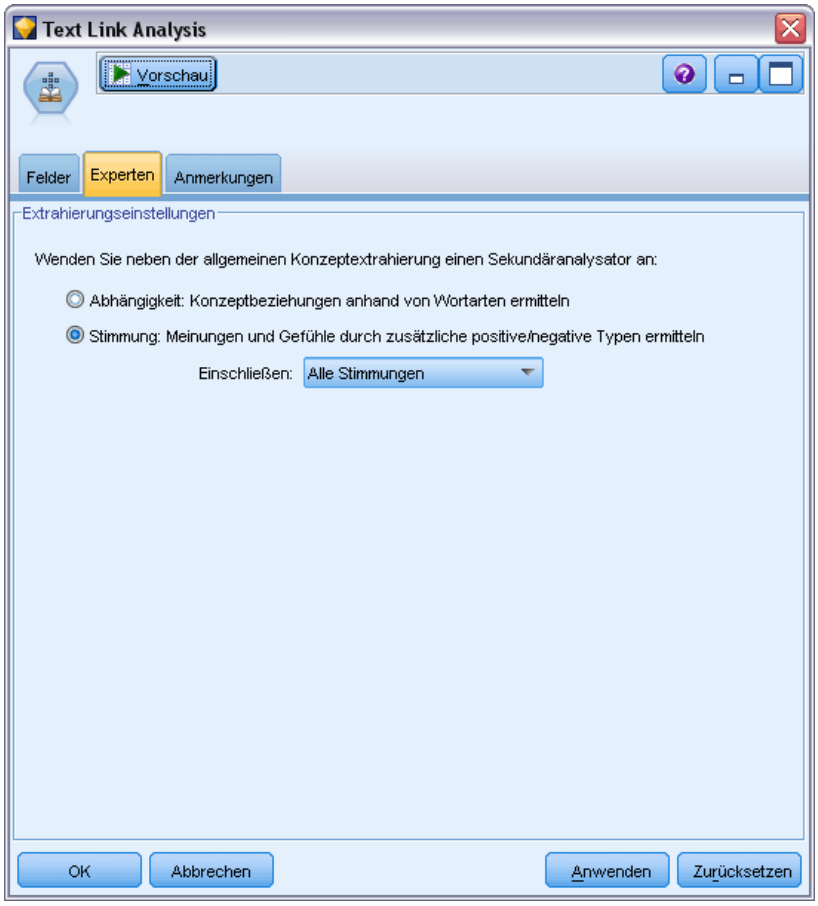
Teilweise und vollständige Personennamen, wenn möglich, gruppieren. Diese Option gruppiert Namen, die zusammen im Text unterschiedlich erscheinen. Diese Funktion ist nützlich, da Namen zu Beginn des Textes oft in voller Länge angegeben werden und später nur noch mit einer Kurzform auf sie verwiesen wird. Diese Option versucht, jeden Uniterm mit dem Typ `<Unknown>` mit dem letzten Wort aller zusammengesetzten Ausdrücke abzugleichen, die dem Typ `<Person>` zugeordnet sind. Wird beispielsweise *doe* gefunden und anfänglich dem Typ `<Unknown>` zugeordnet, überprüft die Extrahierungsengine, ob ein zusammengesetzter Ausdruck

vom Typ `<Person>` als letztes Wort *doe* enthält, z. B. *john doe*. Diese Option wird nicht auf Vornamen angewendet, da sie in den meisten Fällen nicht als Uniterms extrahiert werden.

Maximale Nicht-Funktions-Wortpermutation. Diese Option gibt die maximale Anzahl von Füllwörtern an, die für die Anwendung des Permutationsverfahrens vorhanden sein müssen. Dieses Permutationsverfahren gruppiert ähnliche Phrasen, die sich nur durch die enthaltenen Füllwörter (zum Beispiel von und der) unabhängig von der Beugung unterscheiden. Nehmen wir zum Beispiel an, dass Sie diesen Wert auf höchstens zwei Wörter eingestellt haben und sowohl `Unternehmen des Vertreters` und `Vertreter des Unternehmens` extrahiert wurden. In diesem Fall würden beide extrahierte Ausdrücke in der endgültigen Konzeptliste zusammen gruppiert, da beide Ausdrücke als gleich betrachtet werden, wenn `des` ignoriert wird.

Für japanischen Text

Abbildung 4-3
Dialogfeld des Textlinkanalyseknotens: Registerkarte "Experten" (für japanischen Text)



Bei japanischem Text können Sie wählen, welche sekundäre Analyse angewendet werden soll.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Sekundäre Analyse. Beim Start einer Extrahierung werden anhand des Standardsatzes an Typen grundlegende Stichwörter extrahiert. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text in Anhang A auf S. 387. Wenn Sie jedoch eine sekundäre Analyse wählen, erhalten Sie mehr und vielfältigere Konzepte, da der Extraktor nun Partikel und Hilfsverben als Teil des Konzepts berücksichtigt. Nehmen Sie beispielsweise an, der Satz 肩の荷が下りた wurde in "*Mir wurde eine große Last von den Schultern genommen*". In diesem Beispiel kann die grundlegende Stichwortextrahierung jedes Konzept wie folgt separat extrahieren: 肩 (*Schultern*), 荷 (*Last*), 下りる (*wurde genommen*), jedoch wird die Beziehung zwischen diesen Wörtern nicht extrahiert. Wenn Sie jedoch die Stimmungsanalyse anwenden, können Sie vielfältigere Konzepte hinsichtlich eines Stimmungstyps extrahieren, z. B. das Konzept =肩の荷が下りた, das als "*eine große Last von den Schultern nehmen*", was dem Typ <良い-安心> zugeordnet ist. Im Fall der Stimmungsanalyse wird auch eine große Anzahl an zusätzlichen Typen eingeschlossen. Des Weiteren ermöglicht Ihnen die Wahl einer sekundären Analyse, auch Textlinkanalysis-Ergebnisse zu generieren.

Anmerkung: Wenn ein Sekundär-Analysator genannt wird, dauert der Extrahierungsprozess länger. Für weitere Informationen siehe Thema Wie sekundäre Extrahierung funktioniert in Anhang A auf S. 380.

- **Abhängigkeitsanalyse**. Die Wahl dieser Option führt zu erweiterten Partikeln für die Extrahierungskonzepte aus der grundlegenden Typ- und Stichwortextrahierung. Sie können auch vielfältigere Musterergebnisse aus einer Abhängigkeits-Textlinkanalyse (TLA) gewinnen.
- **Stimmungsanalyse**. Bei dieser Analyse werden zusätzliche Konzepte und – wann immer möglich – TLA-Musterergebnisse extrahiert. Neben den grundlegenden Typen erhalten Sie auch den Vorteil von über 80 Stimmungstypen, z. B. 嬉しい, 吉報, 幸運, 安心, 幸福 usw. Mithilfe dieser Typen werden Konzepte und Muster im Text durch den Ausdruck von Emotionen, Stimmungen und Meinungen aufgedeckt. Es gibt drei Optionen, die den Fokus für die Stimmungsanalyse festlegen: Alle Stimmungen, Nur repräsentative Stimmung und Nur Schlussfolgerungen.

Ausgabe des TLA-Knotens

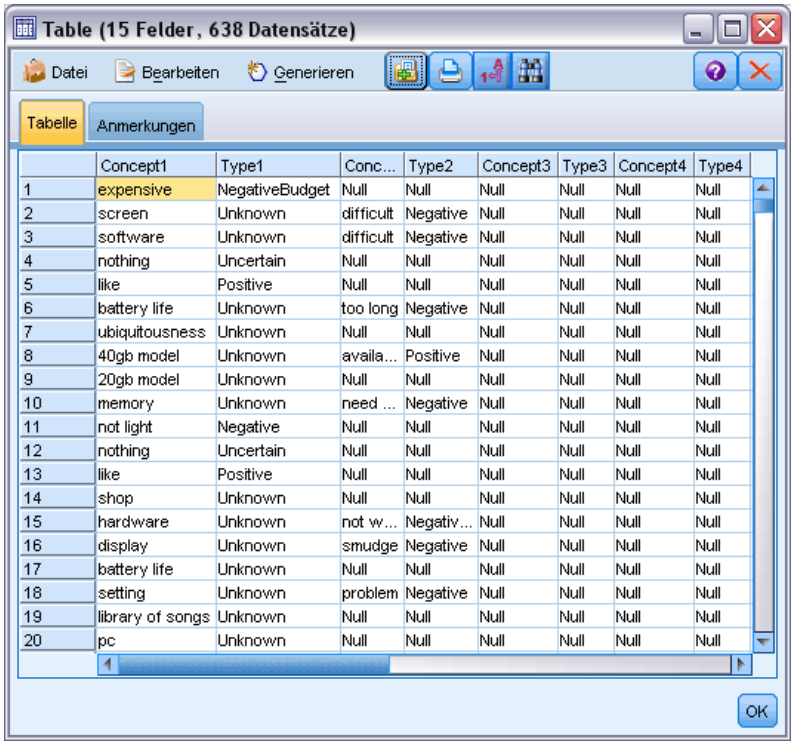
Nach Ausführung des Textlinkanalyseknotens werden die Daten neu strukturiert. Ihnen muss klar sein, wie Text-Mining Ihre Daten umstrukturiert. Wenn Sie eine andere Struktur für das Data-Mining wünschen, können Sie diese mithilfe der Knoten in der Feldoperationen-Palette erreichen. Wenn Sie beispielsweise mit Daten gearbeitet haben, bei denen jede Zeile aus einem Textdatensatz bestand, wird für jedes Muster, das in den Quellentextdaten aufgedeckt wird, eine Zeile erstellt. Es sind 15 Felder für jede Zeile in der Ausgabe vorhanden:

- Sechs Felder (Konzept#, z. B. Konzept1, Konzept2, ... und Konzept6) bezeichnen die beim Musterabgleich ermittelten Konzepte.
- Sechs Felder (Typ#, z. B. Typ1, Typ2, ... und Typ6) geben den Typ des jeweiligen Konzepts an.
- Regelname repräsentiert den Namen der Textlinkregel, mit der der Text abgeglichen und die Ausgabe erzeugt wurde.

- Ein Feld mit dem von Ihnen im Knoten festgelegten Namen des ID-Felds, das die ID des Datensatzes oder Dokuments entsprechend den Eingangsdaten angibt.
- Abgeglichener Text zur Darstellung des Anteils an Textdaten des Ausgangsdatensatzes oder -dokuments, der mit dem TLA-Muster abgeglichen wurde.

Anmerkung: Textlinkanalyse-Musterregeln für japanischen Text erzeugen nur Ergebnisse in einem oder zwei Slots.

Abbildung 4-4
Ausgabe im Tabellenknoten



	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

Anmerkung: Bestehende Streams, die einen Textlinkanalyseknoten aus einer früheren Version als 5.0 enthalten, sind möglicherweise erst nach der Aktualisierung der Knoten vollständig ausführbar. Um von bestimmten Verbesserungen der späteren Version von IBM® SPSS® Modeler zu profitieren, müssen ältere Knoten durch neuere Versionen ersetzt werden, die sich durch eine bessere Bereitstellbarkeit und höhere Leistungsfähigkeit auszeichnen.

Außerdem wird die automatische Übersetzung bestimmter Sprachen ermöglicht. Mit dieser Funktion können Sie Mining auf Dokumente in einer Sprache anwenden, die Sie nicht sprechen und lesen können. Wenn Sie die Übersetzungsfunktion nutzen möchten, müssen Sie Zugriff auf SDL Software as a Service (SaaS) besitzen. Für weitere Informationen siehe Thema Übersetzungseinstellungen in Kapitel 5 auf S. 99.

Caching von TLA-Ergebnissen

Über die Caching-Funktion können Sie die Ergebnisse der Textlinkanalyse im Stream speichern. Wenn Sie vermeiden möchten, dass die Ergebnisse der Textlinkanalyse jedes Mal neu extrahiert werden, wenn der Stream ausgeführt wird, wählen Sie den Textlinkanalyseknoten aus. Anschließend wählen Sie die folgende Befehlsfolge aus den Menüs: Bearbeiten > Knoten > Cache > Aktivieren. Bei der nächsten Ausführung des Streams wird die Ausgabe zwischengespeichert, wobei der Knoten als Cache fungiert. Das Knotensymbol zeigt ein kleines "Dokument", das sich von weiß in grün ändert, wenn der Cache gefüllt wurde. Der Cache bleibt für die Dauer der Sitzung erhalten. Um den Cache einen weiteren Tag zu erhalten (nachdem der Stream geschlossen und erneut geöffnet wurde), wählen Sie den Knoten aus und verwenden Sie folgende Optionen in den Menüs: Bearbeiten > Knoten > Cache > Cache speichern. Wenn Sie den Stream das nächste Mal öffnen, können Sie den gespeicherten Cache neu laden und müssen die Übersetzung nicht erneut ausführen.

Alternativ können Sie einen Knoten-Cache speichern oder aktivieren, indem Sie mit der rechten Maustaste auf den Knoten klicken und im Kontextmenü Cache auswählen.

Verwenden des Textlinkanalyseknotens in einem Stream

Der Textlinkanalyseknoten wird für den Zugriff auf Daten und zum Extrahieren von Konzepten in einem Stream verwendet. Sie können jeden beliebigen Quellenknoten für den Zugriff auf die Daten verwenden.

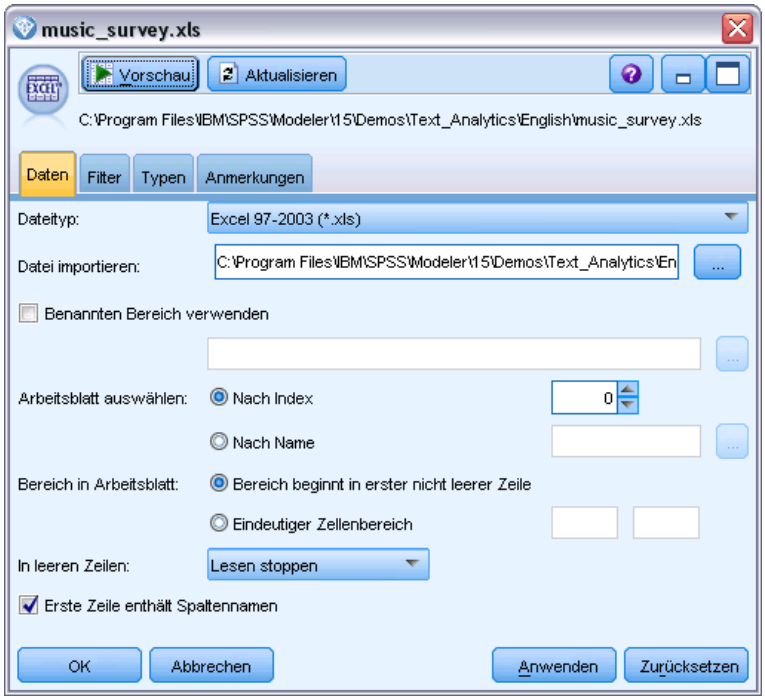
Beispiel: Statistikdateiknoten mit Textlinkanalyseknoten

Das folgende Beispiel zeigt die Verwendung des Textlinkanalyseknotens.

Abbildung 4-5
Beispiel: Statistikdateiknoten mit Textlinkanalyseknoten

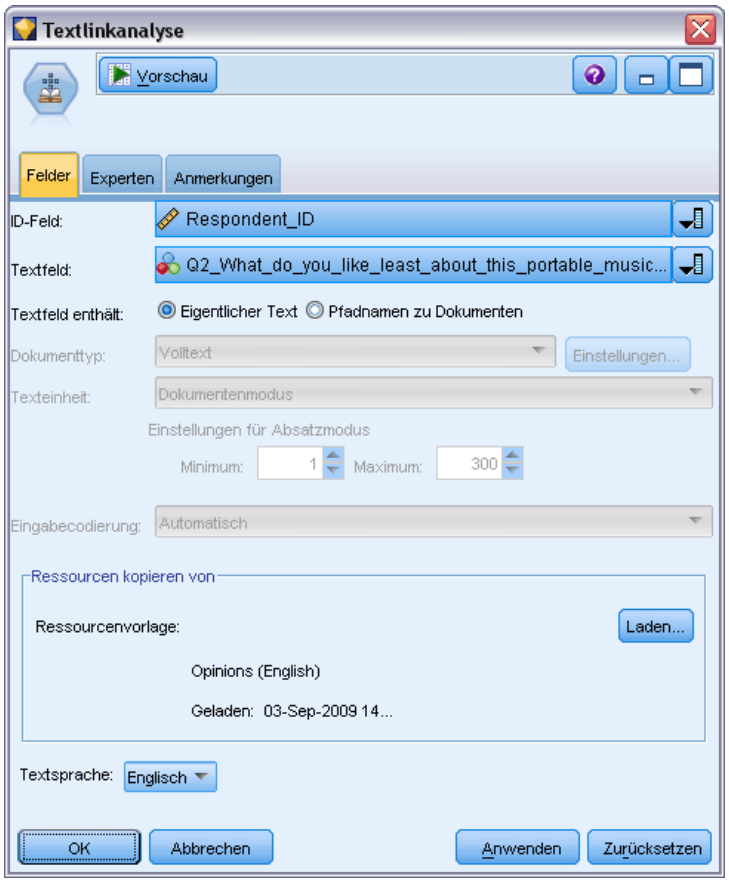
- **Statistikdateiknoten (Registerkarte "Daten").** Zuerst fügten wir diesen Knoten dem Stream hinzu, um anzugeben, wo der Text gespeichert ist.

Abbildung 4-6
Dialogfeld des Knotens für Statistikdateien: Registerkarte "Daten"



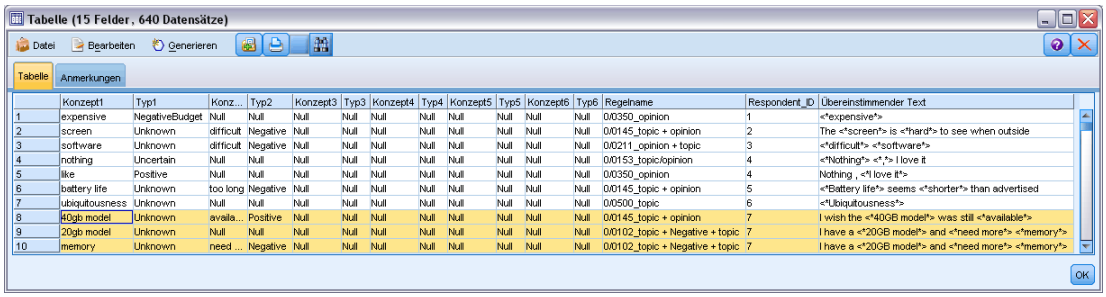
- **Textlinkanalyse-Knoten (Registerkarte "Felder").** Anschließend fügten wir diesen Knoten dem Stream hinzu, um Konzepte für die Downstream-Modellierung bzw. die Anzeige zu extrahieren. Wir gaben das ID-Feld und den Namen des Textfeldes an, das die Daten enthielt, sowie weitere Einstellungen.

Abbildung 4-7
Dialogfeld des Textlinkanalyseknotens: Registerkarte "Felder"



- **Tabellenknoten.** Schließlich fügten wir einen Tabellenknoten hinzu, um die Konzepte anzuzeigen, die aus unseren Textdokumenten extrahiert wurden. Der angezeigten Tabellenausgabe können Sie die Ergebnisse für die TLA-Muster entnehmen, die in den Daten ermittelt wurden, nachdem dieser Stream mit einem Textlinkanalyseknoten ausgeführt wurde. Einige Ergebnisse zeigen, dass nur für ein einziges Konzept bzw. einen einzigen Typ eine Übereinstimmung gefunden wurde. In anderen Fällen fallen die Ergebnisse komplexer aus und enthalten mehrere Typen und Konzepte. Zudem werden mehrere Aspekte der Daten geändert, wenn Daten den Textanalyseknoten durchlaufen und Konzepte extrahiert werden. Die Ausgangsdaten in unserem Beispiel enthielten acht Felder und 405 Datensätze. Nach der Ausführung des Textlinkanalyseknotens gibt es nun 15 Felder und 640 Datensätze. Jetzt ist für jedes erkannte Ergebnis für ein TLA-Muster eine Zeile vorhanden. Für `ID 7` sind gegenüber den Ausgangsdaten drei Zeilen vorhanden, da drei TLA-Musterergebnisse extrahiert wurden. Sie können einen Zusammenführungsknoten verwenden, wenn Sie diese Ausgabedaten wieder mit den Ausgangsdaten zusammenführen möchten.

Abbildung 4-8
Tabellenausgabeknoten



	Konzept1	Typ1	Konz...	Typ2	Konzept3	Typ3	Konzept4	Typ4	Konzept5	Typ5	Konzept6	Typ6	Regelname	Respondent_ID	Übereinstimmender Text
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*>

Kapitel 5

Übersetzen von Text für die Extrahierung

Übersetzungsknoten

Mit dem Übersetzungsknoten können Texte aus unterstützten Sprachen wie dem Arabischen, Chinesischen oder Persischen zur Analyse mit IBM® SPSS® Modeler Text Analytics ins Englische übersetzt werden. Dadurch kann Text-Mining in Dokumenten durchgeführt werden, die in Double-Byte-Sprachen verfasst sind und andernfalls nicht unterstützt würden. Außerdem können Analysten Konzepte aus fremdsprachigen Dokumenten extrahieren, selbst wenn sie die betreffende Sprache nicht beherrschen. Beachten Sie, dass Sie in der Lage sein müssen, eine Verbindung zu Software as a Service (SaaS) von SDL herzustellen, um den Übersetzungsknoten verwenden zu können.

Bei der Durchführung von Text-Mining in einer dieser Sprachen fügen Sie einfach einen Übersetzungsknoten vor dem Text-Mining-Modellierungsknoten in den Stream ein. Außerdem können Sie Caching im Übersetzungsknoten aktivieren, um eine erneute Übersetzung bei jeder Ausführung des Streams zu vermeiden.

Sie finden diesen Knoten auf der SPSS Modeler Text Analytics -Registerkarte der Knotenpalette am unteren Rand des IBM® SPSS® Modeler-Fensters. Für weitere Informationen siehe Thema IBM SPSS Modeler Text Analytics-Knoten in Kapitel 1 auf S. 11.

Cachen der Übersetzung. Wenn Sie die Übersetzung cachen, wird der übersetzte Text in einem Stream gespeichert und nicht in externen Dateien. Um zu vermeiden, dass die Übersetzung bei jeder Ausführung des Streams wiederholt werden muss, wählen Sie den Übersetzungsknoten aus und wählen Sie in den Menüs folgende Optionen: Bearbeiten > Knoten > Cache > Aktivieren. Bei der nächsten Ausführung des Streams wird die Ausgabe aus der Übersetzung zwischengespeichert, wobei der Knoten als Cache fungiert. Das Knotensymbol zeigt ein kleines “Dokument”, das sich von weiß in grün ändert, wenn der Cache gefüllt wurde. Der Cache bleibt für die Dauer der Sitzung erhalten. Um den Cache einen weiteren Tag zu erhalten (nachdem der Stream geschlossen und erneut geöffnet wurde), wählen Sie den Knoten aus und verwenden Sie folgende Optionen in den Menüs: Bearbeiten > Knoten > Cache > Cache speichern. Wenn Sie den Stream das nächste Mal öffnen, können Sie den gespeicherten Cache neu laden und müssen die Übersetzung nicht erneut ausführen.

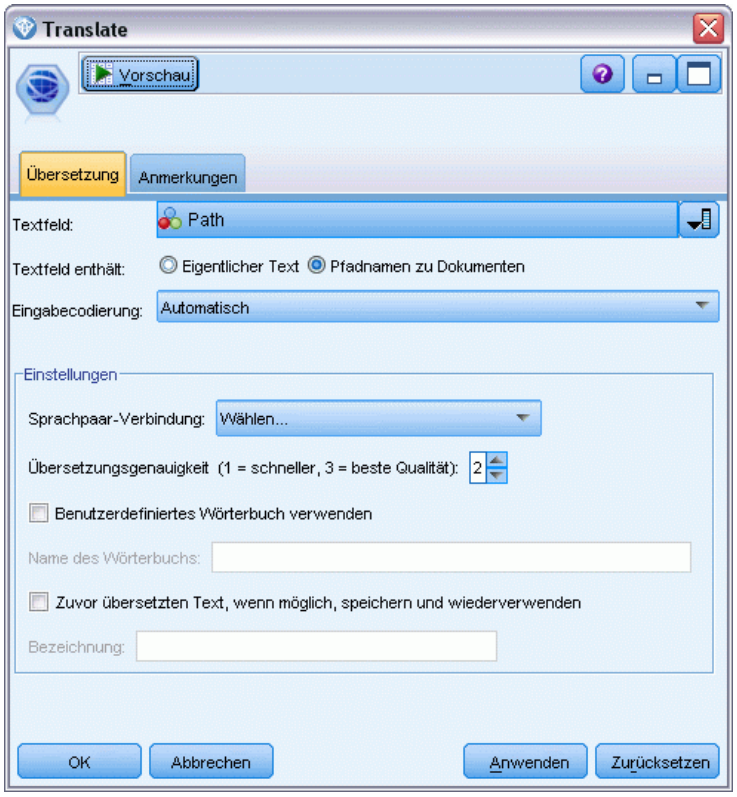
Alternativ können Sie einen Knoten-Cache speichern oder aktivieren, indem Sie mit der rechten Maustaste auf den Knoten klicken und im Kontextmenü Cache auswählen.

Wichtig: Wenn Sie versuchen, Informationen über das Internet durch einen Proxy-Server abzurufen, müssen Sie den Proxy-Server in der Datei `net.properties` für die Client- und Server-Version von SPSS Modeler Text Analytics aktivieren. Befolgen Sie die in der Datei enthaltenen Anweisungen. Dies trifft zu, wenn Sie über den Web-Feed-Knoten auf das Internet zugreifen oder eine Lizenz für SDL Software as a Service (SaaS) abrufen, da diese Verbindungen über Java aufgebaut werden. Beim Client befindet sich diese Datei

standardmäßig im Verzeichnis *C:\Program Files\IBM\SPSS\Modeler\15\jre\lib\net.properties*. Beim Server befindet sich diese Datei standardmäßig im Verzeichnis *C:\Program Files\IBM\SPSS\Modeler\15\ext\bin\spss.TMWBServer\jre\lib\net.properties*.

Übersetzungsknoten: Registerkarte "Übersetzung"

Abbildung 5-1
Dialogfeld des Übersetzungsknotens: Registerkarte "Felder"



Textfeld. Wählen Sie das Feld, das den Text für das Mining enthält, den Dokumentenpfadnamen bzw. den Verzeichnispfadnamen, der zu den Dokumenten führt. Dieses Feld hängt von der Datenquelle ab. Sie können ein beliebiges Zeichenkettenfeld angeben, sogar solche mit `Direction=None` oder `Type=Typeless`.

Textfeld enthält. Geben Sie an, was das Textfeld enthalten soll, das in der vorausgehenden Einstellung festgelegt wurde. Folgende Optionen stehen zur Auswahl:

- **Eigentlicher Text.** Wählen Sie diese Option aus, wenn das Feld genau den Text enthält, aus dem Konzepte extrahiert werden sollen.
- **Pfadnamen zu Dokumenten.** Wählen Sie diese Option aus, wenn das Feld einen oder mehrere Pfadnamen zu externen Dokumenten enthält, deren Textinhalt extrahiert werden soll. Diese Option sollte beispielsweise ausgewählt werden, wenn ein Dateilistenknoten verwendet wird, um eine Liste von Dokumenten einzulesen. Für weitere Informationen siehe Thema Dateilistenknoten in Kapitel 2 auf S. 14.

Eingabekodierung. Wählen Sie die Kodierung des Quelltexts. Sie können mit der Auswahl der Option Automatisch beginnen. Wenn Sie jedoch bemerken, dass einige Dateien nicht korrekt verarbeitet werden, sollten Sie die tatsächliche Kodierung hier aus der Liste wählen. Die Option "Automatisch" erkennt die Kodierung bei kurzen Texten wie kurzen Datensätzen eventuell nicht richtig. Die Textausgabe von diesem Knoten wird in `UTF-8` kodiert.

Einstellungen. Gibt die Übersetzungseinstellungen für den Stream an.

- **Sprachpaar-Verbindung**. Wählen Sie das zu verwendende Sprachpaar aus. Verfügbare Sprachpaare werden automatisch in dieser Liste angezeigt, nachdem Sie in der Registerkarte Übersetzungseinstellungen den Link zum SDL-Dienst eingerichtet haben. Für weitere Informationen siehe Thema Übersetzungseinstellungen auf S. 99.
- **Übersetzungsgenauigkeit**. Geben Sie die gewünschte Genauigkeit an, indem Sie einen Wert von 1 bis 3 wählen, um das gewünschte Verhältnis zwischen Geschwindigkeit und Genauigkeit festzulegen. Ein niedrigerer Wert führt zu einer schnelleren Übersetzung, jedoch auch zu einer geringeren Genauigkeit. Ein höherer Wert führt zu Ergebnissen mit größerer Genauigkeit, jedoch höherer Verarbeitungszeit. Zur Zeitoptimierung wird empfohlen, mit einer niedrigen Stufe zu beginnen und diese nur zu erhöhen, wenn Sie nach einer Überprüfung der Ergebnisse das Gefühl haben, dass größere Genauigkeit erforderlich ist.
- **Benutzerdefiniertes Wörterbuch verwenden**. Wenn Sie zuvor benutzerdefinierte Wörterbücher erstellt haben, die bei SDL gespeichert sind, können Sie sie zusammen mit der Übersetzung verwenden. Aktivieren Sie zur Auswahl eines benutzerdefinierten Wörterbuchs die Option Benutzerdefiniertes Wörterbuch und geben Sie den Namen unter Wörterbuchname ein Zur Verwendung mehrerer Wörterbücher, müssen Sie die Namen mit Kommas trennen.
- **Nach Möglichkeit vorherigen Übersetzungstext speichern und wiederverwenden.** Gibt an, dass die Übersetzungsergebnisse gespeichert werden. Wenn dann bei der nächsten Ausführung des Streams dieselben Datensätze/Dokumente vorhanden sind, wird der Inhalt als identisch angesehen und die Übersetzungsergebnisse werden wiederverwendet, um die Verarbeitung zu beschleunigen. Wenn diese Option bei der Ausführung ausgewählt ist und die Zahl der Datensätze nicht mit der zuletzt gespeicherten Zahl übereinstimmt, wird der Text vollständig übersetzt und anschließend unter dem Labelnamen für die nächste Ausführung gespeichert. Diese Option ist nur verfügbar, wenn Sie eine SDL-Übersetzungssprache ausgewählt haben.

  Anmerkung: Wenn der Text im Stream gespeichert wird, können Sie auch die Caching-Funktion in einem Übersetzungsknoten aktivieren. In diesem Fall wird nicht nur das Übersetzungsergebnis erneut verwendet, sondern alle Eingaben weiter oben im Stream werden ignoriert, wenn der Cache verfügbar ist.
- **Label.** Wenn Sie die Option Nach Möglichkeit vorherigen Übersetzungstext speichern und wiederverwenden auswählen, geben Sie einen Labelnamen für die Ergebnisse an. Dieses Label wird verwendet, um den vorherigen Übersetzungstext zu identifizieren. Wenn kein Label angegeben ist, wird bei Ausführung des Streams eine Warnung zu den Stream-Eigenschaften hinzugefügt. In diesem Fall ist eine Wiederverwendung ausgeschlossen.

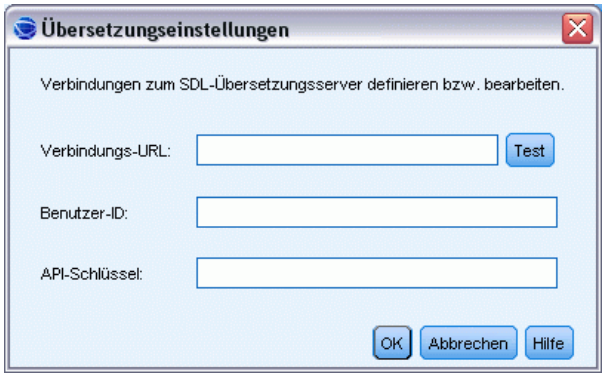
Übersetzungseinstellungen

In diesem Dialogfeld können Sie die Übersetzungsverbindung mit SDL Software as a Service (SaaS) definieren und verwalten, die Sie dann bei jeder Übersetzung wieder verwenden können. Sobald Sie hier eine Verbindung definiert haben, können Sie beim Übersetzen rasch eine Sprachpaar-Verbindung wählen, ohne erneut alle Verbindungseinstellungen eingeben zu müssen.

Eine Sprachpaar-Verbindung identifiziert die Quell- und Zielsprache sowie die Verbindungsdetails für den Server. Beispiel: *Chinesisch – Englisch* bedeutet, dass der Quelltext in Chinesisch ist und die resultierende Übersetzung in Englisch sein wird. Jede Verbindung, auf die Sie über die SDL-Onlinedienste zugreifen möchten, muss manuell definiert werden.

Wichtig: Wenn Sie versuchen, Informationen über das Internet durch einen Proxy-Server abzurufen, müssen Sie den Proxy-Server in der Datei `net.properties` für die Client- und Server-Version von IBM® SPSS® Modeler Text Analytics aktivieren. Befolgen Sie die in der Datei enthaltenen Anweisungen. Dies trifft zu, wenn Sie über den Web-Feed-Knoten auf das Internet zugreifen oder eine Lizenz für SDL Software as a Service (SaaS) abrufen, da diese Verbindungen über Java aufgebaut werden. Beim Client befindet sich diese Datei standardmäßig im Verzeichnis *C:\Program Files\IBM\SPSS\Modeler\15\jre\lib\net.properties*. Beim Server befindet sich diese Datei standardmäßig im Verzeichnis *C:\Program Files\IBM\SPSS\Modeler\15\ext\bin\spss.TMWBServer\jre\lib\net.properties*.

Abbildung 5-2
Dialogfeld "Übersetzungseinstellungen"



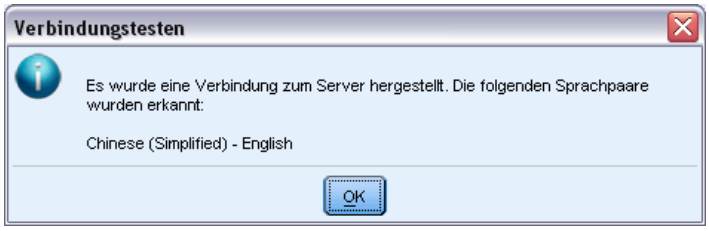
Verbindungs-URL. Geben Sie die URL für die Verbindung mit SDL Software as a Service ein.

Benutzer-ID. Geben Sie die eindeutige ID ein, die Ihnen von SDL bereitgestellt wurde.

API-Schlüssel. Geben Sie den Schlüssel ein, der Ihnen von SDL bereitgestellt wurde.

Test. Klicken Sie auf Test, um sicherzustellen, dass die Verbindung korrekt konfiguriert ist, und um das bzw. die gefundenen Sprachpaare auf dieser Verbindung anzuzeigen.

Abbildung 5-3
Meldung über erfolgreiche Verbindung



Verwenden des Übersetzungsknotens

Fügen Sie einfach vor einem Text-Mining-Knoten im Stream einen Übersetzungsknoten ein, um Konzepte aus unterstützten Übersetzungssprachen wie Arabisch, Chinesisch oder Persisch zu extrahieren.

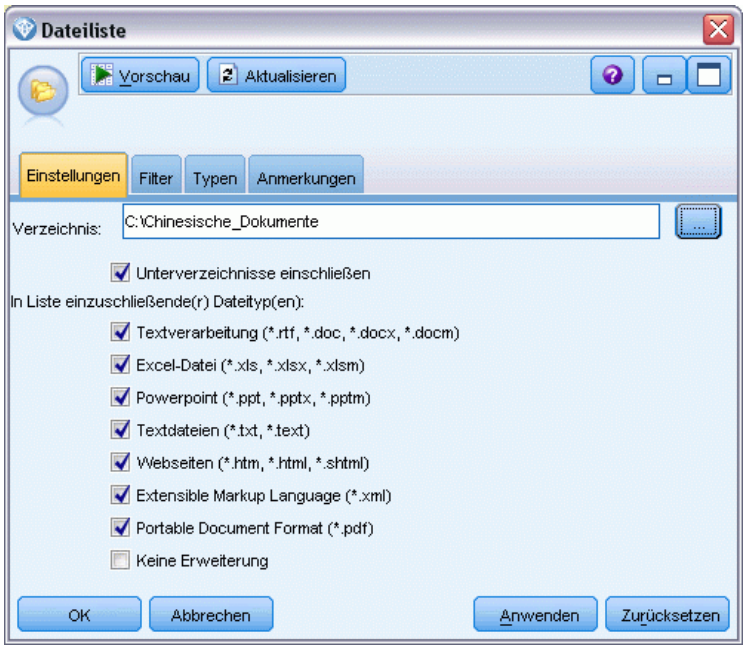
Beispiel: Übersetzen von Text in externen Dokumenten

Wenn der zu übersetzende Text in einer oder mehreren externen Dateien enthalten ist, kann ein Dateilistenknoten zum Einlesen einer Namensliste verwendet werden. In diesem Fall würde der Übersetzungsknoten zwischen dem Dateilistenknoten und allen nachfolgenden Text-Mining-Knoten hinzugefügt und die Ausgabe befände sich in dem Verzeichnis, in dem auch der Übersetzungstext gespeichert ist.

Abbildung 5-4
Beispiel-Stream: Dateilistenknoten mit Übersetzungsknoten

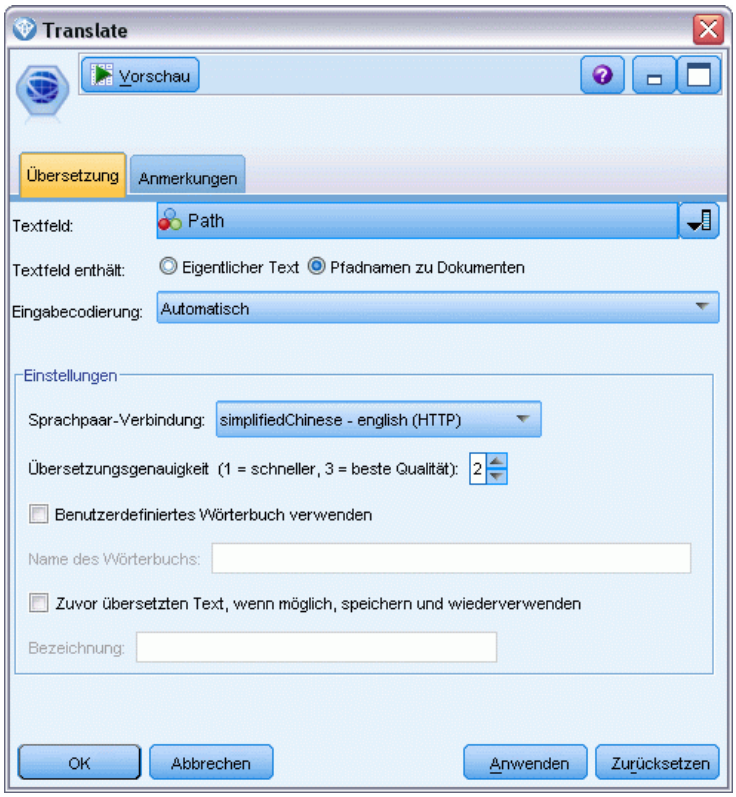
- **Dateilistenknoten (Registerkarte "Einstellungen").** Im Dateilistenknoten haben wir die Ausgangsdokumente ausgewählt.

Abbildung 5-5
Dialogfeld des Dateilistenknotens: Registerkarte "Einstellungen"



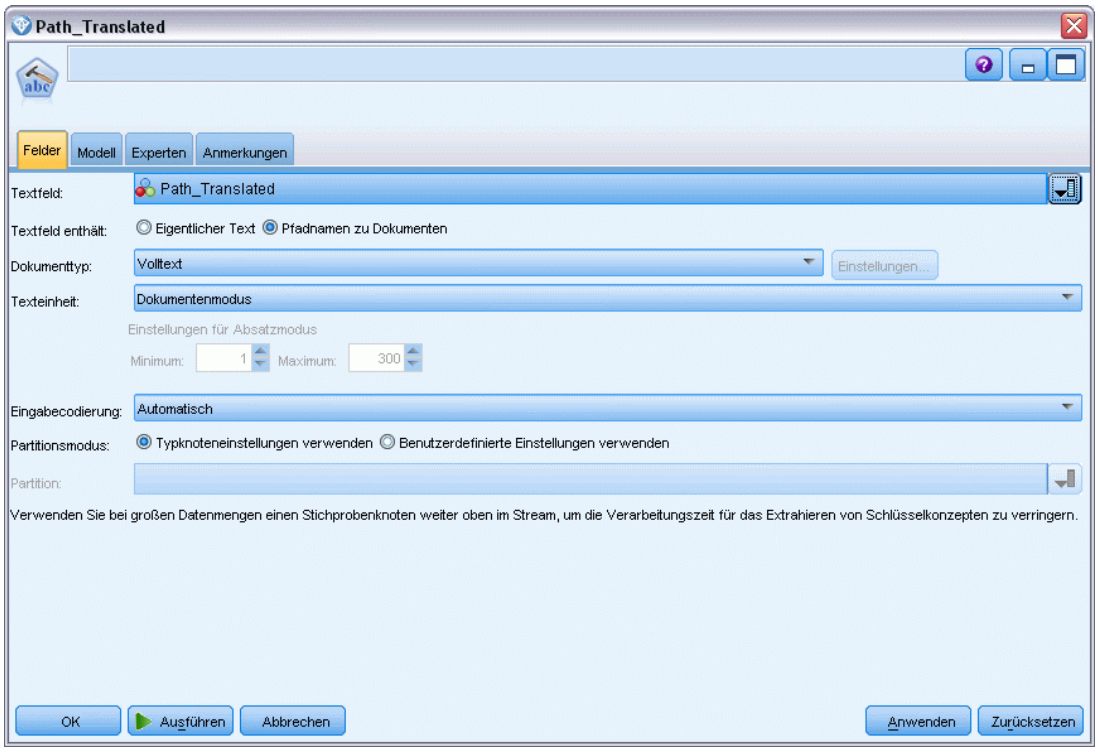
- **Übersetzungsknoten (Registerkarte "Übersetzung").** Als Nächstes haben wir einen Übersetzungsknoten hinzugefügt und angeschlossen. Im Knoten wählten wir das vom Dateilistenknoten erstellte Feld aus – standardmäßig *Pfad* genannt –, das den ursprünglichen Speicherort der Dateien angibt. In derselben Registerkarte wählen wir ein vordefiniertes Sprachpaar aus.

Abbildung 5-6
Dialogfeld des Übersetzungsknotens: Registerkarte "Übersetzung"



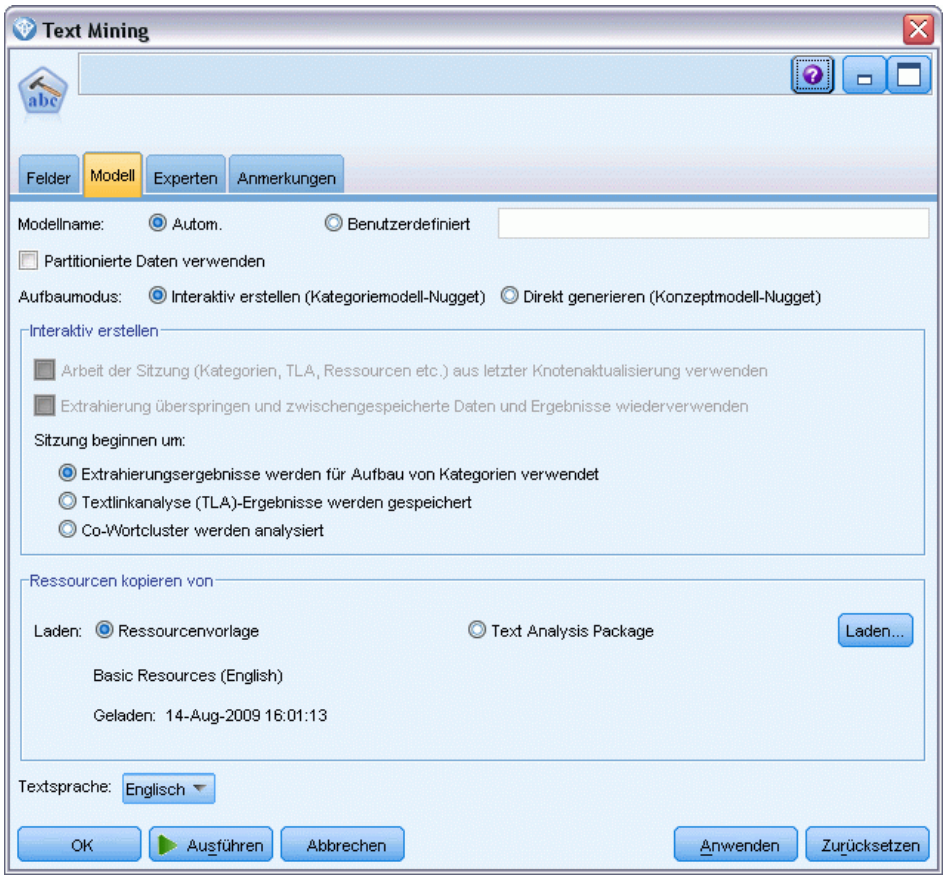
- **Text-Mining-Knoten (Registerkarte "Felder").** In allen nachfolgenden Text-Mining-Knoten wählten wir das neue, durch den Übersetzungsknoten erstellte Feld aus, —benannt nach dem Textfeld aus dem Dateilistenknoten und gefolgt vom Zusatz *_Translated* —, das den Speicherort der übersetzten Dateien angibt.

Abbildung 5-7
Dialogfeld des Text-Mining-Modellierungsknotens: Registerkarte "Felder"



- **Text-Mining-Knoten (Registerkarte "Modell").** Auf der Registerkarte "Modell" haben wir Englisch als Sprache ausgewählt.

Abbildung 5-8
Dialogfeld des Text-Mining-Knotens: Registerkarte "Modell"



Kapitel 6

Durchsuchen von Text aus externen Quellen

Datei-Viewer-Knoten

Wenn Sie ein Text-Mining für mehrere Dokumente durchführen, können Sie die vollständigen Pfadnamen der Dateien direkt in Ihre Text-Mining-Modellierungs- und Übersetzungsknoten eingeben. Wenn die Ausgabe jedoch in einen Tabellenknoten erfolgt, wird nur der vollständige Pfadname des Dokuments und nicht der darin enthaltene Text angezeigt. Der Datei-Viewer-Knoten kann als Analogon für den Tabellenknoten verwendet werden und ermöglicht den Zugriff auf den eigentlichen Text in den einzelnen Dokumenten, ohne sie in eine einzelne Datei zusammenführen zu müssen.

Der Datei-Viewer-Knoten kann ein besseres Verständnis der Ergebnisse aus der Textextrahierung ermöglichen, indem er Ihnen Zugriff auf den Quellentext bzw. auf den unübersetzten Text gewährt, aus dem die Konzepte extrahiert wurden, da er andernfalls im Stream nicht zugänglich wäre. Dieser Knoten wird nach einem Dateilistenknoten zum Stream hinzugefügt, um die Verknüpfungen zu sämtlichen Dateien aufzulisten.

Das Ergebnis des Knotens ist ein Fenster, in dem alle Dokumentenelemente angezeigt werden, die gelesen und zum Extrahieren von Konzepten verwendet wurden. Aus diesem Fenster können Sie auf ein Symbol in der Symbolleiste klicken, um den Bericht in einem externen Browser zu starten, wobei Dokumentennamen als Hyperlinks aufgeführt werden. Sie können auf einen Link klicken, um das entsprechende Dokument in der Sammlung zu öffnen. Für weitere Informationen siehe Thema Verwenden des Datei-Viewer-Knotens auf S. 106.

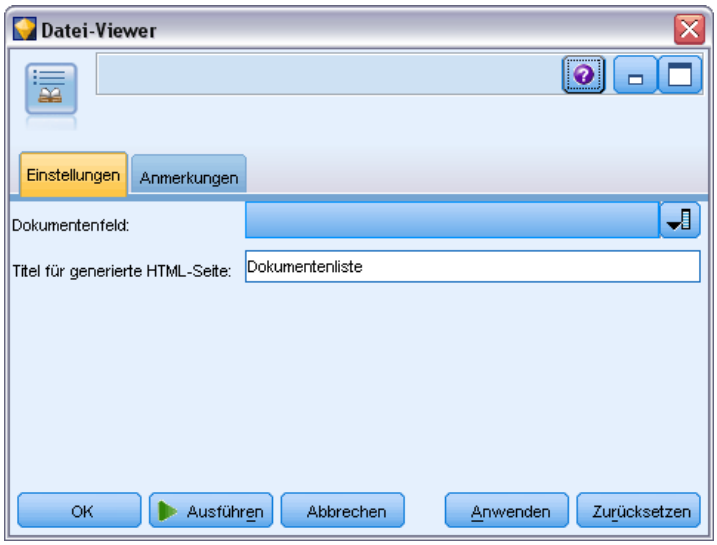
Sie finden diesen Knoten auf der IBM® SPSS® Modeler Text Analytics -Registerkarte der Knotenpalette am unteren Rand des IBM® SPSS® Modeler-Fensters. Für weitere Informationen siehe Thema IBM SPSS Modeler Text Analytics-Knoten in Kapitel 1 auf S. 11.

Anmerkung: Wenn Sie im Client-Server-Modus arbeiten und die Datei-Viewer-Knoten Teil des Streams sind, müssen Dokumentensammlungen in einem Webserver-Verzeichnis auf dem Server gespeichert werden. Da der Text-Mining-Ausgabeknoten eine Liste der im Webserver-Verzeichnis gespeicherten Dokumente erstellt, verwalten die Sicherheitseinstellungen des Webservers die Berechtigungen für diese Dokumente.

Einstellungen für Datei-Viewer-Knoten

Das folgende Dialogfeld wird zur Angabe der Einstellungen für den Datei-Viewer-Knoten verwendet

Abbildung 6-1
Dialogfeld des Datei-Viewer-Knotens: Registerkarte "Einstellungen"



Dokumentenfeld. Wählen Sie in Ihren Daten das Feld aus, das den vollständigen Namen und den Pfad der anzuzeigenden Dokumente enthält.

Titel für generierte HTML-Seite. Dient zum Erstellen eines Titels oben auf der Seite, der die Liste der Dokumente enthält.

Verwenden des Datei-Viewer-Knotens

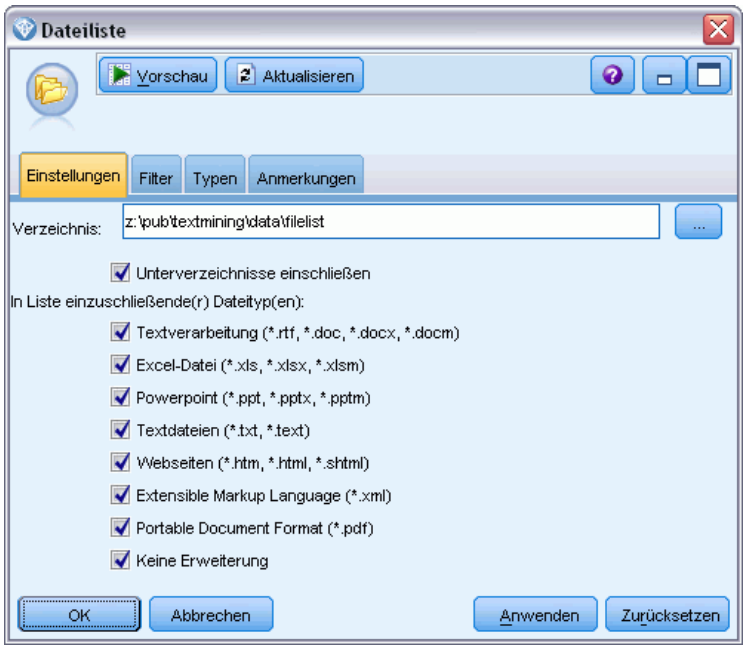
Das folgende Beispiel zeigt die Verwendung des Datei-Viewer-Knotens.

Beispiel: Dateilistenknoten und Datei-Viewer-Knoten

Abbildung 6-2
Stream, der die Verwendung des Datei-Viewer-Knotens erläutert

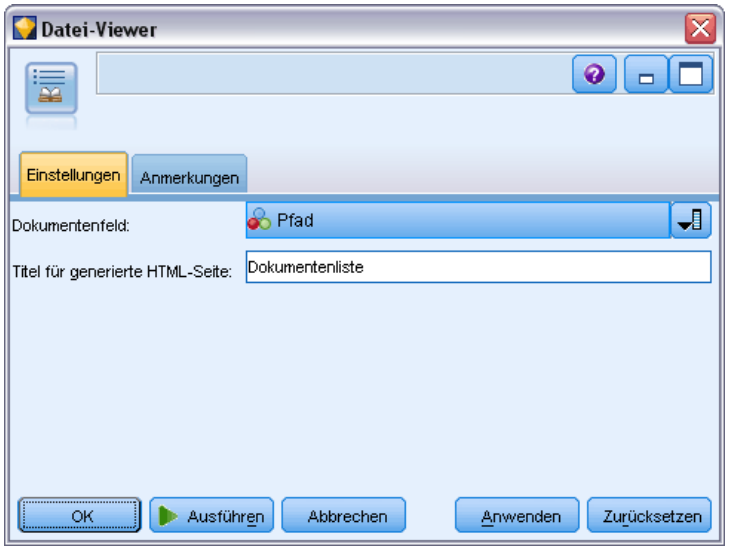
- **Dateilistenknoten (Registerkarte "Einstellungen").** Zuerst fügten wir diesen Knoten hinzu, um anzugeben, wo sich die Dokumente befinden.

Abbildung 6-3
Dialogfeld des Dateilistenknotens: Registerkarte "Einstellungen"



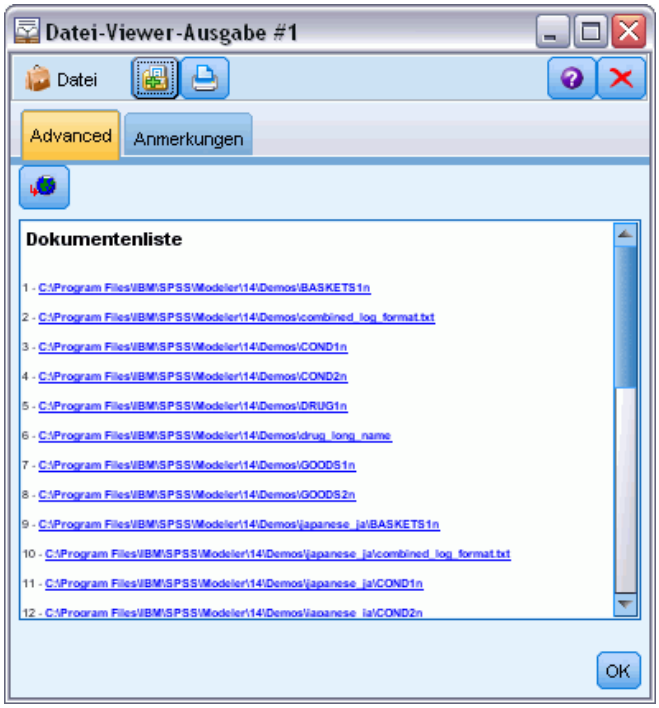
- **Datei-Viewer-Knoten (Registerkarte "Einstellungen").** Anschließend fügten wir den Datei-Viewer-Knoten hinzu, um die Dokumente in der HTML-Ansicht aufzulisten.

Abbildung 6-4
Dialogfeld des Datei-Viewer-Knotens: Registerkarte "Einstellungen"



- **Dialogfeld "Datei-Viewer-Ausgabe".** Als Nächstes führten wir den Stream aus, der die Liste der Dokumente in einem neuen Fenster ausgibt.

Abbildung 6-5
Datei-Viewer-Ausgabe



- Um die Dokumente anzuzeigen, klickten wir auf die Symbolleistenschaltfläche, auf der ein Globus mit einem roten Pfeil dargestellt ist. Dadurch wurde in unserem Browser eine Liste mit Hyperlinks zu Dokumenten angezeigt.

Kapitel
7

Knoteneigenschaften für Skripts

IBM® SPSS® Modeler ist mit einer Skriptsprache ausgestattet, mit der Sie Streams über die Befehlszeile ausführen können. Hier erhalten Sie Informationen über die Knoteneigenschaften jedes der Knoten, die mit IBM® SPSS® Modeler Text Analytics geliefert werden. Weitere Informationen zum Standardknotenset, das mit SPSS Modeler geliefert wird, finden Sie im Skript- und Automatisierungshandbuch.

Dateilistenknoten: filelistnode

Sie können die Eigenschaften in der folgenden Tabelle für Skripts verwenden. Der Knoten als solcher heißt filelistnode.

Tabelle 7-1
Dateilistenknoten – Skript-Eigenschaften

Skript-Eigenschaften	Datentyp
path	*Zeichenkette*
recurse	*Flag*
word_processing	*Flag*
excel_file	*Flag*
powerpoint_file	*Flag*
text_file	*Flag*
web_page	*Flag*
xml_file	*Flag*
pdf_file	*Flag*
no_extension	*Flag*

Anmerkung: Der Parameter “Liste erstellen” ist nicht mehr verfügbar und alle Skripts mit dieser Option werden automatisch in die Ausgabe “Dateien” umgewandelt.

Web-Feed-Knoten: webfeednode

Sie können die Eigenschaften in der folgenden Tabelle für Skripts verwenden. Der Knoten als solcher heißt webfeednode.

Tabelle 7-2
Skript-Eigenschaften des Web-Feed-Knotens

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
urls	*string1* *string2* *...stringn*	Alle URLs werden in der Listenstruktur angegeben. URL-Liste durch “\n” getrennt.
recent_entries	*Flag*	
limit_entries	*Ganze Zahl*	Anzahl der aktuellen, pro URL zu lesenden Einträge.

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
use_previous	*Flag*	Zum Speichern und Wiederverwenden von Web-Feed-Cache.
use_previous_label	*Zeichenkette*	Name des gespeicherten Web-Cache.
start_record	*Zeichenkette*	Nicht-RSS-Anfangs-Tag.
url*n*.title	*Zeichenkette*	Für alle in der Liste enthaltenen URLs muss auch hier ein Eintrag definiert sein. Der erste Eintrag lautet url1.title, wobei die Nummer die Position in der URL-Liste anzeigt. Dies ist der Anfangs-Tag mit dem Titel des Inhalts.
url*n*.short_description	*Zeichenkette*	Genau wie bei url*n*.title.
url*n*.description	*Zeichenkette*	Genau wie bei url*n*.title.
url*n*.authors	*Zeichenkette*	Genau wie bei url*n*.title.
url*n*.contributors	*Zeichenkette*	Genau wie bei url*n*.title.
url*n*.published_date	*Zeichenkette*	Genau wie bei url*n*.title.
url*n*.modified_date	*Zeichenkette*	Genau wie bei url*n*.title.
html_alg	None HTMLCleaner	Methode zur Inhaltsfilterung.
discard_lines	*Flag*	Kurze Zeilen verwerfen. Verwendet mit min_words
min_words	*Ganze Zahl*	Minimale Anzahl an Wörtern.
discard_words	*Flag*	Kurze Zeilen verwerfen. Verwendet mit min_avg_len
min_avg_len	*Ganze Zahl*	
discard_scw	*Flag*	Zeilen mit vielen Wörtern aus einzelnen Zeichen verwerfen. Verwendet mit max_scw
max_scw	*Ganze Zahl*	Maximaler Anteil (0-100 Prozent) an Einzelzeichenwörtern in einer Zeile
discard_tags	*Flag*	Zeilen mit bestimmten Tags verwerfen.
tags	*Zeichenkette*	Nach Sonderzeichen muss ein Schrägstrich (\) stehen.
discard_spec_words	*Flag*	Zeilen mit bestimmten Zeichenketten verwerfen.
words	*Zeichenkette*	Nach Sonderzeichen muss ein Schrägstrich (\) stehen.

Text-Mining-Knoten: TextMiningWorkbench

Sie können die folgenden Parameter verwenden, um einen Knoten über Skripts zu definieren oder zu aktualisieren. Der Knoten als solcher heißt TextMiningWorkbench.

Wichtig: Es ist nicht möglich, eine andere Ressourcenvorlage über Skripts festzulegen. Wenn Sie glauben, dass Sie eine Vorlage benötigen, wählen Sie im Dialogfeld des Knotens eine aus.

Tabelle 7-3
Text-Mining-Modellierungsknoten – Skript-Eigenschaften

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
Text	*Feld*	
Methode	ReadText ReadPath	

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
Dok.typ	*Ganze Zahl*	Zulässige Werte (0,1,2), wobei 0 = Volltext, 1 = Gegliederter Text und 2 = XML
Kodierung	Automatisch "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Beachten Sie, dass Werte mit Sonderzeichen, wie "UTF-8", in Anführungszeichen gesetzt werden sollten, damit keine Verwechslung mit einem mathematischen Operator entsteht.
unity	*Ganze Zahl*	Zulässige Werte (0,1,2), wobei 0 = Absatz und 1 = Dokument
para_min	*Ganze Zahl*	
para_max	*Ganze Zahl*	
mtag	*Zeichenkette*	Enthält sämtliche mtag-Einstellungen (aus dem Dialogfeld “Einstellungen” für XML-Dateien)
mclef	*Zeichenkette*	Enthält sämtliche mclef-Einstellungen (aus dem Dialogfeld “Einstellungen” für Dateien mit gegliedertem Text)
partition	*Feld*	
custom_field	*Flag*	Zeigt an, ob ein Partitionsfeld angegeben wird oder nicht.
use_model_name	*Flag*	
model_name	*Zeichenkette*	
use_partitioned_data	*Flag*	Wenn ein Partitionsfeld definiert ist, werden nur die Trainingsdaten für die Modellerstellung verwendet.
model_output_type	Interactive Model	Interactive führt zu einem Kategoriemodell. Model führt zu einem Konzeptmodell.
use_interactive_info	*Flag*	Nur zum interaktiven Erstellen in einer Workbench-Sitzung.
reuse_extraction_results	*Flag*	Nur zum interaktiven Erstellen in einer Workbench-Sitzung.
interactive_view	Categories TLA Clusters	Nur zum interaktiven Erstellen in einer Workbench-Sitzung.
extract_top	*Ganze Zahl*	Dieser Parameter wird verwendet, wenn gilt: model_type = Concept
use_check_top	*Flag*	
check_top	*Ganze Zahl*	
use_uncheck_top	*Flag*	
uncheck_top	*Ganze Zahl*	

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
language	de en es fr it ja nl pt	
frequency_limit	*Ganze Zahl*	Ab Version 14.0 abgeschafft.
concept_count_limit	*Ganze Zahl*	Extrahierung einschränken auf Konzepte mit globaler Häufigkeit von mindestens diesem Wert Für japanischen Text nicht verfügbar.
fix_punctuation	*Flag*	Für japanischen Text nicht verfügbar.
fix_spelling	*Flag*	Für japanischen Text nicht verfügbar.
spelling_limit	*Ganze Zahl*	Für japanischen Text nicht verfügbar.
extract_uniterm	*Flag*	Für japanischen Text nicht verfügbar.
extract_nonlinguistic	*Flag*	Für japanischen Text nicht verfügbar.
upper_case	*Flag*	Für japanischen Text nicht verfügbar.
group_names	*Flag*	Für japanischen Text nicht verfügbar.
permutation	*Ganze Zahl*	Maximale Anzahl von Füllwörtern in zusammengesetzten Konzepten (Standard ist 3). Für japanischen Text nicht verfügbar.
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	Nur für die Extrahierung von japanischem Text. *Anmerkung*: Verfügbar in IBM® SPSS® Modeler Premium. 0 = Sekundäre Stimmungsextrahierung 1 = Abhängigkeitsextrahierung 2 = Kein sekundäres Analyseset.
jp_algorithm_sense_mode	0 1 2	Nur für die Extrahierung von japanischem Text. *Anmerkung*: Verfügbar in SPSS Modeler Premium. 0 = Nur Schlussfolgerungen 2 = Nur repräsentative Stimmung 3 = Alle Stimmungen

Text-Mining-Modell-Nugget: TMWBModelApplier

Sie können die Eigenschaften in der folgenden Tabelle für Skripts verwenden. Das Nugget als solches heißt TMWBModelApplier.

Tabelle 7-4
Text-Mining-Modell-Nugget – Eigenschaften

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
scoring_mode	Fields Records	
field_values	Flags Counts	Diese Option steht im Kategoriemodell-Nugget nicht zur Verfügung. Für Flags auf TRUE oder FALSE setzen

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
true_value	*Zeichenkette*	Dient bei Flags zur Definition des Werts für "True".
false_value	*Zeichenkette*	Dient bei Flags zur Definition des Werts für "False".
extension_concept	*Zeichenkette*	Dient zur Angabe einer Erweiterung für den Feldnamen. Feldnamen werden unter Verwendung des Konzeptnamens und dieser Erweiterung generiert. Geben Sie mithilfe des Werts add_as an, wo diese Erweiterung gesetzt werden soll.
extension_category	*Zeichenkette*	Feldnamenerweiterung. Sie können ein Erweiterungspräfix/-suffix für den Feldnamen angeben oder die Kategoriecodes verwenden. Feldnamen werden unter Verwendung des Kategorienamens und dieser Erweiterung generiert. Geben Sie mithilfe des Werts add_as an, wo diese Erweiterung gesetzt werden soll.
add_as	Suffix Prefix	
fix_punctuation	*Flag*	
excluded_subcategories_descriptors	RollUpToParent Ignore	Nur für Kategoriemodelle. Wenn eine Unterkategorie nicht ausgewählt ist. Mit dieser Option können Sie angeben, wie die Deskriptoren, die zu nicht für das Scoring ausgewählten Unterkategorien gehören, behandelt werden. Es gibt zwei Optionen. ■ Ignore. Die Option "Deskriptoren vollständig aus Scoring ausschließen" verursacht, dass Deskriptoren von Unterkategorien ohne Häkchen (nicht ausgewählt) beim Scoring ignoriert und nicht verwendet werden. ■ RollUpToParent. Die Option "Deskriptoren mit jenen in der übergeordneten Kategorie verbinden" verursacht, dass Deskriptoren von Unterkategorien ohne Häkchen (nicht ausgewählt) als Deskriptoren für die übergeordnete Kategorie (die Kategorie über dieser Unterkategorie) verwendet werden. Wenn mehrere Ebenen von Unterkategorien nicht ausgewählt sind, werden die Deskriptoren unter der ersten verfügbaren übergeordneten Kategorie zusammengefasst.
check_model	*Flag*	In Version 14 abgeschafft
text	*Feld*	
method	ReadText ReadPath	
docType	*Ganze Zahl*	Mit den möglichen Werten (0,1,2), wobei gilt: 0 = Full Text, 1 = Structured Text und 2 = XML

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
Kodierung	Automatisch "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Beachten Sie, dass Werte mit Sonderzeichen, wie "UTF-8", in Anführungszeichen gesetzt werden sollten, damit keine Verwechslung mit einem mathematischen Operator entsteht.
language	de en es fr it ja nl pt	

Textlinkanalyseknoten: textlinkanalysis

Sie können die Parameter in der folgenden Tabelle verwenden, um einen Knoten über Skripts zu definieren oder zu aktualisieren. Der Knoten als solcher heißt textlinkanalysis.

Wichtig: Es ist nicht möglich, eine Ressourcenvorlage über Skripts festzulegen. Vorlagen können nur im Knotendialogfeld ausgewählt werden.

Tabelle 7-5
Skript-Eigenschaften von Textlinkanalyseknoten (TLA-Knoten)

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
id_field	*Feld*	
Text	*Feld*	
Methode	ReadText ReadPath	
Dok.typ	*Ganze Zahl*	Zulässige Werte (0,1,2), wobei 0 = Volltext, 1 = Gegliederter Text und 2 = XML
Kodierung	Automatisch "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Beachten Sie, dass Werte mit Sonderzeichen, wie "UTF-8", in Anführungszeichen gesetzt werden sollten, damit keine Verwechslung mit einem mathematischen Operator entsteht.
unity	*Ganze Zahl*	Zulässige Werte (0,1,2), wobei 0 = Absatz und 1 = Dokument
para_min	*Ganze Zahl*	
para_max	*Ganze Zahl*	
mtag	*Zeichenkette*	Enthält sämtliche mtag-Einstellungen (aus dem Dialogfeld “Einstellungen” für XML-Dateien)

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
mclef	*Zeichenkette*	Enthält sämtliche mclef-Einstellungen (aus dem Dialogfeld "Einstellungen" für Dateien mit gegliedertem Text)
language	de en es fr it ja nl pt	
concept_count_limit	*Ganze Zahl*	Extrahierung einschränken auf Konzepte mit globaler Häufigkeit von mindestens diesem Wert Für japanischen Text nicht verfügbar.
fix_punctuation	*Flag*	Für japanischen Text nicht verfügbar.
fix_spelling	*Flag*	Für japanischen Text nicht verfügbar.
spelling_limit	*Ganze Zahl*	Für japanischen Text nicht verfügbar.
extract_uniterm	*Flag*	Für japanischen Text nicht verfügbar.
extract_nonlinguistic	*Flag*	Für japanischen Text nicht verfügbar.
upper_case	*Flag*	Für japanischen Text nicht verfügbar.
group_names	*Flag*	Für japanischen Text nicht verfügbar.
permutation	*Ganze Zahl*	Maximale Anzahl von Füllwörtern in zusammengesetzten Konzepten (Standard ist 3). Für japanischen Text nicht verfügbar.
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	Nur für die Extrahierung von japanischem Text. *Anmerkung*: Verfügbar in IBM® SPSS® Modeler Premium. 0 = Sekundäre Stimmungsextrahierung 1 = Abhängigkeitsextrahierung 2 = Kein sekundäres Analyseset.
jp_algorithm_sense_mode	0 1 2	Nur für die Extrahierung von japanischem Text. *Anmerkung*: Verfügbar in SPSS Modeler Premium. 0 = Nur Schlussfolgerungen 2 = Nur repräsentative Stimmung 3 = Alle Stimmungen

Übersetzungsknoten: translatenode

Sie können die Eigenschaften in der folgenden Tabelle für Skripts verwenden. Der Knoten als solcher heißt translatenode.

Tabelle 7-6
Übersetzungsknoteneigenschaften

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
text	*Feld*	
method	ReadText ReadPath	

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
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"	Beachten Sie, dass Werte mit Sonderzeichen, wie "UTF-8", in Anführungszeichen gesetzt werden sollten, damit keine Verwechslung mit einem mathematischen Operator entsteht.
lw_server_type	LOC WAN HTTP	
lw_hostname	*Zeichenkette*	
lw_port	*Ganze Zahl*	
url	*Zeichenkette*	URL des Übersetzungsservers
apiKey	*Zeichenkette*	
user_id	*Zeichenkette*	
lpid	*Ganze Zahl*	Nicht verwendet, falls *language_from* oder *language_from_id* festgelegt ist.
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	

Skript-Eigenschaften	Datentyp	Eigenschaftsbeschreibung
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	*Ganze Zahl*	Dient zur Angabe der gewünschten Genauigkeitsstufe für den Übersetzungsprozess. Wählen Sie einen Wert im Bereich von 1 bis 3.
use_previous_translation	*Flag*	Gibt an, dass bereits Übersetzungsergebnisse aus einer früheren Ausführung vorliegen, die wiederverwendet werden können.
translation_label	*Zeichenkette*	Geben Sie ein Label ein, um die Übersetzungsergebnisse zu kennzeichnen, die wiederverwendet werden können.

Teil II:
Interaktive Workbench

Kapitel
8

Modus "Interaktive Workbench"

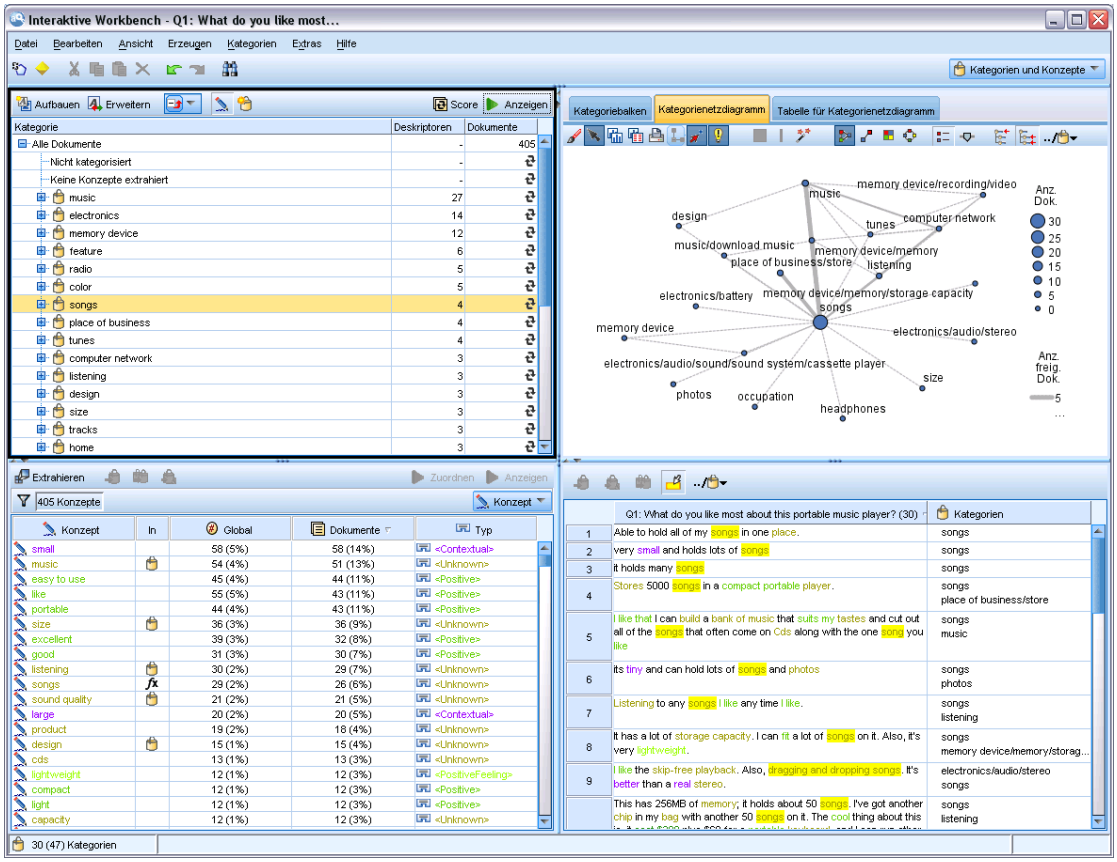
Sie können über einen Text-Mining-Modellierungsknoten eine interaktive Workbenchsitzung während der Stream-Ausführung starten. In dieser Workbench können Sie die wichtigsten Konzepte aus Ihren Textdaten extrahieren, Kategorien aufbauen und Text Link Analysis-Muster und -Cluster untersuchen und Kategoriemodelle generieren. In diesem Kapitel finden Sie einen allgemeinen Überblick über die Workbench-Schnittstelle sowie die wichtigsten Elemente, mit denen Sie arbeiten. Dazu gehören:

- **Extraction results.** Nach der Durchführung einer Extrahierung sind dies die Schlüsselwörter und Ausdrücke, die identifiziert und aus Ihren Textdaten extrahiert wurden. Sie werden auch als *Konzepte* bezeichnet. Diese Konzepte werden zu *Typen* zusammengefasst. Mit diesen Konzepten und Typen können Sie sowohl Ihre Daten untersuchen als auch Kategorien erstellen. Diese Elemente werden in der Ansicht **Kategorien und Konzepte** verwaltet.
- **Kategorien.** Mithilfe von Deskriptoren (wie Extrahierungsergebnissen, Mustern und Regeln) als Definition können Sie manuell oder automatisch ein Set mit Kategorien erstellen, denen Dokumente und Datensätze zugewiesen werden, je nachdem, ob sie einen Teil der Kategoriedefinition enthalten oder nicht. Diese Elemente werden in der Ansicht **Kategorien und Konzepte** verwaltet.
- **Cluster.** *Cluster* sind eine Zusammenstellung von Konzepten, zwischen denen Zusammenhänge erkannt wurden, die auf eine Beziehung zwischen ihnen hinweisen. Die Konzepte werden mithilfe eines komplexen Algorithmus zu Gruppen zusammengefasst, der unter anderem als Faktor dafür verwendet wird, wie häufig zwei Konzepte zusammen vorkommen im Vergleich zu der Häufigkeit, in der sie getrennt voneinander vorkommen. Diese Elemente werden in der Ansicht **Cluster** verwaltet. Außerdem können Sie die Konzepte, die einen Cluster ausmachen, zu den Kategorien hinzufügen.
- **Muster für die Textlinkanalyse.** Wenn Sie Musterregeln für die Textlinkanalyse (TLA) in Ihren linguistischen Ressourcen haben oder eine Ressourcenvorlage verwenden, die bereits TLA-Regeln enthält, können Sie Muster aus Ihren Textdaten extrahieren. Mit diesen Mustern können Sie interessante Beziehungen zwischen einzelnen Konzepten in Ihren Daten aufdecken. Außerdem können Sie diese Muster als Deskriptoren in Ihren Kategorien verwenden. Diese Elemente werden in der Ansicht **Textlinkanalyse** verwaltet. Für japanischen Text müssen Sie eine sekundäre Analyse wählen und TLA-Extrahierung aktivieren. *Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.
- **Linguistische Ressourcen.** Der Extrahierungsvorgang beruht auf einer Menge von Parametern und linguistischen Definitionen, die regeln, wie Text extrahiert und gehandhabt wird. Diese Elemente werden in Form von Vorlagen und Bibliotheken in der Ansicht **Ressourceneditor** verwaltet.

Die Ansicht “Kategorien und Konzepte”

Die Oberfläche der Anwendung besteht aus mehreren Ansichten. Die Ansicht “Kategorien und Konzepte” ist das Fenster, in dem Sie Kategorien erstellen und untersuchen sowie die Extrahierungsergebnisse untersuchen und optimieren können. **Kategorien** beziehen sich auf eine Gruppe von eng miteinander verwandten Ideen und Mustern, denen über einen Scoring-Vorgang Dokumente und Datensätze zugewiesen werden. **Konzepte** beziehen sich hingegen auf die allgemeinste Ebene der verfügbaren Extrahierungsergebnisse, die als Bausteine (sogenannte Deskriptoren) für Ihre Kategorien verwendet werden können.

Abbildung 8-1
Kategorie- und Konzeptansicht



Die Ansicht “Kategorien und Konzepte” gliedert sich in vier Fensterbereiche, die jeweils aus- bzw. eingeblendet werden können, indem Sie ihren Namen im Menü “Ansicht” auswählen. Für weitere Informationen siehe Thema Kategorisieren von Textdaten in Kapitel 10 auf S. 167.

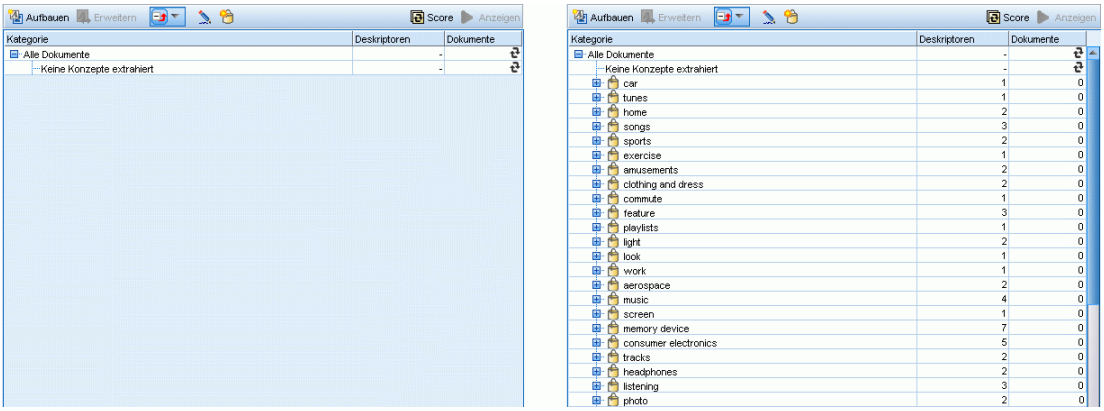
Fensterbereich “Kategorien”

Dieser Bereich befindet sich links oben und zeigt eine Tabelle, in der Sie alle von Ihnen erstellten Kategorien verwalten können. Nachdem Sie die Konzepte und Typen aus Ihren Textdaten extrahiert haben, können Sie Kategorien mithilfe von Verfahren wie semantische Netze und Konzepteinbeziehung oder manuell erstellen. Wenn Sie auf den Namen einer Kategorie

doppelklicken, wird das Dialogfeld "Kategoriedefinition" geöffnet und alle Deskriptoren, die zu seiner Definition gehören (z. B. Konzepte, Typen und Regeln) werden angezeigt. Für weitere Informationen siehe Thema Kategorisieren von Textdaten in Kapitel 10 auf S. 167. Nicht alle automatischen Techniken stehen für alle Sprachen zur Verfügung.

Wenn Sie eine Zeile im Fensterbereich auswählen, können Sie Informationen zu den entsprechenden Dokumenten/Datensätzen bzw. Deskriptoren in den Fensterbereichen "Daten" und "Visualisierung" anzeigen.

Abbildung 8-2
Ansicht "Kategorien und Konzepte": Fensterbereich "Kategorien" mit und ohne Kategorien



Fensterbereich "Extrahierungsergebnisse".

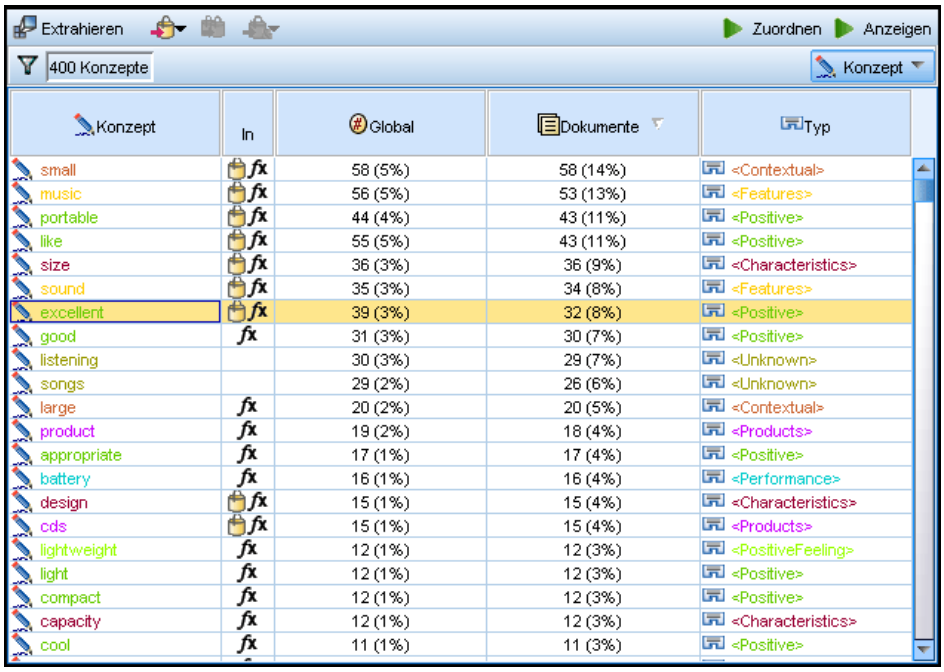
Dieser Bereich befindet sich links unten und zeigt die Ergebnisse der Extrahierung an. Wenn Sie eine Extrahierung ausführen, liest die Extrahierungsengine die Textdaten, identifiziert die relevanten Konzepte und weist jedem davon einen Typ zu. **Konzepte** sind Wörter bzw. Ausdrücke, die aus Ihren Textdaten extrahiert wurden. **Typen** sind semantische Gruppierungen von Konzepten, die in Form von Typ-Wörterbüchern gespeichert sind. Wenn die Extrahierung abgeschlossen ist, werden die Konzepte und Typen im Fensterbereich "Extrahierungsergebnisse" mit Farbkodierung angezeigt. Für weitere Informationen siehe Thema Extrahierungsergebnisse: Konzepte und Typen in Kapitel 9 auf S. 143.

Sie können das Set an zugrundeliegenden Fachausdrücken sehen, indem Sie den Mauszeiger auf einen Konzeptnamen halten. Dadurch wird eine QuickInfo mit dem Konzeptnamen und bis zu mehrere Zeilen mit Ausdrücken angezeigt, die unter diesem Konzept gruppiert sind. Diese zugrundeliegenden Ausdrücke umfassen die Synonyme, die in den linguistischen Ressourcen definiert sind (unabhängig davon, ob sie im Text gefunden wurden) sowie extrahierte Plural-/Singularausdrücke, permutierte Ausdrücke, Ausdrücke aus unscharfen Gruppierungen usw. Sie können diese Ausdrücke kopieren oder das vollständige Set der zugrundeliegenden Ausdrücke anzeigen, indem Sie mit der rechten Maustaste auf den Konzeptnamen klicken und die Option aus dem Kontextmenü wählen.

Text-Mining ist ein iterativer Prozess, in dem Extrahierungsergebnisse dem Kontext der Textdaten gemäß überprüft, für die Gewinnung neuer Ergebnisse verfeinert und dann neu bewertet werden. Die Extrahierungsergebnisse können durch Bearbeiten der linguistischen Ressourcen verfeinert werden. Diese Feinabstimmung kann teilweise direkt im Fensterbereich "Extrahierungsergebnisse" oder "Daten" vorgenommen werden, aber auch direkt in der Ansicht

"Ressourceneditor". Für weitere Informationen siehe Thema Die Ressourceneditoransicht auf S. 133.

Abbildung 8-3
Ansicht "Kategorien und Konzepte": Fensterbereich "Extrahierungsergebnisse" nach einer Extrahierung

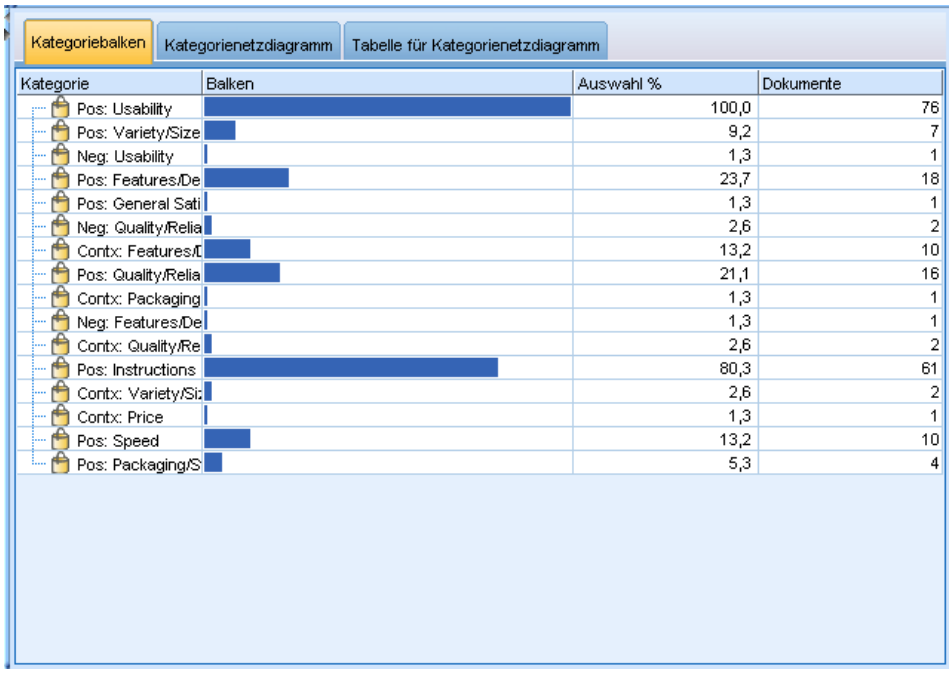


Konzept	In	Global	Dokumente	Typ
small	fx	58 (5%)	58 (14%)	<Contextual>
music	fx	56 (5%)	53 (13%)	<Features>
portable	fx	44 (4%)	43 (11%)	<Positive>
like	fx	55 (5%)	43 (11%)	<Positive>
size	fx	36 (3%)	36 (9%)	<Characteristics>
sound	fx	35 (3%)	34 (8%)	<Features>
excellent	fx	39 (3%)	32 (8%)	<Positive>
good	fx	31 (3%)	30 (7%)	<Positive>
listening		30 (3%)	29 (7%)	<Unknown>
songs		29 (2%)	26 (6%)	<Unknown>
large	fx	20 (2%)	20 (5%)	<Contextual>
product	fx	19 (2%)	18 (4%)	<Products>
appropriate	fx	17 (1%)	17 (4%)	<Positive>
battery	fx	16 (1%)	16 (4%)	<Performance>
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>
cool	fx	11 (1%)	11 (3%)	<Positive>

Fensterbereich "Visualisierung"

Dieser Bereich befindet sich rechts unten und bietet mehrere Perspektiven auf die Gemeinsamkeiten in der Dokument-/Datensatzkategorisierung. Jede Grafik bzw. jedes Diagramm stellt ähnliche Informationen dar, jedoch auf unterschiedliche Weise oder unterschiedlich detailliert. Diese Diagramme und Grafiken können zur Analyse Ihrer Kategorisierungsergebnisse und zur Unterstützung bei der Feinabstimmung von Kategorien oder bei der Berichterstellung verwendet werden. Sie könnten beispielsweise in einer Grafik Kategorien aufdecken, die zu große Ähnlichkeiten aufweisen (z. B. mehr als 75 % ihrer Datensätze gemeinsam haben) oder zu verschieden sind. Die Inhalte in einer Grafik bzw. einem Diagramm entsprechen der Auswahl in den anderen Fensterbereichen. Für weitere Informationen siehe Thema Kategoriendiagramme und Grafiken in Kapitel 13 auf S. 262.

Abbildung 8-4
Ansicht "Kategorien und Konzepte": Visualisierungsbereich

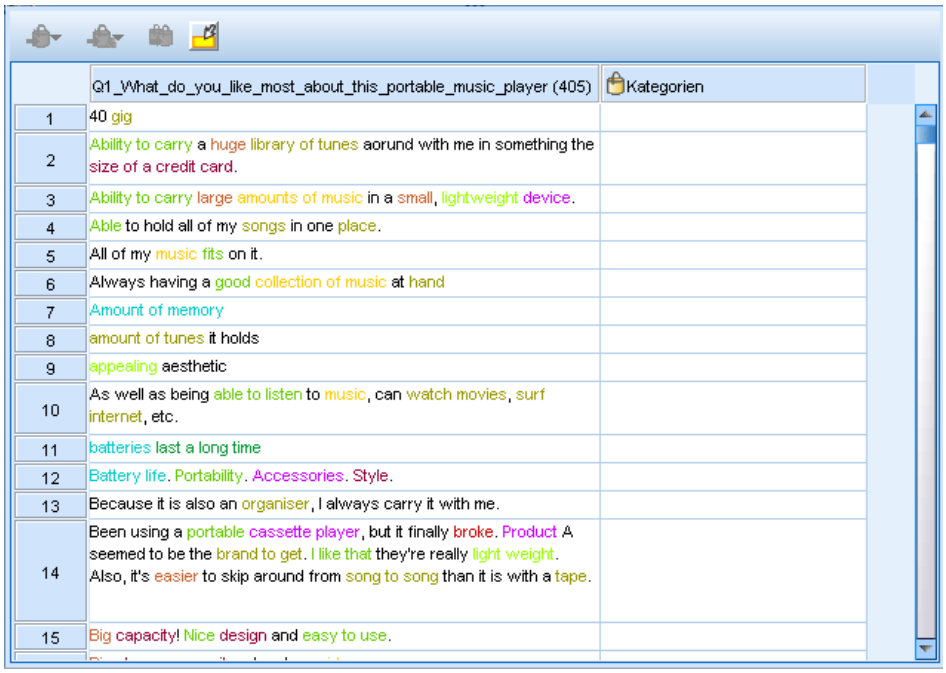


Datenbereich

Der Datenbereich befindet sich in der rechten unteren Ecke. In diesem Bereich wird eine Tabelle mit den Dokumenten oder Datensätzen entsprechend der Auswahl in einem anderen Bereich der Ansicht angezeigt. Je nach Auswahl wird nur der entsprechende Text im Bereich "Daten" angezeigt. Wenn Sie eine Auswahl getroffen haben, klicken Sie auf die Schaltfläche Anzeigen, um den Datenbereich mit dem entsprechenden Text aufzufüllen.

Wenn Sie eine Auswahl in einem anderen Bereich haben, werden in den entsprechenden Dokumenten oder Datensätzen die Konzepte farblich hervorgehoben, damit Sie diese leicht im Text identifizieren können. Alternativ können Sie die Maus über farbkodierte Elemente bewegen, um eine QuickInfo mit dem Namen des Konzepts anzuzeigen, unter dem das betreffende Element extrahiert wurde, und den Typ, dem es zugewiesen wurde. Für weitere Informationen siehe Thema Der Fensterbereich "Daten" in Kapitel 10 auf S. 179.

Abbildung 8-5
Ansicht "Kategorien und Konzepte": Datenbereich



	Q1_What_do_you_like_most_about_this_portable_music_player (405)	Kategorien
1	40 gig	
2	Ability to carry a huge library of tunes aorund with me in something the size of a credit card.	
3	Ability to carry large amounts of music in a small, lightweight device.	
4	Able to hold all of my songs in one place.	
5	All of my music fits on it.	
6	Always having a good collection of music at hand	
7	Amount of memory	
8	amount of tunes it holds	
9	appealing aesthetic	
10	As well as being able to listen to music, can watch movies, surf internet, etc.	
11	batteries last a long time	
12	Battery life. Portability. Accessories. Style.	
13	Because it is also an organiser, I always carry it with me.	
14	Been using a portable cassette player, but it finally broke. Product A seemed to be the brand to get. I like that they're really light weight. Also, it's easier to skip around from song to song than it is with a tape.	
15	Big capacity! Nice design and easy to use.	

Suche in der Ansicht "Kategorien und Konzepte"

In manchen Fällen ist es erforderlich, Informationen in einem bestimmten Abschnitt schnell aufzufinden. Mithilfe der Symbolleiste "Suchen" können Sie die Zeichenfolge eingeben, nach der Sie suchen wollen, und andere Suchkriterien wie Unterscheidung zwischen Groß- und Kleinschreibung oder Suchrichtung definieren. Sie können den Fensterbereich auswählen, in dem die Suche durchgeführt werden soll.

Abbildung 8-6
Symbolleiste "Suchen" in der Ansicht "Kategorien und Konzepte"

So verwenden Sie die Suchfunktion:

- Wählen Sie Bearbeiten > Suchen aus den Menüs in der Ansicht "Kategorien und Konzepte" aus. Die Symbolleiste "Suchen" erscheint über dem Kategoriebereich und den Visualisierungsbereichen.
- Geben Sie die Wortfolge, nach der Sie suchen möchten, in das Textfeld ein. Mit den Schaltflächen der Symbolleiste können Sie festlegen, ob zwischen Groß- und Kleinschreibung unterschieden wird, ob eine teilweise Übereinstimmung zulässig ist und in welche Richtung die Suche durchgeführt wird.
- Klicken Sie in der Symbolleiste auf den Namen des Fensterbereichs, in dem die Suche durchgeführt werden soll. Wenn eine Übereinstimmung gefunden wird, wird der Text im Fenster markiert.

- Um nach der nächsten Übereinstimmung zu suchen, klicken Sie erneut auf den Namen des Fensterbereichs.

Die Clusteransicht

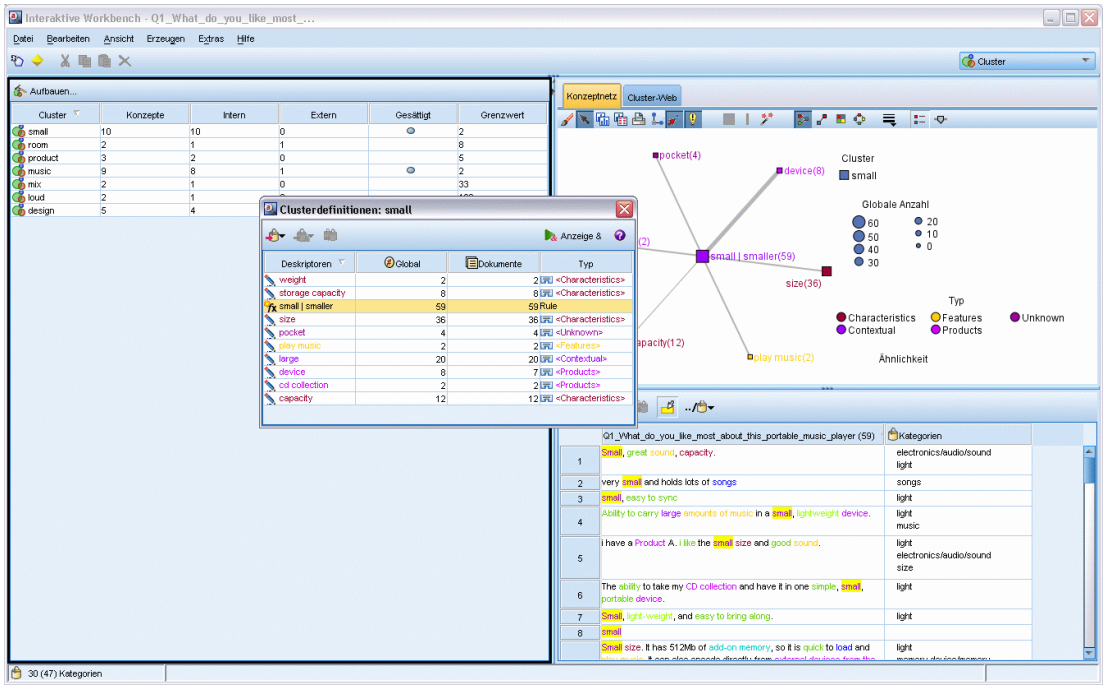
In der Clusteransicht können Sie die in Ihren Textdaten gefundenen Clusterergebnisse erstellen und untersuchen. **Cluster** sind Gruppierungen von Konzepten, die durch Clusteralgorithmen erstellt werden. Als Grundlage für die Erstellung dient die Häufigkeit, mit der die Konzepte auftreten, und die Häufigkeit, mit der sie gemeinsam vorkommen. Cluster zielen darauf ab, Konzepte zu gruppieren, die gemeinsam auftreten. Kategorien hingegen zielen darauf ab, Dokumente oder Datensätze auf der Grundlage dessen zu gruppieren, wie der enthaltene Text den Deskriptoren (Konzepten, Regeln, Mustern) für jede Kategorie entspricht.

Je häufiger die Konzepte in einem Cluster zusammen auftreten und je seltener sie zusammen mit anderen Konzepten vorkommen, desto besser ist der Cluster zur Identifizierung interessanter Konzeptbeziehungen geeignet. Zwei Konzepte treten gemeinsam auf, wenn sie beide (oder eines ihrer Synonyme oder einer ihrer Begriffe) im selben Dokument bzw. Datensatz vorkommen. Für weitere Informationen siehe Thema Analyse von Clustern in Kapitel 11 auf S. 243.

Sie können Cluster erstellen und mithilfe einer Reihe von Diagrammen und Grafiken untersuchen, um Beziehungen zwischen Konzepten aufzudecken, deren Ermittlung ansonsten zu zeitaufwändig wäre. Sie können zwar keine ganzen Cluster zu Ihren Kategorien hinzufügen, aber Sie können mithilfe des Dialogfelds "Clusterdefinitionen" die Konzepte in einem Cluster zu einer Kategorie hinzufügen. Für weitere Informationen siehe Thema Clusterdefinitionen in Kapitel 11 auf S. 249.

Sie können Änderungen an den Einstellungen für das Clustering vornehmen, um die Ergebnisse zu beeinflussen. Für weitere Informationen siehe Thema Cluster aufbauen in Kapitel 11 auf S. 245.

Abbildung 8-7
Clusteransicht



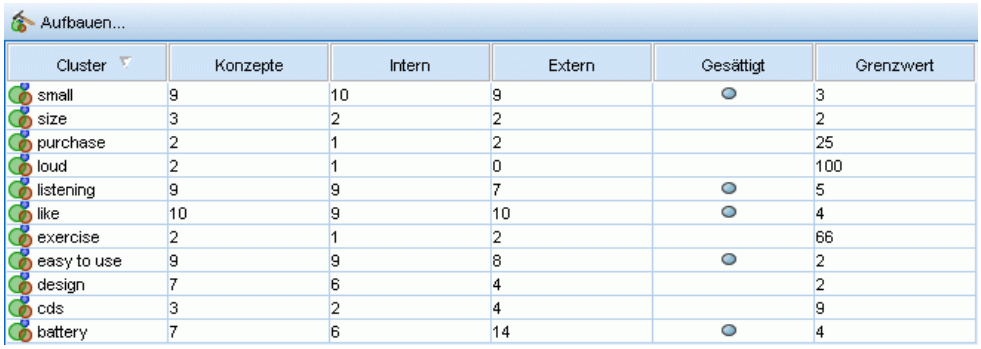
Die Ansicht “Cluster” gliedert sich in drei Fensterbereiche, die jeweils aus- bzw. eingeblendet werden können, indem Sie ihren Namen im Menü “Ansicht” auswählen. Normalerweise sind nur die Fensterbereiche “Cluster” und “Visualisierung” sichtbar.

Fensterbereich “Cluster”

Dieser Fensterbereich auf der linken Seite zeigt die in den Textdaten ermittelten Cluster. Durch Klicken auf die Schaltfläche Aufbauen können Sie Clustering-Ergebnisse erstellen. Cluster werden durch einen Clusteralgorithmus gebildet, der versucht, Konzepte zu identifizieren, die häufig gemeinsam auftreten.

Bei jeder Ausführung der Extrahierung werden die Clusterergebnisse gelöscht und Sie müssen die Cluster erneut erstellen, um die aktuellsten Ergebnisse zu erhalten. Beim Erstellen der Cluster können Sie einige Einstellungen ändern, wie beispielsweise die maximal zu erstellende Anzahl an Clustern, die maximale Anzahl an darin enthaltenen Konzepten bzw. die maximale Anzahl an Zusammenhängen mit externen Konzepten. Für weitere Informationen siehe Thema Untersuchen von Clustern in Kapitel 11 auf S. 248.

Abbildung 8-8
Clusteransicht: Fensterbereich "Cluster"



Cluster	Konzepte	Intern	Extern	Gesättigt	Grenzwert
small	9	10	9	●	3
size	3	2	2		2
purchase	2	1	2		25
loud	2	1	0		100
listening	9	9	7	●	5
like	10	9	10	●	4
exercise	2	1	2		66
easy to use	9	9	8	●	2
design	7	6	4		2
cds	3	2	4		9
battery	7	6	14	●	4

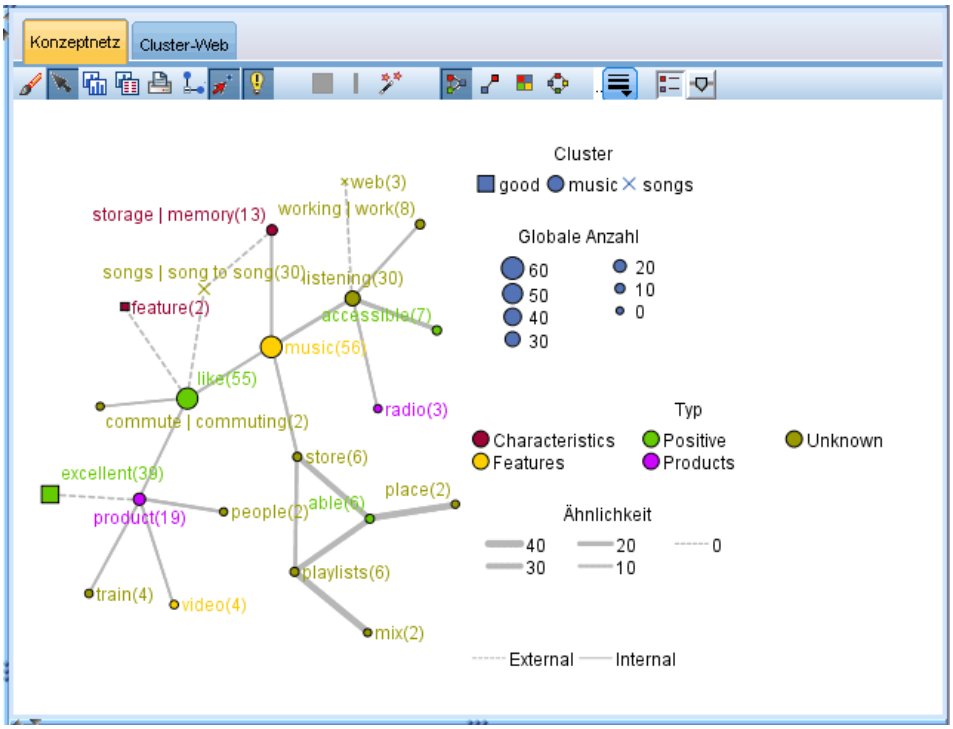
Fensterbereich "Visualisierung"

Dieser Fensterbereich rechts oben stellt zwei Clustering-Ansichten bereit: ein Konzeptnetzdiagramm und ein Clusternetzdiagramm. Falls dieser Fensterbereich nicht sichtbar ist, können Sie ihn im Menü "Ansicht" (Ansicht > Visualisierung) aufrufen. Je nachdem, was im Clusterfensterbereich ausgewählt wurde, können Sie die entsprechenden Interaktionen clusterübergreifend oder innerhalb der einzelnen Cluster anzeigen. Die Ergebnisse werden in mehreren Formaten ausgegeben:

- **Konzeptnetzdiagramm.** Netzdiagramm, das alle Konzepte innerhalb der ausgewählten Cluster anzeigt sowie die verknüpften Konzepte außerhalb des Clusters.
- **Clusternetzdiagramm.** Netzdiagramm, das die Verknüpfungen von den ausgewählten Clustern zu anderen Clustern anzeigt sowie Zusammenhänge zwischen diesen anderen Clustern.

Anmerkung: Um ein Clusternetzdiagramm anzuzeigen, müssen Sie bereits Cluster mit externen Links erstellt haben. Externe Links sind Verknüpfungen zwischen Konzeptpaaren, die sich in unterschiedlichen Clustern befinden (ein Konzept befindet sich in einem und das andere Konzept in einem anderen Cluster). Für weitere Informationen siehe Thema Clusterdiagramme in Kapitel 13 auf S. 266.

Abbildung 8-9
Clusteransicht: Visualisierungsbereich



Datenbereich

Der Fensterbereich “Daten” befindet sich rechts unten und ist standardmäßig ausgeblendet. Über den Fensterbereich “Cluster” können keine Ergebnisse aus dem Fensterbereich “Daten” angezeigt werden, da diese Cluster mehrere Dokumente/Datensätze umfassen, wodurch die Datenergebnisse ihre Aussagekraft verlieren. Sie können jedoch die zu einer Auswahl gehörenden Daten im Dialogfeld “Clusterdefinitionen” anzeigen. Je nachdem, was in diesem Dialogfeld ausgewählt wurde, wird im Bereich “Daten” nur der zugehörige Text angezeigt. Sobald Sie eine Auswahl getroffen haben, können Sie durch Klicken auf die Schaltfläche “&” anzeigen die Dokumente bzw. Datensätzen in den Fensterbereich “Daten” aufnehmen lassen, die alle Konzepte zusammen enthalten.

In den zugehörigen Dokumenten bzw. Datensätzen sind die Konzepte farbig hervorgehoben, damit Sie sie leichter im Text identifizieren können. Alternativ können Sie die Maus über farbkodierte Elemente bewegen, um das Konzept anzuzeigen, unter dem das betreffende Element extrahiert wurde, und den Typ, dem es zugewiesen wurde. Der Fensterbereich “Daten” kann mehrere Spalten enthalten, doch die Spalte “Textfeld” wird immer angezeigt. Sie trägt den Namen des während der Extrahierung verwendeten Textfelds bzw. den Namen eines Dokuments, wenn sich die Textdaten in vielen verschiedenen Dateien befinden. Weitere Spalten sind verfügbar. Für weitere Informationen siehe Thema Der Fensterbereich “Daten” in Kapitel 10 auf S. 179.

Die Textlinkanalysenansicht

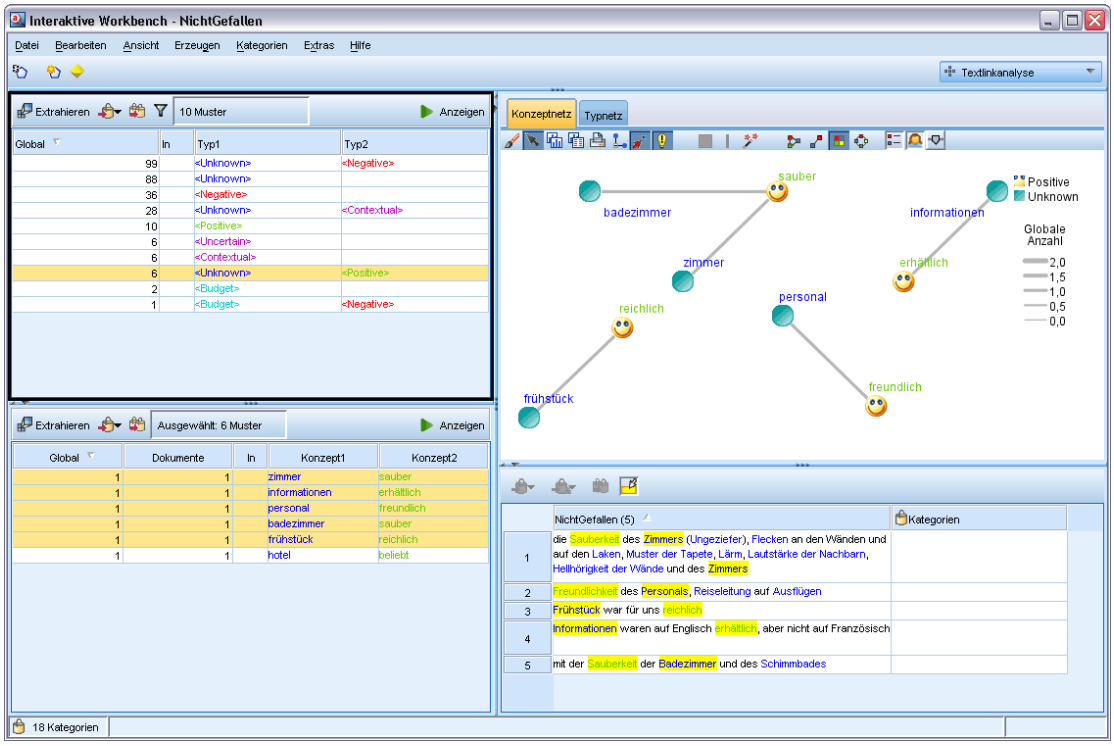
In der Textlinkanalysenansicht können Sie die in Ihren Textdaten gefundenen Textlinkanalysemuster erstellen und untersuchen. Textlinkanalyse (TLA) ist eine Technologie zum Musterabgleich, mit der Sie TLA-Regeln definieren und mit tatsächlichen extrahierten Konzepten und Beziehungen vergleichen können, die in Ihrem Text gefunden wurden.

Muster sind besonders nützlich, wenn Sie versuchen, Beziehungen zwischen Konzepten oder Meinungen zu einem bestimmten Thema zu ermitteln. Bei einigen Beispielen kommt auch das Ziel vor, Meinungen zu Produkten aus Umfragedaten, genomische Beziehungen aus medizinischen Forschungsberichten oder Beziehungen zwischen Personen oder Orten aus Geheimdienstdaten zu extrahieren.

Nachdem Sie einige TLA-Muster extrahiert haben, können Sie sie im Fensterbereich "Daten" bzw. "Visualisierung" untersuchen und sogar zu Kategorien in der Ansicht "Kategorien und Konzepte" hinzufügen. Um TLA-Ergebnisse extrahieren zu können, müssen in der verwendeten Ressourcenvorlage bzw. in den verwendeten Bibliotheken TLA-Regeln definiert sein. Für weitere Informationen siehe Thema Textlinkregeln in Kapitel 19 auf S. 346.

Wenn Sie TLA-Musterergebnisse extrahieren lassen, werden die Ergebnisse in dieser Ansicht gezeigt. Wenn Sie festgelegt haben, dass keine TLA-Musterergebnisse angezeigt werden sollen, müssen Sie die Schaltfläche Extrahieren verwenden und die Option zur Extrahierung von Mustern auswählen.

Abbildung 8-10
Ansicht "Text Link Analysis"

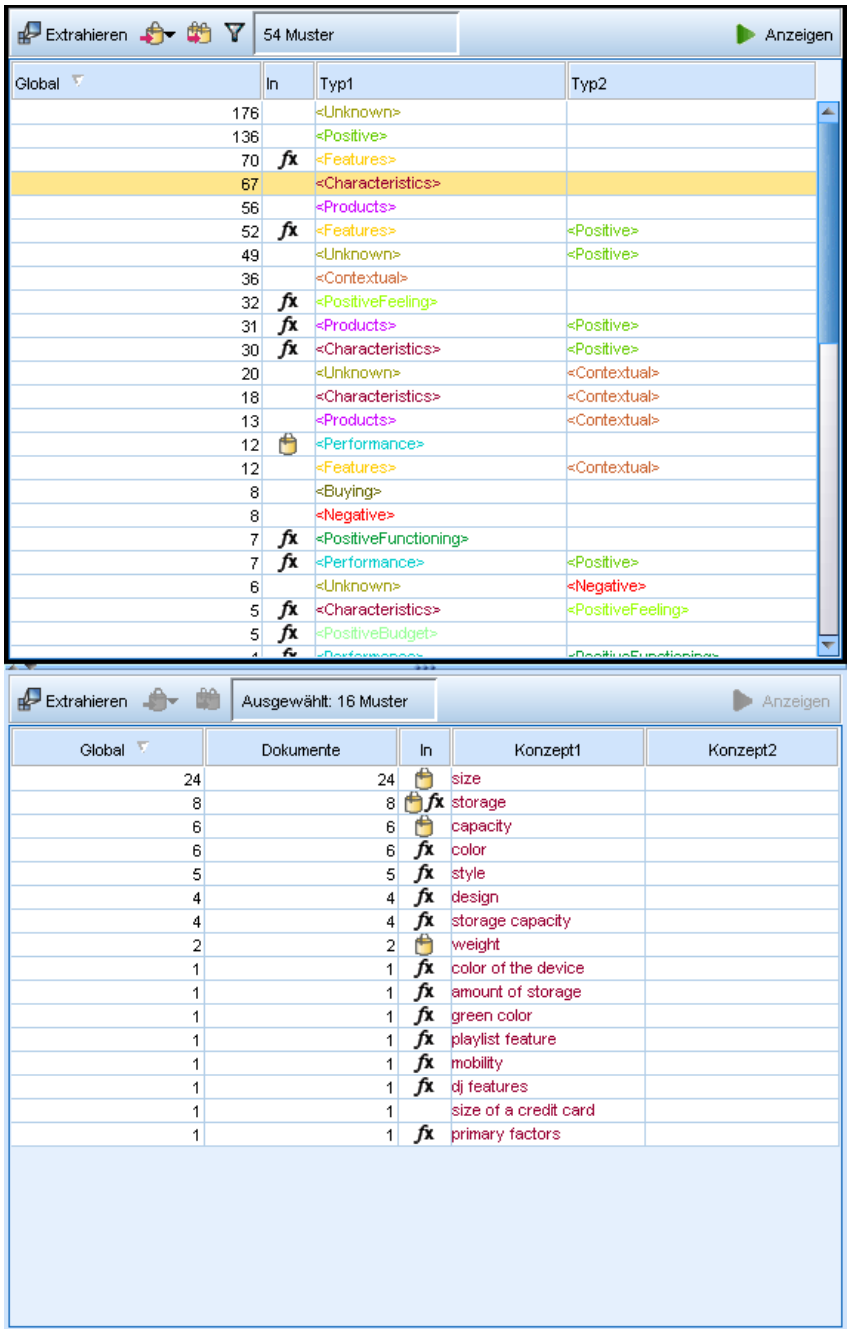


Die Ansicht “Textlinkanalyse” gliedert sich in vier Fensterbereiche, die jeweils aus- bzw. eingeblendet werden können, indem Sie ihren Namen im Menü “Ansicht” auswählen. Für weitere Informationen siehe Thema Untersuchen von Textlinkanalyse in Kapitel 12 auf S. 252.

Fensterbereiche “Typmuster” und “Konzeptmuster”

Auf der linken Seite befinden sich die beiden miteinander verbundenen Fensterbereiche “Typmuster” und “Konzeptmuster”, in denen Sie Ihre TLA-Musterergebnisse untersuchen und auswählen können. Muster bestehen aus Reihen von bis zu sechs Typen bzw. Konzepten. Beachten Sie, dass die Muster für japanischen Text Serien von nur bis zu einem oder beiden Typen oder Fachausdrücken sind. Die in den linguistischen Ressourcen definierte TLA-Musterregel legt die Komplexität der Musterergebnisse fest. Für weitere Informationen siehe Thema Textlinkregeln in Kapitel 19 auf S. 346.*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Abbildung 8-11
Textlinkanalysenansicht: Fensterbereiche "Typmuster" und "Konzeptmuster"



Die Musterergebnisse werden zunächst auf der Typebene gruppiert und anschließend in Konzeptmuster unterteilt. Aus diesem Grund gibt es zwei verschiedene Ergebnisbereiche: Typmuster (links oben) und Konzeptmuster (links unten).

- **Typmuster.** Im Fensterbereich “Typmuster” finden Sie extrahierte Muster, die aus zwei oder mehr verwandten Typen bestehen, die einer TLA-Musterregel entsprechen. Typmuster werden angezeigt als `<Organization>` + `<Location>` + `<Positive>`, was ein positives Feedback zu einem Unternehmen an einem bestimmten Ort bereitstellen könnte.
- **Konzeptmuster.** Im Fensterbereich “Konzeptmuster” finden Sie die extrahierten Muster auf der Konzeptebene für alle aktuell im oberhalb gelegenen Fensterbereich “Typmuster” ausgewählten Typmuster. Konzeptmuster weisen folgende Struktur auf: `Hotel + Paris + herrlich`.

Wie bei den Extrahierungsergebnissen in der Ansicht “Kategorien und Konzepte” können Sie hier die Ergebnisse überprüfen. Wenn Sie die Typen und Konzepte, aus denen diese Muster bestehen, weiter verfeinern möchten, können Sie das im Bereich “Extrahierungsergebnisse” in der Ansicht “Kategorien und Konzepte” oder direkt im Ressourceneditor vornehmen und Ihre Muster erneut extrahieren.

Fensterbereich “Visualisierung”

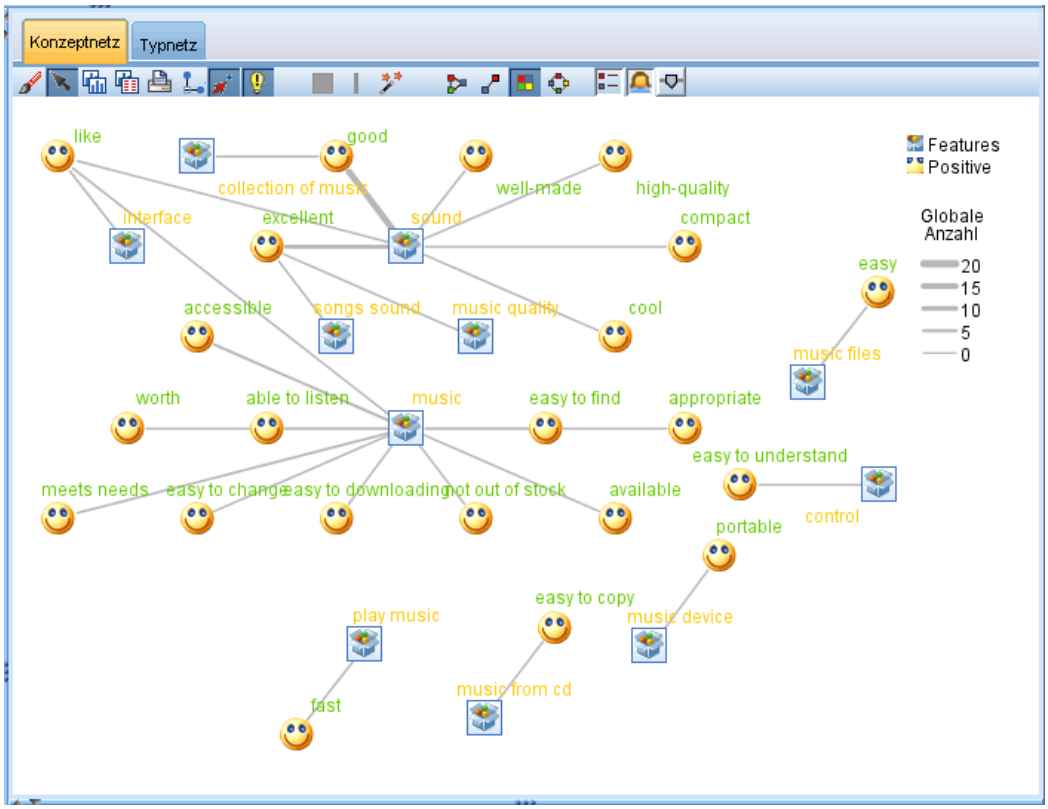
Dieser Fensterbereich befindet sich rechts oben in der Textlinkanalyse-Ansicht und zeigt ein Netzdiagramm der gewählten Muster, entweder als Typmuster oder als Konzeptmuster. Falls dieser Fensterbereich nicht sichtbar ist, können Sie ihn im Menü “Ansicht” (Ansicht > Visualisierung) aufrufen. Je nachdem, was in den anderen Fensterbereichen ausgewählt wurde, können Sie die entsprechenden Interaktionen zwischen Dokumenten/Datensätzen und Mustern anzeigen.

Die Ergebnisse werden in mehreren Formaten ausgegeben:

- **Konzeptdiagramm.** Dieses Diagramm zeigt alle Konzepte in den ausgewählten Mustern. Die Zeilenstärke und die Knotengrößen (sofern keine Typsymbole angezeigt werden) in einem Konzeptdiagramm zeigen die Anzahl der globalen Vorkommnisse in der ausgewählten Tabelle an.
- **Typdiagramm.** Dieses Diagramm zeigt alle Typen in den ausgewählten Mustern. Die Zeilenstärke und die Knotengrößen (sofern keine Typsymbole angezeigt werden) im Diagramm zeigen die Anzahl der globalen Vorkommnisse in der ausgewählten Tabelle an. Knoten werden entweder durch eine Typfarbe oder durch ein Symbol dargestellt.

Für weitere Informationen siehe Thema Textlinkanalyse-Diagramme in Kapitel 13 auf S. 268.

Abbildung 8-12
Textlinkanalyse: Visualisierungsbereich



Datenbereich

Der Datenbereich befindet sich in der rechten unteren Ecke. In diesem Bereich wird eine Tabelle mit den Dokumenten oder Datensätzen entsprechend der Auswahl in einem anderen Bereich der Ansicht angezeigt. Je nach Auswahl wird nur der entsprechende Text im Bereich "Daten" angezeigt. Wenn Sie eine Auswahl getroffen haben, klicken Sie auf die Schaltfläche Anzeigen, um den Datenbereich mit dem entsprechenden Text aufzufüllen.

Wenn Sie eine Auswahl in einem anderen Bereich haben, werden in den entsprechenden Dokumenten oder Datensätzen die Konzepte farblich hervorgehoben, damit Sie diese leicht im Text identifizieren können. Alternativ können Sie die Maus über farbkodierte Elemente bewegen, um eine QuickInfo mit dem Namen des Konzepts anzuzeigen, unter dem das betreffende Element extrahiert wurde, und den Typ, dem es zugewiesen wurde. Für weitere Informationen siehe Thema Der Fensterbereich "Daten" in Kapitel 10 auf S. 179.

Die Ressourceneditoransicht

IBM® SPSS® Modeler Text Analytics erfasst mithilfe einer robusten Extrahierungs-Engine schnell und genau die Schlüsselkonzepte aus Textdaten. Diese Engine ist stark auf linguistische Ressourcen angewiesen, die vorgeben, wie große Mengen an unstrukturierten Textdaten zu analysieren und interpretieren sind.

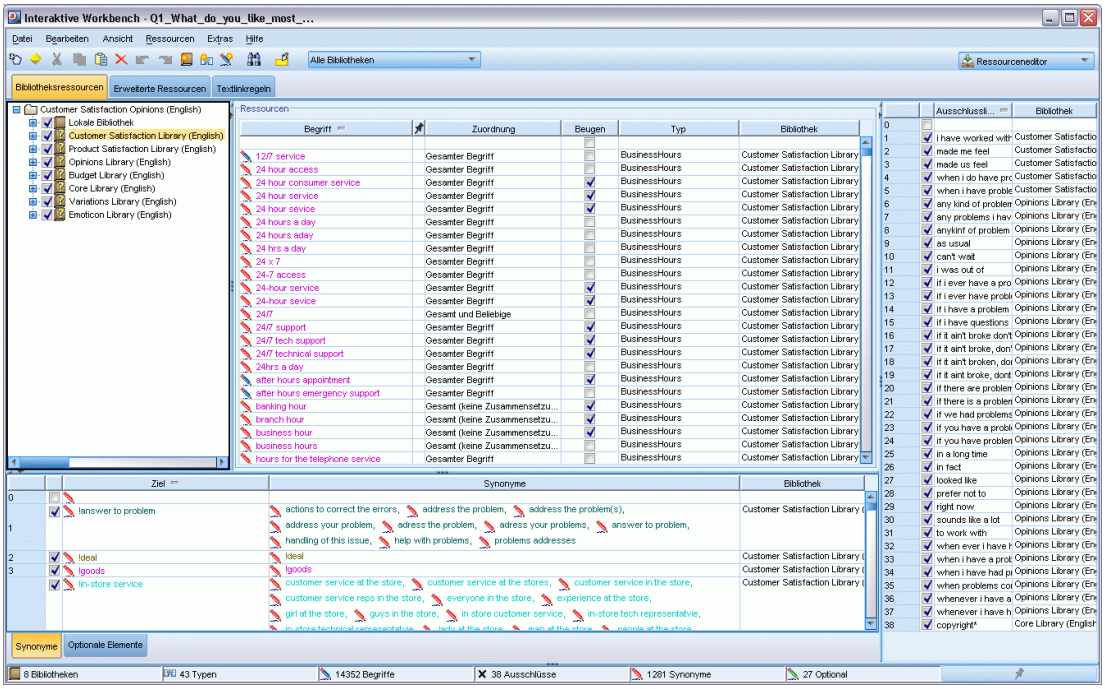
In der Ansicht Resource Editor können Sie die zum Extrahieren von Konzepten verwendeten linguistischen Ressourcen anzeigen und durch Feinabstimmung optimieren, zu Typen zusammenfassen, Muster in den Textdaten ermitteln und vieles mehr. SPSS Modeler Text Analytics bietet verschiedne vorkonfigurierte Ressourcenvorlagen. In einigen Sprachen können Sie auch die Ressourcen in einem Textanalysepaket verwenden. Für weitere Informationen siehe Thema Verwendung von Text Analysis Packages in Kapitel 10 auf S. 230.

Da diese Ressourcen möglicherweise nicht immer perfekt an den Kontext Ihrer Daten angepasst sind, können Sie im Resource Editor Ihre eigenen Ressourcen für einen bestimmten Kontext oder eine bestimmte Domäne erstellen, bearbeiten und verwalten. Für weitere Informationen siehe Thema Mit Bibliotheken arbeiten in Kapitel 16 auf S. 297.

Um die Feinabstimmung Ihrer linguistischen Ressourcen zu vereinfachen, können Sie über Kontextmenüs in den Fensterbereichen “Extrahierungsergebnisse” und “Daten” häufig verwendete Wörterbuchaufgaben direkt aus der Ansicht Kategorien und Konzepte heraus durchführen. Für weitere Informationen siehe Thema Extrahierungsergebnisse verfeinern in Kapitel 9 auf S. 158.

Anmerkung: Die Oberfläche für Ressourcen, die auf japanischen Text eingestellt sind, unterscheidet sich geringfügig. Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung. Für weitere Informationen siehe Thema Bearbeitungsressourcen für japanischen Text in Anhang A auf S. 383.

Abbildung 8-13
Ressourceneditoransicht



Bei den in der Ansicht Resource Editor durchgeführten Operationen geht es um die Verwaltung und Feinabstimmung der linguistischen Ressourcen. Diese Ressourcen werden in Form von Vorlagen und Bibliotheken gespeichert. Die Ansicht “Resource Editor” gliedert sich in vier Bereiche: den Fensterbereich “Bibliotheksbaum”, den Fensterbereich “Typ-Wörterbuch”, den Fensterbereich “Austauschwörterbuch” und den Fensterbereich “Ausschlusswörterbuch”.

Anmerkung: Für weitere Informationen siehe Thema Die Editoroberfläche in Kapitel 15 auf S. 282.

Festlegen von Optionen

Im Dialogfeld "Optionen" können Sie allgemeine Optionen für IBM® SPSS® Modeler Text Analytics festlegen. Dieses Dialogfeld enthält folgende Registerkarten:

- **Sitzung.** Diese Registerkarte enthält allgemeine Optionen und Trennzeichen.
- **Anzeigen.** Diese Registerkarte enthält Optionen für die auf der Benutzeroberfläche verwendeten Farben.
- **Klänge.** Diese Registerkarte enthält Optionen für Tonsignale.

So bearbeiten Sie Optionen:

▶ Wählen Sie in den Menüs die Optionsfolge Extras > Optionen aus. Das Dialogfeld "Optionen" wird geöffnet.

▶ Wählen Sie die Registerkarte mit den zu ändernden Informationen aus.

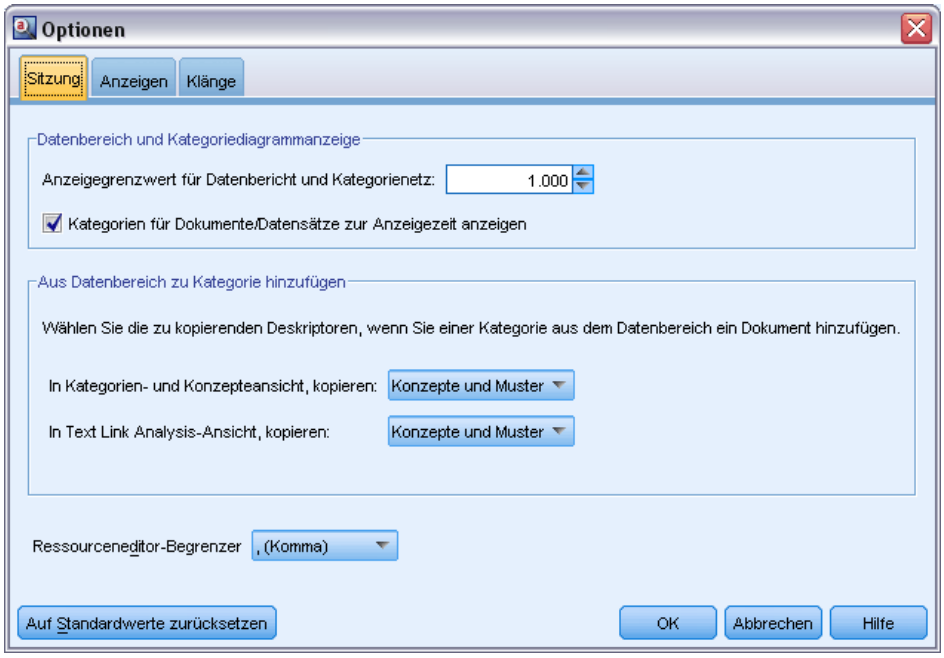
▶ Ändern Sie die Optionen nach Bedarf.

▶ Klicken Sie auf OK, um die Änderungen zu speichern.

Optionen: Registerkarte "Sitzung"

Auf dieser Registerkarte können Sie einige Grundeinstellungen festlegen

Abbildung 8-14
Dialogfeld "Optionen": Registerkarte "Sitzung"



Datenbereich und Kategoriediagrammanzeige. Diese Optionen beeinflussen, wie Daten im Datenbereich im Fensterbereich “Visualisierung” in der Ansicht “Kategorien und Konzepte” dargestellt werden.

- **Einschränkung für Datenbereich und Kategorienetzdiagramm anzeigen.** Diese Option legt fest, wie viele Dokumente maximal angezeigt bzw. zum Ausfüllen der Datenbereiche bzw. der Grafiken und Diagramme in der Ansicht “Kategorien und Konzepte” verwendet werden sollen.
- **Kategorien für Dokumente/Datensätze bei der Anzeige anzeigen.** Wenn diese Option ausgewählt ist, werden Dokumente oder Datensätze gescort, wenn Sie auf die Schaltfläche “Anzeigen” klicken, so dass Kategorien, zu denen sie gehören, im Bereich “Daten” in der Spalte “Kategorien” und in den Kategoriediagrammen angezeigt werden können. In einigen Fällen, insbesondere bei größeren Daten-Sets, kann es sinnvoll sein, diese Option zu deaktivieren, weil dadurch Daten und Grafiken wesentlich schneller angezeigt werden.

Aus Datenbereich zu Kategorie hinzufügen. Diese Optionen beeinflussen, welche Elemente Kategorien hinzugefügt werden, wenn Dokumente und Datensätze aus dem Datenbereich hinzugefügt werden.

- **In der Ansicht “Konzepte und Kategorien”, Kopieren.** Beim Hinzufügen eines Dokuments oder Datensatzes aus dem Datenbereich in dieser Ansicht werden entweder Nur Konzepte oder Konzepte und Muster kopiert.
- **In der Ansicht “Textlinkanalyse”, Kopieren.** Beim Hinzufügen eines Dokuments oder Datensatzes aus dem Datenbereich in dieser Ansicht werden entweder Nur Muster oder Konzepte und Muster kopiert.

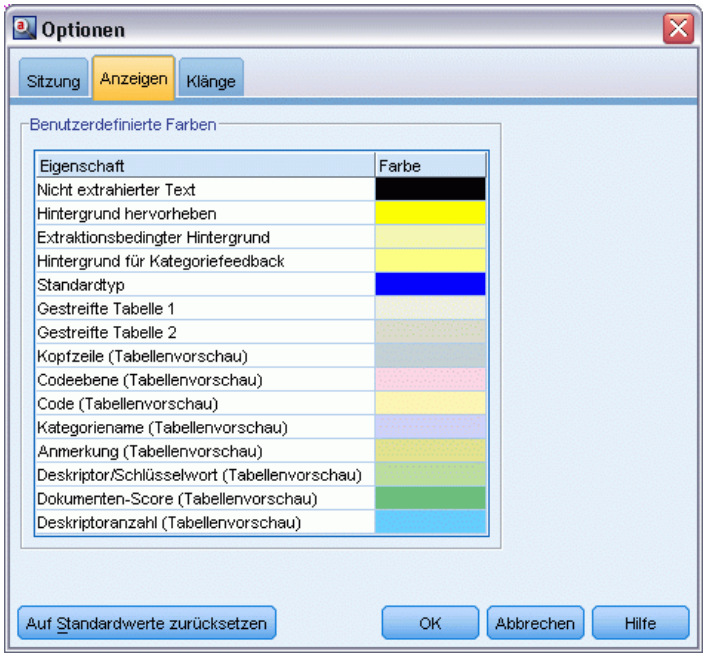
Ressourceneditortrennzeichen. Dient zur Auswahl des Zeichens, das bei der Eingabe von Elementen, wie beispielsweise Konzepten, Synonymen und optionalen Elementen, in der Ressourceneditoransicht als Trennzeichen verwendet werden soll.

Optionen: Registerkarte “Anzeige”

Auf dieser Registerkarte können Sie Optionen bearbeiten, die das allgemeine Erscheinungsbild der Anwendung und die zur Unterscheidung der verschiedenen Elemente verwendeten Farben betreffen.

Anmerkung: Um für das Produkt ein klassisches Erscheinungsbild oder eines aus einer früheren Version zu verwenden, öffnen Sie das Dialogfeld “Benutzeroptionen” im Menü “Extras” im IBM® SPSS® Modeler-Hauptfenster.

Abbildung 8-15
Dialogfeld "Optionen": Registerkarte "Farben"



Benutzerdefinierte Farben. Dient zum Bearbeiten der Farben für Elemente, die auf dem Bildschirm angezeigt werden. Sie können die Farbe für jedes Element in der Tabelle ändern. Um eine benutzerdefinierte Farbe anzugeben, klicken Sie auf den Farbbereich rechts neben dem zu ändernden Element und wählen Sie in der Dropdown-Farbliste die gewünschte Farbe aus.

- **Nicht extrahierter Text.** Textdaten, die nicht extrahiert wurden, aber dennoch im Bereich "Daten" sichtbar sind.
- **Hintergrund hervorheben.** Hintergrundfarbe für die Textauswahl (bei der Auswahl von Elementen in den Fensterbereichen oder von Text im Bereich "Daten").
- **Für Extrahierung erforderlicher Hintergrund.** Hintergrundfarbe der Fensterbereiche "Extrahierungsergebnisse", "Muster" und "Cluster", die anzeigt, dass Änderungen an den Bibliotheken vorgenommen wurden und eine Extrahierung erforderlich ist.
- **Hintergrund für Kategorie-Feedback.** Hintergrundfarbe für Kategorien, die nach einer Operation angezeigt werden.
- **Standardtyp.** Standardfarbe für Typen und Konzepte, die in den Fensterbereichen "Daten" und "Extrahierungsergebnisse" angezeigt werden. Diese Farbe gilt für alle benutzerdefinierten Typen, die Sie im Ressourceneditor erstellen. Sie können diese Standardfarbe für Ihre benutzerdefinierten Typ-Wörterbücher außer Kraft setzen, indem Sie die Eigenschaften für diese Typ-Wörterbücher im Resource Editor bearbeiten. Für weitere Informationen siehe Thema Erstellen von Typen in Kapitel 17 auf S. 314.

- **Gestreifte Tabelle 1.** Die erste der beiden Farben, die sich in der Tabelle im Dialogfeld "Erzwungene Konzepte bearbeiten" abwechseln, damit die einzelnen Zeilen besser auseinandergehalten werden können.
- **Gestreifte Tabelle 2.** Die zweite der beiden Farben, die sich in der Tabelle im Dialogfeld "Erzwungene Konzepte bearbeiten" abwechseln, damit die einzelnen Zeilen besser auseinandergehalten werden können.

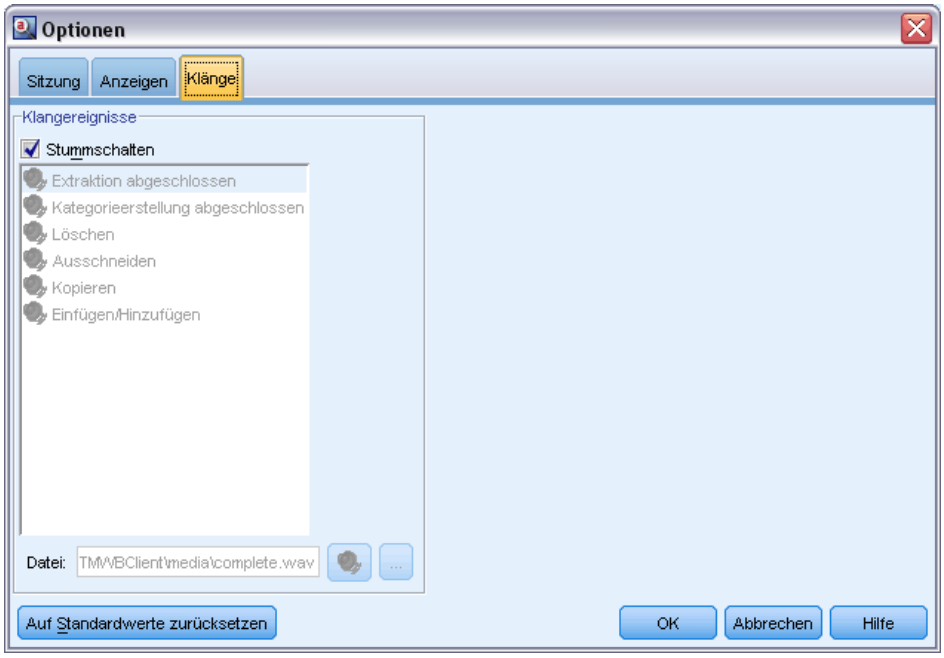
Anmerkung: Wenn Sie auf Auf Standardeinstellungen zurücksetzen klicken, werden alle Optionen in diesem Dialogfeld auf die Werte zurückgesetzt, die sie bei der ursprünglichen Installation des Produkts aufwiesen.

Optionen: Registerkarte "Klänge"

Auf dieser Registerkarte können Sie Optionen bearbeiten, die Klänge betreffen. Unter "Klänge" können Sie einen Klang angeben, der als Signal verwendet werden soll, wenn ein Ereignis eintritt. Es stehen mehrere Klänge zur Auswahl zur Verfügung. Mit der Auslassungsschaltfläche (...) wählen Sie einen Klang aus. Die *.wav*-Dateien, mit denen Klänge für IBM® SPSS® Modeler Text Analytics erstellt werden, sind im Unterverzeichnis *media* des Installationsverzeichnisses gespeichert. Wenn keine Klänge abgespielt werden sollen, wählen Sie die Option Alle Klänge stummschalten. Die Klänge sind standardmäßig stummgeschaltet.

Anmerkung: Wenn Sie auf Auf Standardeinstellungen zurücksetzen klicken, werden alle Optionen in diesem Dialogfeld auf die Werte zurückgesetzt, die sie bei der ursprünglichen Installation des Produkts aufwiesen.

Abbildung 8-16
Dialogfeld "Optionen": Registerkarte "Klänge"



Microsoft Internet Explorer-Einstellungen für die Hilfe

Einstellungen für Microsoft Internet Explorer

Die meisten Hilfefunktionen in dieser Anwendung verwenden Technologie, die auf Microsoft Internet Explorer beruht. Einige Versionen von Internet Explorer (insbesondere die in Microsoft Windows XP, Service Pack 2, bereitgestellte Version) blockieren standardmäßig als "aktiver Inhalt" betrachtete Elemente in Internet Explorer-Fenstern auf dem lokalen Computer. Diese Standardeinstellung kann dazu führen, dass bestimmte Inhalte in Hilfefunktionen blockiert werden. Um alle Hilfe-Inhalte anzuzeigen, können Sie das Standardverhalten von Internet Explorer ändern.

- Wählen Sie im Menü des Internet Explorer folgende Optionen:
  Extras > Internetoptionen...
- Klicken Sie auf die Registerkarte Erweitert.
- Führen Sie einen Bildlauf nach unten zum Abschnitt Sicherheit durch.
- Aktivieren Sie Ausführung aktiver Inhalte in Dateien auf dem lokalen Computer zulassen.

Generieren von Modell-Nuggets und Modellierungsknoten

In einer interaktiven Sitzung können Sie Ihre Arbeit verwenden, um eines der folgenden Elemente zu generieren:

- **Text-Mining-Modellierungsknoten.** Bei einem aus einer interaktiven Workbench-Sitzung generierten Modellierungsknoten handelt es sich um einen Text-Mining-Knoten, dessen Einstellungen und Optionen denen entsprechen, die in der offenen interaktiven Sitzung gespeichert wurden. Dies kann nützlich sein, wenn Sie nicht mehr über den ursprünglichen Text-Mining-Knoten verfügen oder eine neue Version erstellen möchten. Für weitere Informationen siehe Thema Mining nach Konzepten und Kategorien in Kapitel 3 auf S. 30.
- **Kategorie-Modell-Nugget.** Bei einem aus einer interaktiven Workbench-Sitzung generierten Modell-Nugget handelt es sich um ein Kategoriemodell-Nugget. In der Ansicht "Kategorien und Konzepte" muss mindestens eine Kategorie vorliegen, um ein Kategoriemodell-Nugget generieren zu können. Für weitere Informationen siehe Thema Text-Mining-Nugget: Kategoriemodell in Kapitel 3 auf S. 72.

So generieren Sie einen Modellierungsknoten für das Text-Mining:

- Wählen Sie in den Menüs die Optionsfolge Generieren > Auswahlknoten generieren. Ein Text-Mining-Modellierungsknoten wird unter Verwendung aller derzeit in der Workbench-Sitzung gültigen Einstellungen zum Arbeitszeichenbereich hinzugefügt. Der Knoten wird nach dem Textfeld benannt.

So erzeugen Sie ein Kategoriemodell-Nugget:

- Wählen Sie in den Menüs die Optionsfolge Generieren > Modell generieren. Ein Modell-Nugget wird mit dem Standardnamen direkt auf der Modellpalette generiert.

Aktualisieren von Modellierungsknoten und Speichern

Wenn Sie in einer interaktiven Sitzung arbeiten, sollten Sie den Modellierungsknoten von Zeit zu Zeit aktualisieren, um Ihre Änderungen zu speichern. Außerdem sollten Sie den Modellierungsknoten jedes Mal aktualisieren, wenn Sie die Arbeit in der interaktiven Workbench-Sitzung abgeschlossen haben und Ihre Arbeit speichern möchten. Wenn Sie den Modellierungsknoten aktualisieren, wird der Inhalt der Workbench-Sitzung wieder in dem Text-Mining-Knoten gespeichert, von dem die interaktive Workbench-Sitzung ausging. Dabei wird das Ausgabefenster nicht geschlossen.

Wichtig: Durch diese Aktualisierung wird Ihr Stream nicht gespeichert. Speichern Sie Ihren Stream im Hauptbereich von IBM® SPSS® Modeler nach der Aktualisierung des Modellierungsknotens.

So aktualisieren Sie einen Modellierungsknoten

▶ Wählen Sie in den Menüs die Optionsfolge Datei > Modellierungsknoten aktualisieren. Der Modellierungsknoten wird mit den Erstellungs- und Extrahierungseinstellungen aktualisiert sowie mit allen vorliegenden Optionen und Kategorien.

Schließen und Beenden von Sitzungen

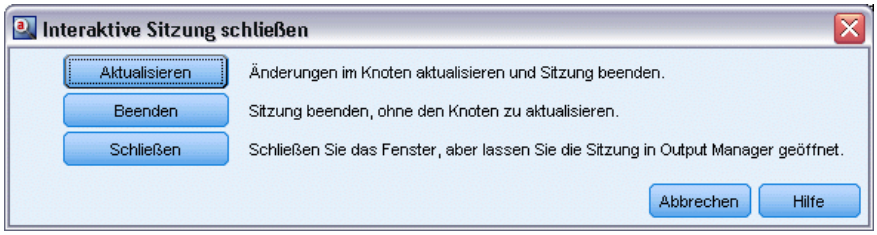
Wenn Sie die Arbeit in Ihrer Sitzung abgeschlossen haben, können Sie die Sitzung auf drei verschiedene Weisen beenden:

- **Speichern.** Mit dieser Option können Sie zunächst Ihre Arbeit für zukünftige Sitzungen wieder im Ausgangs-Modellierungsknoten speichern sowie Bibliotheken zur Wiederverwendung in anderen Sitzungen veröffentlichen. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen in Kapitel 16 auf S. 306. Nach dem Speichern wird das Sitzungsfenster geschlossen und die Sitzung wird aus dem Ausgabe-Manager im IBM® SPSS® Modeler-Fenster gelöscht.
- **Beenden.** Mit dieser Option werden alle nicht gespeicherten Arbeiten verworfen, das Sitzungsfenster wird geschlossen und die Sitzung wird aus dem Ausgabe-Manager im SPSS Modeler-Fenster gelöscht. Um Arbeitsspeicher freizugeben, sollten Sie alle wichtigen Arbeiten speichern und die Sitzung beenden.
- **Schließen.** Bei dieser Option werden keine Arbeiten gespeichert oder verworfen. Bei dieser Option wird das Sitzungsfenster geschlossen, die Sitzung wird jedoch weiterhin ausgeführt. Sie können das Sitzungsfenster erneut öffnen, indem Sie diese Sitzung im Ausgabe-Manager des SPSS Modeler-Fensters auswählen.

So schließen Sie eine Workbench-Sitzung:

▶ Wählen Sie in den Menüs die Optionsfolge Datei > Schließen.

Abbildung 8-17
Dialogfeld "Interaktive Sitzung schließen"



Tastatureingabehilfen

Die interaktive Workbench-Oberfläche enthält Direktzugriffstasten, um den Zugriff auf die Funktionen des Produkts zu erleichtern. So können Sie die Taste "Alt" zusammen mit der entsprechenden Taste drücken, um Fenstermenüs zu aktivieren (z. B. Alt-D, um das Menü "Datei" aufzurufen), oder die Tabulatortaste drücken, um durch die Steuerelemente im Dialogfeld zu blättern. In diesem Abschnitt werden die Direktzugriffstasten für die alternative Navigation behandelt. Es gibt andere Direktzugriffstasten für die IBM® SPSS® Modeler-Oberfläche.

Tabelle 8-1
Allgemeine Direktzugriffstasten

Direktzugriffstaste	**Funktion**
Strg+1	Zeigt die erste Registerkarte in einem Fensterbereich mit Registerkarten an.
Strg+2	Zeigt die zweite Registerkarte in einem Fensterbereich mit Registerkarten an.
Strg+A	Wählt alle Elemente für den Fensterbereich aus, auf dem der Fokus liegt.
Strg+C	Kopiert den ausgewählten Text in die Zwischenablage.
Strg+E	Startet die Extrahierung in den Ansichten "Kategorien und Konzepte" sowie "Textlinkanalyse".
Strg+F	Zeigt die Suchsymbolleiste im Resource Editor/Template Editor an, sofern noch nicht sichtbar, und setzt den Fokus auf diese Symbolleiste.
Strg+I	Startet in der Ansicht "Kategorien und Konzepte" das Dialogfeld "Kategoriedefinitionen" für die ausgewählte Kategorie. Startet in der Ansicht "Cluster" das Dialogfeld "Clusterdefinitionen" für das ausgewählte Cluster.
Strg+R	Öffnet das Dialogfeld "Fachausdrücke hinzufügen" im Resource Editor/Template Editor.
Strg+T	Öffnet das Dialogfeld "Typeigenschaften" zum Erstellen eines neuen Typs im Resource Editor/Template Editor.
Strg+V	Fügt den Inhalt der Zwischenablage ein.
Strg+X	Schneidet die ausgewählten Elemente aus dem Resource Editor/Template Editor aus.
STRG + Y	Wiederholt die letzte Aktion in der Ansicht.
Strg+Z	Macht die letzte Aktion in der Ansicht rückgängig.
F1	Zeigt die Hilfe an; in Dialogfeldern wird die Kontexthilfe zu einem Element angezeigt.
F2	Schaltet den Bearbeitungsmodus in Tabellenzellen ein bzw. aus.
F6	Wechselt den Fokus zwischen den Hauptbereichen in der aktiven Ansicht.

Direktzugriffstaste	Funktion
F8	Verschiebt den Fokus auf die Fensterteiler (zum Ändern der Größe).
F10	Erweitert das Haupt-Dateimenü.
Pfeil-nach-oben, Pfeil-nach-unten	Dient zur vertikalen Größenänderung, wenn der Teilungsbalken ausgewählt ist.
Pfeil-nach-links, Pfeil-nach-rechts	Dient zur horizontalen Größenänderung, wenn der Teilungsbalken ausgewählt ist.
Pos1, Ende	Maximiert bzw. minimiert die Fenstergröße, wenn der Teilungsbalken ausgewählt ist.
Tabulatortaste	Wechselt in Vorwärtsrichtung durch die Elemente im Fenster, Bereich bzw. Dialogfeld.
Umschalt+F10	Zeigt das Kontextmenü für ein Element an.
Umschalt+ Tabulatortaste	Wechselt rückwärts durch die Elemente im Fenster bzw. Dialogfeld.
Umschalt+Pfeil	Wählt die Zeichen im Bearbeitungsfeld aus, wenn Sie sich im Bearbeitungsmodus befinden (F2).
Strg+Tabulatortaste	Verlagert den Fokus auf den nächsten Hauptbereich im Fenster.
Umschalt+Strg+ Tabulatortaste	Verlagert den Fokus auf den vorherigen Hauptbereich im Fenster.

Direktzugriffstasten für Dialogfelder

Einige der Direktzugriffstasten für die Tastatur und das Bildschirm-Lesesystem sind hilfreich, wenn Sie mit Dialogfeldern arbeiten. Beim Aufrufen eines Dialogfelds müssen Sie möglicherweise die Tabulatortaste drücken, um den Fokus auf das erste Steuerelement zu verlagern und das Bildschirm-Lesesystem zu starten. In der folgenden Tabelle finden Sie eine vollständige Liste mit speziellen Direktzugriffstasten für die Tastatur und das Bildschirm-Lesesystem.

Tabelle 8-2
Direktzugriffstasten für Dialogfelder

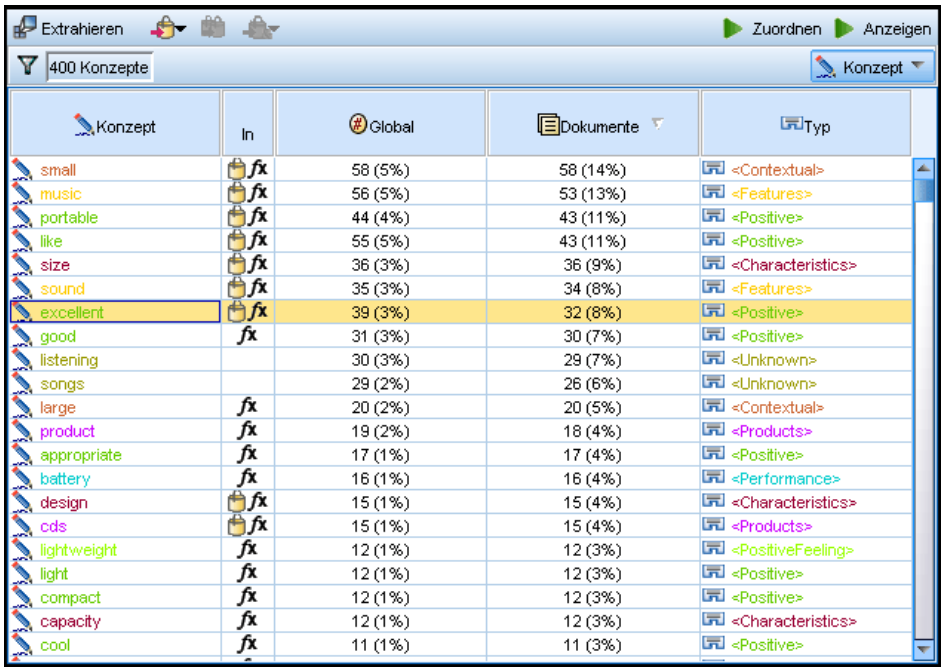
Direktzugriffstaste	Funktion
Tabulatortaste	Wechselt in Vorwärtsrichtung durch die Elemente im Fenster bzw. Dialogfeld.
Strg+Tabulatortaste	Wechselt in Vorwärtsrichtung von einem Textfeld zum nächsten Element.
Umschalt+ Tabulatortaste	Wechselt rückwärts durch die Elemente im Fenster bzw. Dialogfeld.
Umschalt+Strg+ Tabulatortaste	Wechselt von einem Textfeld zum vorherigen Element zurück.
Leertaste	Dient zur Auswahl des Steuerelements bzw. der Schaltfläche, auf der der Fokus liegt.
Esc	Bricht die Änderungen ab und schließt das Dialogfeld.
Eingabetaste	Validiert die Änderungen und schließt das Dialogfeld (entspricht der Schaltfläche “OK”). Wenn Sie sich in einem Textfeld befinden, müssen Sie zunächst das Textfeld mit Strg+Tabulatortaste verlassen.

Kapitel 9

Konzepte und Typen extrahieren

Wenn Sie einen Stream ausführen, der die interaktive Workbench startet, erfolgt automatisch eine Extrahierung der Textdaten in dem Stream. Das Endergebnis dieser Extrahierung ist eine Reihe von Konzepten, Typen und, falls TLA-Muster in den linguistischen Ressourcen vorhanden sind, von Mustern. Im Bereich "Extrahierungsergebnisse" können Sie Konzepte und Typen anzeigen und mit ihnen arbeiten. Für weitere Informationen siehe Thema Wie Extrahierung funktioniert in Kapitel 1 auf S. 7.

Abbildung 9-1
Fensterbereich "Extrahierungsergebnisse" nach einer Extrahierung



Um die Extrahierungsergebnisse zu verfeinern, können Sie die linguistischen Ressourcen verändern und eine erneute Extrahierung durchführen. Für weitere Informationen siehe Thema Extrahierungsergebnisse verfeinern auf S. 158. Die Ressourcen und Parameter im Dialogfeld "Extrahieren" bestimmen, wie die Ergebnisse extrahiert und geordnet werden. Mit den Extrahierungsergebnissen können Sie den Großteil, wenn nicht sogar alle Kategoriedefinitionen festlegen.

Extrahierungsergebnisse: Konzepte und Typen

Während des Extrahierungsprozesses werden sämtliche Textdaten untersucht und die relevanten Konzepte ermittelt, extrahiert und entsprechenden Typen zugewiesen. Nach Abschluss der Extrahierung werden die Ergebnisse in dem Bereich "Extrahierungsergebnisse" in der linken

unteren Ecke der Ansicht “Kategorien und Konzepte” angezeigt. Beim ersten Start der Sitzung wird die Vorlage für linguistische Ressourcen, die Sie in dem Knoten ausgewählt haben, zum Extrahieren und Ordnen dieser Konzepte und Typen verwendet.

Die extrahierten Konzepte, Typen und TLA-Muster werden als **Extrahierungsergebnisse** bezeichnet und dienen als Deskriptoren oder Bausteine für Ihre Kategorien. Sie können auch Konzepte, Typen und Muster in Ihren Kategorieregeln verwenden. Außerdem verwenden die automatischen Verfahren Konzepte und Typen zum Aufbau der Kategorien.

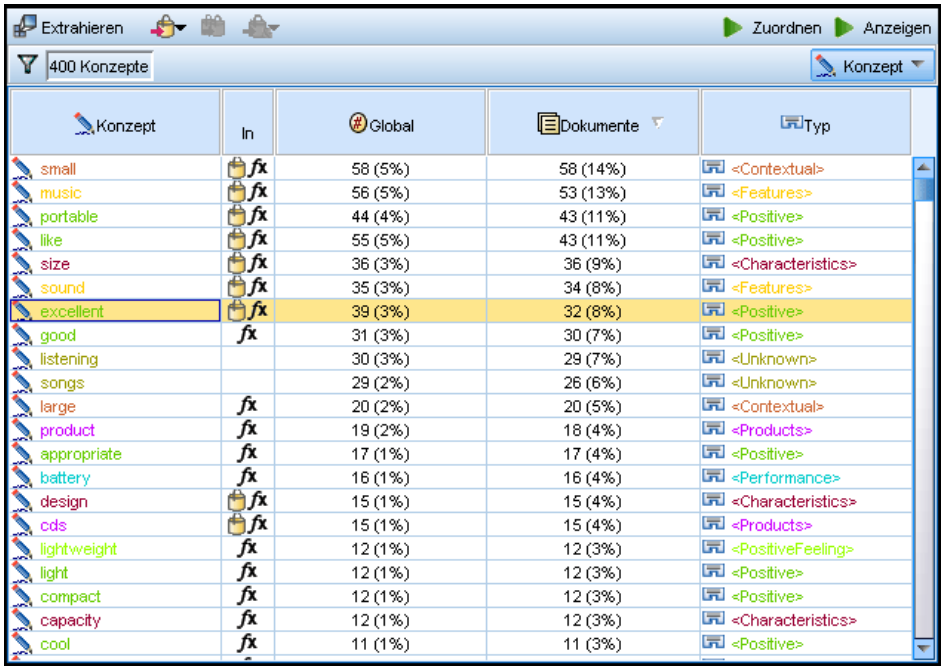
Text-Mining- ist ein iterativer Prozess, in dem Extrahierungsergebnisse dem Kontext der Textdaten gemäß überprüft, für die Gewinnung neuer Ergebnisse verfeinert und dann neu bewertet werden. Überprüfen Sie nach dem Extrahieren die Ergebnisse und nehmen Sie Änderungen, die Sie für erforderlich halten, durch Bearbeiten der linguistischen Ressourcen vor. Sie können die Ressourcen zum Teil direkt vom Bereich “Extrahierungsergebnisse”, vom Datenbereich, vom Dialogfeld “Kategoriedefinitionen” oder vom Dialogfeld “Clusterdefinitionen” aus verfeinern. Für weitere Informationen siehe Thema Extrahierungsergebnisse verfeinern auf S. 158. Außerdem können Sie Verfeinerungen direkt in der Resource Editor-Ansicht vornehmen. Für weitere Informationen siehe Thema Die Ressourceneditoransicht in Kapitel 8 auf S. 133.

Nach der Verfeinerung können Sie erneut extrahieren, um die neuen Ergebnisse anzuzeigen. Wenn Sie Ihre Extrahierungsergebnisse von Anfang an verfeinern, gehen Sie sicher, dass Sie bei jeder neuen Extrahierung dieselben perfekt auf den Kontext der Daten angepassten Ergebnisse in Ihren Kategoriedefinitionen erhalten. So wird die Zuweisung von Dokumenten/Datensätzen zu Ihren Kategoriedefinitionen genauer und wiederholbar.

Konzepte

Beim Extrahieren werden die Textdaten untersucht und analysiert, um interessante oder relevante einzelne Wörter (z. B. `Wahl` oder `Frieden`) und zusammengesetzte Ausdrücke (z. B. `Wahl zur Präsidentschaft`, `Wahl des Präsidenten`, oder `Verhandlungen um Frieden`) im Text zu ermitteln. Diese Wörter und Zusammensetzungen werden auch als *Fachausdrücke* bezeichnet. Die relevanten Fachausdrücke werden mithilfe der linguistischen Ressourcen extrahiert und anschließend werden ähnliche Fachausdrücke unter einem übergeordneten Fachausdruck, einem **Konzept**, gruppiert.

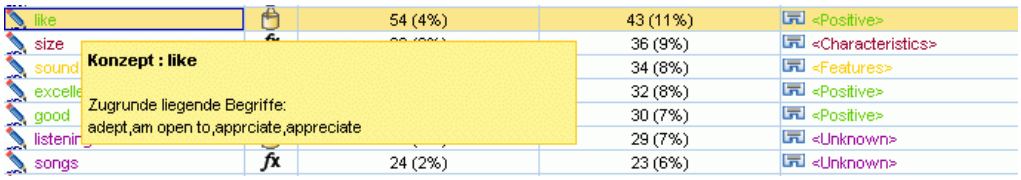
Abbildung 9-2
Fensterbereich "Extrahierungsergebnisse" nach einer Extrahierung



Konzept	In	Global	Dokumente	Typ
small	fx	58 (5%)	58 (14%)	<Contextual>
music	fx	56 (5%)	53 (13%)	<Features>
portable	fx	44 (4%)	43 (11%)	<Positive>
like	fx	55 (5%)	43 (11%)	<Positive>
size	fx	36 (3%)	36 (9%)	<Characteristics>
sound	fx	35 (3%)	34 (8%)	<Features>
excellent	fx	39 (3%)	32 (8%)	<Positive>
good	fx	31 (3%)	30 (7%)	<Positive>
listening		30 (3%)	29 (7%)	<Unknown>
songs		29 (2%)	26 (6%)	<Unknown>
large	fx	20 (2%)	20 (5%)	<Contextual>
product	fx	19 (2%)	18 (4%)	<Products>
appropriate	fx	17 (1%)	17 (4%)	<Positive>
battery	fx	16 (1%)	16 (4%)	<Performance>
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>
cool	fx	11 (1%)	11 (3%)	<Positive>

Sie können das Set an zugrundeliegenden Fachausdrücken sehen, indem Sie den Mauszeiger auf einen Konzeptnamen halten. Dadurch wird eine QuickInfo mit dem Konzeptnamen und bis zu mehrere Zeilen mit Ausdrücken angezeigt, die unter diesem Konzept gruppiert sind. Diese zugrundeliegenden Ausdrücke umfassen die Synonyme, die in den linguistischen Ressourcen definiert sind (unabhängig davon, ob sie im Text gefunden wurden) sowie extrahierte Plural-/Singularausdrücke, permutierte Ausdrücke, Ausdrücke aus unscharfen Gruppierungen usw. Sie können diese Ausdrücke kopieren oder das vollständige Set der zugrundeliegenden Ausdrücke anzeigen, indem Sie mit der rechten Maustaste auf den Konzeptnamen klicken und die Option aus dem Kontextmenü wählen.

Abbildung 9-3
Fensterbereich "Extrahierungsergebnisse" nach einer Extrahierung



Standardmäßig werden die Konzepte in Kleinbuchstaben angezeigt und absteigend entsprechend der Dokumentanzahl(Spalte Doc.), sortiert. Beim Extrahieren wird den Konzepten ein Typ zugewiesen, um das Gruppieren ähnlicher Konzepte zu erleichtern. Sie sind gemäß diesem Typ farblich gekennzeichnet. Die Farben sind in den Typ-Eigenschaften im Resource Editor definiert. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

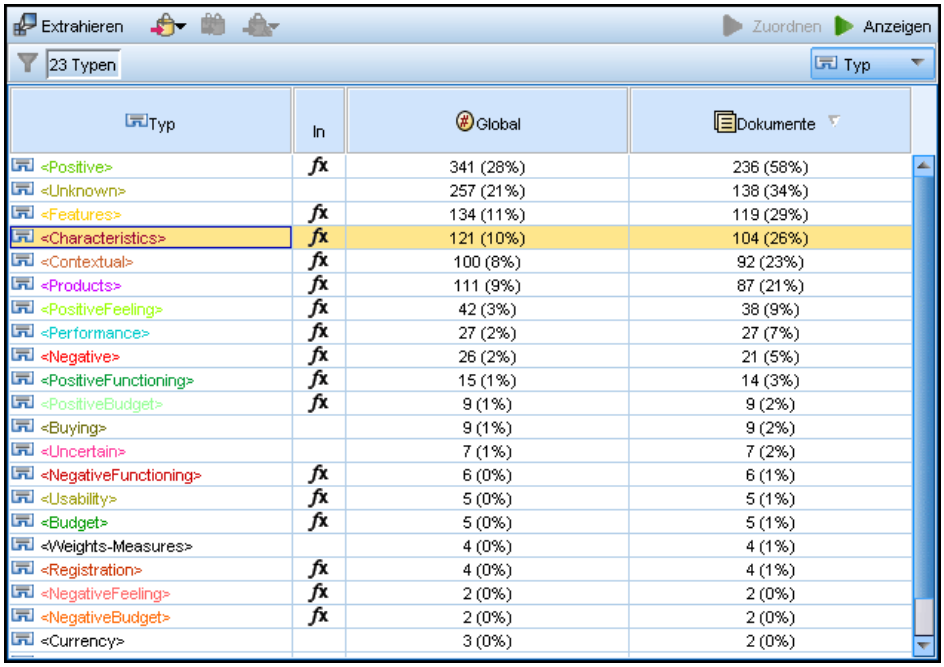
Wenn ein Konzept, Typ oder Muster in einer Kategoriedefinition verwendet wird, wird ein Symbol in der sortierbaren In-Spalte.

Typen

Typen sind semantische Gruppierungen von Konzepten. Beim Extrahieren wird den Konzepten ein Typ zugewiesen, um das Gruppieren ähnlicher Konzepte zu erleichtern. Es sind mehrere integrierte Typen im Lieferumfang von IBM® SPSS® Modeler Text Analytics enthalten, z. B. `<Location>`, `<Organization>`, `<Person>`, `<Positive>`, `<Negative>` usw. Der Typ `<Location>` gruppiert z. B. geografische Stichwörter und Orte. Dieser Typ würde Konzepten wie z. B. `Chicago`, `Paris` und `Tokio` zugewiesen. In den meisten Sprachen gilt: Konzepte, die in keinem Typen-Wörterbuch gefunden, aber aus dem Text extrahiert werden, erhalten automatisch den Typ `<Unbekannt>`. Bei japanischem Text jedoch erhalten sie automatisch den Typ <名詞>*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung. Für weitere Informationen siehe Thema Integrierte Typen in Kapitel 17 auf S. 314.

Wählen Sie die Ansicht "Typ", um die extrahierten Typen standardmäßig in absteigender Reihenfolge nach globaler Häufigkeit anzuzeigen. Sie sehen außerdem, dass die Typen farblich gekennzeichnet sind, um ihre Unterscheidung zu erleichtern. Farben sind Teil der Typeigenschaften. Für weitere Informationen siehe Thema Erstellen von Typen in Kapitel 17 auf S. 314. Sie können auch eigene Typen erstellen.

Abbildung 9-4
Typansicht: Fensterbereich "Extrahierungsergebnisse"



Typ	In	Global	Dokumente
<Positive>	fx	341 (28%)	236 (58%)
<Unknown>		257 (21%)	138 (34%)
<Features>	fx	134 (11%)	119 (29%)
<Characteristics>	fx	121 (10%)	104 (26%)
<Contextual>	fx	100 (8%)	92 (23%)
<Products>	fx	111 (9%)	87 (21%)
<PositiveFeeling>	fx	42 (3%)	38 (9%)
<Performance>	fx	27 (2%)	27 (7%)
<Negative>	fx	26 (2%)	21 (5%)
<PositiveFunctioning>	fx	15 (1%)	14 (3%)
<PositiveBudget>	fx	9 (1%)	9 (2%)
<Buying>		9 (1%)	9 (2%)
<Uncertain>		7 (1%)	7 (2%)
<NegativeFunctioning>	fx	6 (0%)	6 (1%)
<Usability>	fx	5 (0%)	5 (1%)
<Budget>	fx	5 (0%)	5 (1%)
<Weights-Measures>		4 (0%)	4 (1%)
<Registration>	fx	4 (0%)	4 (1%)
<NegativeFeeling>	fx	2 (0%)	2 (0%)
<NegativeBudget>	fx	2 (0%)	2 (0%)
<Currency>		3 (0%)	2 (0%)

Muster

Außerdem können Muster aus Ihren Textdaten extrahiert werden. Es muss jedoch eine Bibliothek vorhanden sein, die einige Musterregeln für die Text Link Analysis (TLA) im Resource Editor enthält. Sie müssen darüber hinaus die Extrahierung dieser Muster in der Knoteneinstellung für SPSS Modeler Text Analytics oder im Dialogfeld "Extrahieren" über die

Option Musterextrahierung für Textlinkanalyse aktivieren auswählen. Für weitere Informationen siehe Thema Untersuchen von Textlinkanalyse in Kapitel 12 auf S. 252.

Daten extrahieren

Wenn eine Extrahierung nötig ist, wird der Bereich "Extrahierungsergebnisse" in gelb dargestellt und die Nachricht Schaltfläche "Extrahieren" drücken, um Konzepte zu extrahieren wird unter der Symbolleiste in diesem Bereich angezeigt.

Eventuell müssen Sie extrahieren, wenn Sie noch keine Extrahierungsergebnisse haben, linguistische Ressourcen geändert haben und die Extrahierungsergebnisse aktualisieren müssen oder eine Sitzung neu geöffnet haben, wo Sie die Extrahierungsergebnisse nicht gespeichert haben (Extras > Optionen).

Anmerkung: Wenn Sie den Quellenknoten für Ihren Stream ändern, nachdem Extrahierungsergebnisse mit der Option Arbeit der Sitzung verwenden... im Cache abgelegt wurden, müssen Sie eine neue Extrahierung ausführen, sobald die interaktive Workbench-Sitzung gestartet wird, damit Sie aktualisierte Extrahierungsergebnisse erhalten.

Wenn Sie eine Extrahierung durchführen, erscheint ein Fortschrittsbalken, der den Status der Extrahierung anzeigt. Währenddessen liest die Extrahierungsengine alle Textdaten, identifiziert die relevanten Ausdrücke und Muster, extrahiert sie und weist sie einem Typ zu. Dann versucht das Modul, synonyme Ausdrücke unter einem Leitausdruck, einem Konzept, zu gruppieren. Wenn der Vorgang abgeschlossen ist, werden die resultierenden Konzepte, Typen und Muster im Bereich "Extrahierungsergebnisse" angezeigt.

Der Extrahierungsprozess liefert eine Reihe von Konzepten und Typen und, sofern aktiviert, Textlinkanalyse-(TLA-)Muster. Sie können diese Konzepte und Typen im Bereich "Extrahierungsergebnisse" in der Ansicht "Kategorien und Konzepte" betrachten und dort mit ihnen arbeiten. Extrahierte TLA-Muster können Sie in der Textlinkanalysenansicht sehen.

Anmerkung: Die für den Extrahierungsprozess benötigte Zeit steht in direkter Beziehung zur Größe Ihrer Datenmenge. Es ist stets möglich, aufsteigend einen Stichprobenknoten einzufügen oder die Konfiguration Ihres Computers zu optimieren.

Daten extrahieren

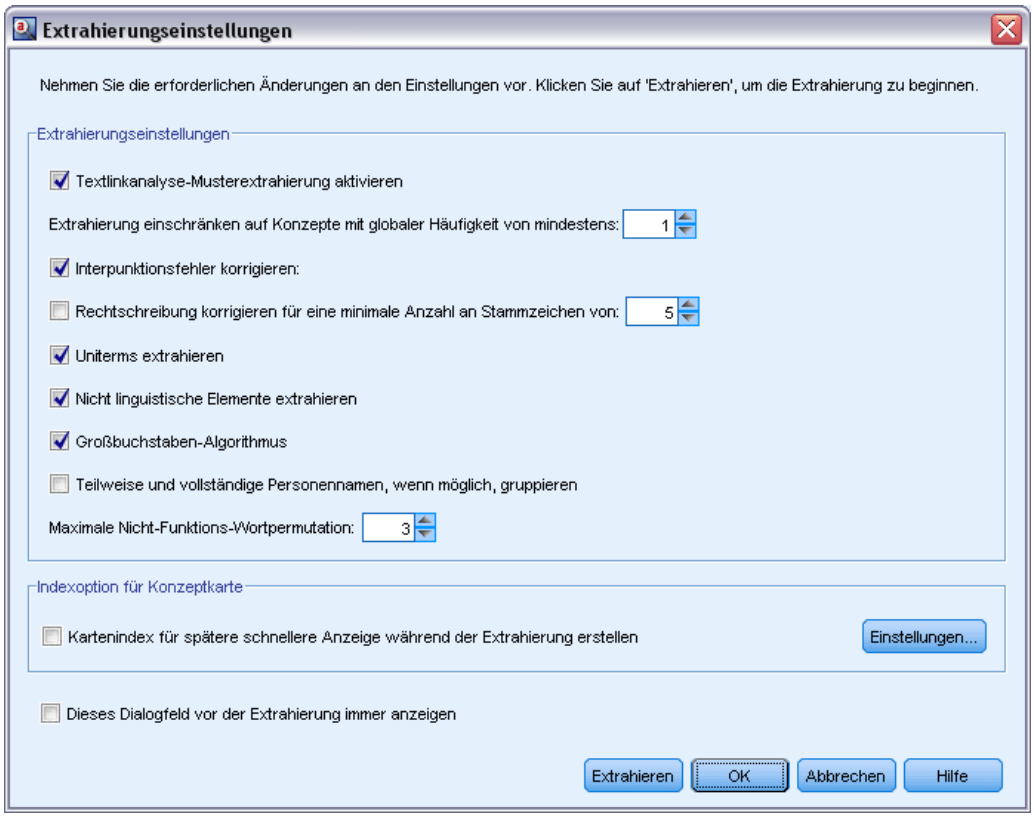
- Wählen Sie in den Menüs die Optionsfolge Extras > Extrahieren aus. Alternativ können Sie auf die Symbolleistenschaltfläche Extrahieren klicken.
- Wenn Sie das Dialogfeld "Extrahierungseinstellungen" immer anzeigen lassen, wird es angezeigt, damit Sie Änderungen vornehmen können. Weitere Informationen zu Deskriptoren für jede Einstellung finden Sie in diesem Thema.
- Klicken Sie auf Extrahieren, um die Extrahierung zu starten. Sobald die Extrahierung beginnt, öffnet sich die Statusanzeige. Nach der Extrahierung werden die Ergebnisse im Bereich "Extrahierungsergebnisse" dargestellt. Standardmäßig werden die Konzepte in Kleinbuchstaben angezeigt und absteigend entsprechend der Dokumentanzahl(Spalte Doc.), sortiert.

Sie können die Ergebnisse überprüfen, indem Sie sie mithilfe der Optionen in der Symbolleiste unterschiedlich sortieren und filtern oder die Ansicht wechseln (Konzepte, oder Typen). Sie können die Ergebnisse auch verfeinern, indem Sie mit den linguistischen Ressourcen arbeiten. Für weitere Informationen siehe Thema Extrahierungsergebnisse verfeinern auf S. 158.

Für niederländischen, englischen, französischen, deutschen, italienischen, portugiesischen und spanischen Text

Das Dialogfeld “Extrahierungseinstellungen” enthält einige grundlegende Extrahierungsoptionen.

Abbildung 9-5
Dialogfeld “Extrahierungseinstellungen”



Musterextrahierung für Textlinkanalyse aktivieren. Gibt an, dass Sie TLA-Muster aus Ihren Textdaten extrahieren möchten. Diese Option setzt außerdem voraus, dass TLA-Musterregeln in einer Ihrer Bibliotheken in dem Ressourceneditor vorhanden sind. Diese Option kann die Extrahierungsdauer erheblich verlängern. Für weitere Informationen siehe Thema Untersuchen von Textlinkanalyse in Kapitel 12 auf S. 252.

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Punktuationsfehlern (zum Beispiel ungeeignete Verwendung) während der Extrahierung, um die Extrahierbarkeit von Konzepten zu verbessern. Diese Option ist besonders nützlich bei kurzem Text und niedriger Textqualität (wie dies beispielsweise bei offenen Antworten bei Umfragen, E-Mails und CRM-Daten der Fall ist) oder wenn der Text viele Abkürzungen enthält.

Rechtschreibung korrigieren für eine minimale Anzahl an Stammzeichen von [n]. Diese Option wendet ein unscharfes Gruppierungsverfahren an, das hilft, häufig falsch geschriebene Wörter oder genau geschriebene Wörter unter einem Konzept zu gruppieren. Der Algorithmus für die unscharfe Gruppierung entfernt alle Vokale (außer den ersten) und doppelte/dreifache Konsonanten temporär aus extrahierten Wörtern und vergleicht sie, um festzustellen, ob sie gleich sind, so dass `Modellierung` und `Modelierung` zusammen gruppiert werden würden. Wenn jedoch jeder Fachausdruck einem anderen Typ (ausschließlich des Typs `<Unknown>`) zugewiesen ist, wird das unscharfe Gruppierungsverfahren nicht angewendet.

Sie können auch die minimal erforderliche Zahl von *Stamm*zeichen definieren, bevor unscharfe Gruppierung eingesetzt wird. Die Anzahl der Stammzeichen in einem Ausdruck berechnet sich aus der Summe aller Zeichen abzüglich aller Zeichen, die Beugungsendungen und – bei zusammengesetzten Ausdrücken – Determinatoren und Präpositionen bilden. So würde beispielsweise der Ausdruck `Aufgaben` durch die Form "Aufgabe" mit 7 Stammzeichen gezählt werden, da der Buchstabe *n* am Ende des Worts eine Beugung darstellt (Pluralform). Gleichermaßen werden für `Apfelmus` 8 Stammzeichen ("Apfelmus") gezählt und `Hersteller von Autos` zählt als 14 Stammzeichen ("Hersteller Auto"). Diese Zählmethode dient nur zur Überprüfung, ob die Fuzzy-Gruppierung angewendet werden soll, hat jedoch keinen Einfluss auf den Abgleich der Wörter.

Hinweis: Wenn Sie feststellen, dass bestimmte Wörter später falsch gruppiert werden, können Sie Wortpaare aus diesem Verfahren ausschließen, indem Sie sie explizit im Abschnitt Unscharfe Gruppierung: Ausnahmen im erweiterten Ressourceneditor deklarieren. Für weitere Informationen siehe Thema Unscharfe Gruppierung in Kapitel 18 auf S. 336.

uniterms extrahieren. Diese Option extrahiert einzelne Wörter (Uniterms), solange das Wort nicht bereits Teil eines zusammengesetzten Worts ist und es entweder ein Nomen oder eine nicht erkannte Wortart ist.

Nicht linguistische Elemente extrahieren. Diese Option extrahiert nicht linguistische Elemente wie beispielsweise Telefonnummern, Personalausweisnummern, Uhrzeiten, Datumsangaben, Währungen, Ziffern, Prozentsätze, E-Mail-Adressen und HTTP-Adressen. Sie können bestimmte Typen von nicht linguistischen Elementen im Abschnitt Nicht linguistische Elemente: Konfiguration des erweiterten Ressourceneditors ein- bzw. ausschließen. Durch Deaktivierung unnötiger Elemente vergeudet die Extrahierungsengine keine Verarbeitungszeit. Für weitere Informationen siehe Thema Konfiguration in Kapitel 18 auf S. 341.

Großbuchstaben-Algorithmus. Diese Option extrahiert einfache und zusammengesetzte Ausdrücke, die sich nicht in den integrierten Wörterbüchern befinden, solange der erste Buchstabe des Begriffs in Großbuchstaben geschrieben ist. Diese Option ist eine gute Möglichkeit, die geeignetsten Substantive zu extrahieren.

Teilweise und vollständige Personennamen, wenn möglich, gruppieren. Diese Option gruppiert Namen, die zusammen im Text unterschiedlich erscheinen. Diese Funktion ist nützlich, da Namen zu Beginn des Textes oft in voller Länge angegeben werden und später nur noch mit einer Kurzform auf sie verwiesen wird. Diese Option versucht, jeden Uniterm mit dem Typ `<Unknown>` mit dem letzten Wort aller zusammengesetzten Ausdrücke abzugleichen, die dem Typ `<Person>` zugeordnet sind. Wird beispielsweise *doe* gefunden und anfänglich dem Typ `<Unknown>` zugeordnet, überprüft die Extrahierungsengine, ob ein zusammengesetzter Ausdruck

vom Typ `<Person>` als letztes Wort *doe* enthält, z. B. *john doe*. Diese Option wird nicht auf Vornamen angewendet, da sie in den meisten Fällen nicht als Uniterms extrahiert werden.

Maximale Nicht-Funktions-Wortpermutation. Diese Option gibt die maximale Anzahl von Füllwörtern an, die für die Anwendung des Permutationsverfahrens vorhanden sein müssen. Dieses Permutationsverfahren gruppiert ähnliche Phrasen, die sich nur durch die enthaltenen Füllwörter (zum Beispiel von und der) unabhängig von der Beugung unterscheiden. Nehmen wir zum Beispiel an, dass Sie diesen Wert auf höchstens zwei Wörter eingestellt haben und sowohl `Unternehmen des Vertreters` und `Vertreter des Unternehmens` extrahiert wurden. In diesem Fall würden beide extrahierte Ausdrücke in der endgültigen Konzeptliste zusammen gruppiert, da beide Ausdrücke als gleich betrachtet werden, wenn `des` ignoriert wird.

Indexoption für Konzeptkarte Gibt an, dass Sie den Kartenindex zur Zeit der Extrahierung erstellen möchten, damit die Konzeptkarten später schneller dargestellt werden können. Um die Indexeinstellungen zu bearbeiten, klicken Sie auf Einstellungen. Für weitere Informationen siehe Thema Erzeugen von Konzeptkartenindizes auf S. 157.

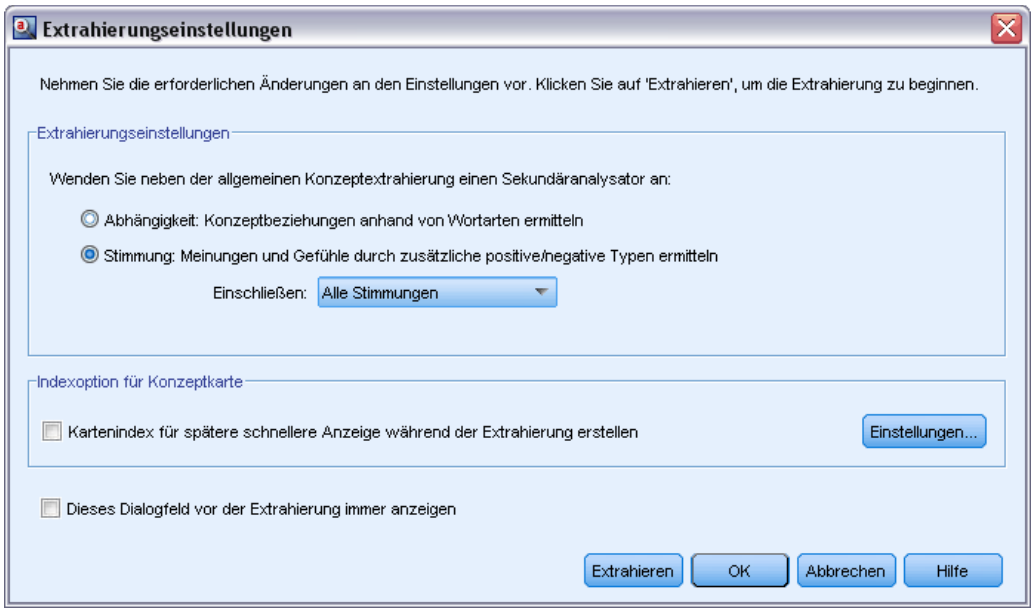
Dieses Dialogfeld vor der Extrahierung immer anzeigen. Legen Sie fest, ob Sie das Dialogfeld “Extrahierungseinstellungen” bei jeder Extrahierung anzeigen möchten, ob Sie es nie anzeigen möchten (außer beim Aufruf über das Menü “Extras”) oder ob Sie bei jeder Extrahierung gefragt werden möchten, ob Sie Änderungen an den Extrahierungseinstellungen vornehmen wollen.

Für japanischen Text

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Das Dialogfeld „Extrahierungseinstellungen“ enthält einige grundlegende Extrahierungsoptionen für die japanische Textsprache. Standardmäßig sind die im Dialogfeld ausgewählten Einstellungen mit denen auf der Registerkarte “Experten” des Text-Mining-Modellierungsknotens identisch. Für die Arbeit mit japanischem Text müssen Sie den Text als Eingabe verwenden sowie eine japanische Sprachvorlage oder ein japanisches Textanalysepaket in der Registerkarte “Modell” des Text-Mining-Knotens wählen. Für weitere Informationen siehe Thema Kopieren von Ressourcen aus TAPs und Vorlagen in Kapitel 3 auf S. 42.

Abbildung 9-6
Dialogfeld "Extrahierungseinstellungen" für japanischen Text



Anmerkung: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Sekundäre Analyse. Beim Start einer Extrahierung werden anhand des Standardsatzes an Typen grundlegende Stichwörter extrahiert. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text in Anhang A auf S. 387. Wenn Sie jedoch eine sekundäre Analyse wählen, erhalten Sie mehr und vielfältigere Konzepte, da der Extraktor nun Partikel und Hilfsverben als Teil des Konzepts berücksichtigt. Nehmen Sie beispielsweise an, der Satz 肩の荷が下りた wurde in "*Mir wurde eine große Last von den Schultern genommen*". In diesem Beispiel kann die grundlegende Stichwortextrahierung jedes Konzept wie folgt separat extrahieren: 肩 (*Schultern*), 荷 (*Last*), 下りる (*wurde genommen*), jedoch wird die Beziehung zwischen diesen Wörtern nicht extrahiert. Wenn Sie jedoch die Stimmungsanalyse anwenden, können Sie vielfältigere Konzepte hinsichtlich eines Stimmungstyps extrahieren, z. B. das Konzept =肩の荷が下りた, das als "*eine große Last von den Schultern nehmen*", was dem Typ <良い-安心> zugeordnet ist. Im Fall der Stimmungsanalyse wird auch eine große Anzahl an zusätzlichen Typen eingeschlossen. Des Weiteren ermöglicht Ihnen die Wahl einer sekundären Analyse, auch Textlinkanalysis-Ergebnisse zu generieren.

Anmerkung: Wenn ein Sekundär-Analysator genannt wird, dauert der Extrahierungsprozess länger. Für weitere Informationen siehe Thema Wie sekundäre Extrahierung funktioniert in Anhang A auf S. 380.

- **Abhängigkeitsanalyse**. Die Wahl dieser Option führt zu erweiterten Partikeln für die Extrahierungskonzepte aus der grundlegenden Typ- und Stichwortextrahierung. Sie können auch vielfältigere Musterergebnisse aus einer Abhängigkeits-Textlinkanalyse (TLA) gewinnen.
- **Stimmungsanalyse**. Bei dieser Analyse werden zusätzliche Konzepte und – wann immer möglich – TLA-Musterergebnisse extrahiert. Neben den grundlegenden Typen erhalten Sie auch den Vorteil von über 80 Stimmungstypen, z. B. 嬉しい, 吉報, 幸運, 安心, 幸福

usw. Mithilfe dieser Typen werden Konzepte und Muster im Text durch den Ausdruck von Emotionen, Stimmungen und Meinungen aufgedeckt. Es gibt drei Optionen, die den Fokus für die Stimmungsanalyse festlegen: Alle Stimmungen, Nur repräsentative Stimmung und Nur Schlussfolgerungen.

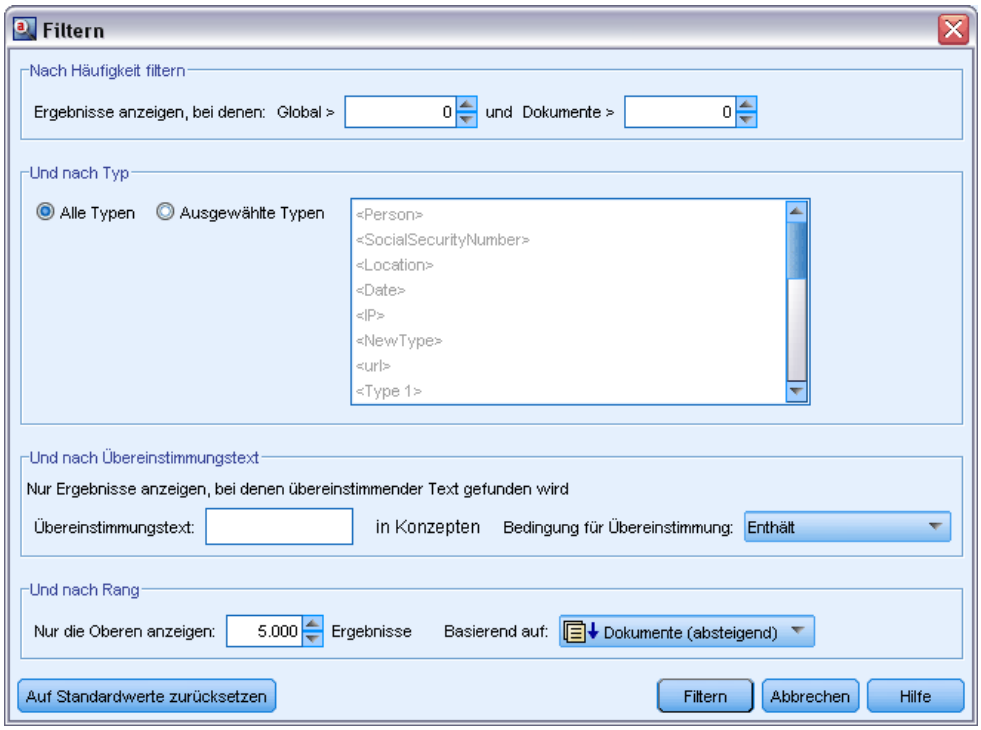
- **Keine sekundäre Analyse**. Diese Option schaltet sämtliche sekundären Analysen aus. Diese Option ist nicht wählbar, wenn die Option Musterextrahierung für Textlinkanalyse aktivieren ausgewählt wurde, da eine sekundäre Analyse erforderlich ist, um TLA-Ergebnisse zu erhalten.

Musterextrahierung für Textlinkanalyse aktivieren. Gibt an, dass Sie TLA-Muster aus Ihren Textdaten extrahieren möchten. Diese Option setzt außerdem voraus, dass TLA-Musterregeln in einer Ihrer Bibliotheken in dem Ressourceneditor vorhanden sind. Diese Option kann die Extrahierungsdauer erheblich verlängern. Zusätzlich muss eine sekundäre Analyse ausgewählt werden, um TLA-Musterergebnisse extrahieren zu können. Für weitere Informationen siehe Thema Untersuchen von Textlinkanalyse in Kapitel 12 auf S. 252.

Extrahierungsrgebnisse filtern

Wenn Sie mit sehr großen Datenmengen arbeiten, kann der Extrahierungsprozess Millionen von Ergebnissen liefern. Durch diese Menge ist eine effektive Überprüfung der Ergebnisse für viele Benutzer mühsam. Um daher auf die interessantesten einzuzoomen, können Sie diese Ergebnisse über das Dialogfeld “Filter” im Bereich “Extrahierungsergebnisse” filtern.

Abbildung 9-7
Dialogfeld “Filter” (im Bereich “Extrahierungsergebnisse”)



Denken Sie daran, dass alle Einstellungen in diesem Dialogfeld “Filter” gemeinsam verwendet werden, um die Extrahierungsergebnisse zu filtern, die für Kategorien verfügbar sind.

Nach Häufigkeit filtern. Mit diesem Filter werden nur Ergebnisse mit einem bestimmten globalen oder Dokumenthäufigkeitswert angezeigt.

- Die **globale Häufigkeit** gibt an, wie oft ein Konzept in der Gesamtmenge der Dokumente bzw. Datensätze vorkommt, und wird in der Spalte Globalwert angezeigt.
- Die **Häufigkeit im Dokument** gibt an, wie oft ein Konzept in der Gesamtmenge der Dokumente bzw. Datensätze vorkommt, und wird in der Spalte Dokumente angezeigt.

Wenn z. B. das Konzept `Nato` 800-mal in 500 Datensätzen vorkommt, hat dieses Konzept eine globale Häufigkeit von 800 und eine Häufigkeit im Dokument von 500.

Und nach Typ. Sie können einen Filter setzen, um nur die Ergebnisse anzuzeigen, die zu bestimmten Typen gehören. Sie können alle Typen oder nur bestimmte Typen auswählen.

Und nach Übereinstimmungstext. Sie können auch einen Filter anwenden, durch den nur Ergebnisse angezeigt werden, die mit den hier definierten Regeln übereinstimmen. Geben Sie die Zeichenfolge, die bei einer Übereinstimmung erkannt werden soll, in das Feld Übereinstimmungstext ein und wählen Sie anschließend die Bedingung aus, bei der die Übereinstimmung erkannt werden soll.

Tabelle 9-1
Bedingungen für Übereinstimmungstext

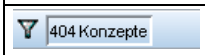
Bedingung	Beschreibung
Enthält	Es liegt eine Übereinstimmung mit dem Text vor, wenn diese Zeichenfolge im Text vorkommt. (Standardauswahl)
Beginnt mit	Text stimmt nur überein, wenn das Konzept oder der Typ mit dem angegebenen Text beginnt.
Endet mit	Text stimmt nur überein, wenn das Konzept oder der Typ mit dem angegebenen Text endet.
Genaue Übere-instimmung	Die gesamte Zeichenfolge muss mit dem Konzept- oder dem Typnamen übereinstimmen.

Und nach Rang. Sie können auch einen Filter setzen, um nur eine oberste Anzahl von Konzepten gemäß der globalen Häufigkeit (global) oder der Häufigkeit im Dokument (Dokumente) in aufsteigender oder in absteigender Reihenfolge anzuzeigen.

Im Bereich “Extrahierungsergebnisse” angezeigte Ergebnisse

Beispiele für die Anzeige der Ergebnisse in der Symbolleiste des Bereichs “Extrahierungsergebnisse” basierend auf den Filtern.

Tabelle 9-2
Beispiele für Filterfeedback

Filterfeedback	Beschreibung
	Die Symbolleiste zeigt die Anzahl der Ergebnisse an. Da kein Textübereinstimmungsfilter gesetzt war und die Höchstzahl nicht erreicht wurde, werden keine weiteren Symbole angezeigt.
	Die Symbolleiste zeigt an, dass die Ergebnisse auf die in dem Filter angegebene Höchstzahl beschränkt wurde, in diesem Fall 300. Wenn ein violettes Symbol angezeigt wird, wurde die Höchstzahl an Konzepten erreicht. Für weitere Informationen bewegen Sie die Maus über das Symbol.
	Die Symbolleiste zeigt, dass die Ergebnisse über einen Übereinstimmungstextfilter eingeschränkt wurden. Diese wird durch ein Vergößerungsglassymbol angezeigt.

Ergebnisse filtern

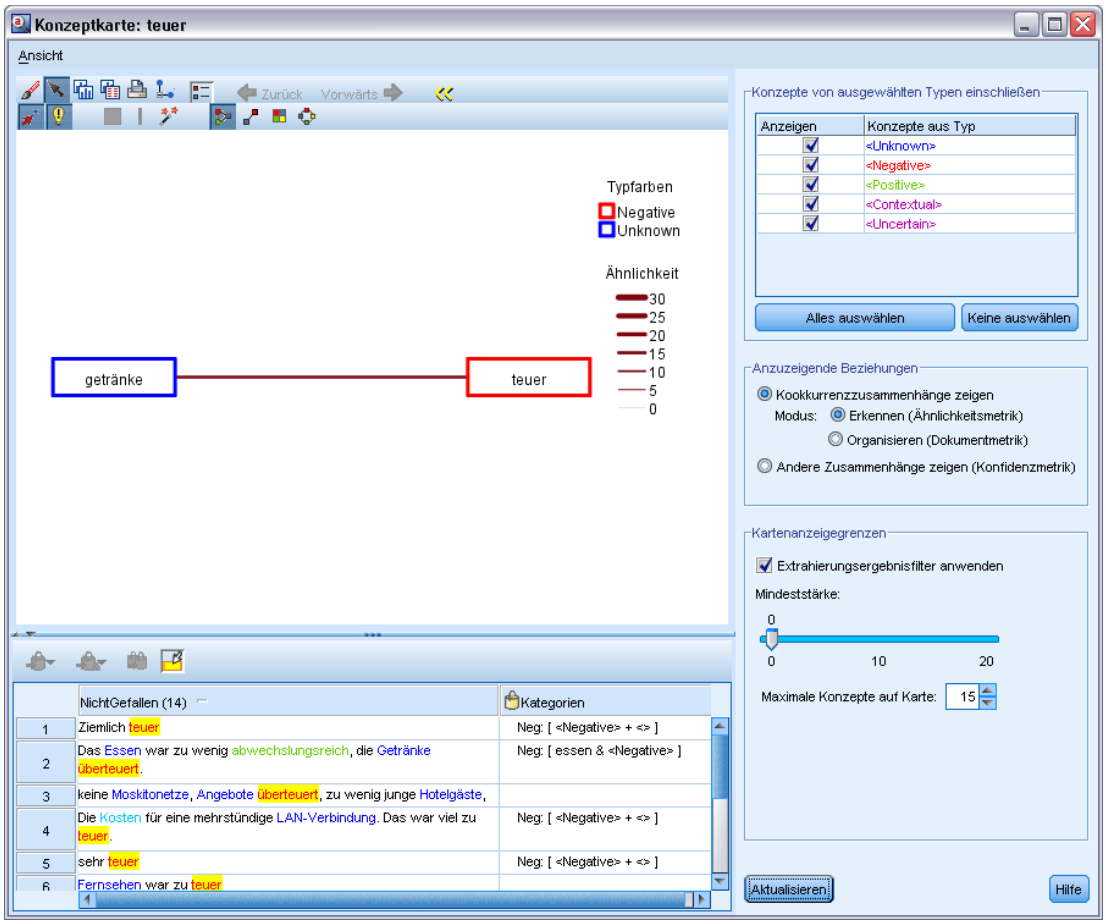
- Wählen Sie in den Menüs die Optionsfolge Extras > Filter aus. Das Filterdialogfeld wird geöffnet.
- Wählen und verfeinern Sie die Filter, die Sie verwenden möchten.
- Klicken Sie auf OK, um die Filter anzuwenden und die neuen Ergebnisse im Bereich “Extrahierungsergebnisse” anzuzeigen.

Untersuchen von Konzeptkarten

Sie können eine Konzeptkarte erstellen, um festzustellen, wie Konzepte miteinerander in Zusammenhang stehen. Wenn Sie ein einzelnes Konzept auswählen und auf Zuordnung klicken, wird ein Fenster mit einer Konzeptkarte geöffnet, über die Sie die Konzepte untersuchen können, die mit dem gewählten Konzept zusammenhängen. Sie können filtern, welche Konzepte angezeigt werden, indem Sie die Einstellungen bearbeiten, also beispielsweise, welche Typen berücksichtigt werden sollen oder nach welchen Beziehungen gesucht werden soll.

Wichtig: Bevor eine Karte erstellt werden kann, muss ein Index erzeugt werden. Dieser Vorgang kann mehrere Minuten in Anspruch nehmen. Wenn Sie den Index erstellt haben, müssen Sie erst wieder einen Index erzeugen, wenn Sie eine neue Extrahierung vornehmen. Wenn der Index bei jeder Extrahierung automatisch erstellt werden soll, wählen Sie in den Extrahierungseinstellungen diese Option. Für weitere Informationen siehe Thema Daten extrahieren auf S. 147.

Abbildung 9-8
Eine Konzeptkarte für das gewählte Konzept



So erstellen Sie eine Konzeptkarte:

- Wählen Sie im Bereich “Extrahierungsergebenisse” ein einzelnes Konzept aus.
- Klicken Sie in der Symbolleiste dieses Bereichs auf die Schaltfläche Zuordnung. Wenn der Kartenindex bereits generiert war, wird die Konzeptkarte in einem separten Dialogfeld geöffnet. Wenn der Kartenindex noch nicht generiert oder veraltet war, muss der Index erneut erstellt werden. Dieser Vorgang kann mehrere Minuten in Anspruch nehmen.
- Klicken Sie an beliebige Punkte in der Karte, um die Ergebnisse zu untersuchen. Wenn Sie auf ein verknüpftes Konzept doppelklicken, wird die Karte erneut erzeugt und zeigt die verknüpften Konzepte für das Konzept an, auf das Sie eben doppelgeklickt haben.
- In der oberen Symbolleiste werden einige allgemeine Zuordnungstools bereitgestellt wie das Wechseln zu einer vorherigen Karte, das Filtern von Verknüpfungen nach der jeweiligen Beziehungsstärke und das Öffnen des Filterdialogfelds, über das gesteuert wird, welche Konzepttypen und Arten von Beziehungen angezeigt werden sollen. Die zweite Symbolleiste

enthält die Tools zum Bearbeiten von Diagrammen. Für weitere Informationen siehe Thema Verwenden von Diagrammsymbolleisten und Paletten in Kapitel 13 auf S. 271.

▶ Wenn Sie mit der Art der gefundenen Verknüpfungen nicht zufrieden sind, untersuchen Sie die Einstellungen für diese Karte rechts von der Karte.

Karteneinstellungen: Konzepte von ausgewählten Typen einschließen

Nur die Konzepte, die zu den ausgewählten Typen in der Tabelle gehören, werden in der Karte angezeigt. Um Konzepte eines bestimmten Typs auszublenden, deaktivieren Sie den Typ in der Tabelle.

Karteneinstellungen: Anzuzeigende Beziehungen

Kookkurrenz-Verknüpfungen anzeigen. Wenn Sie Kookkurrenz-Verknüpfungen anzeigen möchten, wählen Sie diesen Modus. Der Modus wirkt sich darauf aus, wie die Höhe berechnet wurde.

- *Erkennen (Ähnlichkeitsmetrik)*. Mit dieser Metrik wird die Stärke einer Verknüpfung anhand einer komplexeren Berechnung ermittelt, die berücksichtigt, wie häufig zwei Konzepte getrennt voneinander und wie häufig sie gemeinsam auftreten. Ein hoher Wert für die Stärke bedeutet, dass ein Paar von Konzepten häufiger zusammen als getrennt voneinander auftritt. Anhand dieser Formel werden alle Fließpunktwerte in Ganzzahlen umgewandelt.

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

  Dabei ist C_I die Anzahl der Dokumente oder Datensätze, in denen das Konzept `I` vorkommt.

  C_J ist die Anzahl der Dokumente oder Datensätze, in denen das Konzept `J` vorkommt.

  C_{IJ} ist die Anzahl der Dokumente oder Datensätze, in denen das Konzeptpaar `I` und `J` gemeinsam im Dokumentensatz vorkommt.

- *Organisieren (Dokumentmetrik)*. Die Verknüpfungsstärke bei dieser Metrik wird durch die reine Anzahl der Kookkurenzen bestimmt. Im Allgemeinen gilt: Je häufiger die beiden Konzepte sind, desto wahrscheinlicher treten sie gemeinsam auf. Ein hoher Stärkewert bedeutet, dass ein Konzeptpaar häufig zusammen vorkommt.

Andere Zusammenhänge zeigen (Konfidenzmetrik). Sie können andere Zusammenhänge für die Anzeige auswählen. Dabei kann es sich um semantische Zusammenhänge, Ableitung (morphologisch) oder eine Einbeziehung (syntaktisch) handeln, und diese Zusammenhänge sind darauf bezogen, wie viele Schritte ein Konzept von dem Konzept entfernt ist, mit dem es verknüpft ist. Dadurch wird die Feinabstimmung der Ressourcen (insbesondere Synonymie) bzw. die Disambiguierung erleichtert.Kurzbeschreibungen zu den einzelnen Gruppierverfahren finden Sie hier: Erweiterte linguistische Einstellungen auf S. 187

Hinweis: Beachten Sie, dass keine angezeigt werden, wenn diese nicht ausgewählt wurden, als der Index erstellt wurde, oder wenn keine Beziehungen gefunden werden. Für weitere Informationen siehe Thema Erzeugen von Konzeptkartenindizes auf S. 157.

Karteneinstellungen: Kartenanzeigegrenzen

Extrahierungsergebnisfilter anwenden. Wenn Sie nicht alle Konzepte verwenden wollen, können Sie den Filter in den Extrahierungsergebnissen verwenden, um das Gezeigte einzugrenzen. Wählen Sie dann die Option. IBM® SPSS® Modeler Text Analytics sucht nach zugehörigen Konzepten unter Verwendung dieser gefilterten Menge. Für weitere Informationen siehe Thema Extrahierungsrgebnisse filtern auf S. 152.

Mindeststärke. Legen Sie hier die Mindeststärke für den Zusammenhang fest. Zugehörige Konzepte mit einer Beziehungsstärke, die unter dieser Grenze liegt, wird aus der Karte ausgeblendet.

Maximale Konzepte auf Karte. Geben Sie die maximale Zahl an Beziehungen an, die auf der Karte angezeigt werden.

Erzeugen von Konzeptkartenindizes

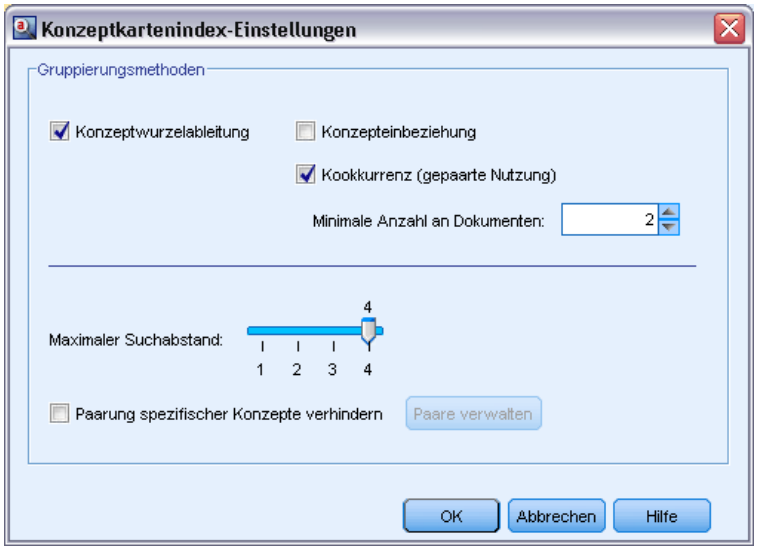
Bevor eine Karte erstellt werden kann, muss ein Index von Konzeptbeziehungen erzeugt werden. Wenn Sie eine Konzeptkarte erstellen, greift IBM® SPSS® Modeler Text Analytics auf diesen Index zurück. Sie können wählen, welche Beziehungen zu indizieren sind, indem Sie die Techniken in diesem Dialogfeld auswählen.

Gruppierverfahren. Wählen Sie ein oder mehrere Verfahren aus. Eine kurze Beschreibung dieser Verfahren finden Sie unter Über linguistische Verfahren auf S. 192. Nicht alle Techniken stehen für alle Sprachen zur Verfügung.

Paarung spezifischer Konzepte verhindern. Markieren Sie dieses Kontrollkästchen, um den Vorgang der Gruppierung oder Paarung von zwei Konzepten in der Ausgabe zu verhindern. Klicken Sie zum Erstellen oder Verwalten von Konzeptpaaren auf Paare verwalten. Für weitere Informationen siehe Thema Verwalten von Verknüpfungsausnahmepaaren in Kapitel 10 auf S. 191.

Dieser Vorgang kann mehrere Minuten in Anspruch nehmen. Wenn Sie den Index erstellt haben, müssen Sie erst wieder einen Index erzeugen, wenn Sie eine neue Extrahierung vornehmen oder wenn Sie die Einstellungen so ändern, dass mehr Beziehungen einbezogen werden. Wenn Sie einen Index generieren möchten, wenn Sie extrahieren, können Sie diese Option in den Extrahierungseinstellungen wählen. Für weitere Informationen siehe Thema Daten extrahieren auf S. 147.

Abbildung 9-9
Indexeinstellungen



Extrahierungsergebnisse verfeinern

Die Extrahierung ist ein iterativer Prozess, in dem Sie extrahieren, die Ergebnisse überprüfen und ändern und dann erneut extrahieren können, um die Ergebnisse zu aktualisieren. Da Genauigkeit und Kontinuität für die erfolgreiche Durchführung von Text-Mining und Kategorisierung unverzichtbar sind, gewährleistet das Verfeinern Ihrer Extrahierungsergebnisse von Anfang an, dass Sie bei jeder erneuten Extrahierung genau dieselben Ergebnisse in Ihren Kategoriedefinitionen erhalten. So wird die Zuweisung von Datensätzen und Dokumenten zu Ihren Kategorien genauer und wiederholbar.

Die Extrahierungsergebnisse dienen als Bausteine für Ihre Kategorien. Beim Erstellen von Kategorien mithilfe der Extrahierungsergebnisse werden Datensätze und Dokumente automatisch Kategorien zugewiesen, wenn sie Text enthalten, der mit einem oder mehreren Kategoriedeskriptoren übereinstimmt. Sie können zwar vor dem Verfeinern der linguistischen Ressourcen mit der Kategorisierung beginnen, aber es ist nützlich, die Extrahierungsergebnisse vorher mindestens einmal zu überprüfen.

Beim Überprüfen Ihrer Ergebnisse stoßen Sie möglicherweise auf Elemente, die die Extrahierungsengine anders verarbeiten soll. Betrachten Sie die folgende Beispiele:

- **Nicht erkannte Synonyme.** Angenommen, Sie entdecken mehrere Konzepte, die Sie als Synonyme betrachten, z. B. `schlau`, `intelligent`, `gescheit` und `klug`, und alle sind als einzelne Konzepte in den Extrahierungsergebnissen enthalten. Dann können Sie eine Synonymdefinition erstellen, nach der `intelligent`, `gescheit` und `klug` unter dem Zielkonzept `schlau` gruppiert werden. So werden diese Konzepte mit `schlau` zusammengefasst und die globale Häufigkeitsanzahl steigt ebenfalls. Für weitere Informationen siehe Thema Synonyme hinzufügen auf S. 160.
- **Den falschen Typen zugeordnete Konzepte.** Angenommen, die Konzepte in Ihren Extrahierungsergebnissen werden bei einem Typ aufgeführt und Sie möchten Sie einem anderen Typen zuweisen. Ein weiteres Beispiel: Sie finden in den Extrahierungsergebnissen

15 Konzepte zu Pflanzen und Sie möchten alle einem neuen Typ mit der Bezeichnung `<Pflanze>` hinzufügen. In den meisten Sprachen gilt: Konzepte, die in keinem Typen-Wörterbuch gefunden, aber aus dem Text extrahiert werden, erhalten automatisch den Typ `<Unbekannt>`. Bei japanischem Text jedoch erhalten sie automatisch den Typ <名詞 >*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung. Sie können Konzepte zu Typen hinzufügen. Für weitere Informationen siehe Thema Konzepte zu Typen hinzufügen auf S. 162.

- **Bedeutungslose Konzepte.** Angenommen, Sie stoßen auf ein extrahiertes Konzept mit einer sehr hohen Häufigkeitsanzahl—, d. h., es kommt in vielen Datensätzen oder Dokumenten vor. Sie halten dieses Konzept aber für bedeutungslos für Ihre Analyse. Dann können Sie es von der Extrahierung ausschließen. Für weitere Informationen siehe Thema Konzepte von der Extrahierung ausschließen auf S. 164.
- **Falsch erkannte Übereinstimmungen.** Angenommen, Sie stellen bei der Überprüfung der Datensätze oder Dokumente mit einem bestimmten Konzept fest, dass zwei Wörter fälschlicherweise gruppiert worden sind, z. B. `Fakultät` und `Faktura`. Diese Übereinstimmung kann durch einen als Fuzzy-Gruppierung bezeichneten internen Algorithmus entstehen, mit dem vorübergehend doppelt/dreifach auftretende Konsonanten und Vokale ignoriert werden, um häufige Schreibfehler zu gruppieren. Nehmen Sie diese Wörter in eine Liste von Wortpaaren auf, die nicht gruppiert werden sollen. Für weitere Informationen siehe Thema Unscharfe Gruppierung in Kapitel 18 auf S. 336. Fuzzy-Gruppierung ist für japanischen Text nicht verfügbar.
- **Nicht extrahierte Konzepte.** Angenommen, Sie erwarten die Extrahierung bestimmter Konzepte, stellen jedoch bei der Überprüfung des Datensatztexts oder Dokumenttexts fest, dass einige Wörter oder Ausdrücke nicht extrahiert worden sind. Häufig handelt es sich bei diesen Wörtern um Verben oder Adjektive, die für Sie uninteressant sind. Manchmal möchten Sie vielleicht trotzdem nicht extrahierte Wörter oder Ausdrücke als Teil einer Kategoriedefinition verwenden. Um das Konzept zu extrahieren, können Sie die Aufnahme eines Fachausdrucks in ein Typ-Wörterbuch erzwingen. Für weitere Informationen siehe Thema Extrahierung von Wörtern erzwingen auf S. 165.

Viele dieser Änderungen können Sie direkt im Bereich “Extrahierungsergebnisse”, im Datenbereich, im Dialogfeld “Kategoriedefinitionen” oder im Dialogfeld “Clusterdefinitionen” vornehmen, indem Sie mit einem Klick auf die rechte Maustaste auf die Kontextmenüs zugreifen.

Wenn Sie die Änderungen vorgenommen haben, wechselt die Hintergrundfarbe des Bereichs und zeigt so, dass zum Anzeigen der Änderungen eine erneute Extrahierung erforderlich ist. Für weitere Informationen siehe Thema Daten extrahieren auf S. 147. Beim Arbeiten mit größeren Datensätzen kann es effizienter sein, erst nach mehreren Änderungen statt nach jeder einzelnen erneut zu extrahieren.

Anmerkung: In der Ansicht Resource Editor (Ansicht > Resource Editor) können Sie die Gesamtmenge der für die Gewinnung der Extrahierungsergebnisse verwendeten editierbaren linguistischen Ressourcen anzeigen. Diese Ressourcen werden in dieser Ansicht als Bibliotheken und Wörterbücher angezeigt. Sie können die Konzepte und Typen direkt in den Bibliotheken und Wörterbüchern anpassen. Für weitere Informationen siehe Thema Mit Bibliotheken arbeiten in Kapitel 16 auf S. 297.

Synonyme hinzufügen

Synonyme verknüpfen zwei oder mehr Wörter mit derselben Bedeutung. Synonyme werden häufig auch verwendet, um Fachausdrücke mit ihren Abkürzungen oder häufig falsch geschriebene Wörter mit der richtigen Schreibweise zu gruppieren. Durch die Verwendung von Synonymen nimmt die Häufigkeit des Zielkonzepts zu, sodass ähnliche Informationen, die in unterschiedlicher Form in den Textdaten vorhanden sind, viel leichter zu erkennen sind.

Die im Lieferumfang enthaltenen Vorlagen für linguistische Ressourcen und Bibliotheken beinhalten bereits viele vordefinierte Synonyme. Sie können jedoch nicht erkannte Synonyme definieren, sodass sie bei der nächsten Extrahierung erkannt werden.

Im ersten Schritt legen Sie das Zielkonzept oder Hauptkonzept fest. Das **Zielkonzept** ist das Wort oder der Ausdruck, unter dem Sie alle synonymen Fachausdrücke in den endgültigen Ergebnissen gruppieren möchten. Während der Extrahierung werden die Synonyme unter diesem Zielkonzept gruppiert. Im zweiten Schritt werden alle Synonyme für dieses Konzept ermittelt. Bei der endgültigen Extrahierung werden alle Synonyme durch das Zielkonzept substituiert. Ein Fachausdruck muss extrahiert werden, um ein Synonym sein zu können. Das Zielkonzept muss jedoch nicht extrahiert werden, damit die Substitution stattfinden kann. Wenn Sie z. B. `intelligent` durch `schlau` ersetzen möchten, dann ist `intelligent` das Synonym und `schlau` das Zielkonzept.

Beim Erstellen einer neuen Synonymdefinition wird ein neues Zielkonzept in das Wörterbuch aufgenommen. Anschließend nehmen Sie Synonyme in das Zielkonzept auf. Wenn Sie Synonyme erstellen oder bearbeiten, werden die Änderungen im Resource Editor in Synonymwörterbücher aufgenommen. Um den gesamten Inhalt der Synonymwörterbücher anzuzeigen oder eine beträchtliche Anzahl von Änderungen vorzunehmen, sollten Sie direkt im Resource Editor arbeiten. Für weitere Informationen siehe Thema Substitutions-/Synonymwörterbücher in Kapitel 17 auf S. 323.

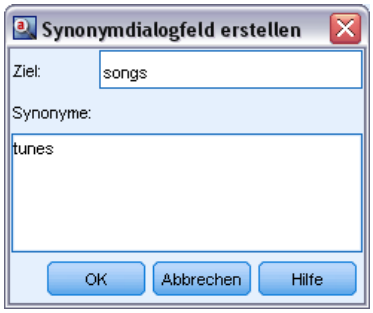
Neue Synonyme werden automatisch in der ersten Bibliothek im Bibliotheksbaum in der Ansicht Resource Editor gespeichert – standardmäßig ist dies die *Local Library*.

Anmerkung: Wenn Sie vergeblich nach einer Synonymdefinition in den Kontextmenüs oder direkt im Resource Editor suchen, könnte eine Übereinstimmung möglicherweise aus einem internen unscharfen Gruppierungsverfahren resultieren. Für weitere Informationen siehe Thema Unscharfe Gruppierung in Kapitel 18 auf S. 336.

So erstellen Sie ein neues Synonym:

- Wählen Sie die Konzepte, für die Sie ein neues Synonym erstellen möchten, im Bereich “Extrahierungsergebnisse”, im Datenbereich, im Dialogfeld “Kategoriedefinitionen” oder im Dialogfeld “Clusterdefinitionen” aus.
- Wählen Sie in den Menüs Bearbeiten > Zu Synonym hinzufügen> Neu. Das Dialogfeld “Synonym erstellen” wird geöffnet.

Abbildung 9-10
Dialogfeld "Synonym erstellen"



- Geben Sie im Textfeld "Ziel" ein Zielkonzept ein. Unter diesem Konzept werden alle Synonyme gruppiert.

- Um weitere Synonyme aufzunehmen, geben Sie sie in das Listenfeld "Synonyme" ein. Trennen Sie die synonymen Fachausdrücke mit dem globalen Trennzeichen. Für weitere Informationen siehe Thema Optionen: Registerkarte "Sitzung" in Kapitel 8 auf S. 135.

- Beim Arbeiten mit japanischem Text legen Sie einen Typen für diese Synonyme fest, indem Sie den Typnamen im Feld Synonyme von Typ auswählen. Das Ziel übernimmt jedoch den Typ, der bei der Extrahierung zugewiesen wurde. Wenn allerdings das Ziel nicht als ein Konzept extrahiert wurde, wird der in dieser Spalte aufgelistete Typ dem Ziel in den Extrahierungsergebnissen zugewiesen.

  Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

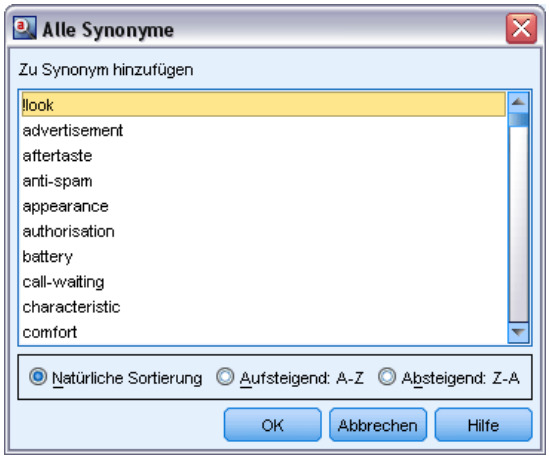
- Klicken Sie auf OK, um die Änderungen anzuwenden. Das Dialogfeld wird geschlossen und die Hintergrundfarbe im Bereich "Extrahierungsergebnisse" wechselt und zeigt so, dass zum Anzeigen der Änderungen eine erneute Extrahierung erforderlich ist. Nehmen Sie alle gewünschten Änderungen vor der erneuten Extrahierung vor.

So fügen Sie Konzepte zu einem Synonym hinzu:

- Wählen Sie die Konzepte, die Sie zu einer vorhandenen Synonymdefinition hinzufügen möchten, im Bereich "Extrahierungsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.

- Wählen Sie in den Menüs Bearbeiten > Zu Synonym hinzufügen>. Im Menü werden die Synonyme angezeigt, wobei das zuletzt erstellte am Anfang der Liste steht. Wählen Sie den Namen des Synonyms aus, zu dem Sie die ausgewählten Konzepte hinzufügen möchten. Wenn Sie das gesuchte Synonym entdecken, dann wählen Sie es aus und die ausgewählten Konzepte werden zu dieser Synonymdefinition hinzugefügt. Wenn Sie es nicht entdecken, wählen Sie Weitere aus, um das Dialogfeld "Alle Synonyme" anzuzeigen.

Abbildung 9-11
Dialogfeld "Alle Synonyme"



- Im Dialogfeld "Alle Synonyme" können Sie die Liste in der natürlichen Reihenfolge (Erstellungsreihenfolge) oder in aufsteigender bzw. absteigender Reihenfolge sortieren. Wählen Sie den Namen des Synonyms aus, zu dem Sie die ausgewählten Konzepte hinzufügen möchten, und klicken Sie auf OK. Das Dialogfeld wird geschlossen und die Konzepte werden zu den Synonymdefinitionen hinzugefügt.

Konzepte zu Typen hinzufügen

Bei der Extrahierung werden die extrahierten Konzepte Typen zugewiesen, damit Fachausdrücke mit Gemeinsamkeiten gruppiert werden. Es sind bereits zahlreiche Typen im Lieferumfang von IBM® SPSS® Modeler Text Analytics enthalten. Für weitere Informationen siehe Thema Integrierte Typen in Kapitel 17 auf S. 314.In den meisten Sprachen gilt: Konzepte, die in keinem Typen-Wörterbuch gefunden, aber aus dem Text extrahiert werden, erhalten automatisch den Typ `<Unbekannt>`. Bei japanischem Text jedoch erhalten sie automatisch den Typ <名詞>*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Wenn Sie Ihre Ergebnisse prüfen, finden Sie eventuell einige Konzepte, die in einem Typ erscheinen, die Sie einem anderen zuweisen möchten, oder Sie stellen fest, dass eine Gruppe von Wörtern eigentlich in einen neuen Typ gehört. In diesen Fällen sollten Sie die Konzepte einem anderen Typ neu zuweisen oder einen neuen Typ erstellen. Für japanischen Text können Sie keine Typen erstellen.

Angenommen, Sie arbeiten z. B. mit Umfragedaten zu Kraftfahrzeugen und möchten mit Fokussierung auf verschiedene Fahrzeugbereiche kategorisieren. Sie können einen Typ `<Armaturenbrett>` erstellen, um alle Konzepte zu Anzeigeinstrumenten und Schaltern von Armaturenbrettern bei Fahrzeugen zu gruppieren. Dann können Sie Konzepte wie `Tankanzeige`, `Heizung`, `Radio` und `Kilometerzähler` diesem neuen Typ zuweisen.

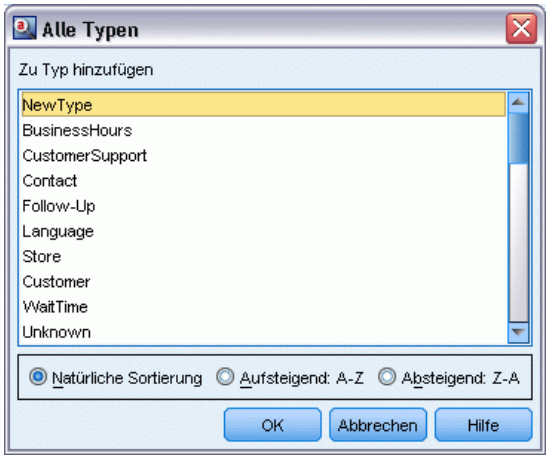
Oder angenommen, Sie arbeiten mit Umfragedaten zu Universitäten und Fachhochschulen und beim Extrahieren hat `Johann Wolfgang Goethe` (die Universität) den Typ `<Person>` statt des Typs `<Organization>` erhalten. In diesem Fall fügen Sie dieses Konzept dem Typ `<Organization>` hinzu.

Wenn Sie einen Typ erstellen oder Konzepte als Fachausdrücke zu einer Ausdruckliste eines Typs hinzufügen, werden die Änderungen in Typ-Wörterbüchern in den Bibliotheken der linguistischen Ressourcen im Resource Editor aufgezeichnet. Um den Inhalt dieser Bibliotheken anzuzeigen oder eine beträchtliche Anzahl von Änderungen vorzunehmen, sollten Sie direkt im Resource Editor arbeiten. Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken in Kapitel 17 auf S. 316.

So fügen Sie ein Konzept zu einem Typ hinzu:

- Wählen Sie die Konzepte, die Sie zu einem vorhandenen Typ hinzufügen möchten, im Bereich "Extrahierungsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.
- Klicken Sie mit der rechten Maustaste, um das Kontextmenü aufzurufen.
- Wählen Sie in den Menüs Bearbeiten > Zu Typ hinzufügen>. Im Menü werden die Typen angezeigt, wobei der zuletzt erstellte am Anfang der Liste steht. Wählen Sie den Namen des Typs aus, zu dem Sie die ausgewählten Konzepte hinzufügen möchten. Wenn Sie den gesuchten Typnamen sehen, dann wählen Sie ihn aus und die ausgewählten Konzepte werden zu diesem Typ hinzugefügt. Wenn Sie ihn nicht sehen, wählen Sie Weitere aus, um das Dialogfeld "Alle Typen" anzuzeigen.

Abbildung 9-12
Dialogfeld "Alle Typen"



- Im Dialogfeld "Alle Typen" können Sie die Liste in der natürlichen Sortierung (Erstellungsreihenfolge) oder in aufsteigender bzw. absteigender Reihenfolge sortieren. Wählen Sie den Namen des Typs aus, dem Sie die ausgewählten Konzepte hinzufügen möchten, und klicken Sie auf OK. Das Dialogfeld wird geschlossen und die Konzepte werden als Fachausdrücke zu den Typen hinzugefügt.

Anmerkung: In japanischem Text gibt es einige Instanzen, bei denen die Typänderung eines Fachausdrucks nicht den Typ ändert, dem er zum Schluss in der endgültigen Extrahierungsliste zugewiesen wird. Dies liegt an internen Wörterbüchern, die bei der Extrahierung für einige grundlegende Fachausdrücke Vorrang haben.

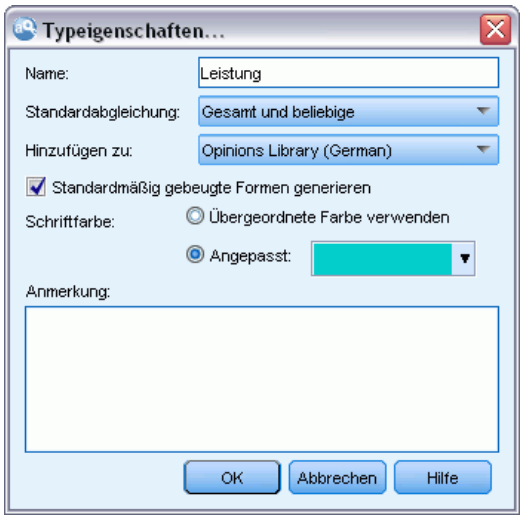
Anmerkung: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

So erstellen Sie einen "Neuen Typ":

- Wählen Sie die Konzepte, für die Sie einen neuen Typ erstellen möchten, im Bereich "Extrahierungsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.

- Wählen Sie in den Menüs Bearbeiten > Zu Typ hinzufügen> Neu. Das Dialogfeld "Typ-Eigenschaften" wird geöffnet.

Abbildung 9-13
Dialogfeld "Typ-Eigenschaften"



- Geben Sie einen neuen Namen für diesen Typ im Textfeld "Name" ein und nehmen Sie die gewünschten Änderungen in den anderen Feldern vor. Für weitere Informationen siehe Thema Erstellen von Typen in Kapitel 17 auf S. 314.

- Klicken Sie auf OK, um die Änderungen anzuwenden. Das Dialogfeld wird geschlossen und die Hintergrundfarbe im Bereich "Extrahierungsergebnisse" wechselt und zeigt so, dass zum Anzeigen der Änderungen eine erneute Extrahierung erforderlich ist. Nehmen Sie alle gewünschten Änderungen vor der erneuten Extrahierung vor.

Konzepte von der Extrahierung ausschließen

Bei der Überprüfung der Ergebnisse entdecken Sie ggf. gelegentlich unerwünschte Konzepte oder Konzepte, die von automatisierten Kategorieaufbauverfahren verwendet werden. In einigen Fällen haben diese Konzepte eine hohe globale Häufigkeitsanzahl, sind aber für Ihre Analyse völlig bedeutungslos. In diesem Fall markieren Sie ein Konzept, um ihn aus der endgültigen Extrahierung auszuschließen. In der Regel sind Konzepte in dieser Liste Füllwörter oder -ausdrücke zur Verbesserung des Textflusses, die keine Bedeutung tragen und die Extrahierungsergebnisse unnötig anfüllen. Wenn Sie diese Konzepte in das Ausschlusswörterbuch aufnehmen, verhindern Sie ihre Extrahierung.

Durch das Ausschließen von Konzepten werden bei der nächsten Extrahierung alle Varianten des ausgeschlossenen Konzepts aus den Extrahierungsergebnissen entfernt. Wenn dieses Konzept bereits als ein Deskriptor in einer Kategorie angezeigt wird, bleibt er nach der erneuten Extrahierung mit der Anzahl null in der Kategorie.

Wenn Sie ausschließen, werden die Änderungen im Resource Editor in einem Ausschlusswörterbuch aufgezeichnet. Um alle ausgeschlossenen Definitionen anzuzeigen und sie direkt zu bearbeiten, sollten Sie direkt im Resource Editor arbeiten. Für weitere Informationen siehe Thema Ausschlusswörterbücher in Kapitel 17 auf S. 329.

Anmerkung: Bei japanischem Text gibt es einige Instanzen, in denen das Ausschließen eines Fachausdrucks oder Typs nicht zu dessen Ausschluss führt. Dies liegt an internen Wörterbüchern, die bei der Extrahierung für einige grundlegende Fachausdrücke für japanische Ressourcen Vorrang haben.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

So schließen Sie Konzepte aus

- Wählen Sie die Konzepte, die Sie von der Extrahierung ausschließen möchten, im Bereich "Extrahierungsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.
- Klicken Sie mit der rechten Maustaste, um das Kontextmenü aufzurufen.
- Wählen Sie Aus Extrahierung ausschließen aus. Das Konzept wird im Resource Editor dem Ausschlusswörterbuch hinzugefügt und die Hintergrundfarbe im Bereich "Extrahierungsergebnisse" wechselt und zeigt so, dass zum Anzeigen der Änderungen eine erneute Extrahierung erforderlich ist. Nehmen Sie alle gewünschten Änderungen vor der erneuten Extrahierung vor.

  Anmerkung. Ausgeschlossene Wörter werden automatisch in der ersten Bibliothek im Bibliotheksbaum im Resource Editor gespeichert – standardmäßig ist dies die *Local Library*.

Extrahierung von Wörtern erzwingen

Wenn Sie nach der Extrahierung die Textdaten im Bereich "Daten" überprüfen, stellen Sie möglicherweise fest, dass einige Wörter oder Ausdrücke nicht extrahiert wurden. Häufig handelt es sich bei diesen Wörtern um Verben oder Adjektive, die für Sie uninteressant sind. Manchmal möchten Sie vielleicht trotzdem nicht extrahierte Wörter oder Ausdrücke als Teil einer Kategoriedefinition verwenden.

Damit diese Wörter und Ausdrücke extrahiert werden, können Sie die Aufnahme eines Fachausdrucks in ein Typ-Wörterbuch erzwingen. Für weitere Informationen siehe Thema Erzwingen von Fachausdrücken in Kapitel 17 auf S. 320.

Wichtig: Einen Fachausdruck in einem Wörterbuch als erzwungen zu markieren, ist kein absolut sicheres Verfahren. Das heißt, obwohl ein Fachausdruck explizit einem Wörterbuch hinzugefügt worden ist, ist er möglicherweise nach einer erneuten Extrahierung nicht im Bereich "Extrahierungsergebnisse" vorhanden oder er wird nicht genau so angezeigt, wie Sie ihn deklariert

haben. Dies kommt selten vor, ist aber möglich, wenn ein Wort oder Ausdruck bereits als Teil eines längeren Ausdrucks extrahiert wurde. Um das zu verhindern, wenden Sie die Abgleichoption Gesamt (keine Zusammensetzungen) auf diesen Ausdruck im Typwörterbuch an. Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken in Kapitel 17 auf S. 316.

Kapitel
10

Kategorisieren von Textdaten

In der Ansicht “Kategorien und Konzepte” können Sie **Kategorien** erstellen, die im Prinzip allgemeinere (auf einer höheren Ebene liegende) Konzepte oder Themen darstellen, die die Grundideen, das Grundwissen bzw. die Grundhaltungen erfassen, die im Text ausgedrückt werden.

Ab IBM® SPSS® Modeler Text Analytics 14 können Kategorien auch eine hierarchische Struktur besitzen, d. h., sie können Unterkategorien enthalten, die wiederum eigene Unterkategorien enthalten können usw. Sie können vordefinierte Kategorien, früher Coderahmen genannt, mit hierarchischen Kategorien importieren und diese hierarchischen Kategorien auch im Produkt aufbauen.

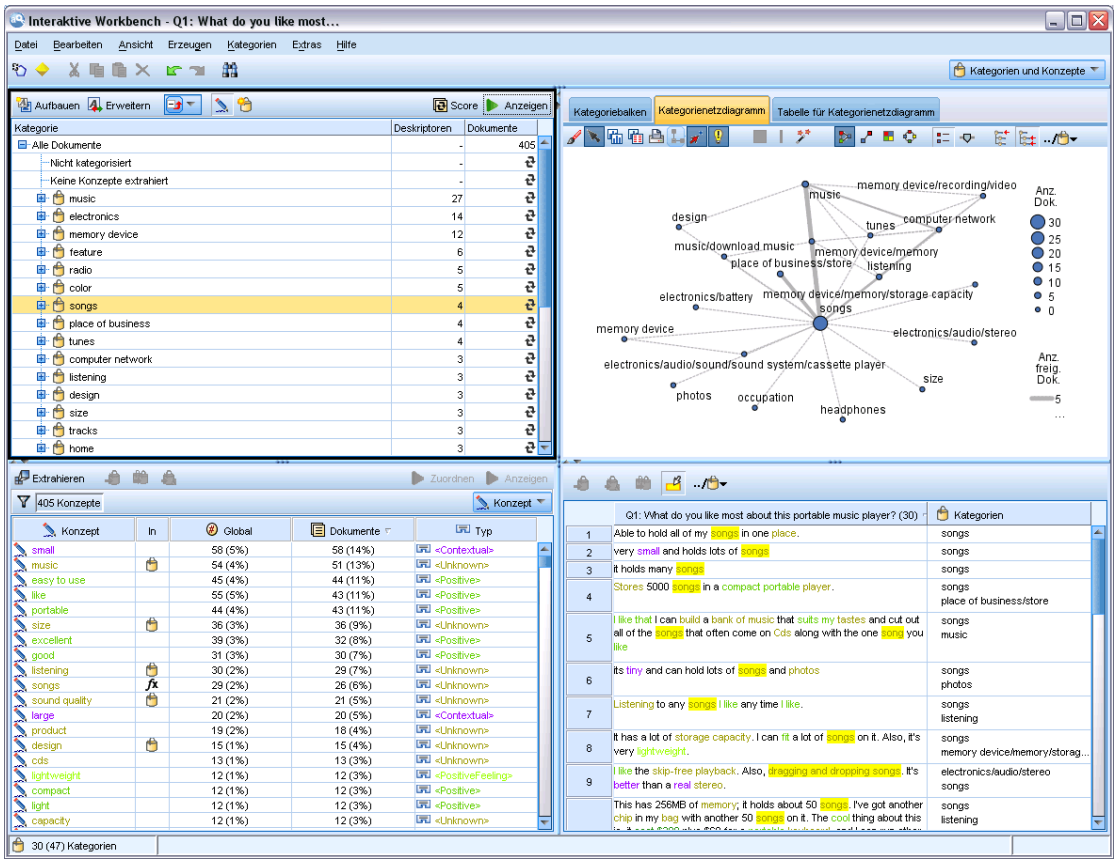
Hierarchische Kategorien ermöglichen Ihnen den Aufbau einer Baumstruktur mit einer oder mehreren Unterkategorien, die eine genauere Gruppierung von Elementen, beispielsweise verschiedenen Konzept- oder Themenbereichen, gestatten. Ein einfaches Beispiel könnte sich auf Freizeitaktivitäten beziehen: Bei der Beantwortung einer Frage wie *Welcher Aktivität würden Sie sich gerne widmen, wenn Sie mehr Zeit hätten?* könnten die Kategorien der ersten Ebene *Sport*, *Kunst und Handarbeiten*, *Angeln* usw. lauten. Auf der Ebene unter *Sport* könnten Sie Unterkategorien einrichten, um zu sehen, ob es sich um *Ballsportarten*, *Wassersportarten* usw. handelt.

Kategorien bestehen aus einer Reihe von Deskriptoren wie *Konzepten*, *Typen*, *Mustern* und *Kategorieregeln*. Diese Deskriptoren werden zusammen verwendet, um zu bestimmen, ob ein Dokument oder Datensatz zu einer gegebenen Kategorie gehört oder nicht. Der Text in einem Dokument oder Datensatz kann gescannt werden, um zu überprüfen, ob es Text gibt, der mit einem Deskriptor übereinstimmt. Liegt eine Übereinstimmung vor, wird das Dokument/der Datensatz dieser Kategorie zugeordnet. Dieser Prozess wird als **Kategorisierung** bezeichnet.

Mithilfe der in den vier Fensterbereichen angezeigten Daten können Sie Kategorien erstellen, damit arbeiten und sie visuell untersuchen. Jeder der vier Fensterbereiche der Ansicht “Kategorien und Konzepte kann durch Auswahl seines Namens im Menü “Ansicht” ein- bzw. ausgeblendet werden.

- **Fensterbereich “Kategorien”.** In diesem Fensterbereich können Sie Kategorien erstellen und verwalten. Für weitere Informationen siehe Thema Der Fensterbereich “Kategorien” auf S. 169.
- **Fensterbereich “Extrahierungsergebnisse”.** In diesem Fensterbereich können Sie mit den extrahierten Konzepten und Typen arbeiten. Für weitere Informationen siehe Thema Extrahierungsergebnisse: Konzepte und Typen in Kapitel 9 auf S. 143.
- **Visualisierungsbereich.** In diesem Fensterbereich können Sie die Kategorien und ihre Interaktionen visuell untersuchen. Für weitere Informationen siehe Thema Kategoriendiagramme und Grafiken in Kapitel 13 auf S. 262.
- **Datenbereich.** In diesem Fensterbereich können Sie den Text untersuchen und überprüfen, der in Dokumenten und Datensätzen enthalten ist, die Ihrer Auswahl entsprechen. Für weitere Informationen siehe Thema Der Fensterbereich “Daten” auf S. 179.

Abbildung 10-1
Kategorie- und Konzeptansicht



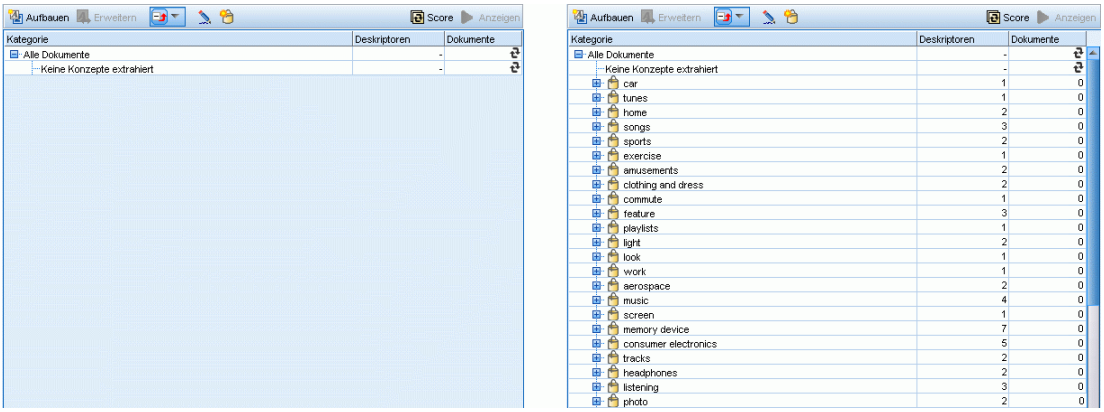
Sie können zwar mit einem Kategorien-Set aus einem Text Analysis Package (TAP) beginnen oder einen Import aus einer vordefinierten Kategoriendatei durchführen, aber eventuell müssen Sie auch Ihre eigenen erstellen. Kategorien können mit den leistungsfähigen automatisierten Methoden des Produkts automatisch erstellt werden, wobei Kategorien und deren Deskriptoren anhand von Extrahierungsergebnissen (Konzepte, Typen und Muster) erzeugt werden. Daneben können Sie Kategorien auch manuell erstellen und dabei zusätzliche Erkenntnisse mit einbeziehen, die Sie hinsichtlich der Datengrundlage möglicherweise gewonnen haben. Die manuelle Erstellung oder Feinabstimmung von Kategorien ist allerdings nur über die interaktive Workbench möglich. Für weitere Informationen siehe Thema Text-Mining-Knoten: Registerkarte “Modell” in Kapitel 3 auf S. 37. Sie können Kategoriedefinitionen manuell erstellen, indem Sie die Extrahierungsergebnisse in die Kategorien ziehen und ablegen. Sie können diese Kategorien oder jede leere Kategorie anreichern, indem Sie einer Kategorie Kategorieregeln hinzufügen.

Diese Verfahren und Methoden eignen sich jeweils gut für bestimmte Arten von Daten und Situationen, häufig ist es jedoch sinnvoll, in einer Analyse mehrere Verfahren zu kombinieren, um das gesamte Spektrum an Dokumenten bzw. Datensätzen zu erfassen. Außerdem können Ihnen im Verlauf der Kategorisierung andere Änderungen auffallen, die an den linguistischen Ressourcen vorgenommen werden sollten.

Der Fensterbereich “Kategorien”

Im Fensterbereich “Kategorien” können Sie Ihre Kategorien erstellen und verwalten. Dieser Fensterbereich befindet sich in der oberen linken Ecke der Ansicht “Kategorien und Konzepte”. Nachdem Sie die Konzepte und Typen aus Ihren Textdaten extrahiert haben, können Sie Kategorien automatisch mithilfe von Verfahren wie Konzepteinbeziehung, Kookkurrenz usw. oder manuell erstellen. Für weitere Informationen siehe Thema Erstellen von Kategorien auf S. 183.

Abbildung 10-2
Fensterbereich “Kategorien” mit und ohne Kategorien



Immer wenn eine Kategorie erstellt oder aktualisiert wird, können die Dokumente bzw. Datensätze durch Klick auf die Schaltfläche Score gescort werden, um zu sehen, ob es Text gibt, der einem Deskriptor in einer bestimmten Kategorie entspricht. Liegt eine Übereinstimmung vor, wird das Dokument/der Datensatz dieser Kategorie zugeordnet. Das Endergebnis besteht darin, dass die meisten, wenn nicht alle Dokumente bzw. Datensätze, anhand der Deskriptoren in den Kategorien bestimmten Kategorien zugewiesen werden.

Kategoriebaumtabelle

Die Baumtabelle in diesem Bereich zeigt das Set von Kategorien, Unterkategorien und Deskriptoren an. Der Baum besitzt auch mehrere Spalten mit Informationen für jedes Baumelement. Folgende Spalten stehen ggf. zur Anzeige zur Verfügung:

- **Code.** Listet den Codewert für jede Kategorie auf. Diese Spalte ist standardmäßig ausgeblendet. Sie können diese Spalte über folgende Optionsfolge anzeigen: Ansicht > Fensterbereich “Kategorien”.
- **Kategorie.** Enthält den Kategoriebaum mit dem Namen der Kategorie und der Unterkategorien. Außerdem wird durch einen Klick auf das Deskriptor-Symbolleistensymbol auch das Set der Deskriptoren angezeigt.
- **Deskriptoren.** Gibt die Anzahl der Deskriptoren an, aus der die Kategorie besteht. Diese Zahl enthält nicht die Anzahl an Deskriptoren in den Unterkategorien. Es wird keine Zahl angezeigt, wenn ein Deskriptorname in der Spalte Kategorien angezeigt wird. Sie können

die Deskriptoren im Baum über folgende Menüfolge anzeigen oder ausblenden: Ansicht > Fensterbereich “Kategorien” > Alle Deskriptoren.

- **Dokumente.** Nach dem Scoring enthält diese Spalte die Anzahl von Dokumenten oder Datensätzen, die einer Kategorie und allen ihren Unterkategorien zugeordnet werden. Wenn also 5 Datensätze mit Ihrer obersten Kategorie auf der Basis ihrer Deskriptoren übereinstimmen und 7 unterschiedliche Datensätze mit einer Unterkategorie auf der Basis ihrer Deskriptoren übereinstimmen, ist die Gesamtanzahl an Dokumenten für die oberste Kategorie die Summe dieser beiden, in diesem Fall also 12. Wenn jedoch der gleiche Datensatz mit der obersten Kategorie und ihrer Unterkategorie übereinstimmt, betrüge die Anzahl 11.

Wenn keine Kategorien vorhanden sind, enthält die Tabelle dennoch zwei Zeilen. Die oberste Zeile, Alle Dokumente, gibt die Gesamtzahl der Dokumente oder Datensätze an. Die zweite Zeile Nicht kategorisiert, zeigt die Anzahl der Dokumente/Datensätze an, die noch kategorisiert werden müssen.

Bei jeder Kategorie im Fensterbereich geht dem Kategorienamen ein kleines Symbol in Form eines gelben Eimers voran. Wenn Sie auf eine Kategorie doppelklicken auswählen oder in den Menüs die Optionsfolge Ansicht> Kategoriedefinitionen wählen, wird das Dialogfeld “Kategoriedefinitionen” geöffnet, in dem alle Elemente (so genannte **Deskriptoren** angezeigt werden, aus denen die Definition besteht, beispielsweise Konzepte, Typen, Muster und Kategorieregeln. Für weitere Informationen siehe Thema Erläuterung von Kategorien auf S. 178. Standardmäßig werden in der Kategoriebaumtabelle die Deskriptoren in den Kategorien nicht angezeigt. Wenn Sie die Deskriptoren direkt in der Tabelle und nicht im Dialogfeld “Kategoriedefinitionen” sehen möchten, klicken Sie auf die Umschaltfläche mit dem Stiftsymbol in der Symbolleiste. Durch Auswahl dieser Umschaltfläche können Sie Ihren Baum erweitern, um auch die Deskriptoren sehen zu können.

Scoren von Kategorien

Die Spalte Dokumente in der Kategoriebaumtabelle zeigt die Anzahl von Dokumenten oder Datensätzen, die dieser bestimmten Kategorie zugeordnet werden. Es erscheint ein Symbol in der Spalte, wenn die Zahlen veraltet sind oder nicht berechnet wurden. Sie können in der Symbolleiste des Bereichs auf Score klicken, um die Anzahl von Dokumenten neu zu berechnen. Denken Sie daran, dass der Scoring-Vorgang bei der Arbeit mit größeren Daten-Sets einige Zeit in Anspruch nehmen kann.

Auswahl von Kategorien im Baum

Wenn Sie eine Auswahl im Baum treffen, können Sie nur Schwesterkategorien auswählen, d. h., wenn Sie Kategorien der obersten Ebene auswählen, können Sie nicht auch eine Unterkategorie auswählen. Oder wenn Sie zwei Unterkategorien einer bestimmten Kategorie auswählen, können Sie nicht gleichzeitig eine Unterkategorie einer anderen Kategorie auswählen. Durch die Auswahl einer nicht zusammenhängenden Kategorie geht die vorherige Auswahl verloren.

Anzeige in den Fensterbereichen "Daten" und "Visualisierung"

Wenn Sie eine Zeile in der Tabelle auswählen, können Sie auf die Schaltfläche Anzeigen klicken, um die Fensterbereiche "Visualisierung" entsprechend Ihrer Auswahl mit Informationen zu aktualisieren. Wenn ein Fensterbereich nicht sichtbar ist, kann er durch Klicken auf Anzeigen aufgerufen werden.

Verfeinern der Kategorien

Die Kategorisierung liefert möglicherweise nicht beim ersten Versuch ideale Ergebnisse für Ihre Daten und es kann Kategorien geben, die Sie löschen oder mit anderen Kategorien kombinieren möchten. Außerdem können Sie durch eine Überprüfung der Extrahierungsergebnisse herausfinden, dass einige Kategorien, die Sie für sinnvoll halten würden, nicht erstellt wurden. In diesem Fall können Sie manuelle Änderungen an den Ergebnissen vornehmen, um sie für den jeweiligen Kontext zu optimieren. Für weitere Informationen siehe Thema Bearbeiten und Verfeinern von Kategorien auf S. 237.

Methoden und Strategien zur Erstellung von Kategorien

Wenn Sie noch keine Extrahierung durchgeführt haben oder die Extrahierungsergebnisse nicht auf dem neuesten Stand sind, werden Sie automatisch bei Verwendung einer dieser Verfahren zum Aufbau oder zur Erweiterung einer Kategorie zu einer Extrahierung aufgefordert. Nachdem Sie ein Verfahren angewendet haben, stehen die Konzepte und Typen, die in eine Kategorie gruppiert wurden, weiterhin für die Kategorieerstellung mit anderen Verfahren zur Verfügung. Das bedeutet, dass Sie unter Umständen ein Konzept in mehreren Kategorien sehen, es sei denn, Sie beschließen, diese nicht wiederzuverwenden.

Als Unterstützung zur Ermittllung der besten Kategorien, beachten Sie bitte die folgenden Abschnitte:

- **Methoden für die Kategorieerstellung**
- **Strategien für die Kategorieerstellung**
- **Tipps zur Erstellung von Kategorien**

Methoden für die Kategorieerstellung

Da jedes Daten-Set anders ist, können sich die Anzahl der Kategorieerstellungsmethoden und die Reihenfolge, in der sie angewendet werden, im Laufe der Zeit ändern. Da zusätzlich Ihre Text-Mining-Ziele von Daten-Set zu Daten-Set unterschiedlich sein können, müssen Sie möglicherweise mit den verschiedenen Methoden experimentieren, um zu sehen, welches die besten Ergebnisse für die jeweiligen Textdaten bringt. Keines der automatischen Verfahren ergibt eine perfekte Kategorisierung Ihrer Daten, weshalb wir empfehlen, ein oder mehrere automatische Verfahren anzuwenden, die gut mit Ihren Daten funktionieren.

Neben der Verwendung von Text Analysis Packages (TAPs, **.tap*) mit vordefinierten Kategoriensets können Sie Ihre Antworten auch mit jeder Kombination der folgenden Methoden kategorisieren:

- **Automatische Erstellungsverfahren.** Verschiedene linguistisch basierte und häufigkeitsbasierte Kategorieoptionen stehen zur Verfügung, um automatisch Kategorien für Sie zu erstellen. Für weitere Informationen siehe Thema Erstellen von Kategorien auf S. 183.
- **Automatische Erweiterungsverfahren.** Verschiedene linguistische Verfahren stehen zur Verfügung, um vorhandene Kategorien durch Hinzufügen und Verbessern von Deskriptoren zu erweitern, so dass sie mehr Datensätze erfassen. Für weitere Informationen siehe Thema Erweitern von Kategorien auf S. 200.
- **Manuelle Verfahren.** Es gibt verschiedene manuelle Methoden wie Drag-and-Drop. Für weitere Informationen siehe Thema Manuelle Erstellung von Kategorien auf S. 205.

Strategien für die Kategorieerstellung

Die folgende Liste der Strategien ist keinesfalls vollständig, kann Ihnen aber einige Ideen geben, sich dem Aufbau Ihrer Kategorien zu nähern.

- Wenn Sie den Text-Mining-Knoten definieren, wählen Sie ein Kategorienset aus einem Text Analysis Package (TAP), um Ihre Analyse mit einigen vordefinierten Kategorien zu beginnen. Diese Kategorien kategorisieren Ihren Text vielleicht gleich zu Beginn in ausreichender Form. Wenn Sie jedoch weitere Kategorien hinzufügen möchten, können Sie die Kategorieaufbaueinstellungen (Kategorien > Aufbaueinstellungen) bearbeiten. Öffnen Sie das Dialogfeld Erweiterte Einstellungen: Linguistik, wählen Sie die Kategorieeingabeoption Nicht verwendete Extrahierungsergebnisse und bauen Sie die zusätzlichen Kategorien auf.
- Wenn Sie den Knoten definieren, wählen Sie ein Kategorienset aus einem TAP in der Ansicht "Kategorien und Konzepte" in der interaktiven Workbench. Ziehen Sie dann nicht verwendete Konzepte oder Muster nach Bedarf in die Kategorien. Erweitern Sie dann die bestehenden Kategorien, die Sie eben bearbeitet haben (Kategorien > Kategorien erweitern), um weitere Deskriptoren zu erhalten, die mit den vorhandenen Kategoriedeskriptoren in Bezug stehen.
- Bauen Sie Kategorien automatisch über die erweiterten linguistischen Einstellungen (Kategorien > Kategorien aufbauen) auf. Verfeinern Sie dann die Kategorien manuell durch Löschen von Deskriptoren, Löschen von Kategorien oder Zusammenführen ähnlicher Kategorien, bis Sie mit den resultierenden Kategorien zufrieden sind. Zusätzlich können Sie, wenn die Kategorien ursprünglich **ohne** die Option Mit Platzhaltern, wenn möglich, verallgemeinern erzeugt wurden, auch versuchen, die Kategorien automatisch über "Kategorien erweitern" mit der Option Verallgemeinern zu vereinfachen.
- Importieren Sie eine vordefinierte Kategoriedatei mit sehr beschreibenden Kategorienamen und/oder Anmerkungen. Zusätzlich können Sie, wenn der Import ursprünglich durchgeführt wurde, **ohne** dass die Option zum Import oder Erzeugen von Deskriptoren aus Kategorienamen gewählt wurde, später das Dialogfeld "Kategorien erweitern" verwenden und die Option Leere Kategorien mit aus dem Kategorienamen erzeugten Deskriptoren erweitern

auswählen. Erweitern Sie dann diese Kategorien ein zweites Mal, aber verwenden Sie dieses Mal die Gruppierverfahren.

- Erstellen Sie manuell ein erstes Set an Kategorien, indem Sie Konzepte oder Konzeptmuster nach Häufigkeit sortieren und dann die interessantesten in den Kategoriebereich ziehen. Wenn Sie dieses erste Set an Kategorien haben, verwenden Sie die Funktion "Erweitern" (Kategorien > Kategorien erweitern), um alle ausgewählten Kategorien zu erweitern und zu verfeinern, so dass sie andere zugehörige Deskriptoren einschließen und so mehr Datensätzen entsprechen.

Es wird empfohlen, dass Sie nach Anwenden dieser Verfahren die resultierenden Kategorien prüfen und manuelle Verfahren verwenden, um kleinere Anpassungen vorzunehmen, etwaige Fehlklassifizierungen zu beheben oder Datensätze oder Wörter hinzuzufügen, die nicht erfasst wurden. Da die Verwendung verschiedener Verfahren auch zu redundanten Kategorien führen kann, können Sie zusätzlich Kategorien zusammenführen bzw. löschen. Für weitere Informationen siehe Thema Bearbeiten und Verfeinern von Kategorien auf S. 237.

Tipps zur Erstellung von Kategorien

Um Sie dabei zu unterstützen, bessere Kategorien zu erstellen, finden Sie hier einige Tipps, die Ihnen helfen, Entscheidungen zu Ihrem Ansatz zu treffen.

Tipps zum Verhältnis zwischen Kategorien und Dokumenten

Die Kategorien, denen Dokumente und Datensätze zugewiesen werden, schließen sich bei der qualitativen Textanalyse nicht oft gegenseitig aus. Dafür gibt es mindestens zwei Gründe:

- Zunächst gilt eine allgemeine Faustregel, die besagt: Je länger Textdokument bzw. Datensatz, desto eindeutiger sind die darin ausgedrückten Ideen und Meinungen. Dadurch wird die Wahrscheinlichkeit, dass ein Dokument oder Datensatz mehreren Kategorien zugewiesen werden kann, stark erhöht.
- Zum anderen gibt es häufig verschiedene Möglichkeiten zur Gruppierung und Interpretation von Textdokumenten oder Datensätzen, die nicht logisch getrennt sind. Bei einer Umfrage mit offenen Fragen zur politischen Haltung des Befragten können wir Kategorien erstellen wie *rechts* und *links* oder *konservativ* und *sozialdemokratisch* sowie speziellere Kategorien wie *sozialliberal*, *fiskalkonservativ* usw. Diese Kategorien müssen sich nicht gegenseitig ausschließen und brauchen nicht erschöpfend zu sein.

Tipps zur Anzahl der zu erstellenden Kategorien

Die Kategorieerstellung sollte sich direkt aus den Daten ergeben: Wenn Sie interessante Aspekte in Bezug auf Ihre Daten feststellen, können Sie eine Kategorie erstellen, die diese Informationen widerspiegelt. Im Allgemeinen gibt es keine empfohlene Obergrenze für die Anzahl der zu erstellenden Kategorien. Es ist jedoch durchaus möglich, so viele Kategorien zu erstellen, dass sie nicht mehr handhabbar sind. Es gelten zwei Prinzipien:

- **Kategoriehäufigkeit.** Damit eine Kategorie sinnvoll eingesetzt werden kann, muss sie eine Mindestanzahl an Dokumenten bzw. Datensätzen enthalten. Ein oder zwei Dokumente können sehr interessante Aspekte enthalten, doch wenn es sich dabei um ein oder zwei von 1.000

Dokumenten handelt, sind die darin enthaltenen Informationen mit großer Wahrscheinlichkeit nicht häufig genug in der Grundgesamtheit vorhanden, um von praktischem Nutzen zu sein.

- **Komplexität.** Je mehr Kategorien Sie erstellen, desto mehr Informationen müssen Sie nach Abschluss der Analyse überprüfen und zusammenfassen. Sehr viele Kategorien vergrößern zwar die Komplexität, führen jedoch nicht unbedingt zu zusätzlichen nützlichen Details.

Leider gibt es keine Regeln zur Ermittlung, ab wann zu viele Kategorien vorliegen oder wie viele Datensätze mindestens pro Kategorie vorhanden sein sollen. Sie müssen diese Festlegungen je nach den Anforderungen der jeweiligen Situation treffen.

Wir können jedoch einige Ratschläge für den Anfang geben. Die Anzahl der Kategorien sollte zwar nicht übermäßig hoch sein, in den frühen Phasen einer Analyse ist es jedoch besser, eher zu viele als zu wenige Kategorien zu verwenden. Es ist einfacher, Kategorien, die sich relativ ähnlich sind, zusammenzufassen, als Fälle in neue Kategorien abzuspalten. Daher ist eine Strategie, die von relativ vielen Kategorien ausgeht und zu einer Verringerung ihrer Zahl führt, normalerweise die beste Vorgehensweise. Angesichts des iterativen Charakters des Text-Minings und der Leichtigkeit, mit der es mithilfe dieses Softwareprogramms durchgeführt werden kann, stellt eine hohe Kategorienzahl für den Anfang kein Problem dar.

Auswahl der besten Deskriptoren

Die folgenden Informationen enthalten einige Richtlinien zur Auswahl oder zum Erstellen der besten Deskriptoren (Konzepte, Typen, TLA-Muster und Kategorieregeln) für Ihre Kategorien. Deskriptoren sind die Bausteine von Kategorien. Wenn Text in einem Dokument oder Datensatz ganz oder teilweise mit einem Deskriptor übereinstimmt, wird das Dokument oder der Datensatz mit der Kategorie abgeglichen.

Ein Deskriptor wird nur dann mit Dokumenten oder Datensätzen abgeglichen, wenn er ein extrahiertes Konzept oder Muster enthält oder ihm entspricht. Verwenden Sie daher Konzepte, Typen, Muster und Kategorieregeln wie beschrieben in den folgenden Absätzen.

Da Konzepte nicht nur sich selbst, sondern auch eine Reihe zugrundeliegender Fachausdrücke darstellen, von Plural- und Singularformen über Synonyme bis hin zu Rechtschreibvariationen, sollte nur das Konzept selbst als Deskriptor oder Teil eines Deskriptors verwendet werden. Um mehr über die zugrunde liegenden Fachausdrücke für ein bestimmtes Konzept zu erfahren, klicken Sie auf den Konzeptnamen in den Extrahierungsergebnissen der Ansicht “Kategorien und Konzepte”. Wenn Sie mit der Maus über den Konzeptnamen fahren, erscheint eine QuickInfo, in der alle bei der letzten Extrahierung in Ihrem Text gefundenen, zugrunde liegenden Fachausdrücke angezeigt werden. Nicht alle Konzepte haben zugrunde liegende Fachausdrücke. Falls beispielsweise `Auto` und `Fahrzeug` als Synonyme gelten, aber `Auto` als Konzept und `Fahrzeug` als zugrunde liegender Fachausdruck extrahiert wurde, sollten Sie nur `Auto` in einem Deskriptor verwenden, da dieser automatisch mit Dokumenten oder Datensätzen abgeglichen wird, die `Fahrzeug` enthalten.

Konzepte und Typen als Deskriptoren

Verwenden Sie ein Konzept als Deskriptor, wenn Sie alle Dokumente oder Datensätze mit diesem Konzept (oder seine zugrunde liegenden Fachausdrücke) finden möchten. In diesem Fall ist es nicht notwendig, eine komplexere Kategorieregel zu verwenden, da der exakte Konzeptname ausreicht. Denken Sie daran, dass sich Konzepte bei der Verwendung von Ressourcen, die

Meinungen extrahieren, manchmal während der Extrahierung von TLA-Mustern verändern können, um den wahren Inhalt des Satzes zu erfassen (siehe dazu das Beispiel im nächsten Abschnitt über TLA).

Beispielsweise könnte die Antwort in einer Umfrage über das Lieblingsobst der Teilnehmer, etwa "*Äpfel und Ananas sind die Besten*", zur Extrahierung von `Apfel` und `Ananas` führen. Indem Sie das Konzept `Apfel` Ihrer Kategorie als Deskriptor hinzufügen, werden alle Antworten mit dem Konzept `Apfel` (oder seinen zugrunde liegenden Fachausdrücken) mit dieser Kategorie abgeglichen.

Wenn Sie jedoch einfach nur wissen möchten, welche Antworten in irgendeiner Form das Wort *Apfel* enthalten, können Sie eine Kategorieregel mit `* Apfel *` erstellen, wodurch alle Antworten erfasst werden, die Konzepte wie `Apfel`, `Apfelsauce` oder `französischer Apfelkuchen` enthalten.

Sie können auch alle Dokumente oder Datensätze mit Konzepten mit der gleichen Schreibweise erfassen, indem Sie einen Typ direkt als Deskriptor verwenden, z. B. `<Obst>`. Hinweis: Sternchen (*) können nicht mit Typen verwendet werden.

Für weitere Informationen siehe Thema Extrahierungsergebnisse: Konzepte und Typen in Kapitel 9 auf S. 143.

Textlinkanalyse- (TLA-) Muster als Deskriptoren

Verwenden Sie ein TLA-Musterergebnis als Deskriptor, wenn Sie feinere, nuancierte Ideen erfassen möchten. Wenn während einer TLA-Extrahierung Text analysiert wird, wird der Text satz- oder absatzweise verarbeitet, anstatt den Text als Ganzes zu betrachten (das Dokument oder der Datensatz). Indem alle Teile eines einzelnen Satzes zusammen betrachtet werden, kann die TLA beispielsweise Meinungen, Beziehungen zwischen zwei Elementen oder eine Negation identifizieren und so den wahren Sinn des Satzes erfassen. Sie können Konzeptmuster oder Typmuster als Deskriptoren verwenden. Für weitere Informationen siehe Thema Typ- und Konzeptmuster in Kapitel 12 auf S. 254.

Beispielsweise könnten in dem Satz "*Das Zimmer war nicht sonderlich sauber.*" die folgenden Konzepte extrahiert werden: `Zimmer` und `sauber`. Falls jedoch die TLA-Extrahierung in den Extrahierungseinstellungen deaktiviert wurde, könnte TLA erkennen, dass `sauber` negativ verwendet wurde und eigentlich `nicht sauber` entspricht, was als Synonym zu dem Konzept `dreckig` gilt. Hier zeigt sich, dass die Verwendung des Konzepts `sauber` als eigenständiger Deskriptor mit diesem Text übereinstimmen würde, jedoch auch andere Dokumente oder Datensätze mit dem Inhalt Sauberkeit erfasst werden könnten. Aus diesem Grund ist es besser, das TLA-Konzeptmuster mit `dreckig` als Ausgabekonzept zu verwenden, da dieses mit dem Text übereinstimmen würde und wahrscheinlich ein passenderer Deskriptor wäre.

Kategoriegeschäftsregeln als Deskriptoren

Kategorieregeln sind Anweisungen, mit denen Dokumente oder Datensätze auf Basis eines logischen Ausdrucks mithilfe von extrahierten Konzepten, Typen und Mustern sowie von Boole'schen Operatoren automatisch einer Kategorie zugewiesen werden. Sie könnten beispielsweise einen Ausdruck schreiben, der bedeutet *Schließe alle Datensätze, die das extrahierte Konzept* `Botschaft` *enthalten, nicht jedoch* `Argentinien`*, in diese Kategorie ein*.

Sie können Kategorieregeln in Ihren Kategorien als Deskriptoren erstellen und verwenden, um mit den Boole'schen Operatoren `&`, `|` und `!()` unterschiedliche Ideen auszudrücken. Nähere Informationen über die Syntax dieser Regeln und wie sie geschrieben und bearbeitet wird finden Sie unter Verwenden von Kategorieregeln auf S. 207

- Verwenden Sie eine Kategorieregel mit dem Boole'schen Operator `&` (UND), um Dokumente oder Datensätze zu finden, in denen zwei oder mehrere Konzepte vorkommen. Die zwei oder mehr durch `&`-Operatoren verbundenen Konzepte müssen nicht im selben Satz oder Satzteil auftreten, sondern können an beliebiger Stelle im selben Dokument oder Datensatz auftreten, um als Übereinstimmung mit einer Kategorie zu gelten. Wenn Sie beispielsweise die Kategorieregel `Essen & günstig` als Deskriptor erstellen, würde dadurch der Datensatz "*Das Essen war ziemlich teuer, aber das Zimmer war günstig*" als Übereinstimmung gelten, obwohl das Nomen `Essen` nicht als `günstig` bezeichnet wurde, da der Text sowohl `Essen` als auch `günstig` enthielt.
- Verwenden Sie eine Kategorieregel mit dem Boole'schen Operator `!()` (NICHT) als Deskriptor, um Dokumente oder Datensätze zu finden, in denen manche Ausdrücke vorkommen, andere jedoch nicht. So können Sie vermeiden, dass Informationen gruppiert werden, deren Wörter zwar ohne Zusammenhang erscheinen, nicht jedoch deren Kontext. Wenn Sie beispielsweise die Kategorieregel `<Unternehmen> & !(ibm)` als Deskriptor erstellen, würde der Text *SPSS Inc. ist ein 1967 gegründetes Unternehmen.* als Übereinstimmung gefunden werden, der Text *Das Softwareunternehmen wurde von IBM aufgekauft.* jedoch nicht.
- Verwenden Sie eine Kategorieregel mit dem Boole'schen Operator `|` (ODER) als Deskriptor, um Dokumente oder Datensätze mit einem oder mehreren Konzepten oder Typen zu finden. Wenn Sie beispielsweise die Kategorieregel `(Personal|Belegschaft|Team|Kollegen) & schlecht` als Deskriptor erstellen, würden alle Dokumente oder Datensätze als Übereinstimmung gelten, in denen mindestens eines dieser Nomen mit dem Konzept `schlecht` gefunden wurde.
- Verwenden Sie Typen in Kategorieregeln, um diese allgemeiner und möglicherweise anwendbarer zu gestalten. Beispielsweise möchten Sie bei der Arbeit mit Hoteldaten erfahren, was Ihre Kunden von dem Hotelpersonal halten. Verwandte Begriffe enthalten unter Umständen Wörter wie Rezeptionist, Kellner, Kellnerin, Hotelrezeption, Empfang usw. In diesem Fall könnten Sie einen neuen Typ namens `<HotelStaff>` erstellen und diesem Typ alle oben erwähnten Begriffe hinzufügen. Es ist zwar möglich, eine Kategorieregel für jede Personalart zu erstellen, etwa `[* Kellnerin * & nett]`, `[* Empfang * & freundlich]`, `[* Rezeptionist * & entgegenkommend]`, Sie können jedoch auch eine einzelne, allgemeinere Kategorieregel mit dem Typ `<HotelStaff>` erstellen, um alle Antworten zu erfassen, die sich positiv über das Hotelpersonal äußern, und zwar in der Form `[<Hotelpersonal> & <Positive>]`.

Anmerkung: Sie können sowohl + als auch `&` in Kategorieregeln verwenden, wenn Sie TLA-Muster in diese Regeln einschließen. Für weitere Informationen siehe Thema Verwenden von TLA-Mustern in Kategorieregeln auf S. 209.

Beispiele für unterschiedliche Übereinstimmungen bei Konzepten, TLA- oder Kategorieregeln

Das folgende Beispiel zeigt, wie sich die Verwendung eines Konzepts als Deskriptor, einer Kategorieregel als Deskriptor oder eines TLA-Musters als Deskriptor auf die Kategorisierung von Dokumenten oder Datensätzen auswirkt. Nehmen wir an, Sie haben folgende fünf Datensätze.

- A: *"Hervorragendes Restaurantpersonal, köstliches Essen und die Zimmer bequem und sauber."*
- B: *"Das Restaurantpersonal war schrecklich, aber die Zimmer waren sauber."*
- C: *"Die Zimmer waren bequem und sauber."*
- D: *"Mein Zimmer war nicht sonderlich sauber."*
- E: *"Sauber."*

Da die Datensätze das Wort *sauber* enthalten und Sie diese Informationen erfassen möchten, könnten Sie einen der in der folgenden Tabelle gezeigten Deskriptoren erstellen. Auf der Basis des wahren Inhalts, den Sie erfassen möchten, können Sie sehen, wie die Verwendung einer Deskriptorart anstatt einer anderen zu unterschiedlichen Ergebnissen führen kann.

Tabelle 10-1
Übereinstimmungen zwischen Beispieldatensätzen und Deskriptoren

Deskriptor	A	B	C	D	E	Erklärung
`sauber`	*Übereinstimmung*	*Übereinstimmung*	*Übereinstimmung*	*Übereinstimmung*	*Übereinstimmung*	Deskriptor ist ein extrahiertes Konzept. Jeder Datensatz enthielt das Konzept `sauber`, selbst Datensatz D, denn ohne TLA ist nicht automatisch bekannt, dass "*nicht sauber*" laut den TLA-Regeln `dreckig` bedeutet.
`sauber + .`	-	-	-	-	*Übereinstimmung*	Deskriptor ist ein TLA-Muster, das selbst für `sauber` steht. Nur Übereinstimmung mit dem Datensatz, in dem `sauber` während der TLA-Extrahierung ohne zugehöriges Konzept extrahiert wurde.
`[sauber]`	*Übereinstimmung*	*Übereinstimmung*	*Übereinstimmung*	-	*Übereinstimmung*	Deskriptor ist eine Kategorieregel, die nach einer TLA-Regel sucht, die das Wort `sauber` allein stehend oder in Verbindung mit anderen Wörtern enthält. Übereinstimmung mit dem Datensatz, in dem eine TLA-Ausgabe mit `sauber` gefunden wurde, unabhängig davon, ob `sauber` mit einem anderen Konzept wie `Zimmer` und in einer beliebigen Slot-Position verknüpft wurde.

Erläuterung von Kategorien

Kategorien bezeichnen eine Gruppe von eng verwandten Konzepten, Meinungen oder Haltungen. Um nützlich zu sein, sollte sich eine Kategorie auch leicht durch einen kurzen Ausdruck oder eine kurze Bezeichnung beschreiben lassen, der bzw. die ihre grundlegende Bedeutung erfasst.

Wenn Sie beispielsweise Umfrageantworten von Verbrauchern zu einem neuen Waschmittel analysieren, können Sie eine Kategorie mit dem Label *Duft* erstellen, das alle Antworten enthält, die den Geruch des Produkts beschreiben. Eine solche Kategorie würde jedoch nicht zwischen den Personen unterscheiden, die den Duft angenehm fanden, und den Personen, denen der Duft unangenehm war. Da IBM® SPSS® Modeler Text Analytics mithilfe der geeigneten Ressourcen Meinungen extrahieren kann, könnten Sie dann zwei andere Kategorien definieren, um Befragte zu identifizieren, die *den Duft mochten*, und Befragte, die *den Duft nicht mochten*.

Sie können Ihre Kategorien im Bereich "Kategorien" im oberen linken Bereich der Ansicht "Kategorien und Konzepte" erstellen und mit ihnen arbeiten. Die einzelnen Kategorien sind durch ein oder mehrere Deskriptoren definiert. **Deskriptoren** sind Konzepte, Typen und Muster sowie Kategorieregeln, die zum Definieren einer Kategorie verwendet wurden.

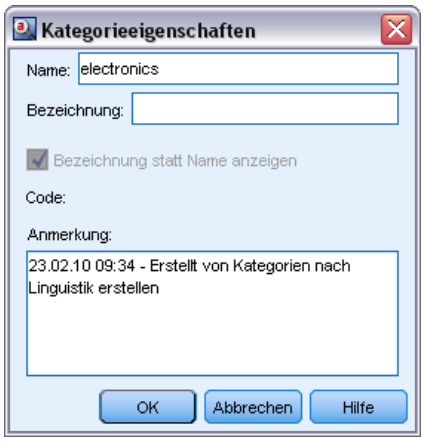
Wenn Sie die Deskriptoren sehen möchten, aus denen eine bestimmte Kategorie besteht, können Sie auf das Stiftsymbol in der Symbolleiste des Kategoriebereichs klicken und dann den Baum erweitern, um die Deskriptoren anzuzeigen. Alternativ können Sie die Kategorie auswählen und das Dialogfeld "Kategoriedefinitionen" öffnen (Ansicht > Kategoriedefinitionen).

Wenn Sie Kategorien automatisch mithilfe von Kategorieerstellungsverfahren (z. B. semantische Netze oder Konzepteinbeziehung) erstellen, verwenden die Verfahren Konzepte und Typen als Deskriptoren zum Erstellen der Kategorien. Beim Extrahieren von TLA-Mustern können Sie außerdem diese Muster oder Teile davon als Kategoriedeskriptoren hinzufügen. Für weitere Informationen siehe Thema Untersuchen von Textlinkanalyse in Kapitel 12 auf S. 252. Wenn Sie Cluster erstellen, können Sie die Konzepte in einem Cluster neuen oder bestehenden Kategorien hinzufügen. Und schließlich können Sie manuell Kategorieregeln erstellen, die als Deskriptoren in den Kategorien verwendet werden sollen. Für weitere Informationen siehe Thema Verwenden von Kategorieregeln auf S. 207.

Kategorieeigenschaften

Neben Deskriptoren verfügen Kategorien zudem über Eigenschaften, die Sie bearbeiten können, um Kategorien umzubenennen oder eine Bezeichnung oder Anmerkung hinzuzufügen.

Abbildung 10-3
Dialogfeld "Kategorieeigenschaften"



Folgende Eigenschaften sind vorhanden:

- **Name.** Dieser Name wird standardmäßig im Baum angezeigt. Wenn eine Kategorie mithilfe eines automatisierten Verfahrens erstellt wird, erhält sie automatisch einen Namen.
- **Bezeichnung.** Die Verwendung von Bezeichnungen ist nützlich beim Erstellen aussagekräftigerer Kategoriebeschreibungen zur Verwendung in anderen Produkten oder anderen Tabellen bzw. Grafiken. Wenn Sie die Option zur Anzeige der Bezeichnung auswählen, wird die Bezeichnung auf der Benutzeroberfläche zur Angabe der Kategorie verwendet.
- **Code.** Die Codenummer entspricht dem Codewert für diese Kategorie. .
- **Anmerkung.** In diesem Feld können Sie eine kurze Beschreibung für die einzelnen Kategorien hinzufügen. Wenn eine Kategorie über das Dialogfeld "Kategorien aufbauen" erstellt wird, wird dieser Anmerkung automatisch ein Hinweis hinzugefügt. Sie können auch direkt über den Datenbereich Text in eine Anmerkung einfügen, indem Sie den Text und anschließend aus den Menüs die Optionen Kategorien > Zu Anmerkung hinzufügen auswählen.

Der Fensterbereich "Daten"

Beim Erstellen von Kategorien kann es vorkommen, dass Sie einen Teil der Textdaten, mit denen Sie gerade arbeiten, überprüfen möchten. Wenn Sie beispielsweise eine Kategorie erstellen, in der 640 Dokumente kategorisiert sind, kann es erforderlich sein, einen Blick auf einige oder alle diese Dokumente zu werfen, um zu sehen, was dort tatsächlich geschrieben wurde. Sie können Datensätze oder Dokumente im Datenbereich überprüfen, der sich unten rechts befindet. Wird dieser nicht standardmäßig angezeigt, wählen Sie die Befehlsfolge Ansicht > Bereiche > Daten.

Der Fensterbereich "Daten" zeigt eine Zeile pro Dokument bzw. Datensatz entsprechend der Auswahl im Fensterbereich "Kategorien", "Extrahierungsergebnisse" bzw. im Dialogfeld "Kategoriedefinitionen". Die Anzeige erfolgt bis zu einer bestimmten Anzeigegrenze. Standardmäßig ist die Anzahl der im Fensterbereich "Daten" angezeigten Dokumente bzw. Datensätze begrenzt, damit Sie die Daten schneller sehen können. Sie können diese Einstellung jedoch im Optionsdialogfeld ändern. Für weitere Informationen siehe Thema Optionen: Registerkarte "Sitzung" in Kapitel 8 auf S. 135.

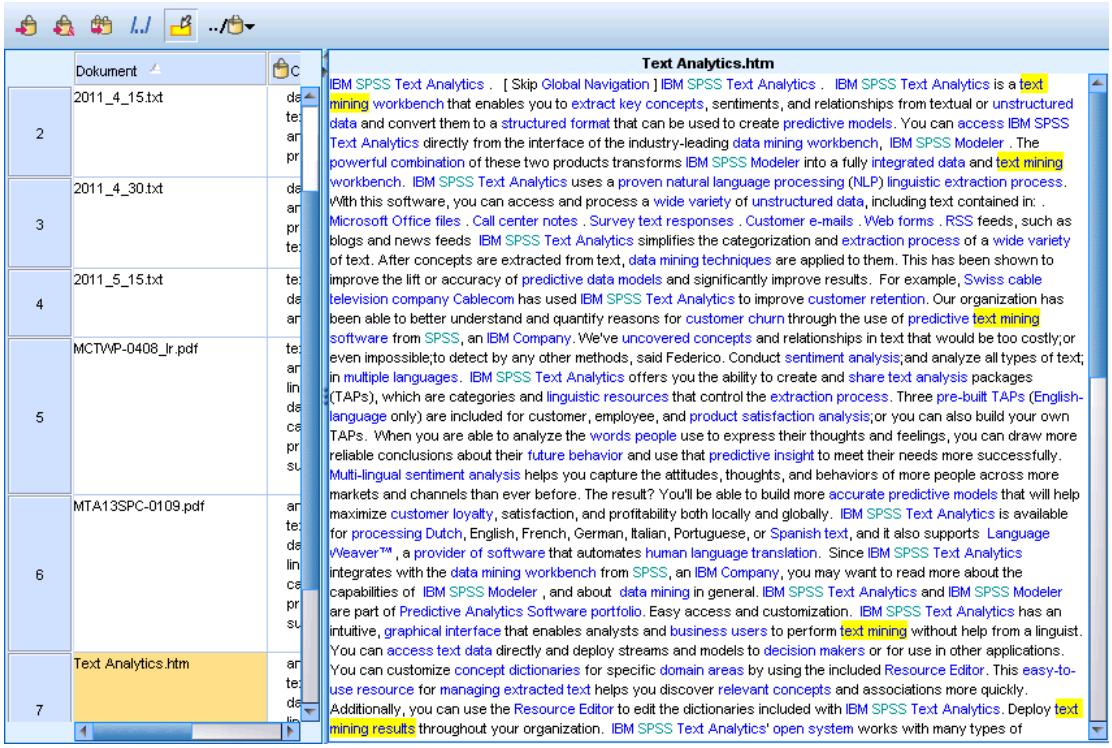
Datenbereich anzeigen und aktualisieren

Die Anzeige des Fensterbereichs “Daten” wird nicht automatisch aktualisiert, da die automatische Datenaktualisierung bei größeren Daten-Sets recht viel Zeit in Anspruch nehmen kann. Wenn Sie eine Auswahl in einem anderen Fensterbereich in dieser Ansicht oder im Dialogfeld “Kategoriedefinitionen” treffen, klicken Sie daher auf Anzeigen, um den Inhalt des Fensterbereichs “Daten” zu aktualisieren.

Textdokumente oder Datensätze

Wenn Ihre Textdaten als Datensätze vorliegen und der Text relativ kurz ist, zeigt das Textfeld im Datenbereich die Textdaten vollständig an. Wenn Sie jedoch mit Datensätzen und größeren Datenmengen arbeiten, zeigt die Textfeldspalte einen kurzen Abschnitt des Texts und öffnet einen Textvorschaubereich auf der rechten Seite, in dem ein größerer Teil oder der ganze Text des in der Tabelle markierten Datensatzes angezeigt wird. Wenn Ihre Textdaten als einzelne Dokumente vorliegen, wird im Datenbereich der Dateiname des Dokuments angezeigt. Wenn Sie ein Dokument markieren, wird der Textvorschaubereich geöffnet und der Text des ausgewählten Dokuments angezeigt.

Abbildung 10-4
Datenbereich mit Textvorschaubereich



Farben und Hervorheben

Wenn Sie die Daten anzeigen, werden die in diesen Dokumenten oder Datensätzen gefundenen Konzepte und Deskriptoren farblich hervorgehoben, damit Sie diese leicht im Text identifizieren können. Die Farbkodierung entspricht den Typen, die den Konzepten zugewiesen sind. Alternativ können Sie die Maus über farbkodierte Elemente bewegen, um das Konzept anzuzeigen, unter dem das betreffende Element extrahiert wurde, und den Typ, dem es zugewiesen wurde. Nicht extrahierter Text wird schwarz angezeigt. Bei diesen nicht extrahierten Wörtern handelt es sich meistens um Verbindungselemente (*und* oder *mit*), Pronomen (*mich* oder *sie*) und Verben (*ist*, *haben* oder *nehmen*).

Datenbereichsspalten

Die Textfeldspalte wird immer angezeigt. Sie können jedoch auch andere Spalten anzeigen. Wählen Sie für die Anzeige anderer Spalten die Optionsfolge Ansicht > Datenbereich und anschließend die Spalte aus, die im Datenbereich angezeigt werden soll. Folgende Spalten stehen ggf. zur Anzeige zur Verfügung:

- **"Textfeldname" (Anzahl)/Dokumente.** Fügt eine Spalte für die Textdaten hinzu, aus denen Konzepte und Typ extrahiert wurden. Wenn sich Ihre Daten in Dokumenten befinden, trägt die Spalte den Titel "Dokumente" und nur der Dateiname des Dokuments oder der vollständige Pfad ist sichtbar. Um den Text für diese Dokumente einzusehen, müssen Sie den Fensterbereich "Textvorschau" betrachten. Die Anzahl der Zeilen im Fensterbereich "Daten" wird in Klammern nach diesem Spaltennamen angezeigt. Es kann vorkommen, dass aufgrund einer Einschränkung im Dialogfeld "Optionen", die zur Beschleunigung des Ladevorgangs dient, nicht alle Dokumente bzw. Datensätze angezeigt werden. Wenn die maximale Anzahl erreicht wurde, steht nach der Zahl die Angabe - Maximum.Für weitere Informationen siehe Thema Optionen: Registerkarte "Sitzung" in Kapitel 8 auf S. 135.
- **Kategorien.** Führt jede Kategorie auf, der ein Datensatz angehört. Wenn diese Spalte angezeigt wird, kann die Aktualisierung des Datenbereichs ein wenig länger dauern, da jeweils die aktuellsten Informationen angezeigt werden.
- **Relevanzrang.** Führt den Rang jedes Datensatzes einer einzelnen Kategorie auf. Dieser Rang gibt an, wie gut der Datensatz im Verhältnis zu den anderen Datensätzen in der Kategorie zu der Kategorie passt. Wählen Sie eine Kategorie im Fensterbereich "Kategorien" (oben links) aus, um den Rang anzuzeigen. Für weitere Informationen siehe Thema Kategorierelevanz auf S. 181.
- **Kategorienanzahl.** Führt die Anzahl der Kategorien auf, denen ein Datensatz angehört.

Kategorierelevanz

Um bessere Kategorien aufzubauen, können Sie die Relevanz der Dokumente Datensätze in jeder Kategorie sowie die Relevanz aller Kategorien überprüfen, zu der ein Dokument oder Datensatz gehört.

Relevanz einer Kategorie für einen Datensatz

Wenn ein Dokument oder Datensatz im Datenbereich angezeigt wird, werden alle zugehörigen Kategorien in der Spalte “Kategorien” aufgeführt. Wenn ein Dokument oder Datensatz zu mehreren Kategorien gehört, werden die Kategorien in dieser Spalte von der relevantesten absteigend bis zur am wenigsten relevanten angezeigt. Die erste Kategorie stimmt also am besten mit dem Dokument oder Datensatz überein. Für weitere Informationen siehe Thema Der Fensterbereich “Daten” auf S. 179.

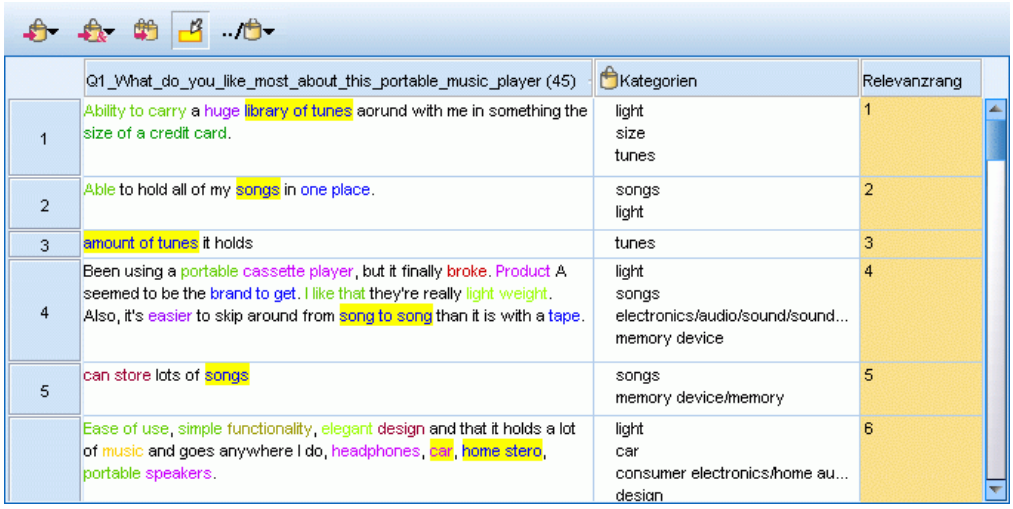
Relevanz eines Datensatzes für eine Kategorie

Wenn Sie eine Kategorie auswählen, können Sie die Relevanz der jeweiligen Datensätze im Datenbereich in der Spalte “Relevanzrang” überprüfen. Dieser Relevanzrang gibt an, wie gut das Dokument oder der Datensatz im Verhältnis zu anderen Datensätzen in dieser Kategorie in die gewählte Kategorie passt. Um den Rang der Datensätze für eine einzelne Kategorie anzuzeigen, wählen Sie diese Kategorie im Fensterbereich “Kategorien” (oben links) aus und der Rang für das Dokument oder den Datensatz wird in der Spalte angezeigt. Diese Spalte wird standardmäßig nicht eingeblendet. Sie können jedoch festlegen, dass sie eingeblendet werden soll. Für weitere Informationen siehe Thema Der Fensterbereich “Daten” auf S. 179.

Je niedriger die Rangnummer eines Datensatzes, desto besser passt er zur gewählten Kategorie, das heißt, eine 1 steht für die beste Entsprechung. Wenn mehrere Datensätze dieselbe Relevanz aufweisen, werden alle mit demselben Rang gefolgt von einem Gleichheitszeichen (=) angezeigt, um zu kennzeichnen, dass sie dieselbe Relevanz besitzen. Sie könnten beispielsweise folgende Ränge haben: 1=, 1=, 3, 4 usw., d. h., es gibt zwei Datensätze, von denen beide die beste Übereinstimmung für diese Kategorie aufweisen.

Tipp: Sie können den Text des relevantesten Datensatzes in die Anmerkung zur Kategorie einfügen, um eine bessere Beschreibung für diese Kategorie zu erstellen. Fügen Sie den Text direkt über den Datenbereich ein, indem Sie den Text und anschließend die Menüoptionen Kategorien > Zu Anmerkung hinzufügen auswählen.

Abbildung 10-5
Datenbereich mit Kategorien und Relevanzrang

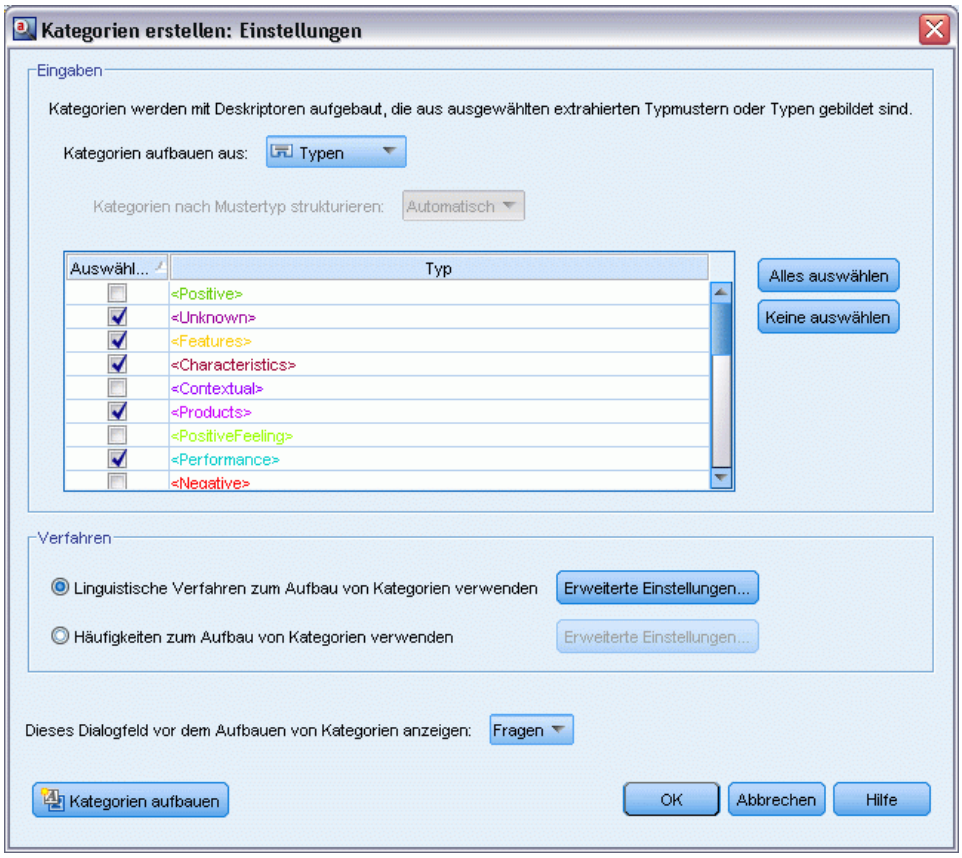


	Q1_What_do_you_like_most_about_this_portable_music_player (45)	Kategorien	Relevanzrang
1	Ability to carry a huge library of tunes aorund with me in something the size of a credit card.	light size tunes	1
2	Able to hold all of my songs in one place.	songs light	2
3	amount of tunes it holds	tunes	3
4	Been using a portable cassette player, but it finally broke. Product A seemed to be the brand to get. I like that they're really light weight. Also, it's easier to skip around from song to song than it is with a tape.	light songs electronics/audio/sound/sound... memory device	4
5	can store lots of songs	songs memory device/memory	5
	Ease of use, simple functionality, elegant design and that it holds a lot of music and goes anywhere I do, headphones, car, home stero, portable speakers.	light car consumer electronics/home au... design	6

Erstellen von Kategorien

Sie können zwar über Kategorien aus einem Textanalysepaket verfügen, aber Sie können auch mit einigen linguistischen und häufigkeitsbasierten Verfahren automatisch Kategorien erstellen. Über das Dialogfeld Kategorien aufbauen: Einstellungen können Sie die automatisierten linguistischen und häufigkeitsbasierten Verfahren anwenden, um Kategorien aus Konzepten oder Konzeptmustern zu erstellen.

Abbildung 10-6
Dialogfeld "Kategorien aufbauen"



In der Regel können Kategorien aus unterschiedlichen Deskriptoren (Typen, Konzepte, TLA-Muster, Kategorieregeln) aufgebaut werden. Wenn Sie Kategorien mit automatisierten Verfahren zum Aufbauen von Kategorien erstellen, wird jede erstellte Kategorie nach einem Konzept oder Konzeptmuster (je nach Auswahl) benannt und enthält eine Reihe an Deskriptoren. Diese Deskriptoren können aus Kategorieregeln oder Konzepten bestehen und enthalten alle zugehörigen Konzepte, die von den Verfahren erkannt werden.

Nach dem Aufbau der Kategorien können Sie viel über die Kategorien erfahren, indem Sie sie im Fensterbereich "Kategorien" überprüfen oder in den Diagrammen und Tabellen untersuchen. Anschließend können Sie mithilfe von manuellen Verfahren kleinere Anpassungen vornehmen, etwaige Fehlklassifizierungen beheben oder Datensätze oder Wörter hinzufügen, die nicht erfasst wurden. Nachdem Sie ein Verfahren angewendet haben, stehen die Konzepte, Typen und Muster, die in eine Kategorie gruppiert wurden, weiterhin für andere Verfahren zur Verfügung. Da

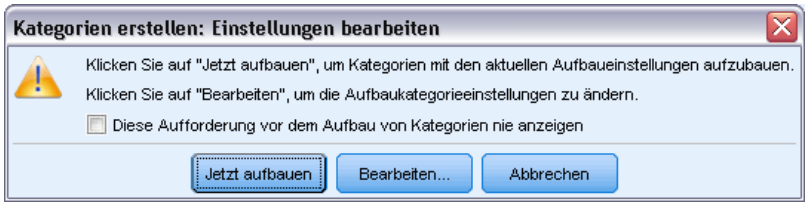
die Verwendung verschiedener Verfahren auch zu redundanten oder ungeeigneten Kategorien führen kann, können Sie außerdem Kategorien zusammenführen bzw. löschen. Für weitere Informationen siehe Thema Bearbeiten und Verfeinern von Kategorien auf S. 237.

Wichtig: In früheren Versionen wurden Kookkurenz- und Synonymregeln von eckigen Klammern umgeben. In dieser neuen Version zeigen nun eckige Klammern ein Musterergebnis für Text-Link-Analyse an. Stattdessen stehen Kookkurrenz- und Synonymregeln in runden Klammern, z. B. `(Lautsprechersysteme|Lautsprecher)`.

So erstellen Sie Kategorien

▶ Wählen Sie in den Menüs die Optionsfolge Kategorien > Kategorien aufbauen. Wenn Sie nicht das Kontrollkästchen markiert haben, dass diese Aufforderung nie angezeigt werden soll, wird ein Dialogfeld geöffnet.

Abbildung 10-7
Aufforderung vor dem Aufbau



▶ Wählen Sie aus, ob Sie die Kategorie jetzt aufbauen oder zuerst die Einstellungen bearbeiten möchten.

- Klicken Sie auf Jetzt aufbauen, um mit den aktuellen Einstellungen den Aufbau einer Kategorie zu beginnen. Die standardmäßig ausgewählten Einstellungen sind für den Beginn der Kategorisierung oft ausreichend. Der Kategorieaufbauprozess wird gestartet und ein Dialogfeld über den Fortschritt wird angezeigt.
- Klicken Sie auf Bearbeiten, um die Aufbaueinstellungen zu überprüfen und zu ändern.

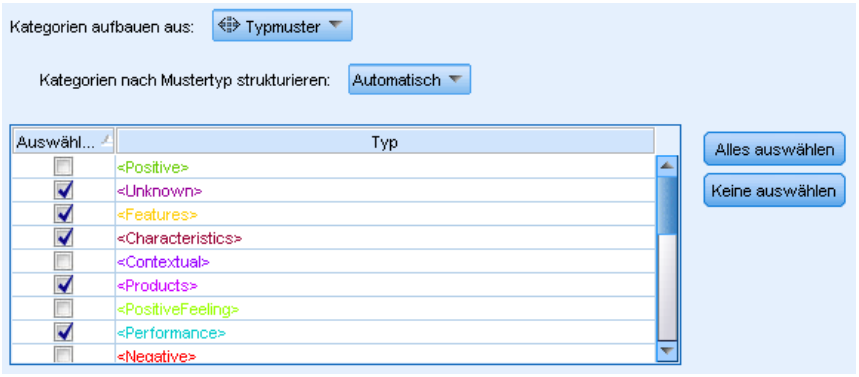
Anmerkung: Es können maximal 10.000 Kategorien angezeigt werden. Wenn diese Zahl erreicht oder überschritten wird, wird eine Warnung angezeigt. Wenn dies geschieht, sollten Sie die Option “Kategorien aufbauen” bzw. “Kategorien erweitern” ändern, um die Anzahl der erstellten Kategorien zu verringern.

Eingaben

Die Kategorien werden aus Deskriptoren aufgebaut, die aus Typmuster oder Typen abgeleitet werden. In der Tabelle können Sie die einzelnen Typen oder Muster auswählen, die in den Kategorieaufbauprozess aufgenommen werden.

Typmuster. Wenn Sie Typmuster auswählen, werden die Kategorien aus Mustern anstelle von Typen und aus einzelnen Konzepten aufgebaut. So werden Datensätze oder Dokumente kategorisiert, die ein Konzeptmuster enthalten, das zum gewählten Typmuster gehört. Wenn Sie in der Tabelle also das Typmuster `<Budget>` und `<Positive>` auswählen, können Kategorien wie `Kosten & <Positive>` oder `Sätze & Hervorragend` erzeugt werden.

Abbildung 10-8
Dialogfeld "Kategorien aufbauen" mit den verfügbaren Typmustern



Wenn Typmuster als Eingabe zum automatisierten Kategorieaufbau verwendet werden, können die Techniken manchmal mehrere Möglichkeiten zur Bildung der Kategoriestruktur identifizieren. Technisch gesehen gibt es nicht nur eine richtige Lösung zum Erzeugen von Kategorien, Sie finden jedoch eventuell eine Struktur besser für Ihre Analyse geeignet als andere. Um in diesem Fall die Ausgabe besser anpassen zu können, können Sie einen Typ als bevorzugt kennzeichnen. Alle erzeugten Kategorien höchster Ebene entstammen einem Konzept des Typs, den Sie hier auswählen (keines anderen Typs). Jede Unterkategorie enthält ein Textlinkmuster aus diesem Typ. Wählen Sie diesen Typ im Feld Kategorien nach Mustertyp strukturieren: aus und die Tabelle wird aktualisiert und zeigt nur die jeweiligen Muster mit dem ausgewählten Typ an. In den meisten Fällen ist `<Unbekannt>` vorausgewählt. Das führt dazu, dass alle Muster, die den Typ `<Unbekannt>` (für nichtjapanischen Text) enthalten, ausgewählt werden. Für japanischen Text wird <名詞> durch das Programm vorausgewählt. *Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung. Die Tabelle zeigt die Typen in absteigender Reihenfolge, beginnend mit der größten Anzahl an Datensätzen oder Dokumenten (Anzahl Doc.).

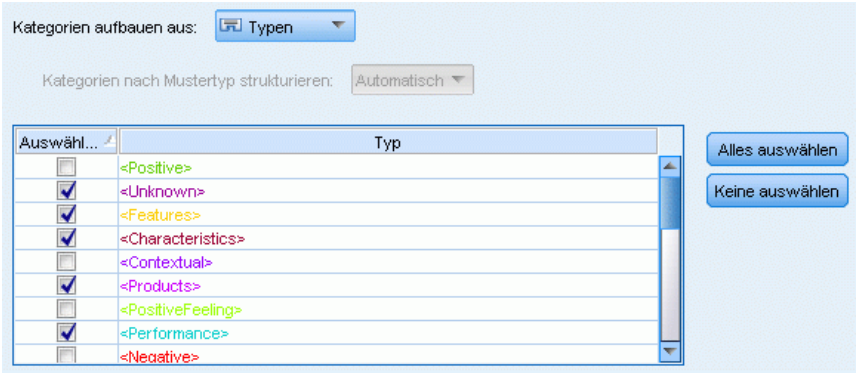
Typen. Wenn Sie Typen auswählen, werden die Kategorien aus Konzepten aufgebaut, die zu den gewählten Typen gehören. Wenn Sie also den Typ `<Budget>` in der Tabelle auswählen, können Kategorien wie `Kosten` oder `Preis` erzeugt werden, da `Kosten` und `Preis` Konzepte sind, die dem Typ `<Budget>` zugewiesen wurden.

Standardmäßig werden nur die Typen ausgewählt, die die meisten Datensätze oder Dokumente erfassen. Dank dieser Vorauswahl können Sie sich schnell auf die interessantesten Typen konzentrieren und den Aufbau uninteressanter Kategorien vermeiden. Die Tabelle zeigt die Typen in absteigender Reihenfolge, beginnend mit der größten Anzahl an Datensätzen oder Dokumenten (Anzahl Doc.). Typen aus der Bibliothek `Opinions` werden standardmäßig in der Typtabelle abgewählt.

Die sich ergebenden Kategorien hängen von der ausgewählten Eingabe ab. Wenn Sie Typen als Eingabe verwenden, können Sie die eindeutig verwandten Konzepte leichter erkennen. Wenn Sie beispielsweise Kategorien mit Typen als Eingaben erstellen, könnten Sie eine Kategorie `Obst` erhalten, die, die Konzepte `Apfel`, `Birne`, `Zitrusfrüchte`, `Orange` usw. enthält. Wenn Sie stattdessen Typmuster als Eingabe wählen und beispielsweise das Muster `<Unbekannt> + <Positiv>` auswählen, erhalten Sie möglicherweise die Kategorie `Obst + <Positiv>` mit ein oder zwei Arten von Obst, wie beispielsweise `Obst + lecker` und `Apfel + gut`. Bei

diesem zweiten Ergebnis werden nur 2 Konzeptmuster angezeigt, da die anderen Vorkommnisse von Obst nicht unbedingt positiv qualifiziert sind. Und während dies möglicherweise für Ihre aktuellen Textdaten brauchbar ist, kann es für Longitudinalstudien, in denen verschiedene Mengen von Dokumenten verwendet werden, sinnvoll sein, von Hand weitere Deskriptoren hinzuzufügen, wie beispielsweise `Zitrusfrucht + positiv`, oder aber Typen zu verwenden. Wenn Sie ausschließlich Typen als Eingabe verwenden, können Sie alle möglichen Arten von Obst finden.

Abbildung 10-9
Dialogfeld "Kategorien aufbauen" mit den verfügbaren Typen



Verfahren

Da jedes Daten-Set anders ist, kann sich die Anzahl der Methoden und die Reihenfolge, in der sie angewendet werden, im Laufe der Zeit ändern. Da Ihre Text-Mining-Ziele von Daten-Set zu Daten-Set unterschiedlich sein können, müssen Sie möglicherweise mit den verschiedenen Verfahren experimentieren, um zu sehen, welches die besten Ergebnisse für die jeweiligen Textdaten bringt.

Sie müssen nicht besonders gut mit diesen Einstellungen vertraut sein, um sie verwenden zu können. Standardmäßig sind die gängigsten Einstellungen bereits ausgewählt. Daher können Sie die Dialogfelder für die erweiterten Einstellungen überspringen und gleich mit dem Aufbau der Kategorien beginnen. Wenn Sie hier Änderungen vornehmen, müssen Sie diese nicht mit jedem Öffnen des Dialogfelds erneut vornehmen, da immer die neuesten Einstellungen beibehalten werden.

Wählen Sie entweder die linguistischen oder häufigkeitsbasierten Verfahren und klicken Sie auf die Schaltfläche "Erweiterte Einstellungen", um die Einstellungen für die gewählten Verfahren anzuzeigen. Keines der automatischen Verfahren ergibt eine perfekte Kategorisierung Ihrer Daten, weshalb wir empfehlen, ein oder mehrere automatische Verfahren anzuwenden, die gut mit Ihren Daten funktionieren. Ein gleichzeitiger Einsatz von linguistischen und häufigkeitsbasierenden Verfahren ist beim Aufbau nicht möglich.

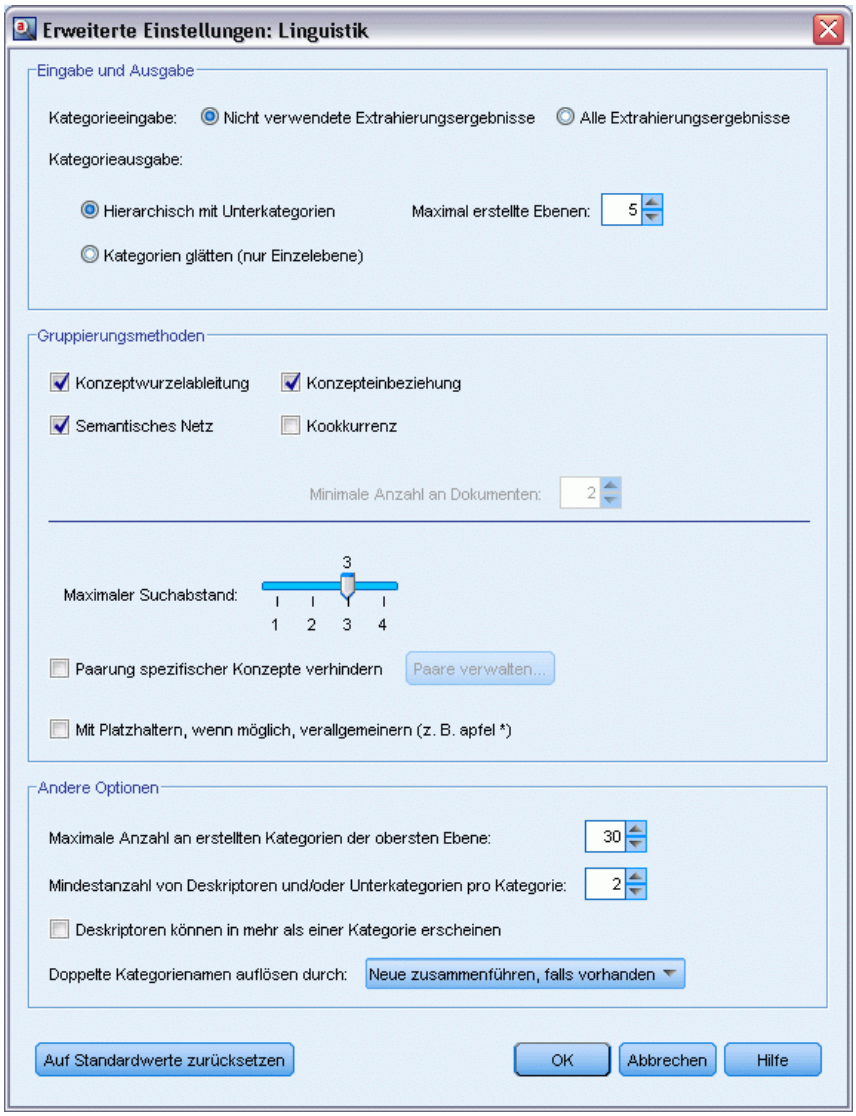
- **Erweiterte linguistische Verfahren**. Weitere Informationen finden Sie unter auf S. 187.
- **Erweiterte häufigkeitsbasierte Verfahren**. Weitere Informationen finden Sie unter auf S. 198.

Erweiterte linguistische Einstellungen

Beim Aufbau von Kategorien haben Sie die Auswahl zwischen einigen erweiterten linguistischen Verfahren für den Kategorieaufbau, darunter *Konzeptwurzelableitung*(in Japanisch nicht verfügbar), *Konzepteinbeziehung*, *semantische Netze* (nur für englischen Text) und *Kookkurrenzregeln*. Zum Erstellen von Kategorien können diese Verfahren einzeln oder in Kombination verwendet werden.

Beachten Sie, dass jedes Daten-Set anders ist und sich die Anzahl der Methoden und die Reihenfolge, in der sie angewendet werden, daher im Laufe der Zeit ändern kann. Da Ihre Text-Mining-Ziele von Daten-Set zu Daten-Set unterschiedlich sein können, müssen Sie möglicherweise mit den verschiedenen Verfahren experimentieren, um zu sehen, welches die besten Ergebnisse für die jeweiligen Textdaten bringt. Keines der automatischen Verfahren ergibt eine perfekte Kategorisierung Ihrer Daten, weshalb wir empfehlen, ein oder mehrere automatische Verfahren anzuwenden, die gut mit Ihren Daten funktionieren.

Abbildung 10-10
Erweiterte Einstellungen: Dialogfeld "Linguistik" für die Erstellung von Kategorien



Eingabe und Ausgabe

Kategorieeingabe. Auswahl, woraus die Kategorien aufgebaut werden:

- Nicht verwendete Extrahierungsergebnisse. Diese Option aktiviert Kategorien, die aus Extrahierungsergebnissen erstellt werden, die nicht in vorhandenen Kategorien verwendet werden. So wird die Tendenz für Datensätze minimiert, mehrere Kategorien abzugleichen und die Anzahl der erzeugten Kategorien zu begrenzen.
- Alle Extrahierungsergebnisse. Diese Option aktiviert unter Verwendung der Extrahierungsergebnisse zu erstellende Kategorien. Dies ist am sinnvollsten, wenn noch keine oder nur sehr wenige Kategorien vorhanden sind.

Kategorieausgabe. Auswahl der allgemeinen Struktur für die aufzubauenden Kategorien:

- Hierarchisch mit Unterkategorien. Diese Option aktiviert das Erstellen von Unterkategorien und Unter-Unterkategorien. Sie können die Tiefe Ihrer Kategorien einstellen, indem Sie die maximale Anzahl an zu erstellenden Ebenen (Feld Maximal erstellte Ebenen) auswählen. Wenn Sie 3 auswählen, können Kategorien Unterkategorien enthalten, welche wiederum ebenfalls Unterkategorien enthalten können.
- Flache Kategorien (nur eine Ebene). Mit dieser Option wird nur eine Kategorieebene aufgebaut, d. h., es werden keine Unterkategorien erzeugt.

Gruppierverfahren

Die einzelnen verfügbaren Verfahren sind für bestimmte Datentypen und Situationen jeweils sehr gut geeignet, doch ist es oftmals nützlich, bei einer Analyse mehrere Verfahren miteinander zu verbinden, um die Dokumente oder Datensätze vollständig zu erfassen. Sie können ein Konzept in mehreren Kategorien erkennen oder redundante Kategorien vorfinden.

Konzeptwurzelableitung. Mit diesem Verfahren werden Kategorien erstellt, indem ausgehend von einem Konzept andere verwandte Konzepte ermittelt werden (durch Analyse, ob bestimmte Konzeptkomponenten morphologisch verwandt sind oder gemeinsame Wurzeln haben). Dieses Verfahren ist sehr nützlich bei der Identifizierung von bedeutungsgleichen Konzepten aus zusammengesetzten Wörtern, da die Konzepte in jeder generierten Kategorie die gleiche oder ähnliche Bedeutung haben. Das Verfahren funktioniert mit Daten unterschiedlicher Länge und erzeugt eine geringere Anzahl an kompakten Kategorien. So wird beispielsweise das Konzept `Möglichkeiten zum Aufstieg` mit den Konzepten`Möglichkeit des Aufstiegs` und `Aufstiegsmöglichkeit` zu einer Kategorie zusammengefasst. Für weitere Informationen siehe Thema Konzeptwurzelableitung auf S. 193.Diese Option ist nicht für japanischen Text verfügbar.

Semantisches Netz. Bei diesem Verfahren wird zunächst auf der Grundlage eines umfassenden Index von Wortbeziehungen jedes Konzept auf seine möglichen Bedeutungen untersucht. Anschließend werden Kategorien durch Gruppieren zusammenhängender Konzepte erstellt. Diese Technik empfiehlt sich, wenn die Konzepte dem semantischen Netz bekannt und nicht zu zweideutig sind. Sie ist weniger hilfreich, wenn der Text spezielle Terminologie oder Sprache enthält, die dem Netz unbekannt ist. Das Konzept `Granny Smith Apfel` würde zum Beispiel mit `Gala Apfel` und `Winesap Apfel` gruppiert, da es sich um Geschwister von Granny Smith handelt. In einem anderen Beispiel würde das Konzept `Tier` mit `Katze` und `Känguru` gruppiert, da dies Hyponyme von `Tier` sind. Dieses Verfahren ist in dieser Version nur für englischen Text verfügbar. Für weitere Informationen siehe Thema Semantische Netze auf S. 195.

Konzepteinbeziehung. Diese Technik erstellt Kategorien durch die Gruppierung von Multiterm-Konzepten (zusammengesetzte Wörter) basierend darauf, ob sie Wörter enthalten, die Unter- oder Übermengen eines Worts in dem anderen sind. So wird beispielsweise `Sitz` mit `Kindersitz`, `Sitzheizung` und `Kindersitzgurt` zu einer Gruppe zusammengefasst. Für weitere Informationen siehe Thema Konzepteinbeziehung auf S. 194.

Kookkurrenz. Diese Technik erstellt Kategorien aus Kookkurrenz im Text. Dahinter steht folgende Idee: Wenn Konzepte oder Konzeptmuster häufig in Dokumenten bzw. Datensätzen gefunden werden, ist diese Kookkurrenz Ausdruck einer zugrunde liegenden Beziehung, die wahrscheinlich in Ihren Kategoriedefinitionen von Nutzen ist. Wenn Wörter eine signifikante

Kookkurrenz aufweisen, wird eine Kookkurrenzregel erstellt, die als Kategoriedeskriptor für eine neue Unterkategorie verwendet werden kann. Wenn beispielsweise viele Datensätze die Wörter `Preis` und `Verfügbarkeit` enthalten (wenige jedoch nur eines von beiden), könnten diese Konzepte in eine Kookkurrenzregel zusammengefasst (`Preis &verfügbar`) und beispielsweise einer Unterkategorie der Kategorie `Preis` zugewiesen werden.Für weitere Informationen siehe Thema Kookkurrenzregeln auf S. 197.

- **Minimale Anzahl anDokumenten.** Um festzustellen, wie interessant Kookkurrenzen sind, definieren Sie die minimale Anzahl an Dokumenten oder Datensätzen, die eine bestimmte Kookkurrenz enthalten muss, um als Deskriptor in einer Kategorie verwendet zu werden.

Maximaler Suchabstand. Legen Sie fest, wie weit die Verfahren suchen sollen, bevor Kategorien erstellt werden. Je niedriger der Wert, desto weniger Ergebnisse erhalten Sie. Allerdings sind die Ergebnisse weniger verrauscht und mit größerer Wahrscheinlichkeit auf signifikante Weise miteinander verknüpft oder verbunden. Je höher der Wert, desto mehr Ergebnisse erhalten Sie. Allerdings sind diese Ergebnisse möglicherweise weniger zuverlässig oder relevant. Während diese Option global auf alle Verfahren angewendet wird, hat sie die größte Auswirkung auf Kookkurrenzen und semantische Netze.

Paarung spezifischer Konzepte verhindern. Markieren Sie dieses Kontrollkästchen, um den Vorgang der Gruppierung oder Paarung von zwei Konzepten in der Ausgabe zu verhindern. Klicken Sie zum Erstellen oder Verwalten von Konzeptpaaren auf Paare verwalten.... Für weitere Informationen siehe Thema Verwalten von Verknüpfungsausnahmepaaren auf S. 191.

Nach Möglichkeit mit Platzhaltern verallgemeinern. Wählen Sie diese Option, um dem Produkt zu ermöglichen, für Kategorien, in denen ein Sternchen als Platzhalter verwendet wird, allgemeine Regeln aufzustellen. Beispielsweise könnte anstelle der Erzeugung mehrerer Deskriptoren wie `[Apfel vom Bioladen + .]` und `[Apfelsauce + .]` der Einsatz von Platzhaltern `[Apfel * + .]` erzeugen. Wenn Sie mit Platzhaltern verallgemeinern, erhalten Sie oft genau die gleiche Anzahl an Datensätzen oder Dokumenten wie zuvor. Diese Option hat jedoch den Vorteil, die Zahl zu verringern und die Kategoriedeskriptoren zu vereinfachen. Zusätzlich erhöht diese Option die Möglichkeit, mehr Datensätze oder Dokumente unter Verwendung dieser Kategorien zu neuen Textdaten (zum Beispiel bei Langzeit-/Wellenstudien) zu kategorisieren.

Weitere Optionen für den Aufbau von Kategorien

Neben der Auswahl der anzuwendenden Gruppierungstechniken können Sie folgende weitere Optionen bearbeiten:

Maximale Anzahl an erstellten Kategorien der obersten Ebene**.** Verwenden Sie diese Option zur Beschränkung der Anzahl an Kategorien, die erstellt werden können, wenn Sie als Nächstes auf die Schaltfläche “Kategorien aufbauen” klicken. In einigen Fällen erzielen Sie bessere Ergebnisse, wenn Sie diesen Wert hoch setzen und dann etwaige uninteressante Kategorien löschen.

Mindestanzahl an Deskriptoren und/oder Unterkategorien pro Kategorie. Verwenden Sie diese Option, um die Mindestanzahl an Deskriptoren und Unterkategorien zu definieren, die eine Kategorie enthalten muss, um erstellt zu werden. Durch diese Option kann das Erstellen von Kategorien eingeschränkt werden, die keine hohe Zahl von Datensätzen oder Dokumenten erfassen.

Deskriptoren können in mehr als einer Kategorie angezeigt werden. Wenn ausgewählt, ermöglicht diese Option, dass Deskriptoren in mehr als einer der Kategorien verwendet werden, die als nächste erstellt werden. Diese Option ist allgemein ausgewählt, da Elemente häufig oder "natürlich" in zwei oder mehr Kategorien fallen, so dass sie in der Regel zu Kategorien höherer Qualität führen. Wenn Sie diese Option nicht auswählen, verringern Sie die Überschneidung von Datensätzen in mehreren Kategorien und abhängig von dem vorhandenen Datentyp. Dies könnte gewünscht sein. Bei den meisten Datentypen jedoch führt die Einschränkung von Deskriptoren auf eine einzelne Kategorie zu einem Verlust an Qualität oder Kategorieabdeckung. Angenommen, Sie hätten das Konzept `Autositzhersteller`. Mit dieser Option könnte dieses Konzept in einer Kategorie basierend auf dem Text `Autositz` und in einem anderen auf `Hersteller` basieren. Wenn diese Option aber nicht ausgewählt ist, wird das Konzept `Autositzhersteller`, auch wenn Sie noch beide Kategorien erhalten, nur als Deskriptor in der Kategorie angezeigt, in der es basierend auf verschiedenen Faktoren einschließlich der Anzahl an Datensätzen, in denen `Autositz` und `Hersteller` jeweils auftreten, am besten passt.

Konflikte mit doppelten Namen lösen. Wählen Sie, wie mit neuen Kategorien oder Unterkategorien verfahren werden soll, deren Namen mit denen von bestehenden Kategorien identisch wären. Sie können entweder die neuen Kategorien (und ihre Deskriptoren) mit den bestehenden Kategorien desselben Namens zusammenführen. Alternativ können Sie die Erstellung jeglicher Kategorien überspringen, wenn in den bestehenden Kategorien ein Namensduplikat gefunden wird.

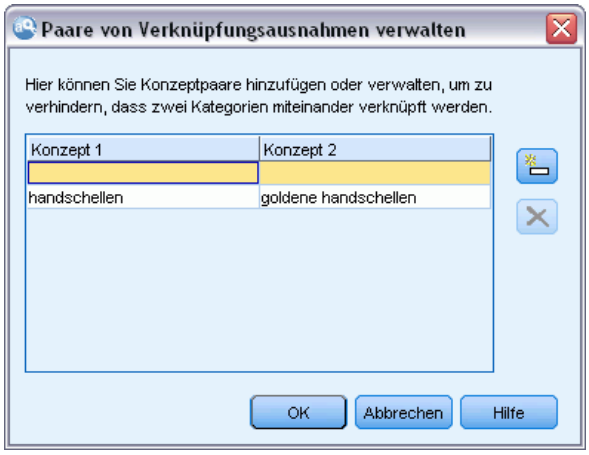
Verwalten von Verknüpfungsausnahmepaaren

Bei der automatischen Kategorieerstellung, Clusterbildung und Zuordnen von Konzepten gruppieren die internen Algorithmen Wörter anhand bekannter Zuweisungen. Damit zwei Konzepte nicht gepaart oder miteinander verknüpft werden können, aktivieren Sie diese Funktion im Dialogfeld Kategorien aufbauen: Erweiterte Einstellungen, im Dialogfeld Cluster aufbauen und im Dialogfeld Konzeptkartenindex: Einstellungen und klicken auf die Schaltfläche Paare verwalten.

Im anschließend gezeigten Dialogfeld Verknüpfungsausnahmen verwalten können Sie Konzeptpaare hinzufügen, bearbeiten oder löschen. Geben Sie ein Paar pro Zeile ein. Durch die Eingabe von Paaren an dieser Stelle wird verhindert, dass die Paarbildung beim Erstellen oder Erweitern von Kategorien, beim Cluster-Aufbau und beim Zuordnen von Konzepten erfolgt. Geben Sie Wörter exakt wie gewünscht ein, z. B. unterscheidet sich die mit einem Akzent versehene Version eines Wortes von der Wortversion ohne Akzent.

Wenn Sie z. B. sicherstellen möchten, dass `Hot Dog` und `Dog` nicht gruppiert werden, können Sie das Set als separate Zeile in der Tabelle hinzufügen.

Abbildung 10-11
Dialogfeld zum Verwalten von Verknüpfungsausnahmepaaren



Über linguistische Verfahren

Beim Aufbau bzw. der Erweiterung von Kategorien haben Sie die Auswahl zwischen einigen erweiterten linguistischen Verfahren für den Kategorieaufbau, darunter *Konzeptwurzelableitung*(in Japanisch nicht verfügbar), *Konzepteinbeziehung*, *semantische Netze* (nur für Englisch) und *Kookkurrenzregeln*. Zum Erstellen von Kategorien können diese Verfahren einzeln oder in Kombination verwendet werden.

Sie müssen nicht besonders gut mit diesen Einstellungen vertraut sein, um sie verwenden zu können. Standardmäßig sind die gängigsten Einstellungen bereits ausgewählt. Sie können die Dialogfelder für die erweiterten Einstellungen überspringen und gleich mit dem Aufbau oder der Erweiterung der Kategorien beginnen. Wenn Sie hier Änderungen vornehmen, müssen Sie diese nicht mit jedem Öffnen des Dialogfelds erneut vornehmen, da immer die neuesten Einstellungen beibehalten werden.

Beachten Sie jedoch, dass jedes Daten-Set anders ist und sich die Anzahl der Methoden und die Reihenfolge, in der sie angewendet werden, daher im Laufe der Zeit ändern kann. Da Ihre Text-Mining-Ziele von Daten-Set zu Daten-Set unterschiedlich sein können, müssen Sie möglicherweise mit den verschiedenen Verfahren experimentieren, um zu sehen, welches die besten Ergebnisse für die jeweiligen Textdaten bringt. Keines der automatischen Verfahren ergibt eine perfekte Kategorisierung Ihrer Daten, weshalb wir empfehlen, ein oder mehrere automatische Verfahren anzuwenden, die gut mit Ihren Daten funktionieren.

Die wichtigsten automatischen linguistischen Verfahren für den Aufbau von Kategorien:

- **Konzeptwurzelableitung.**Bei diesem Verfahren werden Kategorien erstellt, indem mit einem Konzept verwandte Konzepte durch die Analyse einzelner Konzeptkomponenten hinsichtlich ihrer morphologischen Verwandtschaft gefunden werden. Für weitere Informationen siehe Thema Konzeptwurzelableitung auf S. 193.Diese Option ist nicht für japanischen Text verfügbar.
- **Konzepteinbeziehung.** Dieses Verfahren erstellt Kategorien, indem es ausgehend von einem Konzept andere Konzepte ermittelt, die dieses Konzept enthalten. Für weitere Informationen siehe Thema Konzepteinbeziehung auf S. 194.

- **Semantisches Netz.** Bei diesem Verfahren wird zunächst auf der Grundlage eines umfassenden Index von Wortbeziehungen jedes Konzept auf seine möglichen Bedeutungen untersucht. Anschließend werden Kategorien durch Gruppieren zusammenhängender Konzepte erstellt. Für weitere Informationen siehe Thema Semantische Netze auf S. 195. Diese Option ist nur für englischen Text verfügbar.
- **Kookkurrenz.** Diese Technik erstellt Kookkurrenzregeln, die verwendet werden können, um eine neue Kategorie zu erstellen oder eine Kategorie zu erweitern, oder als Eingabe einer anderen Kategorietechnik verwendet werden. Für weitere Informationen siehe Thema Kookkurrenzregeln auf S. 197.

Konzeptwurzelableitung

Anmerkung: Diese Technik ist nicht für japanischen Text verfügbar.

Beim Konzeptwurzelableitungsverfahren werden Kategorien erstellt, indem mit einem Konzept verwandte Konzepte durch die Analyse einzelner Konzeptkomponenten hinsichtlich ihrer morphologischen Verwandtschaft gefunden werden. Eine Komponente ist ein Wort. Das Verfahren versucht, Konzepte durch Untersuchung der Endungen (Suffixe) der einzelnen Komponenten und durch Ermittlung anderer Konzepte, die daraus abgeleitet werden können, zusammenzufassen. Dahinter steht die Idee, dass Wörter, die voneinander abgeleitet werden, mit hoher Wahrscheinlichkeit auch dieselbe oder eine ähnliche Bedeutung aufweisen. Zur Ermittlung der Endungen werden interne sprachspezifische Regeln angewendet. So wird beispielsweise das Konzept `Möglichkeiten zum Aufstieg` mit den Konzepten`Möglichkeit des Aufstiegs` und `Aufstiegsmöglichkeit` zu einer Kategorie zusammengefasst.

Sie können die Konzeptwurzelableitung bei allen Textsorten einsetzen. Für sich genommen führt sie zu relativ wenigen Kategorien und die einzelnen Kategorien enthalten in der Regel nur wenige Konzepte. Die Konzepte in den einzelnen Kategorien sind entweder synonym oder situationsbezogen verwandt. Es kann nützlich sein, diesen Algorithmus zu verwenden, selbst bei einer manuellen Erstellung der Kategorien; die mithilfe dieses Algorithmus gefundenen Synonyme sind möglicherweise Synonyme der Konzepte, an denen Sie besonders interessiert sind.

Anmerkung: Sie können verhindern, dass Konzepte miteinander gruppiert werden, indem Sie sie explizit angeben. Für weitere Informationen siehe Thema Verwalten von Verknüpfungsausnahmepaaren auf S. 191.

Begriffskomponentenbildung und Bildung der Grundform

Bei Anwendung der Verfahren zur Konzeptwurzelableitung bzw. Konzepteinbeziehung werden die Ausdrücke zunächst in Komponenten (Wörter) gegliedert und anschließend wird die Grundform der Komponenten gebildet. Bei der Anwendung eines Verfahrens werden die Konzepte und die einem Konzept zugeordneten Fachausdrücke geladen und auf der Grundlage von Trennzeichen wie Leerzeichen, Bindestriche und Apostrophe in Komponenten aufgespalten. Der Ausdruck `Systemadministrator` beispielsweise wird wie folgt in Komponenten aufgespalten `{Administrator, System}`.

Einige Teile des ursprünglichen Ausdrucks werden jedoch möglicherweise nicht verwendet. Diese werden als "Stoppwörter" bezeichnet. Im Deutschen gehören beispielsweise folgende Wörter zu diesen ignorierbaren Komponenten: `ein`, `und`, `als`, `durch`, `für`, `von`, `in`, `auf`, `oder`, `der`, `bis` und `mit`.

So weist der Ausdruck `examination of the data` das Komponenten-Set `{data, examination}` auf; `von` und `der` werden als ignorierbar betrachtet. Die Reihenfolge der Komponenten in einem Komponenten-Set ist nicht von Bedeutung. Daher können folgende drei Ausdrücke äquivalent sein: `Hustenmedikament für Kinder`, `Kindermedikament gegen Husten` und `Medikament gegen Husten bei Kindern`, da alle dasselbe Komponenten-Set `{Kind, Husten, Medikament}` aufweisen. Die einzelnen Begriffspaare werden als äquivalent identifiziert. Die zugehörigen Konzepte werden zu einem neuen Konzept zusammengeführt, das alle drei Ausdrücke referenziert.

Da die Komponenten eines Terms außerdem flektiert sein können, werden intern sprachspezifische Regeln angewendet, um äquivalente Ausdrücke unabhängig von ihren Flexionsabweichungen zu ermitteln. Somit können die Ausdrücke `Waldschutz` und `Schutz der Wälder` als äquivalent identifiziert werden, da die Grundform im Singular `Wald` lautet.

Funktionsprinzipien der Konzeptwurzelableitung

Nachdem die Ausdrücke in Komponenten zerlegt und ihre Grundform gebildet wurde (siehe vorheriger Abschnitt), analysiert der Konzeptwurzelableitungsalgorithmus die Komponentenendungen (Suffixe), um den Stamm der Komponente zu ermitteln, und fasst anschließend die Konzepte mit anderen Konzepten, die denselben oder einen ähnlichen Stamm besitzen, zusammen. Die Endungen werden mithilfe einer Reihe linguistischer Ableitungsregeln für die jeweilige Textsprache identifiziert. So gibt es beispielsweise eine Ableitungsregel für englischsprachige Texte, die besagt, dass eine Konzeptkomponente, die auf das Suffix `ical` endet, von einem Konzept abgeleitet sein könnte, das denselben Stamm aufweist und auf das Suffix `ic` endet. Mit dieser Regel (und der Bildung der Grundform) kann der Algorithmus die Konzepte `epidemiologic study` und `epidemiological studies` zusammenfassen.

Da die Begriffe bereits in Komponenten zerlegt sind und die ignorierbaren Konzepte (z. B. `in` und `von`) identifiziert wurden, kann der Konzeptwurzelableitungsalgorithmus auch das Konzept `studies in epidemiology` mit `epidemiological studies` in diese Kategorie einordnen.

Der Satz an Komponentenableitungsregeln wurde so gewählt, dass die meisten durch diesen Algorithmus zusammengefassten Konzepte synonym sind: Die Konzepte `epidemiologic studies`, `epidemiological studies`, `studies in epidemiology` sind alle äquivalente Begriffe. Zur größeren Vollständigkeit gibt es auch Ableitungsregeln, die dem Algorithmus die Zusammenfassung von Konzepten gestatten, die situationsbezogen verwandt sind. So kann der Algorithmus beispielsweise Konzepte wie `Gründer des Reichs` und `Gründung des Reichs` zusammenfassen.

Konzepteinbeziehung

Beim Konzepteinbeziehungsverfahren werden Kategorien aufgebaut, indem mit Algorithmen für lexikalische Reihen Konzepte identifiziert werden, die in anderen Konzepten enthalten sind. Dahinter steht folgende Idee: Wenn Wörter in einem Konzept eine Untergruppe eines anderen

Konzepts bilden, ist dies Ausdruck einer zugrunde liegenden semantischen Beziehung. Die Einbeziehung ist ein leistungsstarkes Verfahren, das auf Texte aller Art angewendet werden kann.

Dieses Verfahren funktioniert am besten in Verbindung mit semantischen Netzen, kann aber auch getrennt verwendet werden. Die Konzepteinbeziehung kann bessere Ergebnisse erzielen, wenn die Dokumente bzw. Datensätze einen großen Anteil an domänenspezifischer Terminologie oder an Fachjargon enthalten. Dies gilt insbesondere dann, wenn die Wörterbücher zuvor abgestimmt wurden, sodass die speziellen Fachausdrücke extrahiert und entsprechend gruppiert werden (mit Synonymen).

Funktionsprinzipien der Konzepteinbeziehung

Vor der Anwendung des Konzepteinbeziehungsalgorithmus werden die Terme in Komponenten zerlegt und auf ihre Grundform zurückgeführt. Für weitere Informationen siehe Thema Konzeptwurzelableitung auf S. 193. Anschließend analysiert der Einbeziehungsalgorithmus die Komponenten-Sets. Bei jedem Komponenten-Set sucht der Algorithmus nach einem weiteren Komponenten-Set, bei dem es sich um eine Untergruppe des ersten Komponenten-Sets handelt.

Wenn Ihnen beispielsweise das Konzept `kontinentales Frühstück`, das das Komponenten-Set `{Frühstück, kontinental}` aufweist, und das Konzept `Frühstück`, das das Komponenten-Set `{Frühstück}` aufweist, vorliegen, folgert der Algorithmus, dass `kontinentales Frühstück` eine Art `Frühstück` ist, und fasst die beiden Konzepte zusammen.

Oder ein umfangreicheres Beispiel: Wenn Ihnen im Fensterbereich “Extrahierungsergebnisse” das Konzept `Sitz` vorliegt und Sie diesen Algorithmus anwenden, werden Konzepte wie `Kindersitz`, `Ledersitz`, `Sitzheizung`, `Elektro-Sitzheizung`, `Kindersitzgurt` und `Kindersitzvorschriften` ebenfalls in diese Kategorie eingeordnet.

Da die Begriffe bereits in Komponenten zerlegt sind und die ignorierbaren Konzepte (z. B. `in` und `von`) identifiziert wurden, würde der Konzepteinbeziehungsalgorithmus erkennen, dass das Konzept `Fortgeschrittenenkurs in Spanisch` das Konzept `Spanischkurs` beinhaltet.

Anmerkung: Sie können verhindern, dass Konzepte miteinander gruppiert werden, indem Sie sie explizit angeben. Für weitere Informationen siehe Thema Verwalten von Verknüpfungsausnahmepaaren auf S. 191.

Semantische Netze

In dieser Version ist das Verfahren mit semantischen Netzen nur für englischsprachige Texte verfügbar.

Bei diesem Verfahren werden Kategorien mithilfe eines integrierten Netzes von Wortbeziehungen erstellt. Aus diesem Grund können mit diesem Verfahren sehr gute Ergebnisse erzielt werden, wenn die Ausdrücke konkret sind und nur einen geringen Grad an Mehrdeutigkeit aufweisen. Es ist jedoch nicht zu erwarten, dass dieses Verfahren viele Zusammenhänge zwischen sehr technischen/spezialisierten Konzepten findet. Beim Umgang mit solchen Konzepten sind das Konzepteinbeziehungs- und das Konzeptwurzelableitungsverfahren zumeist von größerem Nutzen.

Funktionsprinzipien semantischer Netze

Hinter dem Verfahren mit semantischen Netzen steht die Idee, bekannte Wortbeziehungen zu nutzen, um Kategorien von Synonymen bzw. Hyponymen zu erzeugen. Ein **Hyponym** liegt vor, wenn ein Konzept eine Sorte eines zweiten Konzepts ist, dergestalt, dass eine hierarchische Beziehung (auch als ISA-Beziehung bezeichnet) vorliegt. Beispiel: Wenn `animal` ein Konzept ist, dann sind `cat` und `kangaroo` Hyponyme von `animal`, da es sich dabei jeweils um eine Art Tier (animal) handelt.

Neben Synonym- und Hyponymbeziehungen untersucht das semantische Netz auch Teilzusammenhänge und vollständige Zusammenhänge zwischen Konzepten aus dem Typ `<Location>` (Ort). Beispielsweise ordnet das Verfahren die Konzepte `normandy`, `provence` und `france` in dieselbe Kategorie ein, da Normandie (Normandy) und Provence Teile von Frankreich (France) sind.

Bei dem Verfahren mit semantischen Netzen werden zunächst die möglichen Bedeutungen der einzelnen Konzepte im semantischen Netz ermittelt. Wenn Konzepte als Synonyme oder Hyponyme identifiziert werden, werden sie alle in dieselbe Kategorie eingeordnet. Beispielsweise erstellt dieses Verfahren eine einzelne Kategorie, die die folgenden drei Konzepte enthält: `Speiseapfel`, `Dessertapfel` und `Granny Smith`, da das semantische Netz folgende Informationen enthält: 1) `Dessertapfel` ist ein Synonym von `Speiseapfel` und 2) `Granny Smith` ist eine Sorte von `Speiseapfel` (d. h. ist ein Hyponym von `Speiseapfel`).

Isoliert betrachtet sind viele Konzepte, insbesondere Uniterms, mehrdeutig. Das Konzept `buffet` beispielsweise kann eine Art Mahlzeit oder ein Möbelstück bezeichnen. Wenn die Menge der Konzepte `Mahlzeit`, `Möbel` und `Schrank` beinhaltet, ist der Algorithmus gezwungen zu entscheiden, ob `Schrank` in dieselbe Kategorie eingeordnet werden soll wie `Mahlzeit` oder `Möbel`. Beachten Sie, dass es vorkommen kann, dass die vom Algorithmus getroffene Wahl im Kontext eines bestimmten Sets an Datensätzen oder Dokumenten nicht angemessen ist.

Das Verfahren mit semantischen Netzen führt bei bestimmten Arten von Daten zu besseren Ergebnissen als die Konzepteinbeziehung. Sowohl das semantische Netz als auch die Konzepteinbeziehung erkennen, dass `Apfelkuchen` eine Art von `Kuchen` ist, aber nur das semantische Netz erkennt, dass `Tarte` auch eine Art von `Kuchen` ist.

Semantische Netze können auch zusammen mit anderen Verfahren eingesetzt werden. Beispielsweise angenommen, Sie haben sowohl das Verfahren mit semantischen Netzen als auch das Einbeziehungsverfahren ausgewählt und das semantische Netz hat das Konzept `teacher` in dieselbe Kategorie eingeordnet wie das Konzept `tutor` (da ein Tutor eine Art von Lehrer (teacher) ist. Der Einbeziehungsalgorithmus kann das Konzept `graduate tutor` mit `tutor` zusammenfassen und als Ergebnis erstellen die beiden Algorithmen zusammen eine Ausgabekategorie, die alle drei Konzepte enthält: `tutor`, `graduate tutor` und `teacher`.

Optionen für semantische Netze

Es sind einige zusätzliche Einstellungen vorhanden, die für dieses Verfahren interessant sein könnten.

- Ändern der Einstellung Maximaler Suchabstand. Legen Sie fest, wie weit die Verfahren suchen sollen, bevor Kategorien erstellt werden. Je niedriger der Wert, desto weniger Ergebnisse erhalten Sie. Allerdings sind die Ergebnisse weniger verrauscht und mit größerer Wahrscheinlichkeit auf signifikante Weise miteinander verknüpft oder verbunden. Je

höher der Wert, desto mehr Ergebnisse erhalten Sie. Allerdings sind diese Ergebnisse möglicherweise weniger zuverlässig oder relevant.

Abhängig vom Abstand sucht der Algorithmus zum Beispiel von `Dänischem Gebäck` bis zu `Blätterteiggebäck` (übergeordnet), dann `Teiggebäck` (über-übergeordnet) und aufwärts bis `Gebäck`.

Durch Verringern des Suchabstands führt dieses Verfahren zu kleineren Kategorien, mit denen sich möglicherweise leichter arbeiten lässt, wenn Sie den Eindruck haben, dass die erstellten Kategorien zu groß sind oder zu viele Elemente zu einer Gruppe zusammenfassen.

Wichtig: Außerdem wird empfohlen, bei Verwendung dieses Verfahrens die Option Rechtschreibung korrigieren für eine minimale Anzahl an Stammzeichen von (definiert auf der Registerkarte "Experten" des Knotens oder im Dialogfeld "Extrahieren") für Fuzzy-Gruppierung nicht anzuwenden, da Fehlgruppierungen große negative Auswirkungen auf die Ergebnisse haben können.

Kookkurrenzregeln

Mit Kookkurrenzregeln können Sie Konzepte ermitteln und zusammenfassen, die im Set an Dokumenten bzw. Datensätzen eng miteinander verwandt sind. Dahinter steht folgende Idee: Wenn Konzepte häufig in Dokumenten und Datensätzen gefunden werden, ist diese Kookkurrenz Ausdruck einer zugrunde liegenden Beziehung, die wahrscheinlich in Ihren Kategoriedefinitionen von Nutzen ist. Diese Technik erstellt Kookkurrenzregeln, die verwendet werden können, um eine neue Kategorie zu erstellen oder eine Kategorie zu erweitern, oder als Eingabe einer anderen Kategorietechnik verwendet werden. Zwei Konzepte weisen eine starke Kookkurrenz auf, wenn sie häufig zusammen in einer Menge von Datensätzen vorkommen und sehr selten einzeln in den anderen Datensätzen. Dieses Verfahren kann bei größeren Daten-Sets mit mindestens mehreren Hundert Dokumenten bzw. Datensätzen zu guten Ergebnissen führen.

Wenn beispielsweise viele Datensätze die Wörter `Preis` und `Verfügbarkeit` enthalten, könnten diese Konzepte in eine Kookkurrenzregel zusammengefasst werden `(Preis & verfügbar)` Wenn die Konzepte `Erdnussbutter`, `Gelee` und `Sandwich` häufiger zusammen als getrennt vorkommen, werden sie gemeinsam in eine Konzeptkookkurrenzregel `(Erdnussbutter&Gelee & Sandwich)` aufgenommen.

Wichtig: In früheren Versionen wurden Kookkurenz- und Synonymregeln von eckigen Klammern umgeben. In dieser neuen Version zeigen nun eckige Klammern ein Musterergebnis für Text-Link-Analyse an. Stattdessen stehen Kookkurrenz- und Synonymregeln in runden Klammern, z. B. `(Lautsprechersysteme|Lautsprecher)`.

Funktionsprinzipien von Kookkurrenzregeln

Bei diesem Verfahren werden die Dokumente bzw. Datensätze durchsucht, um zwei oder mehr Konzepte zu finden, die häufig gemeinsam vorkommen. Zwei oder mehr Konzepte weisen eine starke Kookkurrenz auf, wenn sie häufig zusammen in einer Menge von Dokumenten bzw. Datensätzen vorkommen und selten einzeln in den anderen Dokumenten und Datensätzen.

Wenn kookkurrierende Konzepte gefunden werden, wird eine Kategorieregel erstellt. Diese Regeln bestehen aus zwei oder mehr Konzepten, die über den Boole'schen Operator & verbunden sind. Bei diesen Regeln handelt es sich um logische Aussagen, die ein -Dokument oder einen

Datensatz automatisch in eine Kategorie einordnen, wenn die in der Regel enthaltenen Konzepte alle in dem betreffenden -Dokument bzw. Datensatz kookkurrieren.

Optionen für Kookkurrenzregeln

Wenn Sie das Kookkurrenzregelverfahren verwenden, können Sie mehrere Einstellungen optimieren, die Einfluss auf die resultierenden Regeln haben:

- Ändern der Einstellung Maximaler Suchabstand. Legen Sie fest, wie weit das Verfahren nach Kookkurrenzen suchen soll. Je größer der Suchabstand desto geringer ist der Mindestwert für die Ähnlichkeit, der für jede Kookkurrenz erforderlich ist. Infolgedessen werden möglicherweise sehr viele Kookkurrenzregeln erstellt, diejenigen mit einem niedrigen Ähnlichkeitswert sind jedoch häufig nur von geringer Bedeutung. Wenn Sie den Suchabstand verringern, erhöht sich der Mindestwert für die Ähnlichkeit; infolgedessen werden weniger Kookkurrenzregeln erstellt, die jedoch tendenziell signifikanter (stärker) sind.
- Minimale Anzahl anDokumenten. Die Mindestanzahl an Datensätzen oder Dokumenten , die ein bestimmtes Konzeptpaar enthalten muss, um als Kookkurrenz zu gelten. Je niedriger Sie diese Option setzen, desto einfacher ist es, Kookkurrenzen zu finden. Die Erhöhung des Werts führt zu weniger, jedoch signifikanteren Kookkurrenzen. Nehmen Sie beispielsweise an, dass die Konzepte “Apfel” und “Birne” gemeinsam in 2 Datensätzen gefunden werden (und keines der beiden Konzepte in irgendeinem anderen Datensatz vorkommt). Wenn Minimale Anzahl an Dokumenten. auf 2 (Standardwert) gesetzt ist, erstellt das Kookkurrenzverfahren eine Kategorieregel (Apfel und Birne). Wenn der Wert auf 3 erhöht wird, wird die Regel nicht mehr erstellt.

Anmerkung: Bei kleinen Daten-Sets (< 1000 Antworten) finden Sie möglicherweise mit den Standardeinstellungen keinerlei Kookkurrenzen. Versuchen Sie in diesem Fall den Wert für den Suchabstand zu erhöhen.

Anmerkung: Sie können verhindern, dass Konzepte miteinander gruppiert werden, indem Sie sie explizit angeben. Für weitere Informationen siehe Thema Verwalten von Verknüpfungsausnahmepaaren auf S. 191.

Erweiterte Einstellungen für Häufigkeit

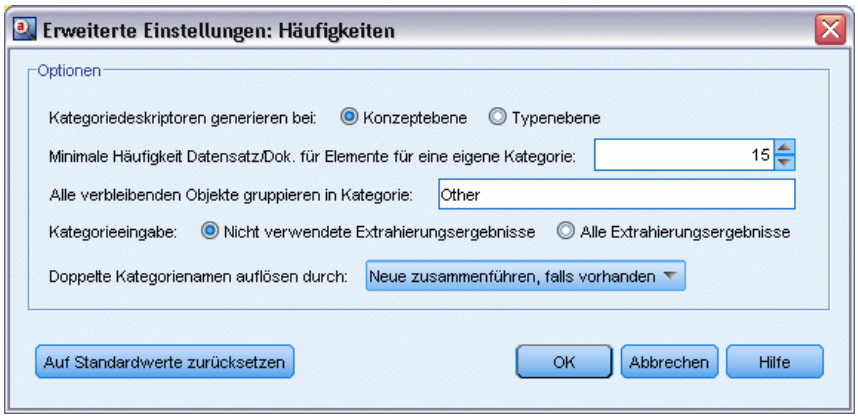
Sie können Kategorien basierend auf einem direkten und mechanischen Häufigkeitsverfahren erstellen. Mit diesem Verfahren können Sie eine Kategorie für jedes Element (Typ, Konzept oder Muster) aufbauen, das über einem gegebenen Datensatz oder Dokument gefunden wurde. Zusätzlich können Sie eine einzelne Kategorie für jedes der weniger häufig auftretenden Elemente erstellen. Häufigkeit bezieht sich auf die Anzahl an Datensätzen oder Dokumenten, die das extrahierte Konzept (und seine Synonyme), den Typ oder das Muster enthalten. Es geht nicht um die Gesamtanzahl an Vorkommnissen im gesamten Text.

Das Gruppieren häufig auftretender Objekte kann interessante Ergebnisse bringen, da dies eine übliche oder signifikante Antwort angeben kann. Das Verfahren ist sehr nützlich auf die unbenutzten Extrahierungsergebnisse anzuwenden, nachdem andere Verfahren angewendet wurden. Eine andere Anwendung besteht in der direkten Ausführung dieses Verfahrens nach der Extrahierung, wenn keine anderen Kategorien vorhanden sind, der Bearbeitung der Ergebnisse, um nicht interessante Kategorien zu löschen und dann der Erweiterung dieser Kategorien, damit

sie noch mehr Datensätze oder Dokumente abdecken. Für weitere Informationen siehe Thema Erweitern von Kategorien auf S. 200.

Anstelle dieser Technik könnten Sie die Konzepte und Konzeptmuster nach absteigenden Nummern von Datensätzen oder Dokumenten in den Extrahierungsergebnissen sortieren und dann die obersten in den Kategorienbereich ziehen, um die entsprechenden Kategorien zu erstellen.

Abbildung 10-12
Erweiterte Einstellungen: Dialogfeld "Häufigkeiten"



Kategoriedeskriptoren generieren bei. Wählen Sie die Eingabeart für Deskriptoren. Für weitere Informationen siehe Thema Erstellen von Kategorien auf S. 183.

- Konzeptebene. Wenn Sie diese Option auswählen, werden Konzept- oder Konzeptmusterhäufigkeiten verwendet. Konzepte werden verwendet, wenn Typen als Eingabe für den Kategorieaufbau gewählt wurden. Konzeptmuster werden verwendet, wenn Typmuster ausgewählt wurden. Im Allgemeinen ergibt dieses Verfahren für die Konzeptebene spezifischere Ergebnisse, da Konzepte und Konzeptmuster eine geringere Messungsebene repräsentieren.
- Typenebene. Wenn Sie diese Option auswählen, werden Typen- oder Typenmusterhäufigkeiten verwendet. Typen werden verwendet, wenn Typen als Eingabe für den Kategorieaufbau gewählt wurden. Typmuster werden verwendet, wenn Typmuster ausgewählt wurden. Wenn Sie dieses Verfahren auf die Typenebene anwenden, erhalten Sie einen raschen Einblick in die Art der vorhandenen Informationen.

Minimale Häufigkeit Dok. für Elemente für eine eigene Kategorie. Mithilfe dieser Option können Sie Kategorien aus häufig auftretenden Elementen aufbauen. Diese Option beschränkt die Ausgabe auf nur die Kategorien mit einem Deskriptor, die in mindestens der Anzahl X von Datensätzen oder Dokumenten aufgetreten sind, wobei X den Wert angibt, der für diese Option eingegeben werden soll.

Alle verbleibenden Objekte gruppieren in Kategorie. Mithilfe dieser Option können Sie alle Konzepte oder Typen, die selten auftreten, in einer einzigen Kategorie für alle mit einem Namen Ihrer Wahl zusammenfassen. Standardmäßig heißt diese Kategorie *Andere*.

Kategorieeingabe. Wählen Sie die Gruppe aus, auf die Sie die Verfahren anwenden möchten:

- Nicht verwendete Extrahierungsergebnisse. Diese Option aktiviert Kategorien, die aus Extrahierungsergebnissen erstellt werden, die nicht in vorhandenen Kategorien verwendet werden. So wird die Tendenz für Datensätze minimiert, mehrere Kategorien abzugleichen und die Anzahl der erzeugten Kategorien zu begrenzen.
- Alle Extrahierungsergebnisse. Diese Option aktiviert unter Verwendung der Extrahierungsergebnisse zu erstellende Kategorien. Dies ist am sinnvollsten, wenn noch keine oder nur sehr wenige Kategorien vorhanden sind.

Konflikte mit doppelten Namen lösen. Wählen Sie, wie mit neuen Kategorien bzw. Unterkategorien verfahren werden soll, deren Namen mit denen von bestehenden Kategorien identisch wären. Sie können entweder die neuen Kategorien (und ihre Deskriptoren) mit den bestehenden Kategorien desselben Namens zusammenführen. Alternativ können Sie die Erstellung jeglicher Kategorien überspringen, wenn in den bestehenden Kategorien ein Namensduplikat gefunden wird.

Erweitern von Kategorien

Das Erweitern ist ein Prozess, durch den Deskriptoren automatisch hinzugefügt oder verbessert werden, um vorhandene Kategorien auszubauen. Ziel ist es, eine bessere Kategorie zu erzeugen, die verwandte Datensätze oder Dokumente erfasst, die der Kategorie ursprünglich nicht zugeordnet waren.

Die automatischen Gruppierungsverfahren, die Sie auswählen, versuchen, Konzepte, TLA-Muster und Kategorieregeln zu den vorhandenen Kategoriedeskriptoren zu identifizieren. Diese neuen Konzepte, Muster und Kategorieregeln werden dann als neue Deskriptoren hinzugefügt oder vorhandenen Deskriptoren hinzugefügt. Zu den Erweiterungsgruppierungsverfahren gehören *Konzeptwurzelableitung*(für Japanisch nicht verfügbar), *Konzepteinbeziehung*, *semantische Netze* (nur Englisch) sowie *Kookkurrenzregeln*. Die Methode Leere Kategorien mit aus dem Kategorienamen erzeugten Deskriptoren erweitern erzeugt Deskriptoren mit den Wörtern in den Kategorienamen. Je beschreibender ein Kategoriename, um so besser daher die Ergebnisse.

Anmerkung: Die häufigkeitsbasierten Verfahren stehen beim Erweitern von Kategorien nicht zur Verfügung.

Erweitern ist ideal, um Ihre Kategorien interaktiv zu verbessern. Hier einige Beispiele für das Erweitern einer Kategorie:

- Nach dem Ziehen von Konzeptmustern, um Kategorien im Bereich “Kategorien” zu erstellen
- Nach dem Erstellen von Kategorien per Hand und dem Hinzufügen einfacher Kategorieregeln und Deskriptoren
- Nach dem Import einer vordefinierten Kategoriedatei, bei dem die Kategorien sehr aussagekräftige Namen hatten
- Nach dem Verfeinern der Kategorien aus dem TAP, das Sie gewählt hatten

Sie können eine Kategorie mehrfach erweitern. Wenn Sie zum Beispiel eine vordefinierte Kategoriedatei mit sehr beschreibenden Namen importiert haben, könnten Sie mit der Option Leere Kategorien mit aus dem Kategorienamen erzeugten Deskriptoren erweitern erweitern, um eine erste Menge an Deskriptoren zu erhalten, und dann diese Kategorien erneut erweitern.

In anderen Fällen könnte das mehrfache Erweitern jedoch zu zu generischen Kategorien führen, wenn die Deskriptoren mehr und mehr erweitert werden. Da den Aufbau- und Erweiterungsgruppierverfahren ähnliche Algorithmen zugrunde liegen, führt das Erweitern direkt nach dem Aufbau von Kategorien mit nur geringer Wahrscheinlichkeit zu interessanteren Ergebnissen.

Tipps:

- Wenn Sie versuchen zu erweitern und die Ergebnisse nicht verwenden möchten, können Sie den Vorgang stets direkt nach dem Erweitern widerrufen (Bearbeiten > Rückgängig).
- Das Erweitern kann zwei oder mehr Kategorieregeln in einer Kategorie erzeugen, die mit genau dem gleichen Satz an Dokumenten übereinstimmen, da Regeln während des Prozesses unabhängig erzeugt werden. Wenn gewünscht, können Sie die Kategorien prüfen und Redundanzen durch manuelle Bearbeitung der Kategoriebeschreibung entfernen. Für weitere Informationen siehe Thema Bearbeiten von Kategoriedeskriptoren auf S. 239.

So erweitern Sie Kategorien

▶ Wählen Sie im Fensterbereich “Kategorien” die Kategorien aus, die Sie erweitern möchten.

▶ Wählen Sie in den Menüs die Optionsfolge Kategorien > Kategorien erweitern. Wenn Sie nicht die Option markiert haben, dass diese Aufforderung nie angezeigt werden soll, wird ein Dialogfeld geöffnet.

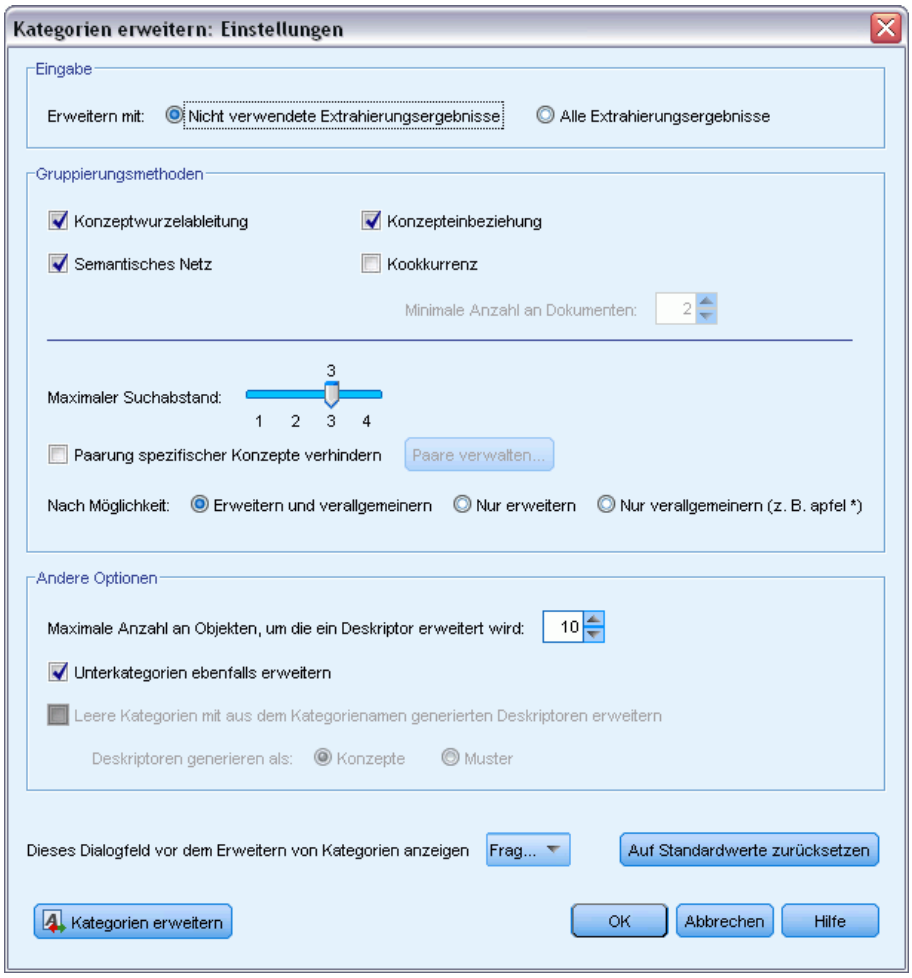
▶ Wählen Sie aus, ob Sie die Kategorie jetzt aufbauen oder zuerst die Einstellungen bearbeiten möchten.

- Klicken Sie auf Jetzt erweitern, um mit den aktuellen Einstellungen mit der Erweiterung von Kategorien zu beginnen. Der Prozess wird gestartet und ein Dialogfeld über den Fortschritt angezeigt.
- Klicken Sie auf Bearbeiten, um die Einstellungen zu überprüfen und zu ändern.

Nach dem Versuch der Erweiterung werden alle Kategorien, für die neue Deskriptoren gefunden werden, mit dem Wort Erweitert im Bereich “Kategorien” gekennzeichnet, so dass Sie sie schnell erkennen. Der Text “Erweitert” bleibt, bis Sie erneut erweitern oder die Kategorie anderweitig bearbeiten oder über das Kontextmenü löschen.

Anmerkung: Es können maximal 10.000 Kategorien angezeigt werden. Wenn diese Zahl erreicht oder überschritten wird, wird eine Warnung angezeigt. Wenn dies geschieht, sollten Sie die Option “Kategorien aufbauen” bzw. “Kategorien erweitern” ändern, um die Anazahl der erstellten Kategorien zu verringern.

Abbildung 10-13
Dialogfeld "Kategorien erweitern"



Diese Verfahren, die für den Aufbau und die Erweiterung von Kategorien verfügbar sind, eignen sich jeweils gut für bestimmte Arten von Daten und Situationen, häufig ist es jedoch sinnvoll, in einer Analyse mehrere Verfahren zu kombinieren, um das gesamte Spektrum an Dokumenten bzw. Datensätzen zu erfassen. Es gibt in der interaktiven WorkbenchKonzepte und Typen, die unter einer Kategorie gruppiert wurden und bei der nächsten Erstellung von Kategorien ebenfalls zur Verfügung stehen. Das bedeutet, dass Sie ein Konzept in mehreren Kategorien erkennen oder redundante Kategorien vorfinden können.

Erweitern mit. Auswahl, welche Eingabe zur Erweiterung der Kategorien verwendet wird:

- Nicht verwendete Extrahierungsergebnisse. Diese Option aktiviert Kategorien, die aus Extrahierungsergebnissen erstellt werden, die nicht in vorhandenen Kategorien verwendet werden. So wird die Tendenz für Datensätze minimiert, mehrere Kategorien abzugleichen und die Anzahl der erzeugten Kategorien zu begrenzen.
- Alle Extrahierungsergebnisse. Diese Option aktiviert unter Verwendung der Extrahierungsergebnisse zu erstellende Kategorien. Dies ist am sinnvollsten, wenn noch keine oder nur sehr wenige Kategorien vorhanden sind.

Gruppierverfahren

Eine kurze Beschreibung dieser Verfahren finden Sie unter Erweiterte linguistische Einstellungen auf S. 187. Zu diesen Techniken zählen:

- Konzeptwurzelableitung*(nicht für Japanisch verfügbar)*
- Semantisches Netz (*Nur für englischen Text, und nicht verwendet, wenn die Option "Nur verallgemeinern" ausgewählt ist.*)
- Konzepteinbeziehung
- Kookkurrenz und Unteroption Mindestanzahl an Dokumenten.

Eine Reihe von Typen wird dauerhaft aus dem Verfahren mit semantischen Netzen ausgeschlossen, da diese Typen nicht zu relevanten Ergebnissen führen. Sie umfassen `<Positive>`, `<Negative>`, `<IP>`, andere nichtlinguistische Typen usw.

Maximaler Suchabstand. Legen Sie fest, wie weit die Verfahren suchen sollen, bevor Kategorien erstellt werden. Je niedriger der Wert, desto weniger Ergebnisse erhalten Sie. Allerdings sind die Ergebnisse weniger verrauscht und mit größerer Wahrscheinlichkeit auf signifikante Weise miteinander verknüpft oder verbunden. Je höher der Wert, desto mehr Ergebnisse erhalten Sie. Allerdings sind diese Ergebnisse möglicherweise weniger zuverlässig oder relevant. Während diese Option global auf alle Verfahren angewendet wird, hat sie die größte Auswirkung auf Kookkurrenzen und semantische Netze.

Paarung spezifischer Konzepte verhindern. Markieren Sie dieses Kontrollkästchen, um den Vorgang der Gruppierung oder Paarung von zwei Konzepten in der Ausgabe zu verhindern. Klicken Sie zum Erstellen oder Verwalten von Konzeptpaaren auf Paare verwalten.... Für weitere Informationen siehe Thema Verwalten von Verknüpfungsausnahmepaaren auf S. 191.

Wenn möglich: Wählen Sie, ob Sie die Deskriptoren einfach erweitern, mithilfe von Platzhaltern verallgemeinern oder beides tun möchten.

- **Erweitern und verallgemeinern.** Mit dieser Option werden die ausgewählten Kategorien erweitert und dann die Deskriptoren verallgemeinert. Wenn Sie verallgemeinern möchten, erstellt das Produkt allgemeine Kategorieregeln in Kategorien mithilfe eines Sternchens als Platzhalter. Beispielsweise könnte anstelle der Erzeugung mehrerer Deskriptoren wie `[Apfel vom Bioladen + .]` und `[Apfelsauce + .]` der Einsatz von Platzhaltern `[Apfel * + .]` erzeugen. Wenn Sie mit Platzhaltern verallgemeinern, erhalten Sie oft genau die gleiche Anzahl an Datensätzen oder Dokumenten wie zuvor. Diese Option hat jedoch den Vorteil, die Zahl zu verringern und die Kategoriedeskriptoren zu vereinfachen. Zusätzlich erhöht diese Option die Möglichkeit, mehr Datensätze oder Dokumente unter Verwendung dieser Kategorien zu neuen Textdaten (zum Beispiel bei Langzeit-/Wellenstudien) zu kategorisieren.
- **Nur erweitern**. Mit dieser Option werden Ihre Kategorien erweitert, ohne die Deskriptoren zu verallgemeinern. Es kann hilfreich sein, für manuell erstellte Kategorien zunächst die Option Nur erweitern zu wählen und dann die gleichen Kategorien mit der Option Erweitern und verallgemeinern noch einmal zu erweitern.
- **Nur verallgemeinern.** Mit dieser Option werden die Deskriptoren verallgemeinert, ohne Ihre Kategorien auf andere Weise zu erweitern.

Anmerkung: Bei Auswahl dieser Option wird die Option Semantisches Netz deaktiviert: Der Grund hierfür ist, dass die Option Semantisches Netz nur verfügbar ist, wenn eine Beschreibung erweitert werden soll.

Weitere Optionen für die Erweiterung von Kategorien

Neben der Auswahl der anzuwendenden Verfahren können Sie folgende weitere Optionen bearbeiten:

Maximale Anzahl an Objekten, um die ein Deskriptor erweitert wird. Definieren Sie bei der Erweiterung eines Deskriptors um Objekte (Konzepte, Typen und andere Ausdrücke) die maximale Anzahl an Objekten, die einem einzelnen Deskriptor hinzugefügt werden können. Wenn Sie als Limit 10 wählen, können einem vorhandenen Deskriptor höchstens 10 zusätzliche Objekte hinzugefügt werden. Wenn mehr als 10 Objekte hinzugefügt werden sollen, beendet das Verfahren das Hinzufügen neuer Objekte nach dem zehnten Objekt. Dies kann eine Deskriptorliste verkürzen, garantiert aber nicht, dass die interessantesten Objekte zuerst verwendet wurden. Eventuell möchten Sie die Größe der Erweiterung ohne Abstriche an der Qualität verringern, indem Sie die Option Mit Platzhaltern, wenn möglich, verallgemeinern. Diese Option gilt nur für Deskriptoren, die die Boole'schen Operatoren & (UND) bzw. ! (NICHT) enthalten.

Unterkategorien ebenfalls erweitern. Mit dieser Option werden auch alle Unterkategorien unter den ausgewählten Kategorien erweitert.

Leere Kategorien mit aus dem Kategorienamen erzeugten Deskriptoren erweitern. Diese Methode wird nur auf leere Kategorien mit null Deskriptoren angewendet. Wenn eine Kategorie bereits Deskriptoren enthält, kann Sie nicht auf diese Art erweitert werden. Diese Option versucht, Deskriptoren für jede importierte Kategorie, basierend auf den Wörtern, die den Namen der Kategorie ausmachen, automatisch zu erstellen. Der Kategoriename wird gescannt, um festzustellen, ob Wörter im Namen einem extrahierten Konzept entsprechen. Wenn ein Konzept erkannt wird, wird es verwendet, um nach passenden Konzeptmustern zu suchen. Diese werden dann beide herangezogen, um Deskriptoren für die Kategorie zu bilden. Diese Option erzeugt die besten Ergebnisse, wenn die Kategorienamen lang und beschreibend sind. Dies ist eine schnelle Methode, um Kategoriedeskriptoren zu erzeugen, die es wiederum der Kategorie ermöglichen, Datensätze zu erfassen, die diese Deskriptoren enthalten. Diese Option ist vor allem dann hilfreich, wenn Sie Kategorien importieren oder Kategorien mit langen beschreibenden Namen manuell erstellen.

Deskriptoren erzeugen als. Diese Option ist nur verfügbar, wenn die vorherige Option ausgewählt wurde.

- **Konzepte.** Wählen Sie diese Option, um die resultierenden Deskriptoren in der Form von Konzepten zu erzeugen, ungeachtet dessen, ob sie aus dem Quelltext extrahiert wurden.
- **Muster.** Wählen Sie diese Option, um die resultierenden Deskriptoren in der Form von Mustern zu erzeugen, ungeachtet dessen, ob die resultierenden Muster oder ein beliebiges Muster extrahiert wurde.

Manuelle Erstellung von Kategorien

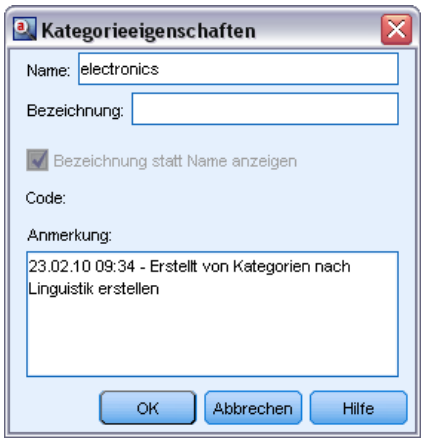
Neben dem Erstellen von Kategorien mithilfe der automatisierten Verfahren und des Regeleditors können Sie Kategorien auch manuell definieren. Es gibt folgende manuelle Methode:

- Erstellen einer leeren Kategorie, der Sie nacheinander Elemente hinzufügen. Für weitere Informationen siehe Thema Erstellen neuer Kategorien bzw. Umbenennen von Kategorien auf S. 205.
- Ziehen von Begriffen, Typen und Mustern in den Kategorienbereich. Für weitere Informationen siehe Thema Erstellen von Kategorien durch Ziehen und Ablegen (Drag&Drop) auf S. 206.

Erstellen neuer Kategorien bzw. Umbenennen von Kategorien

Sie können leere Kategorien erstellen, um Konzepte und Typen hinzuzufügen. Außerdem können Sie die Kategorien umbenennen.

Abbildung 10-14
Dialogfeld "Kategorieeigenschaften"



So erstellen Sie eine neue, leere Kategorie:

▶ Rufen Sie den Fensterbereich "Kategorien" auf.

▶ Wählen Sie in den Menüs die Optionsfolge Kategorien > Leere Kategorie erstellen. Das Dialogfeld wird geöffnet.

▶ Geben Sie im Namensfeld einen Namen für diese Kategorie ein.

▶ Klicken Sie auf OK, um den Namen zu übernehmen und das Dialogfeld zu schließen. Das Dialogfeld wird geschlossen und im Fensterbereich wird ein neuer Kategoriename angezeigt.

Sie können nun weitere Elemente in diese Kategorie aufnehmen. Für weitere Informationen siehe Thema Hinzufügen von Deskriptoren zu Kategorien auf S. 238.

So benennen Sie eine Kategorie um:

▶ Wählen Sie eine Kategorie aus und wählen Sie die Optionsfolge Kategorien > Kategorie umbenennen. Das Dialogfeld wird geöffnet.

▶ Geben Sie im Namensfeld einen neuen Namen für diese Kategorie ein.

▶ Klicken Sie auf OK, um den Namen zu übernehmen und das Dialogfeld zu schließen. Das Dialogfeld wird geschlossen und im Fensterbereich wird ein neuer Kategoriename angezeigt.

Erstellen von Kategorien durch Ziehen und Ablegen (Drag&Drop)

Das Drag&Drop-Verfahren erfolgt manuell und beruht nicht auf Algorithmen. Sie können Kategorien im Bereich “Kategorien” erstellen, indem Sie folgende Elemente ziehen:

- Extrahierte Konzepte, Typen oder Muster aus dem Bereich “Extrahierungsergebnisse” in den Bereich “Kategorien”.
- Extrahierte Konzepte aus dem Bereich “Daten” in den Bereich “Kategorien”.
- Ganze Zeilen aus dem Bereich “Daten” in den Bereich “Kategorien”. Damit wird eine Kategorie erstellt, die aus allen extrahierten Konzepten und Mustern in dieser Zeile besteht.

Anmerkung: Der Bereich “Extrahierungsergebnisse” unterstützt die Mehrfachauswahl, um das Ziehen und Ablegen mehrerer Elemente zu vereinfachen.

Wichtig: Sie können keine Konzepte aus dem Bereich “Daten” ziehen, wenn diese nicht aus dem Text extrahiert wurden. Wenn Sie die Extrahierung eines Konzepts erzwingen möchten, das Sie in Ihren Daten gefunden haben, müssen Sie das Konzept einem Typ hinzufügen. Führen Sie dann die Extrahierung erneut aus. Die neuen Extrahierungsergebnisse enthalten das soeben hinzugefügte Konzept. Sie können es dann in Ihrer Kategorie verwenden. Für weitere Informationen siehe Thema Konzepte zu Typen hinzufügen in Kapitel 9 auf S. 162.

So erstellen Sie Kategorien durch Ziehen und Ablegen:

▶ Wählen Sie im Bereich “Extrahierungsergebnisse” oder im Datenbereich eines oder mehrere Konzepte, Muster, Typen, Datensätze oder Datensatzteile aus.

▶ Ziehen Sie das Element bei gedrückter Maustaste in eine bestehende Kategorie oder in den Fensterbereich, um eine neue Kategorie anzulegen.

▶ Wenn Sie den Bereich erreicht haben, in dem Sie das Element ablegen möchten, lassen Sie die Maustaste los. Das Element wird dem Bereich “Kategorien” hinzugefügt. Die geänderten Kategorien werden mit einer besonderen Hintergrundfarbe gekennzeichnet. Diese Farbe ist der Hintergrund für Kategoriefeedback. Für weitere Informationen siehe Thema Festlegen von Optionen in Kapitel 8 auf S. 135.

Anmerkung: Die resultierende Kategorie wurde automatisch benannt. Wenn Sie einen Namen ändern möchten, können Sie die Kategorie umbenennen.

Um anzuzeigen, welche Datensätze einer Kategorie zugewiesen wurden, wählen Sie die betreffende Kategorie im Fensterbereich “Kategorien” aus. Der Datenbereich wird automatisch aktualisiert und zeigt alle Datensätze für diese Kategorie an.

Verwenden von Kategorieregeln

Es gibt verschiedene Methoden zum Erstellen von Kategorien. Eine dieser Methoden ist die Definition von Kategorieregeln, die Ideen ausdrücken. Kategorieregeln sind Anweisungen, mit denen Dokumente oder Datensätze auf Basis eines logischen Ausdrucks mithilfe von extrahierten Konzepten, Typen und Mustern sowie von Boole’schen Operatoren automatisch einer Kategorie zugewiesen werden. Sie könnten beispielsweise einen Ausdruck schreiben, der bedeutet *Schließe alle Datensätze, die das extrahierte Konzept* `Botschaft` *enthalten, nicht jedoch* `Argentinien`*, in diese Kategorie ein*.

Während manche Kategorieregeln beim Aufbau von Kategorien mithilfe von Gruppierungstechniken wie *Kookkurrenz* und *Konzeptwurzelableitung* (Kategorien > Aufbaueinstellungen > Erweiterte Einstellungen: Linguistik) automatisch erzeugt werden, können Sie Kategorieregeln auch manuell im Regeleditor erstellen, indem Sie Ihr Kategorieverständnis der Daten und des Kontexts zu Rate ziehen. Jede Regel wird an eine einzelne Kategorie angehängt, so dass jedes Dokument oder jeder Datensatz, der mit dieser Regel übereinstimmt, über das Scoring dieser Kategorie zugewiesen wird.

Kategorieregeln helfen dabei, die Qualität und Produktivität Ihrer Text-Mining-Ergebnisse und weiterer quantitativer Analysen zu verbessern, indem sie es Ihnen ermöglichen, Antworten mit höherer Spezifität zu kategorisieren. Ihre Erfahrung und Ihr geschäftliches Wissen ermöglichen Ihnen unter Umständen ein spezifisches Verständnis Ihrer Daten und des Kontexts. Dieses Verständnis können Sie nutzen, um dieses Wissen in Kategorieregeln umzusetzen, um so Ihre Dokumente bzw. Datensätze noch effizienter und genauer zu kategorisieren, indem Sie extrahierte Elemente mit Boole’scher Logik kombinieren.

Die Möglichkeit zur Erstellung dieser Regeln verbessert die Kodierungsgenauigkeit, Effizienz und Produktivität, indem sie es Ihnen ermöglicht, Ihr Geschäftswissen in die Extrahierungstechnologie des Produkts einzubringen.

Anmerkung: Beispiele für die Übereinstimmung von Regeln mit Text finden Sie unter Beispiele für Kategorieregeln auf S. 214

Kategorieregelsyntax

Während manche Kategorieregeln beim Aufbau von Kategorien mithilfe von Gruppierungstechniken wie *Kookkurrenz* und *Konzeptwurzelableitung* (Kategorien > Aufbaueinstellungen > Erweiterte Einstellungen: Linguistik) automatisch erzeugt werden, können Sie Kategorieregeln auch manuell im Regeleditor erstellen. Jede Regel ist ein Deskriptor einer einzelnen Kategorie, daher wird jedes Dokument oder jeder Datensatz, der mit dieser Regel übereinstimmt, automatisch über das Scoring dieser Kategorie zugewiesen.

Anmerkung: Beispiele für die Übereinstimmung von Regeln mit Text finden Sie unter Beispiele für Kategorieregeln auf S. 214

Beim Erstellen oder Bearbeiten einer Regel muss diese im Regeleditor geöffnet sein. Sie können Konzepte, Typen oder Muster hinzufügen oder Platzhalter verwenden, um die Übereinstimmungsmöglichkeiten zu vergrößern. Wenn Sie extrahierte Konzepte, Typen und Muster verwenden, haben Sie den Vorteil, dass alle verwandten Konzepte gefunden werden.

Wichtig: Um häufig vorkommende Fehler zu vermeiden, sollten Sie die Konzepte direkt per Drag-and-Drop aus dem Fensterbereich "Extrahierungsergebnisse", Bereichen zur Textverknüpfungsanalyse oder dem Bereich "Daten" in den Regeleditor übertragen oder sie über die Kontextmenüs hinzufügen.

Wenn Konzepte, Typen und Muster erkannt werden, erscheint neben dem Text ein Symbol.

 Extrahiertes Konzept  Extrahierter Typ  Extrahiertes Muster

Regelsyntax und Operatoren

Die folgende Tabelle enthält die Zeichen, mit deren Hilfe Sie Ihre Regelsyntax definieren. Verwenden Sie diese Zeichen zusammen mit den Konzepten, Typen und Mustern, um Ihre Regel zu erstellen.

Tabelle 10-2
Unterstützte Syntax

Zeichen	Beschreibung
`&`	Der Boole'sche Operator "und" Beispielsweise enthält `a & b` sowohl `a`*als auch*`b` wie etwa bei: - `Invasion & Vereinigte Staaten` - `2016 & Olympiade` - `gut & Apfel`
`|`	Der Boole'sche Operator "oder" ist eingeschlossen. Wenn also ein Element oder alle Elemente gefunden werden, erfolgt eine Übereinstimmung. Beispielsweise enthält `a | b` entweder `a`*oder*`b` wie etwa bei: - `Angriff | Frankreich` - `Eigentumswohnung | Apartment`
`!()`	Der Boole'sche Operator "nicht" Beispielsweise enthält `!(a)` nicht `a` wie etwa bei `!(gutes & Hotel), Anschlag & !(Österreich)` oder `!(Gold) & !(Kupfer)`
`*`	Ein Platzhalter, der je nach Verwendung für alles zwischen einem einzelnen Zeichen und einem ganzen Wort stehen kann. Für weitere Informationen siehe Thema Platzhalter in Kategorieregeln auf S. 212.
`()`	Ein Trennzeichen für Ausdrücke. Jeder Ausdruck innerhalb der Klammern wird zuerst ausgewertet.

Zeichen	Beschreibung
+	Der Musterkonnektor, der zur Bildung eines reihenfolgespezifischen Musters verwendet wird. Wenn vorhanden, müssen eckige Klammern verwendet werden. Für weitere Informationen siehe Thema Verwenden von TLA-Mustern in Kategorieregeln auf S. 209.
[]	Das Trennzeichen für Muster wird benötigt, wenn Sie nach Übereinstimmungen auf der Basis eines extrahierten TLA-Musters innerhalb einer Kategorieregel suchen. Der Inhalt in den eckigen Klammern verweist auf TLA-Muster und wird niemals mit Konzepten oder Typen auf der Basis einfacher Kookkurrenz übereinstimmen. Wenn Sie dieses TLA-Muster nicht extrahiert haben, ist keine Übereinstimmung möglich. Für weitere Informationen siehe Thema Verwenden von TLA-Mustern in Kategorieregeln auf S. 209. Verwenden Sie keine eckigen Klammern, wenn Sie nach Übereinstimmungen von Konzepten und Typen anstelle von Mustern suchen. *Anmerkung*: In älteren Versionen wurden mithilfe von Kategorieerstellungsmethoden generierte Kookkurrenz- und Synonymregeln in eckige Klammern eingefasst. In allen neuen Versionen zeigen eckige Klammern das Vorhandensein eines TLA-Musters an. Stattdessen stehen durch die Kookkurrenztechnik erzeugte Regeln und Synonyme in runden Klammern, z. B. `(Lautsprechersysteme|Lautsprecher)`.

Die Operatoren `&` und `|` sind kommutativ, das heißt `a & b` = `b & a` und `a | b` = `b | a`.

Maskieren von Zeichen mit einem umgekehrten Schrägstrich

Falls Sie ein Konzept haben, das ein Zeichen enthält, das auch ein Syntaxzeichen ist, müssen Sie diesem Zeichen einen umgekehrten Schrägstrich voranstellen, damit die Regel korrekt interpretiert wird. Mit dem umgekehrten Schrägstrich (\) werden Zeichen maskiert, die anderenfalls eine besondere Bedeutung hätten. Bei Drag-and-Drop-Verschiebungen in den Editor werden die umgekehrten Schrägstriche automatisch hinzugefügt.

Den folgenden Regelsyntaxzeichen muss ein umgekehrter Schrägstrich vorangestellt werden, wenn diese nicht als Regelsyntax behandelt werden sollen:

& ! | + < > ( ) [ ] *

Da beispielsweise das Konzept `r&d` den Operator “und” enthält (`&`), ist der umgekehrte Schrägstrich bei einer Eingabe in den Regeleditor erforderlich, also: `r\&d`.

Verwenden von TLA-Mustern in Kategorieregeln

Text Link Analysis-Muster können explizit in Kategorienregeln definiert werden, damit Sie sogar noch spezifischere und kontextabhängige Ergebnisse erhalten. Wenn Sie ein Muster in einer Kategorienregel definieren, umgehen Sie die einfacheren Konzeptextrahierungsergebnisse und vergleichen Dokumente und Datensätze auf der Basis der extrahierten Ergebnissen von Text Link Analysis-Mustern.

Wichtig: Um Dokumente mithilfe von TLA-Mustern in Ihren Kategorieregeln zu vergleichen, müssen Sie eine Extrahierung mit aktivierter Textlinkanalyse durchgeführt haben. Die Kategorieregel sucht nach den bei diesem Prozess gefundenen Übereinstimmungen. Wenn Sie die Untersuchung von TLA-Ergebnissen auf der Registerkarte “Modell” Ihres Text-Mining-Knotens nicht aktiviert haben, können Sie die TLA-Extrahierung in den Extrahierungseinstellungen während der interaktiven Sitzung aktivieren und anschließend eine neue Extrahierung vornehmen. Für weitere Informationen siehe Thema Daten extrahieren in Kapitel 9 auf S. 147.

Trennung durch eckige Klammern. Ein TLA-Muster muss in eckigen Klammern `[ ]` stehen, wenn Sie es innerhalb einer Kategorieregel verwenden. Das Trennzeichen für Muster wird benötigt, wenn Sie nach Übereinstimmungen auf der Basis eines extrahierten TLA-Musters suchen. Da Kategorieregeln Typen, Konzepte oder Muster enthalten können, verdeutlichen die Klammern für die Regel, dass der Inhalt in den Klammern auf extrahierte TLA-Muster verweist. Wenn Sie dieses TLA-Muster nicht extrahiert haben, ist keine Übereinstimmung möglich. Falls Sie ein Muster ohne Klammern sehen, z. B. `Apfel + gut` im Bereich "Kategorien", bedeutet das höchstwahrscheinlich, dass das Muster außerhalb des Kategorieregeleditors der Kategorie direkt hinzugefügt wurde. Wenn Sie beispielsweise ein Konzeptmuster aus der Text Link Analysis-Ansicht direkt einer Kategorie hinzufügen, erscheint es nicht in eckigen Klammern. Wird ein Muster jedoch innerhalb einer Kategorienregel verwendet, müssen Sie das Muster innerhalb der Kategorieregel in eckige Klammern setzen, z. B. `[Banane + !(gut)]`.

Verwendung des Pluszeichens in Mustern. In IBM® SPSS® Modeler Text Analytics sind Muster mit bis zu sechs Teilen (Slots) möglich. Wenn die Reihenfolge wichtig ist, verwenden Sie das +-Zeichen, um alle Elemente miteinander zu verbinden, z. B. `[Firma1 + übernahm + Firma2]`. Hier ist die Reihenfolge wichtig, da sie anzeigt, welche Firma die andere übernimmt. Die Reihenfolge wird nicht durch die Satzstruktur bestimmt, sondern durch die Art der Strukturierung der TLA-Musterausgabe. Wenn Sie zum Beispiel die Idee des Satzes "*Ich liebe Paris*" extrahieren möchten, so ist das TLA-Muster wahrscheinlich `[Paris + mag]` oder `[<Location> + <Positive>]` und nicht `[<Positive> + <Location>]`, da die Standard-Meinungsressourcen im Allgemeinen Meinungen in zweiteiligen Mustern an die zweite Stelle setzen. Es kann daher nützlich sein, zur Vermeidung von Problemen das Muster direkt als Deskriptor in Ihrer Kategorie zu verwenden. Wenn Sie jedoch ein Muster als Teil einer komplexeren Aussage verwenden müssen, sollten Sie besonders auf die Reihenfolge der Elemente innerhalb der in der Textlinkanalyseansicht angezeigten Muster achten, da die Reihenfolge darüber entscheidet, ob eine Übereinstimmung gefunden werden kann.

Angenommen, Sie hätten die folgenden Ausdrücke: "Ich *mag Ananas*" und "Ich mag keine *Ananas*. Aber ich *mag* Erdbeeren". Der Ausdruck `mag & Ananas` weist eine Übereinstimmung mit beiden Texten auf, da es sich um einen Konzeptausdruck und nicht um eine Textlinkregel (nicht in Klammern eingeschlossen) handelt. Der Ausdruck `Ananas + mag` entspricht nur "Ich *mag Ananas*", da im zweiten Text das Wort *mag* stattdessen mit *Erdbeeren* verknüpft ist.

Gruppierung mit Mustern. Sie können Ihre Regeln mit Ihren eigenen Mustern vereinfachen. Angenommen, Sie möchten die folgenden drei Ausdrücke festhalten: `Roter Pfeffer + mag`, `Grüner Pfeffer + mag` und `Pfeffer + mag`. Sie können sie zu einer einzigen Kategorieregel gruppieren, z. B. `[* Pfeffer & mag]`. Falls Sie einen anderen Ausdruck hätten, `scharfer Pfeffer + gut`, könnten Sie alle vier mit einer Regel wie `[* Pfeffer + <Positive>]` gruppieren.

Reihenfolge in Mustern. Zur besseren Organisation von Ausgaben versuchen die Text Link Analysis-Regeln aus den mit dem Produkt installierten Vorlagen, die grundlegenden Ausgabemuster in derselben Reihenfolge unabhängig von der Wortfolge im Satz auszugeben. Wenn beispielsweise ein Datensatz den Text "*gute Präsentationen.*" und ein anderer Datensatz den Text "*die Präsentationen hatten gute Inhalte*" enthält, werden beide Texte durch dieselbe Regel erfasst und in derselben Reihenfolge wie `Präsentation + gute` in swn Konzeptmusterergebnissen ausgegeben, nicht als `Präsentation + gute` und gleichzeitig

`gute + Präsentation`. Und in Zwei-Slot-Mustern wie im Beispiel werden die Konzepte, die Typen in der Meinungsbibliothek zugeordnet sind, standardmäßig in der Ausgabe als letzte präsentiert, wie z. B. `Apfel + schlecht`.

Tabelle 10-3
Mustersyntax und Verwendung von Boole'schen Operatoren

Ausdruck	**Resultiert in der Übereinstimmung mit Dokumenten oder Datensätzen, auf die Folgendes zutrifft:**
`[]`	Enthält ein beliebiges TLA-Muster. Das Trennzeichen für Muster wird *in Kategorieregeln* benötigt, wenn Sie nach Übereinstimmungen auf der Basis eines extrahierten TLA-Musters suchen. Der Inhalt in den Klammern verweist auf TLA-Muster und nicht auf einfache Konzepte und Typen. Wenn Sie dieses TLA-Muster nicht extrahiert haben, ist keine Übereinstimmung möglich. Wenn Sie eine Regel erstellen wollten, die keinerlei Muster enthält, könnten Sie Folgendes verwenden: `!([ ])`.
`[a]`	Enthält ein Muster, in dem mindestens ein Element `a` ist, unabhängig von seiner Position im Muster. Beispielsweise kann `[Geschäft][Geschäft+ gut]` oder nur `[Geschäft + .]` als Übereinstimmung finden.
`[a + b]`	Enthält ein Konzeptmuster. Zum Beispiel `[Geschäft + gut]`. *Anmerkung:* Wenn Sie nur dieses Muster festhalten und keine weiteren Elemente hinzufügen möchten, ist es empfehlenswert, das Muster direkt Ihrer Kategorie hinzuzufügen, anstatt eine Regel damit zu erstellen.
`[a + b + c]`	Enthält ein Konzeptmuster. Das +-Zeichen zeigt an, dass die Reihenfolge der übereinstimmenden Elemente wichtig ist. Zum Beispiel `[Firma1 + übernahm + Firma2]`.
`[<A> + <B>]`	Enthält ein Muster vom Typ `<A>` im ersten Slot und vom Typ `<B>` im zweiten Slot, wobei es genau zwei Slots gibt. Das +-Zeichen zeigt an, dass die Reihenfolge der übereinstimmenden Elemente wichtig ist. Zum Beispiel `[<Budget> + <Negative>]`. *Anmerkung:* Wenn Sie nur dieses Muster festhalten und keine weiteren Elemente hinzufügen möchten, ist es empfehlenswert, das Muster direkt Ihrer Kategorie hinzuzufügen, anstatt eine Regel damit zu erstellen.
`[<A> & <B>]`	Enthält ein Typmuster mit Typ `<A>` und Typ `<B>`. Zum Beispiel `[<Budget> & <Negative>]`. Dieses TLA-Muster wird nie extrahiert; bei dieser Schreibweise gleicht es jedoch tatsächlich `[<Budget> + <Negative>]|[<Negative> + <Budget>]`. Die Reihenfolge der übereinstimmenden Elemente ist unwichtig. Es können sich zusätzliche Elemente im Muster befinden, mindestens jedoch `<Budget>` und `<Negative>`.
`[a + .]`	Enthält ein Muster, in dem `a` das einzige Konzept ist und keine anderen Slots für dieses Muster einen Inhalt haben. Zum Beispiel `[deal + .]` findet als Übereinstimmung das Konzeptmuster, bei dem die einzige Ausgabe das Konzept `Geschäft` ist. Wenn Sie das Konzept `Geschäft` als Kategoriedeskriptor hinzufügen, erhalten Sie alle Datensätze mit "Geschäft" als Konzept, enschließlich positive Aussagen über ein Geschäft. Bei Verwendung von `[Geschäft + .]` werden jedoch nur die Datensatzmusterergebnisse abgeglichen, die `Geschäft` präsentieren und keine anderen Beziehungen oder Meinungen, d. h. sie würden nicht mit `Geschäft + großartig` übereinstimmen. *Anmerkung:* Wenn Sie nur dieses Muster festhalten und keine weiteren Elemente hinzufügen möchten, ist es empfehlenswert, das Muster direkt Ihrer Kategorie hinzuzufügen, anstatt eine Regel damit zu erstellen.

Ausdruck	Resultiert in der Übereinstimmung mit Dokumenten oder Datensätzen, auf die Folgendes zutrifft:
`[<A> + <>]`	Enthält ein Muster, in dem `<A>` der einzige Typ ist. `[<Budget> + <>]` findet das Muster als Übereinstimmung, dessen einzige Ausgabe ein Konzept vom Typ `<Budget>` ist. *Anmerkung:* Sie können das `<>` nur dann verwenden, um einen leeren Typ zu kennzeichnen, wenn Sie es hinter das +-Symbol im Typmuster stellen, z. B. `[<Budget> + <>]`, nicht jedoch `[price + <>]`. *Anmerkung:* Wenn Sie nur dieses Muster festhalten und keine weiteren Elemente hinzufügen möchten, ist es empfehlenswert, das Muster direkt Ihrer Kategorie hinzuzufügen, anstatt eine Regel damit zu erstellen.
`[a + !(b)]`	Enthält mindestens ein Muster, das das Konzept `a` einschließt, nicht jedoch das Konzept `b`. Muss mindestens ein Muster einschließen. Zum Beispiel `[Preis + !(hoch)]` oder für Typen `[!(<Obst>|<Gemüse>) + <Positive>]`
`!([<A> & <B>])`	Enthält kein spezifisches Muster. Zum Beispiel `!([<Budget> & <Negative>])`.

Anmerkung: Beispiele für die Übereinstimmung von Regeln mit Text finden Sie unter Beispiele für Kategorieregeln auf S. 214

Platzhalter in Kategorieregeln

Platzhalter können Konzepten in Regeln hinzugefügt werden, um die Übereinstimmungsmöglichkeiten zu erweitern. Der Platzhalter Stern (*) kann vor und/oder nach ein Wort gestellt werden, um anzuzeigen, wie nach Übereinstimmungen in Konzepten gesucht wird. Es gibt zwei Arten der Verwendung von Platzhaltern:

- **Affix-Platzhalter**. Diese Platzhalter werden einer Zeichenfolge direkt voran- oder nachgestellt, ohne dass ein Leerzeichen zwischen der Zeichenfolge und dem Stern steht. Übereinstimmungen für `operat*` können beispielsweise *operat*, *Operator*, *Operation*, *Operationen*, *operativ* usw. sein.
- **Wort-Platzhalter**. Diese Platzhalter stehen vor oder nach einem Konzept, wobei ein Leerzeichen zwischen dem Konzept und dem Stern steht. Übereinstimmungen für `* operation` können beispielsweise *Operation*, *medizinische Operation*, *geschäftliche Operation* usw. sein. Außerdem kann ein Wort-Platzhalter neben einem Affix-Platzhalter verwendet werden, z. B. `* operat* *`, wobei mögliche Übereinstimmungen *Operation*, *medizinische Operation*, *erfahrener Operateur*, *operative Maßnahme* usw. sein können. Wie das letzte Beispiel zeigt, sollte man mit Platzhaltern vorsichtig umgehen, um zu vermeiden, dass ein zu weites Feld abgedeckt wird und ungewollte Übereinstimmungen entstehen.

Ausnahmen!

- Ein Platzhalter kann nie alleine stehen. `(Apfel | * )` zum Beispiel wäre nicht möglich.
- Ein Platzhalter kann nie für Übereinstimmungen mit Typnamen verwendet werden. `<Negative*>` stimmt mit überhaupt keinen Typnamen überein.

- Sie können bestimmte Typen durch Filtern nicht aus der Suche nach Übereinstimmungen in Konzepten ausschließen, die über Platzhalter gefunden wurden. Der Typ, dem das Konzept zugewiesen ist, wird automatisch verwendet.
- Ein Platzhalter kann sich nie in der Mitte einer Wortfolge befinden, weder am Ende noch am Beginn eines Worts (Konto* eröffnen) oder als eigene Komponente (Konto * eröffnen). Sie können Platzhalter außerdem nicht in Typnamen verwenden. Beispiel: `word* word` wie `Apfel* Rezept` entspricht nicht Apfelcreme-Rezept und auch keinem anderen Begriff. Jedoch würde `Apfel* *` den Einträgen *Apfelcreme-Rezept*, *Apfel im Schlafrock*, *Apfel* usw. entsprechen. In einem anderen Beispiel stimmt `word * word` wie `Apfel * Toast` nicht mit *Apfel auf Toast* und auch keinem anderen Begriff überein, da das Sternchen zwischen zwei anderen Wörtern erscheint. Jedoch würde `Apfel *` den Einträgen *Apfel auf Toast*, *Apfel*, *Apfel im Schlafrock* usw. entsprechen.

Tabelle 10-4
Verwendung von Wildcards

Ausdruck	**Resultiert in der Übereinstimmung mit Dokumenten oder Datensätzen, auf die Folgendes zutrifft:**
`*apfel`	Enthält ein Konzept, das mit der angegebenen Zeichenfolge endet, jedoch beliebig viele Buchstaben als Präfix haben kann. Beispiel: `*apfel` endet mit den Buchstaben *apfel*, kann aber beispielsweise folgende Präfixe enthalten: - `Apfel` - `Granatapfel` - `Augapfel`
`apfel*`	Enthält ein Konzept, das mit der angegebenen Zeichenfolge beginnt, jedoch beliebig viele Buchstaben als Suffix haben kann. Beispiel: `apfel*` beginnt mit den Buchstaben *apfel*, kann aber beispielsweise folgende Suffixe (oder kein Suffix) enthalten: - `Apfel` - `Apfelkompott` - `Apfelwein.` `Apfel* & !(Birne* | Quitte)` beispielsweise enthält ein Konzept, das mit den Buchstaben `Apfel` beginnt, jedoch kein Konzept, das mit den Buchstaben *Birne* beginnt und auch nicht das Konzept `Quitte`. Daher ergäbe Folgendes KEINE Übereinstimmung: `Apfel & Quitte` Folgendes ergäbe hingegen durchaus eine Übereinstimmung: - `Apfelkompott` - `Apfel & Orange`
`*produkt*`	Enthält ein Konzept mit der Buchstabenfolge `produkt`, das aber beliebig viele Buchstaben als Präfix, Suffix oder beides aufweisen kann. Beispiel: Übereinstimmungen für `*produkt*` wären zum Beispiel: - `Produkt` - `Nebenprodukt` - `unproduktiv`
`* Darlehen`	Enthält ein Konzept mit dem Wort `Darlehen`, das aber mit einem anderen Wort, das davor steht, eine Zusammensetzung bilden kann. Übereinstimmungen für `* Darlehen` wären zum Beispiel: - `Darlehen` - `zinsloses Darlehen` - `nicht gewährtes Darlehen.` Übereinstimmende Konzeptmuster für `[* Lieferung + <Negative>]`, das an erster Stelle ein Konzept enthält, das mit dem Wort `Lieferung` endet und an zweiter Stelle einen Typ `<Negative>`, wären zum Beispiel: - `erste Lieferung + langsam` - `wichtige Lieferung + spät`

Ausdruck	Resultiert in der Übereinstimmung mit Dokumenten oder Datensätzen, auf die Folgendes zutrifft:
`Veranstal-tung *`	Enthält ein Konzept mit dem Wort `Veranstaltung`, das aber mit einem anderen Wort, das dahinter steht, eine Zusammensetzung bilden kann. Übereinstimmungen für `* Veranstaltung` wären zum Beispiel: - `Veranstaltung` - `Veranstaltung Berlin` - `Veranstaltung am Fluss`
`* Apfel *`	Enthält ein Konzept, das mit einem beliebigen Wort beginnen kann, nach dem das Wort `Apfel` steht, auf das ein anderes Wort folgen kann. `*` bedeutet 0 oder n, daher stimmt es auch mit `Apfel` überein. Übereinstimmungen für `* Apfel *` wären zum Beispiel: - `warmer Apfelkompott` - `Granny Smith Apfel marktfrisch` - `roter Apfel kandiert.` - `Apfel` Übereinstimmende Konzeptmuster für `[* Reservierung* * + <Positive>]`, das an erster Stelle ein Konzept enthält, das das Wort `Reservierung` enthält (unabhängig davon, an welcher Stelle es im Konzept steht) und an zweiter Stelle einen Typ `<Positive>`, wären zum Beispiel: - `Reservierung Hotel + gut` - `Online Reservierung + gut`

Anmerkung: Beispiele für die Übereinstimmung von Regeln mit Text finden Sie unter Beispiele für Kategorieregeln auf S. 214

Beispiele für Kategorieregeln

Werfen Sie einen Blick auf folgende Beispiele, um zu veranschaulichen, wie Regeln mit Datensätzen verglichen werden, die auf unterschiedliche Weise auf der Syntax basieren, die für deren Ausdruck verwendet wird.

Beispieldatensätze

Angenommen, Sie hatten zwei Datensätze:

- **Datensatz A**: *"Als ich in meinen Geldbeutel sah, stellte ich fest, dass mir fünf Euro fehlten."*
- **Datensatz B**: *"Ich fand die fünf Euro im Picknickbereich, aber die Decke fehlte."*

Die folgenden zwei Tabellen zeigen mögliche extrahierte Konzepte und Typen sowie Konzept- und Typmuster.

Aus dem Beispiel extrahierte Konzepte und Typen

Tabelle 10-5
Aus Beispiel extrahierte Konzepte und Typen

Extrahiertes Konzept	Art des Konzepts
`Geldbeutel`	`<Unbekannt>`
`fehlend`	`<Negativ>`
`5 Euro`	`<Währung>`
`Decke`	`<Unbekannt>`
`Picknickbereich`	`<Unbekannt>`

Aus dem Beispiel extrahierte TLA-Muster

Tabelle 10-6
Aus Beispiel extrahierte TLA-Musterausgabe

Extrahierte Konzeptmuster	Extrahierte Typmuster	Aus Datensatz
`Picknickbereich + .`	`<Unbekannt> + <>`	Datensatz B
`Geldbeutel + .`	`<Unbekannt> + <>`	Datensatz A
`Decke + fehlend`	`<Unbekannt> + <Negativ>`	Datensatz B
`5 Euro + .`	`<Währung> + <>`	Datensatz B
`5 Euro + fehlend`	`<Währung> + <Negativ>`	Datensatz A

Mögliche Übereinstimmung von Kategorieregeln

Die folgende Tabelle enthält Syntaxeinträge, die im Kategorieregeleditor eingegeben werden könnten. Nicht alle hier aufgeführten Regeln funktionieren und nicht alle stimmen mit denselben Datensätzen überein. Beachten Sie, wie sich die unterschiedliche Syntax auf die verglichenen Datensätze auswirkt.

Tabelle 10-7
Beispielregeln

Regelsyntax	Ergebnis
`5 Euro & fehlend`	Übereinstimmung mit Datensatz A und B, da beide das extrahierte Konzept `fehlend` und das extrahierte Konzept `5 Euro` enthalten. Dies entspricht: `(5 Euro & fehlend)`
`fehlend & 5 Euro`	Übereinstimmung mit Datensatz A und B, da beide das extrahierte Konzept `fehlend` und das extrahierte Konzept `5 Euro` enthalten. Dies entspricht: `(fehlend & 5 Euro)`
`fehlend & <Währung>`	Übereinstimmung mit Datensatz A und B, da beide das extrahierte Konzept `fehlend` und ein Konzept enthalten, das mit dem Typ `<Währung>` übereinstimmt. Dies entspricht: `(fehlend & <Währung>)`
`<Währung> & fehlend`	Übereinstimmung mit Datensatz A und B, da beide das extrahierte Konzept `fehlend` und ein Konzept enthalten, das mit dem Typ `<Währung>` übereinstimmt. Dies entspricht: `(<Währung> & fehlend)`
`[5 Euro + fehlend]`	Übereinstimmung mit A, aber **nicht** mit B, da Datensatz B keine TLA-Musterausgabe mit `5 Euro + fehlend` erzeugt hat (siehe vorherige Tabelle). Dies entspricht folgender TLA-Musterausgabe: `5 Euro + fehlend`
`[fehlend + 5 Euro]`	Weder Übereinstimmung mit A noch mit B, da kein extrahiertes TLA-Muster (siehe vorherige Tabelle) mit der hier ausgedrückten Reihenfolge mit `fehlend` an erster Stelle übereinstimmt. Dies entspricht folgender TLA-Musterausgabe: `5 Euro + fehlend`
`[fehlend & 5 Euro]`	Übereinstimmung mit A, aber **nicht** mit B, da kein derartiges TLA-Muster aus Datensatz B extrahiert wurde. Das Symbol `&` zeigt an, dass die Reihenfolge beim Vergleich unwichtig ist; aus diesem Grund sucht diese Regel nach einer Musterübereinstimmung entweder mit `[fehlend + 5 Euro]` oder `[5 Euro + fehlend]`. Es gibt nur eine Übereinstimmung bei `[5 Euro + fehlend]` aus Datensatz A.

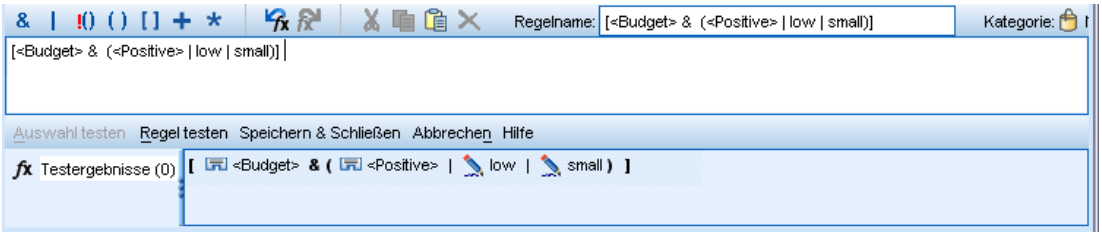
Regelsyntax	Ergebnis
`[fehlend + <Währung>]`	Weder Übereinstimmung mit Datensatz A noch mit Datensatz B, da kein extrahiertes TLA-Muster mit dieser Reihenfolge übereinstimmte. Hier gibt es keine Entsprechung, da eine TLA-Ausgabe nur auf Begriffen `(5 Euro + fehlend)` oder auf Typen `(<Währung> + <Negativ>)` basiert, Konzepte und Typen jedoch nicht vermischt.
`[<Währung> + <Negativ>]`	Übereinstimmung mit Datensatz A, aber nicht B, da kein TLA-Muster aus Datensatz B extrahiert wurde. Dies entspricht folgender TLA-Ausgabe: `<Währung> + <Negativ>`
`[<Negativ> + <Währung>]`	Weder Übereinstimmung mit Datensatz A noch mit Datensatz B, da kein extrahiertes TLA-Muster mit dieser Reihenfolge übereinstimmte. Standardmäßig belegt in der Vorlage `Meinungen`, wenn ein *Thema* mit einer *Meinung*gefunden wird, das *Thema* (`<Währung>`) die erste Slot-Position und die *Meinung* (`<Negativ>`) die zweite Slot-Position.

Erstellen von Kategorienregeln

Beim Erstellen oder Bearbeiten einer Regel muss diese im Regeleditor geöffnet sein. Sie können Konzepte, Typen oder Muster hinzufügen oder Platzhalter verwenden, um die Übereinstimmungsmöglichkeiten zu vergrößern. Wenn Sie erkannte Konzepte, Typen und Muster verwenden, haben Sie den Vorteil, dass alle verwandten Konzepte gefunden werden. Wenn Sie beispielsweise ein Konzept verwenden, werden alle damit verbundenen Begriffe, Pluralformen und Synonyme ebenfalls mit der Regel in Übereinstimmung gebracht. Ebenso werden, wenn Sie einen Typ verwenden, all seine Konzepte ebenfalls von der Regel erfasst.

Sie können den Regeleditor öffnen, indem Sie eine vorhandene Regel bearbeiten oder mit der rechten Maustaste auf den Namen der Kategorie klicken und die Option Regel erstellen wählen.

Abbildung 10-15
Regeleditorbereich



Sie können Kontextmenüs und Drag-and-Drop verwenden oder Konzepte, Typen und Muster manuell in den Editor eingeben. Danach kombinieren Sie diese mit Boole'schen Operatoren (`&`, `!()`, `|`) und Klammern, um Ihre Regelausdrücke zu formulieren. Um häufig vorkommende Fehler zu vermeiden, sollten Sie die Konzepte direkt per Drag-and-Drop aus dem Fensterbereich "Extrahierungsergebnisse" oder "Daten" in den Regeleditor übertragen. Achten Sie genau auf die Syntax dieser Regeln, um Fehler zu vermeiden. Für weitere Informationen siehe Thema Kategorieregelsyntax auf S. 207.

Anmerkung: Beispiele für die Übereinstimmung von Regeln mit Text finden Sie unter Beispiele für Kategorieregeln auf S. 214

So erstellen Sie eine Regel:

- Wenn Sie noch keine Daten extrahiert haben oder Ihre Extrahierung veraltet ist, tun Sie es jetzt. Für weitere Informationen siehe Thema Daten extrahieren in Kapitel 9 auf S. 147.

- Wählen Sie im Fensterbereich “Kategorien” die Kategorie aus, die Ihrer Regel hinzugefügt werden soll.

- Wählen Sie in den Menüs die Optionsfolge Kategorien > Regel erstellen. Im Fenster wird der Kategorienregel-Editorbereich geöffnet.

- Geben Sie im Feld “Regelname” einen Namen für Ihre Regel ein. Wenn Sie keinen Namen angeben, wird der Ausdruck automatisch als Name verwendet. Sie können diese Regel später umbenennen.

- Im größeren Textfeld für Ausdrücke haben Sie folgende Möglichkeiten:
  - Geben Sie Text direkt in das Feld ein oder verschieben Sie ihn per Drag-and-Drop aus einem anderen Bereich dorthin. Verwenden Sie nur extrahierte Konzepte, Typen und Muster Wenn Sie beispielsweise das Wort `Katzen` eingeben, im Fensterbereich “Extrahierungsergebnisse” aber nur die Singularform, `Katze`, angezeigt wird, kann der Editor `Katzen` nicht erkennen. In diesem letzten Fall könnte die Singularform automatisch den Plural einschließen, ansonsten könnten Sie einen Platzhalter verwenden. Für weitere Informationen siehe Thema Kategorieregelsyntax auf S. 207.
  - Wählen Sie die Konzepte, Typen oder Muster aus, die Sie Regeln hinzufügen möchten, und verwenden Sie die Menüs.
  - Boole’sche Operatoren hinzufügen, um Elemente in Ihrer Regel miteinander zu verknüpfen. Über die Schaltflächen der Symbolleiste können Sie Ihrer Regel die Boole’schen Operatoren “und”, &, “oder”, | und “nicht” !() sowie runde Klammern () und eckige Klammern für Muster [ ] hinzufügen.

- Klicken Sie auf die Schaltfläche Regel testen, um zu überprüfen, ob Ihre Regel korrekt formuliert ist. Für weitere Informationen siehe Thema Kategorieregelsyntax auf S. 207. Die Anzahl der gefundenen Dokumente bzw. Datensätze wird in Klammern neben dem Text Testergebnis angezeigt. Rechts neben dem Text können Sie die Elemente in Ihrer Regel sehen, die erkannt wurden, bzw. etwaige Fehlermeldungen. Ein rotes Fragezeichen neben dem Typ, Muster oder Konzept zeigt an, dass das Element mit keinen bekannten Extrahierungen übereinstimmt. Falls dies der Fall ist, wird die Regel keine Datensätze finden.

- Um einen Teil der Regel zu testen, wählen Sie den betreffenden Teil aus und klicken Sie auf Auswahl testen.

- Nehmen Sie ggf. die erforderlichen Änderungen vor und testen Sie die Regel erneut, wenn Sie Probleme festgestellt haben.

- Klicken Sie abschließend auf Speichern & Schließen, um Ihre Regel erneut zu speichern und den Editor zu schließen. Der neue Regelname wird in der Kategorie angezeigt.

Bearbeiten und Löschen von Regeln

Nachdem Sie eine Regel erstellt und gespeichert haben, können Sie sie jederzeit bearbeiten. Für weitere Informationen siehe Thema Kategorieregelsyntax auf S. 207.

Wenn Sie eine Regel nicht mehr benötigen, können Sie sie löschen.

So bearbeiten Sie Regeln:

- Wählen Sie die Regel im Dialogfeld “Kategoriedefinition” auf der Registerkarte “Deskriptoren” aus.
- Wählen Sie in den Menüs die Optionsfolge Kategorien > Regel bearbeiten oder doppelklicken Sie auf den Regelnamen. Der Editor wird geöffnet und die ausgewählte Regel wird angezeigt.
- Nehmen Sie Änderungen an der Regel über Extrahierungsergebnisse und die Schaltflächen der Symbolleiste vor.
- Testen Sie die Regel erneut, um sicherzustellen, dass sie die erwarteten Ergebnisse liefert.
- Klicken Sie auf Speichern & Schließen, um Ihre Regel erneut zu speichern und den Editor zu schließen.

So löschen Sie eine Regel

- Wählen Sie die Regel im Dialogfeld “Kategoriedefinition” auf der Registerkarte “Deskriptoren” aus.
- Wählen Sie im Menü Bearbeiten > Löschen. Die Regel wird aus der Kategorie gelöscht.

Importieren und Exportieren von vordefinierten Kategorien

Wenn Sie Ihre eigenen Kategorien in einer Microsoft Excel-Datei (**.xls*, **.xlsx*) gespeichert haben, können Sie sie in IBM® SPSS® Modeler Text Analytics importieren.

Sie können die vorhandenen Kategorien in einer geöffneten interaktiven Workbench-Sitzung in eine Microsoft Excel-Datei (**.xls*, **.xlsx*) exportieren. Wenn Sie Ihre Kategorien exportieren, können Sie wie erforderlich einige zusätzliche Informationen einschließen oder ausschließen, z. B. Deskriptoren und Werte. Für weitere Informationen siehe Thema Exportieren von Kategorien auf S. 227.

Wenn Ihre vordefinierten Kategorien keine Codes haben bzw. Sie neue Codes wünschen, können Sie automatisch ein neues Set an Codes für das Kategorienset im Bereich “Kategorien” generieren, indem Sie aus den Menüs die Befehle Kategorien > Kategorien verwalten > Codes automatisch generieren wählen. Damit werden alle bestehenden Codes entfernt und automatisch neu nummeriert.

Importieren vordefinierter Kategorien

Sie können vordefinierte Kategorien in IBM® SPSS® Modeler Text Analytics importieren. Stellen Sie vor dem Import sicher, dass sich die vordefinierte Kategoriedatei in einer Microsoft Excel-Datei (**.xls*, **.xlsx*) befindet und in einem der unterstützten Formate strukturiert ist. Sie

können auch wählen, dass das Produkt das Format automatisch erkennen soll. Die folgenden Formate werden unterstützt:

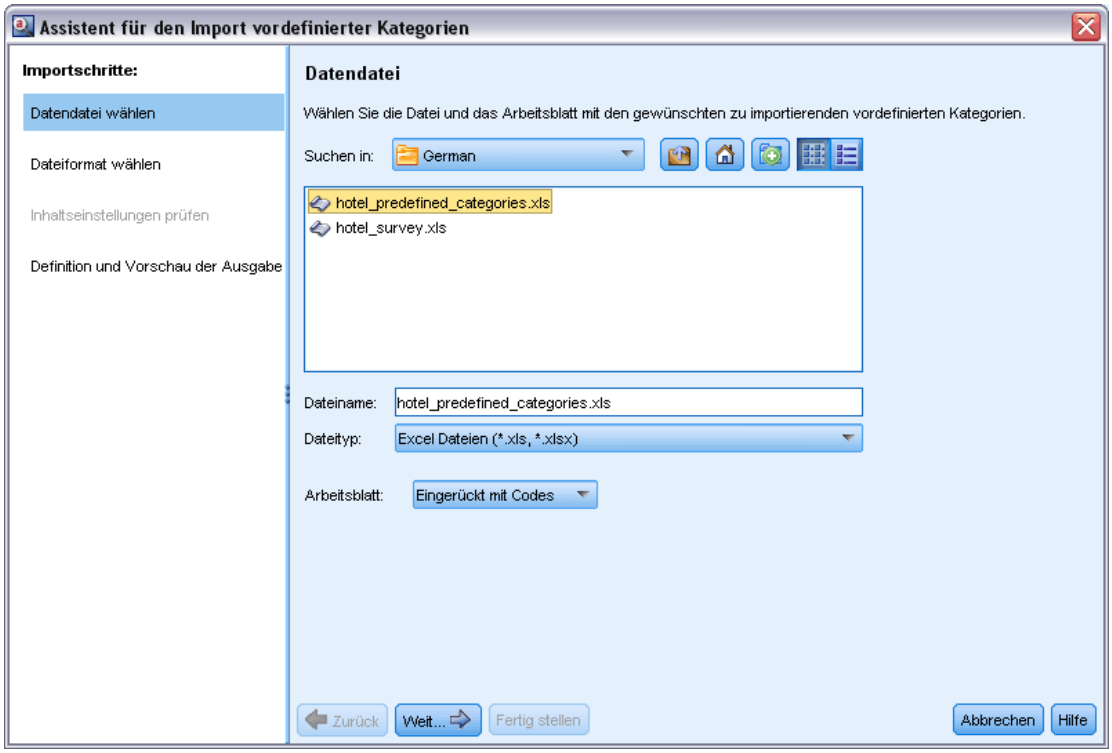
- Einfaches Listenformat: Für weitere Informationen siehe Thema Einfaches Listenformat auf S. 223.
- Kompaktes Format: Für weitere Informationen siehe Thema Kompaktes Format auf S. 224.
- Eingerücktes Format: Für weitere Informationen siehe Thema Eingerücktes Format auf S. 225.

Anmerkung: Für die meisten Sprachen stehen ein Demo-Stream und eine Demo-Datendatei zur Verfügung, die illustrieren, wie vordefinierte Kategorien importiert werden. Sehen Sie im Unterverzeichnis `<Modeller-Installationsverzeichnis>\Demos\Text_Analytics\` für Ihre Sprache nach.

So importieren Sie vordefinierte Kategorien:

▶ Wählen Sie in den Menüs die Optionsfolge Kategorien > Kategorien verwalten > Vordefinierte Kategorien importieren. Ein Assistent für den Import vordefinierter Kategorien wird angezeigt.

Abbildung 10-16
Assistent für den Import vordefinierter Kategorien



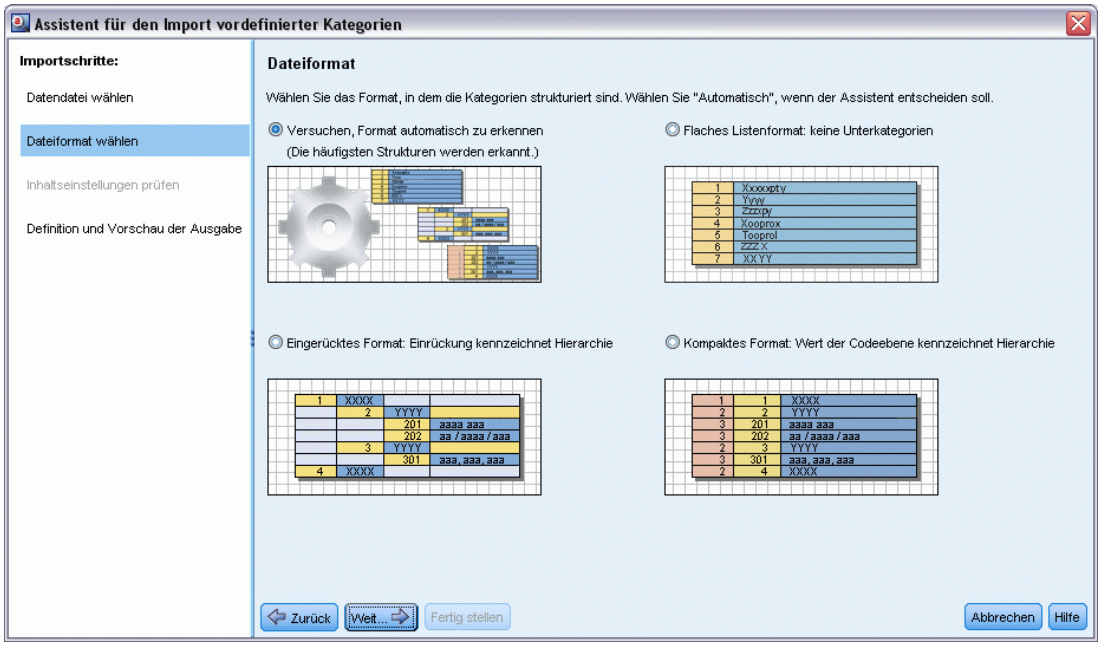
▶ Wählen Sie in der Dropdown-Liste “Suchen in” das Laufwerk und den Ordner aus, in dem sich die Datei befindet.

▶ Wählen Sie die Datei aus der Liste aus. Der Name der Datei wird im Feld “Dateiname” angezeigt.

▶ Wählen Sie aus der Liste das Arbeitsblatt aus, das die vordefinierten Kategorien enthält. Der Name des Arbeitsblatts wird im Feld “Arbeitsblatt” angezeigt.

▶ Klicken Sie auf Weiter, um zu beginnen, das Datenformat zu wählen.

Abbildung 10-17
Dialogfeld "Vordefinierte Kategorien importieren", Schritt "Datenformat"

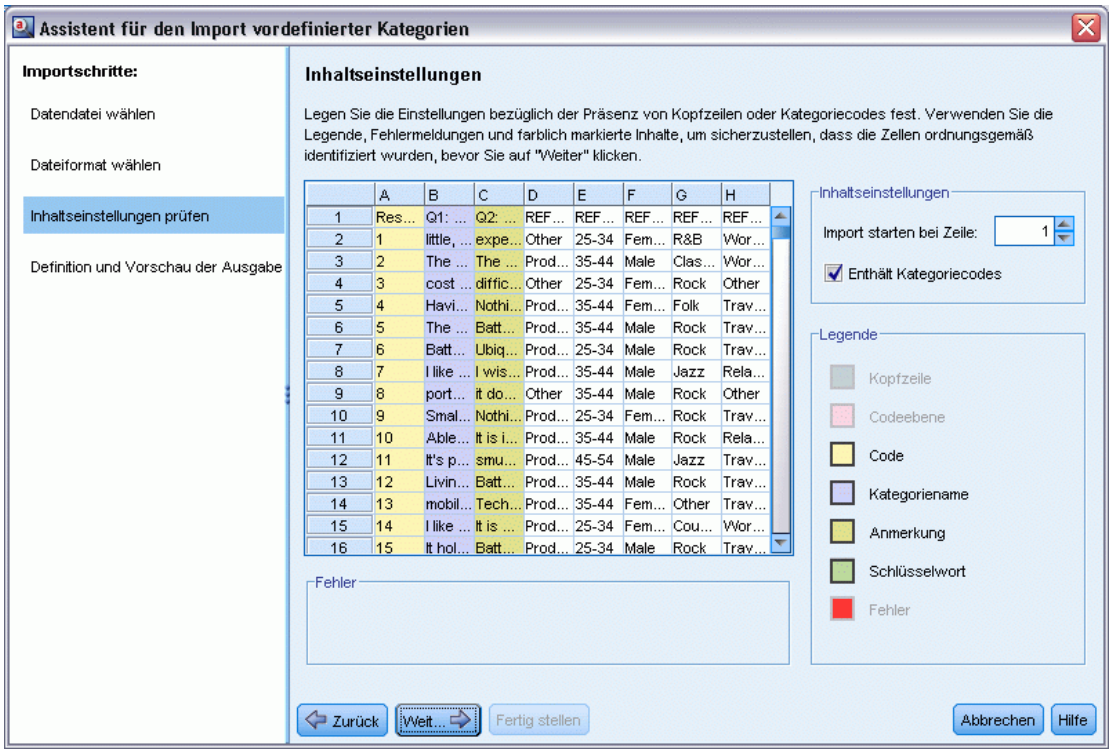


▶ Wählen Sie das Format für Ihre Datei oder wählen Sie die Option, damit das Produkt versucht, das Format automatisch zu erkennen. Die automatische Erkennung funktioniert am besten bei den häufigsten Formaten.

- Einfaches Listenformat: Für weitere Informationen siehe Thema Einfaches Listenformat auf S. 223.
- Kompaktes Format: Für weitere Informationen siehe Thema Kompaktes Format auf S. 224.
- Eingerücktes Format: Für weitere Informationen siehe Thema Eingerücktes Format auf S. 225.

▶ Klicken Sie auf Weiter, um die zusätzlichen Importoptionen zu definieren. Wenn Sie die automatische Formaterkennung wählen, werden Sie zum letzten Schritt geführt.

Abbildung 10-18
"Vordefinierte Kategorien importieren", Schritt "Importoptionen"



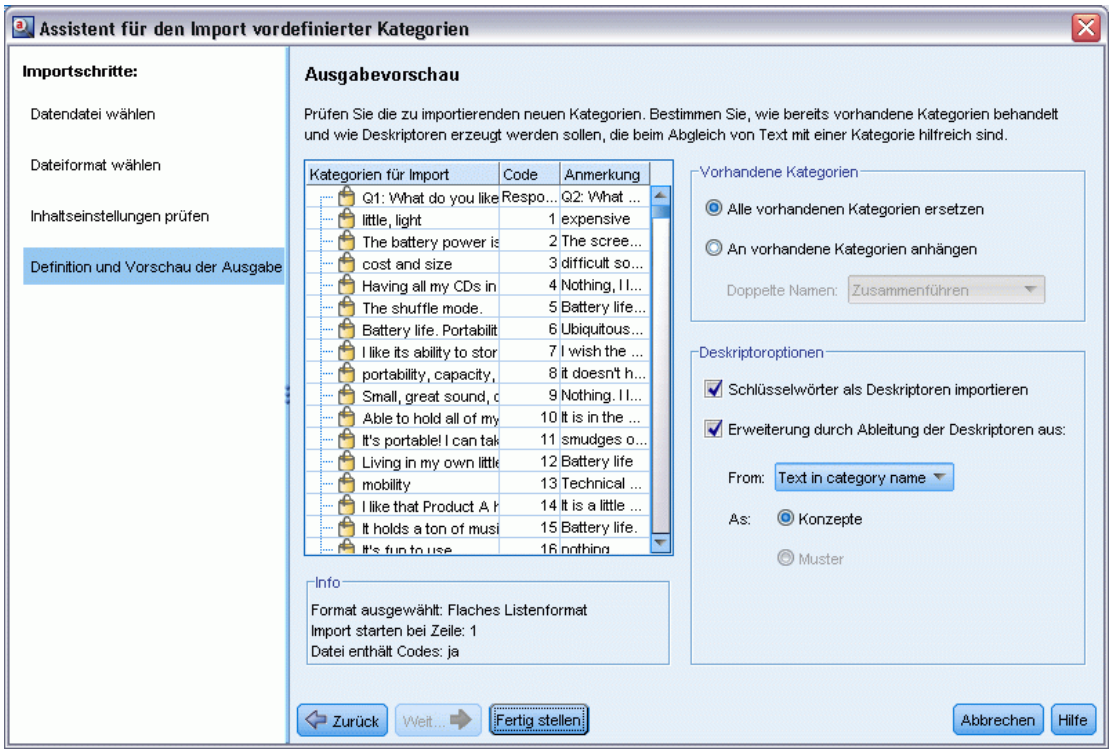
- Wenn eine oder mehrere Zeilen Spaltentitel oder andere irrelevante Informationen enthalten, wählen Sie in der Option Import beginnen bei Zeile die Zeilennummer aus, bei der Sie mit dem Import beginnen möchten. Wenn Ihre Kategorienamen beispielsweise in Zeile 7 beginnen, müssen Sie die Ziffer 7 für diese Option eingeben, damit die Datei korrekt importiert wird.

- Wenn Ihre Datei Kategoriecodes enthält, wählen Sie die Option Enthält Kategoriecodes. Damit unterstützen Sie den Assistenten bei der korrekten Erkennung Ihrer Daten.

- Überprüfen Sie die farbkodierten Zellen und die Legende, um sicherzustellen, dass die Daten korrekt identifiziert wurden. Etwaige in der Datei erkannte Fehler werden in Rot angezeigt und erhalten einen Verweis unter der Formatvorschautabelle. Wenn das falsche Format ausgewählt wurde, gehen Sie zurück und wählen Sie ein anderes Format aus. Falls Sie Korrekturen in Ihrer Datei durchführen müssen, korrigieren Sie die Datei und starten Sie dann den Assistenten neu, indem Sie die Datei erneut auswählen. Sie müssen alle Fehler korrigieren, bevor Sie den Assistenten beenden können.

- Klicken Sie auf Weiter, um das Set an Kategorien und Unterkategorien zu sehen, das importiert wird, und um zu definieren, wie Deskriptoren für diese Kategorien erstellt werden.

Abbildung 10-19
Dialogfeld "Vordefinierte Kategorien importieren", Schritt "Vorschau"



- Prüfen Sie das Kategorienset, das in die Tabelle importiert wird. Wenn Sie die als Deskriptoren erwarteten Stichwörter nicht sehen, wurden diese eventuell beim Import nicht erkannt. Stellen Sie sicher, dass sie über das korrekte Präfix verfügen und in der korrekten Zelle erscheinen.

- Wählen Sie, wie bereits existierende Kategorien in Ihrer Sitzung gehandhabt werden.
  - Alle vorhandenen Kategorien ersetzen. Diese Option löscht alle existierenden Kategorien. Anschließend werden die importierten Kategorien alleine an deren Stelle verwendet.
  - An vorhandene Kategorien anhängen. Diese Option importiert die Kategorien und führt alle häufigen Kategorien mit den vorhandenen Kategorien zusammen. Beim Hinzufügen zu vorhandenen Kategorien müssen Sie festlegen, wie Duplikate behandelt werden sollen. Eine Option (Zusammenführen) ist es, alle importierten Kategorien mit bestehenden Kategorien zusammenzuführen, wenn sie einen gemeinsamen Kategorienamen haben. Eine andere Option (Aus Import ausschließen) ist es, den Import von Kategorien zu untersagen, wenn eine Kategorie mit demselben Namen vorhanden ist.

- Stichwörter als Deskriptoren importieren ist eine Option für den Import der Stichwörter, die in Ihren Daten als Stichwörter für die verknüpfte Kategorie identifiziert werden.

- Kategorien durch abgeleitete Deskriptoren erweitern ist eine Option, die Deskriptoren aus den Wörtern erzeugt, die den Namen der Kategorie oder Unterkategorie und/oder die Wörter repräsentieren, aus denen die Anmerkung besteht. Wenn die Wörter extrahierten Ergebnissen entsprechen, werden diese als Deskriptoren zur Kategorie hinzugefügt. Diese Option erzeugt die besten Ergebnisse, wenn die Kategorienamen oder Anmerkungen lang und beschreibend

sind. Dies ist eine schnelle Methode, um Kategoriedeskriptoren zu erzeugen, mit deren Hilfe die Kategorie Datensätze erfassen kann, die diese Deskriptoren enthalten.

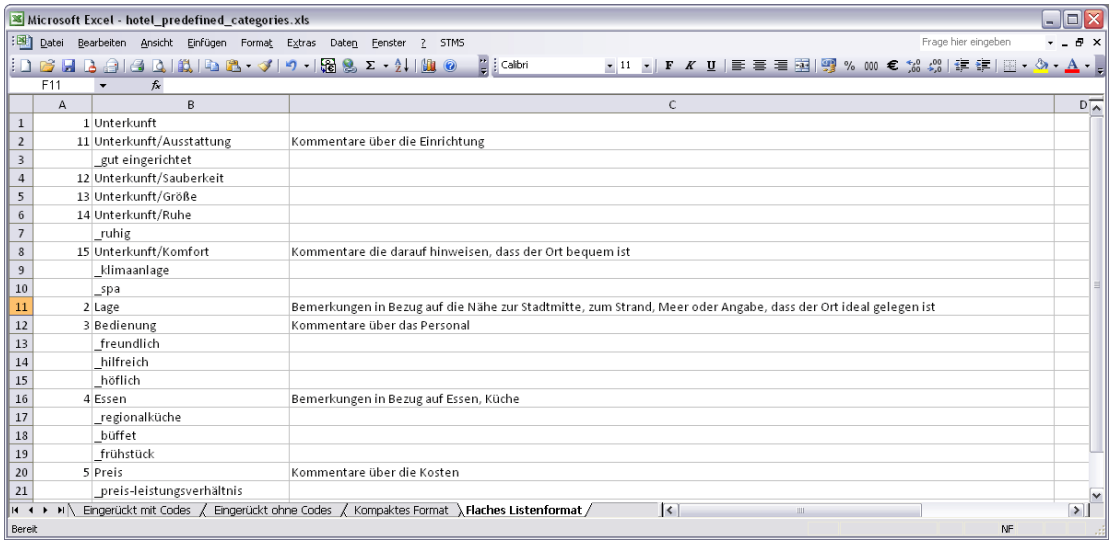
- Im Feld Von können Sie wählen, von welchem Text die Deskriptoren abgeleitet werden, den Namen der Kategorien und Unterkategorien und/oder den Wörtern in den Anmerkungen.
- Im Feld As können Sie wählen, ob diese Deskriptoren in der Form von Konzepten oder TLA-Mustern erstellt werden. Wenn keine TLA-Extrahierung erfolgt ist, sind die Optionen für Muster in diesem Assistenten deaktiviert.

▶ Klicken Sie auf Fertigstellen, um die vordefinierten Kategorien in den Bereich "Kategorien" zu importieren.

Einfaches Listenformat

In diesem einfachen Listenformat befindet sich nur eine erste Ebene von Kategorien ohne Unterkategorien oder Unternetzen. Kategoriennamen befinden sich in einer einzelnen Spalte.

Abbildung 10-20
Beispiel für ein einfaches Listenformat



	A	B	C
1	1	Unterkunft	
2	11	Unterkunft/Ausstattung	Kommentare über die Einrichtung
3		_gut eingerichtet	
4	12	Unterkunft/Sauberkeit	
5	13	Unterkunft/Größe	
6	14	Unterkunft/Ruhe	
7		_ruhig	
8	15	Unterkunft/Komfort	Kommentare die darauf hinweisen, dass der Ort bequem ist
9		_klimaanlage	
10		_spa	
11	2	Lage	Bemerkungen in Bezug auf die Nähe zur Stadtmitte, zum Strand, Meer oder Angabe, dass der Ort ideal gelegen ist
12	3	Bedienung	Kommentare über das Personal
13		_freundlich	
14		_hilfreich	
15		_höflich	
16	4	Essen	Bemerkungen in Bezug auf Essen, Küche
17		_regionalküche	
18		_büffet	
19		_frühstück	
20	5	Preis	Kommentare über die Kosten
21		_preis-leistungsverhältnis	

Die folgenden Informationen können sich in einer Datei dieses Formats befinden:

- Die optionale Spalte **Codes** enthält numerische Werte, die eine Kategorie eindeutig identifizieren. Wenn Sie angeben, dass die Datendatei Codes enthält (Option Enthält Kategoriecodes im Schritt Inhaltseinstellungen), muss eine Spalte mit eindeutigen Codes für jede Kategorie in der Zelle direkt links neben dem Kategorienamen vorhanden sein. Wenn Ihre Daten keine Codes enthalten, Sie jedoch später Codes erstellen möchten, können Sie diese jederzeit tun (Kategorien > Kategorien verwalten > Codes automatisch erzeugen).
- Eine *erforderliche* Spalte **Kategorienamen** enthält alle Namen der Kategorien. Diese Spalte ist erforderlich, um einen Import mit diesem Format durchzuführen.

- Optionale **Anmerkungen** in der Zelle direkt rechts neben dem Kategorienamen. Diese Anmerkung besteht aus Text, der Ihre Kategorien/Unterkategorien beschreibt.
- Optionale **Stichwörter** können als Deskriptoren für Kategorien importiert werden. Damit sie erkannt werden, müssen diese Stichwörter in der Zelle direkt unter dem verknüpften Kategorie- bzw. Unterkategorienamen vorhanden sein und der Liste der Stichwörter muss ein Unterstrich (_) vorangestellt sein, wie beispielsweise "`_Feuerwaffen, Waffen/Gewehre`". Die Stichwortzelle kann ein oder mehrere Wörter enthalten, die zur Beschreibung jeder Kategorie verwendet werden. Diese Wörter werden als Deskriptoren importiert oder ignoriert, abhängig von Ihrer Angabe im letzten Schritt des Assistenten. Später werden Deskriptoren mit den extrahierten Ergebnissen aus dem Text verglichen. Liegt eine Übereinstimmung vor, wird der Datensatz bzw. das Dokument der Kategorie zugewiesen, die diesen Deskriptor enthält.

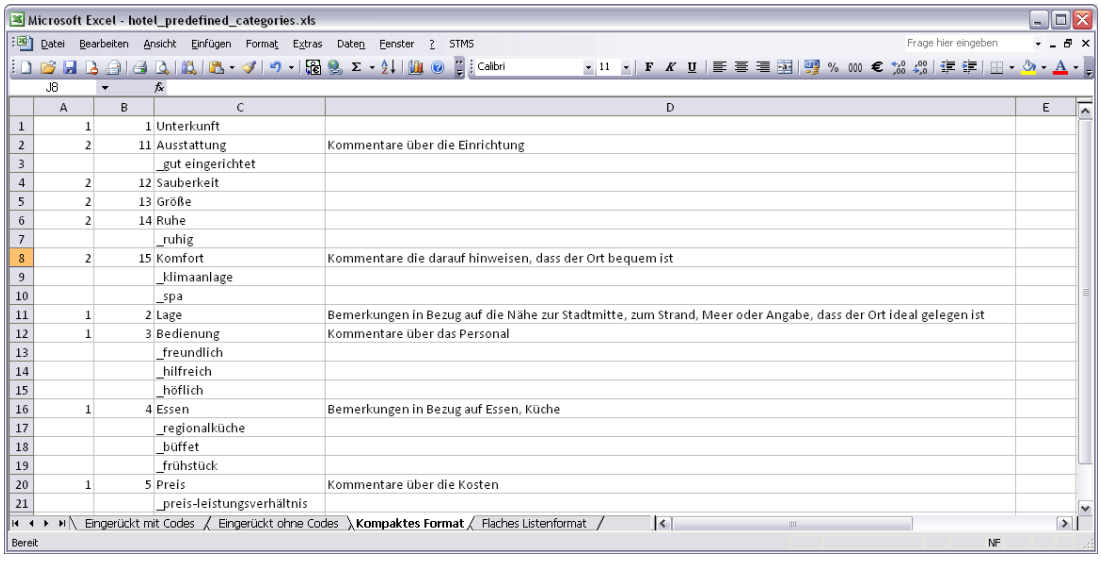
Tabelle 10-8
Einfaches Listenformat mit Codes, Stichwörtern und Anmerkungen

Spalte A	**Spalte B**	**Spalte C**
Kategoriecode (*optional*)	Kategoriename	Anmerkung
	_Deskriptor-/Stichwortliste (*optional*)	

Kompaktes Format

Das kompakte Format ist ähnlich strukturiert wie das einfache Listenformat, allerdings wird das kompakte Format mit hierarchischen Kategorien verwendet. Daher ist eine Spalte für die Codeebene erforderlich, um die hierarchische Stufe jeder Kategorie und Unterkategorie zu definieren.

Abbildung 10-21
Beispiel einer kompakten, vordefinierten Kategoriedatei in Microsoft Excel



Die folgenden Informationen können sich in einer Datei dieses Formats befinden:

- Die *erforderliche* Spalte **Code-Ebene** enthält Zahlen, die die hierarchische Position für die nachfolgenden Informationen in dieser Zeile enthalten. Beispiel: Wenn die Werte 1, 2 oder 3 angegeben sind und Sie sowohl Kategorien als auch Unterkategorien haben, steht 1 für Kategorien, 2 für Unterkategorien und 3 für Unter-Unterkategorien. Wenn Sie nur Kategorien und Unterkategorien haben, steht 1 für Kategorien und 2 für Unterkategorien. Das kann beliebig weitergeführt werden bis zur gewünschten Kategorientiefe.
- Die optionale Spalte **Codes** enthält Werte, die eine Kategorie eindeutig identifizieren. Wenn Sie angeben, dass die Datendatei Codes enthält (Option Enthält Kategoriecodes im Schritt Inhaltseinstellungen), muss eine Spalte mit eindeutigen Codes für jede Kategorie in der Zelle direkt links neben dem Kategorienamen vorhanden sein. Wenn Ihre Daten keine Codes enthalten, Sie jedoch später Codes erstellen möchten, können Sie diese jederzeit tun (Kategorien > Kategorien verwalten > Codes automatisch erzeugen).
- Eine *erforderliche* Spalte **Kategorienamen** enthält alle Namen der Kategorien und Unterkategorien. Diese Spalte ist erforderlich, um einen Import mit diesem Format durchzuführen.
- Optionale **Anmerkungen** in der Zelle direkt rechts neben dem Kategorienamen. Diese Anmerkung besteht aus Text, der Ihre Kategorien/Unterkategorien beschreibt.
- Optionale **Stichwörter** können als Deskriptoren für Kategorien importiert werden. Damit sie erkannt werden, müssen diese Stichwörter in der Zelle direkt unter dem verknüpften Kategorie- bzw. Unterkategorienamen vorhanden sein und der Liste der Stichwörter muss ein Unterstrich (_) vorangestellt sein, wie beispielsweise "`_Feuerwaffen, Waffen/Gewehre`". Die Stichwortzelle kann ein oder mehrere Wörter enthalten, die zur Beschreibung jeder Kategorie verwendet werden. Diese Wörter werden als Deskriptoren importiert oder ignoriert, abhängig von Ihrer Angabe im letzten Schritt des Assistenten. Später werden Deskriptoren mit den extrahierten Ergebnissen aus dem Text verglichen. Liegt eine Übereinstimmung vor, wird der Datensatz bzw. das Dokument der Kategorie zugewiesen, die diesen Deskriptor enthält.

Tabelle 10-9
Beispiel für kompaktes Format mit Codes

Spalte A	**Spalte B**	**Spalte C**
Hierarchische Code-Ebene	Kategoriecode (*optional*)	Kategoriename
Hierarchische Code-Ebene	Unterkategoriecode (*optional*)	Unterkategoriename

Tabelle 10-10
Beispiel für kompaktes Format ohne Codes

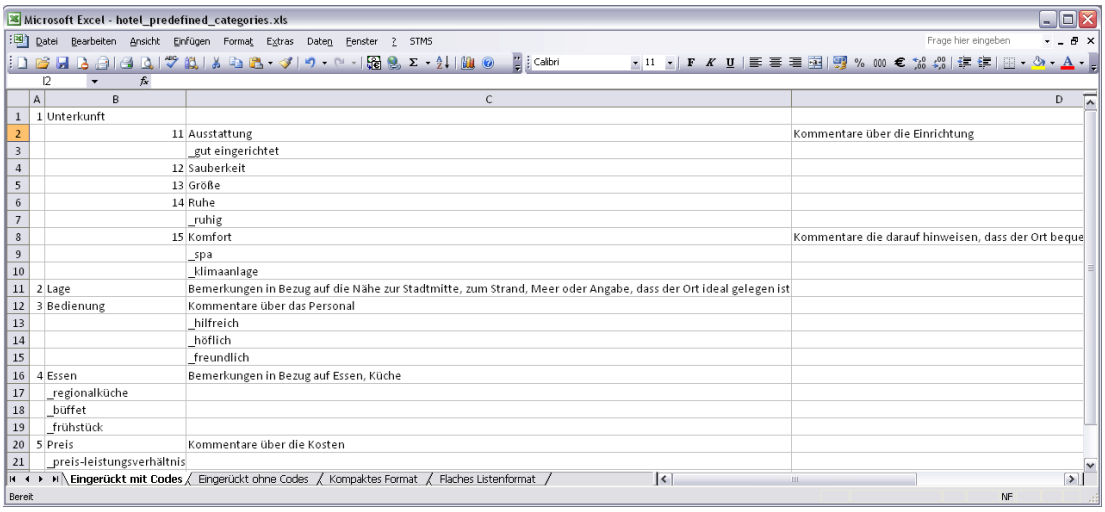
Spalte A	**Spalte B**
Hierarchische Code-Ebene	Kategoriename
Hierarchische Code-Ebene	Unterkategoriename

Eingerücktes Format

Im Dateiformat "Eingerückt" ist der Inhalt hierarchisch strukturiert, d. h., die Datei enthält Kategorien und eine oder mehrere Ebenen von Unterkategorien. Darüber hinaus ist die Struktur ihrer Hierarchie entsprechend eingerückt. Jede Zeile in der Datei enthält entweder eine Kategorie oder eine Unterkategorie, dabei sind Unterkategorien weiter eingerückt als die Kategorien,

Unter-Unterkategorien sind weiter eingerückt als die Unterkategorien usw. Sie können diese Struktur manuell in Microsoft Excel erstellen oder eine Struktur verwenden, die aus einem anderen Produkt exportiert und in einem Microsoft Excel-Format gespeichert wurde.

Abbildung 10-22
Beispiel einer eingerückten Kategorie in Microsoft Excel



	A	B	C	D
1	1	Unterkunft		
2		11	Ausstattung	Kommentare über die Einrichtung
3			_gut eingerichtet	
4		12	Sauberkeit	
5		13	Größe	
6		14	Ruhe	
7			_ruhig	
8		15	Komfort	Kommentare die darauf hinweisen, dass der Ort beque
9			_spa	
10			_klimaanlage	
11	2	Lage	Bemerkungen in Bezug auf die Nähe zur Stadtmitte, zum Strand, Meer oder Angabe, dass der Ort ideal gelegen ist	
12	3	Bedienung	Kommentare über das Personal	
13			_hilfreich	
14			_höflich	
15			_freundlich	
16	4	Essen	Bemerkungen in Bezug auf Essen, Küche	
17		_regionalküche		
18		_büffet		
19		_frühstück		
20	5	Preis	Kommentare über die Kosten	
21		_preis-leistungsverhältnis		

- **Kategoriecodes und Kategorienamen der ersten Ebene** belegen die Spalten A bzw. B. Falls keine Codes vorhanden sind, befindet sich der Kategoriename in Spalte A.
- **Unterkategoriecodes und Unterkategorienamen** belegen die Spalten B bzw. C. Falls keine Codes vorhanden sind, befindet sich der Unterkategoriename in Spalte B. Die Unterkategorie gehört einer Kategorie an. Unterkategorien können nur vorhanden sein, wenn Kategorien in der ersten Ebene vorhanden sind.

Tabelle 10-11
Eingerückte Struktur mit Codes

Spalte A	**Spalte B**	**Spalte C**	**Spalte D**
Kategoriecode (*optional*)	Kategoriename		
	Unterkategoriecode (*optional*)	Unterkategoriename	
		Unter-Unterkategoriecode (*optional*)	Unter-Unterkategoriename

Tabelle 10-12
Eingerückte Struktur ohne Codes

Spalte A	**Spalte B**	**Spalte C**
Kategoriename		
	Unterkategoriename	
		Unter-Unterkategoriename

Die folgenden Informationen können sich in einer Datei dieses Formats befinden:

- Optionale **Codes** müssen Werte sein, die jede Kategorie oder Unterkategorie eindeutig identifizieren. Wenn Sie angeben, dass die Datendatei Codes enthält (Option Enthält Kategoriecodes im Schritt Inhaltseinstellungen), muss in der Zelle direkt links neben dem

Kategorie-/Unterkategorienamen ein eindeutiger Code für jede Kategorie bzw. Unterkategorie vorhanden sein. Wenn Ihre Daten keine Codes enthalten, Sie jedoch später Codes erstellen möchten, können Sie diese jederzeit tun (Kategorien > Kategorien verwalten > Codes automatisch erzeugen).

- Ein *erforderlicher* **Name** für jede Kategorie und Unterkategorie. Unterkategorien müssen um eine Zelle nach rechts unter den Kategorien in einer separaten Zeile eingerückt sein.
- Optionale **Anmerkungen** in der Zelle direkt rechts neben dem Kategorienamen. Diese Anmerkung besteht aus Text, der Ihre Kategorien/Unterkategorien beschreibt.
- Optionale **Stichwörter** können als Deskriptoren für Kategorien importiert werden. Damit sie erkannt werden, müssen diese Stichwörter in der Zelle direkt unter dem verknüpften Kategorie- bzw. Unterkategorienamen vorhanden sein und der Liste der Stichwörter muss ein Unterstrich (_) vorangestellt sein, wie beispielsweise "`_Feuerwaffen, Waffen/Gewehre`". Die Stichwortzelle kann ein oder mehrere Wörter enthalten, die zur Beschreibung jeder Kategorie verwendet werden. Diese Wörter werden als Deskriptoren importiert oder ignoriert, abhängig von Ihrer Angabe im letzten Schritt des Assistenten. Später werden Deskriptoren mit den extrahierten Ergebnissen aus dem Text verglichen. Liegt eine Übereinstimmung vor, wird der Datensatz bzw. das Dokument der Kategorie zugewiesen, die diesen Deskriptor enthält.

Wichtig: Wenn Sie auf einer Ebene einen Code verwenden, müssen Sie einen Code für jede Kategorie und Unterkategorie angeben. Andernfalls schlägt der Importvorgang fehl.

Exportieren von Kategorien

Sie können die vorhandenen Kategorien in einer geöffneten interaktiven Workbench-Sitzung in ein Microsoft Excel-Dateiformat (**.xls*, **.xlsx*) exportieren. Die exportierten Daten stammen im Wesentlichen aus dem aktuellen Inhalt des Bereichs "Kategorie" bzw. aus den Kategorieeigenschaften. Daher wird ein erneutes Scoring empfohlen, wenn Sie auch den Score-Wert Docs. exportieren möchten.

Immer exportieren...

- Kategoriecodes, falls vorhanden
- Kategorienamen (und Unterkategorienamen)
- Codeebenen, falls vorhanden (*Flaches*/*Kompaktes* Format)
- Spaltenüberschriften (*Flaches*/*Kompaktes* Format)

Optional exportieren...

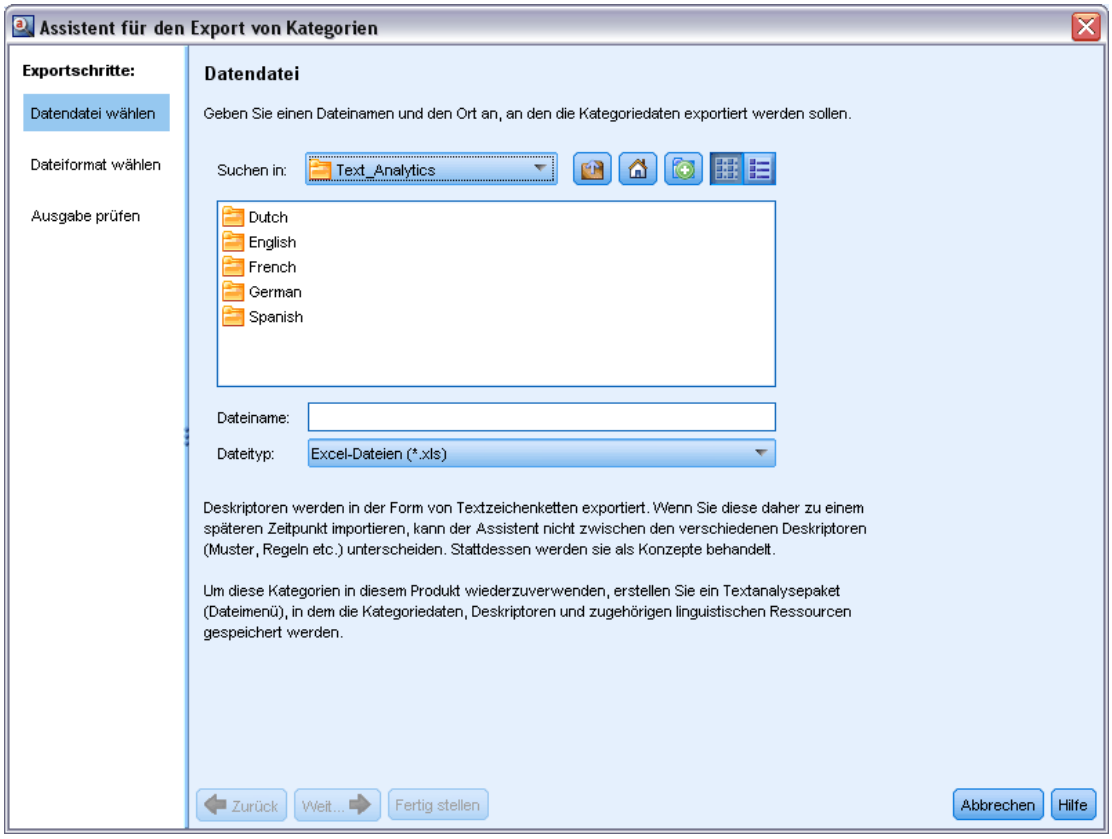
- Docs.-Scores
- Kategorieanmerkungen
- Deskriptornamen
- Deskriptoranzahl

Wichtig: Wenn Sie Deskriptoren exportieren, werden diese in Textzeichenfolgen umgewandelt und ihnen wird ein Unterstrich vorangestellt. Wenn Sie diese erneut in dieses Produkt importieren, kann nicht mehr zwischen Deskriptoren, bei denen es sich um Muster, Kategorieregeln oder reguläre Konzepte handelt, unterschieden werden. Wenn Sie diese Kategorien in diesem Produkt erneut verwenden möchten, wird dringend empfohlen, stattdessen eine Textanalysepaket- (TAP-) Datei zu erstellen, da das TAP-Format alle aktuell definierten Deskriptoren sowie all Ihre Kategorien, Codes und die verwendeten linguistischen Ressourcen speichert. TAP-Dateien können sowohl in IBM® SPSS® Modeler Text Analytics als auch in IBM® SPSS® Text Analytics for Surveys verwendet werden. Für weitere Informationen siehe Thema Verwendung von Text Analysis Packages auf S. 230.

So exportieren Sie vordefinierte Kategorien:

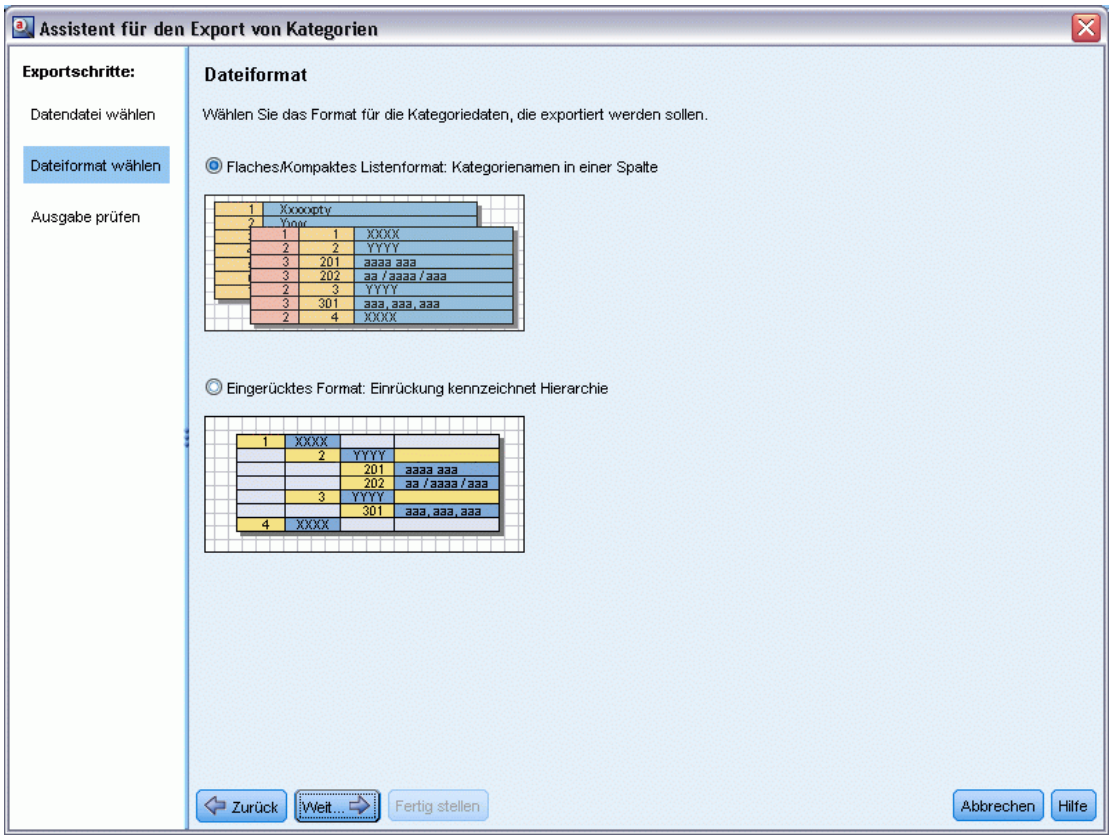
- Wählen Sie in den Menüs die Optionsfolge Kategorien > Kategorien verwalten > Kategorien exportieren. Ein Assistent für den Export von Kategorien wird angezeigt.

Abbildung 10-23
Assistent "Kategorien exportieren", Schritt 1



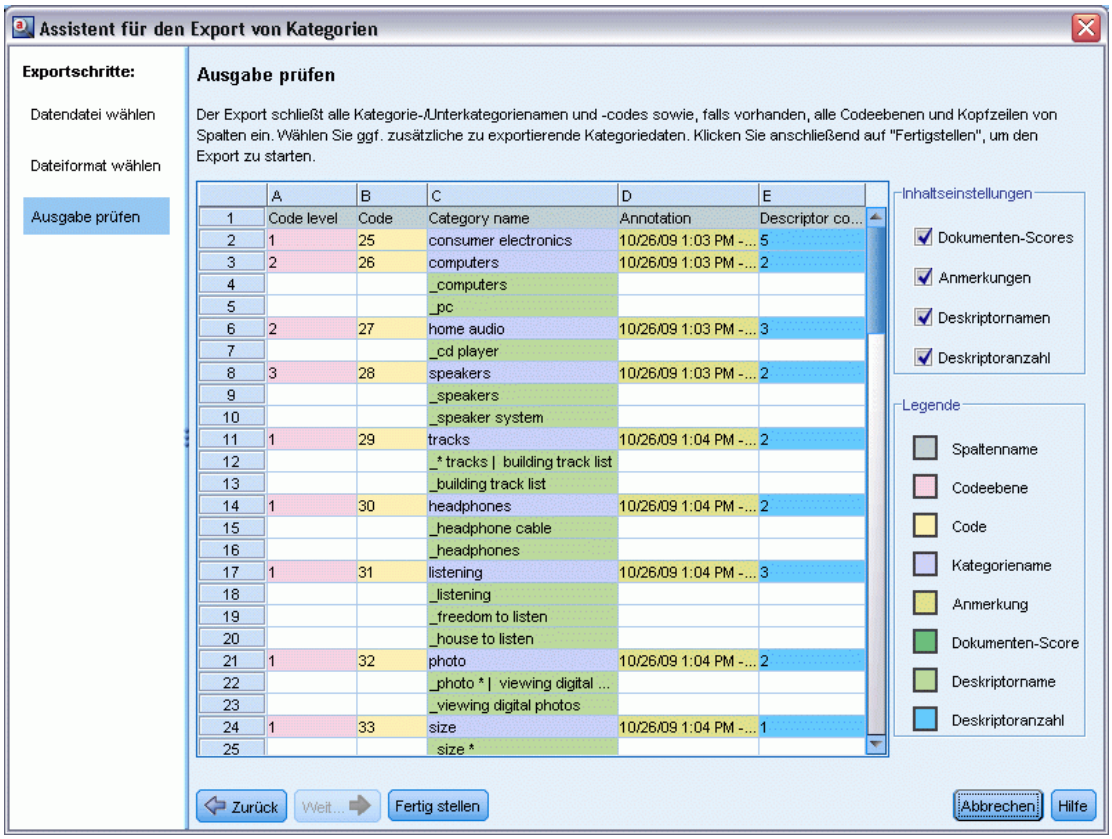
- Wählen Sie den Speicherort aus und geben Sie den Namen der Datei ein, die exportiert wird.
- Geben Sie einen Namen für die Ausgabedatei in das Textfeld "Dateiname" ein.
- Klicken Sie auf Weiter, um das Format zu wählen, in das Sie Ihre Kategoriendaten exportieren möchten.

Abbildung 10-24
Assistent "Kategorien exportieren", Schritt 2



- Wählen Sie das Format aus den folgenden Optionen:
  - Einfaches oder kompaktes Listenformat: Für weitere Informationen siehe Thema Einfaches Listenformat auf S. 223. Einfache Liste enthält keine Unterkategorien. Für weitere Informationen siehe Thema Kompaktes Format auf S. 224. Kompaktes Listenformat enthält hierarchische Kategorien.
  - Eingerücktes Format: Für weitere Informationen siehe Thema Eingerücktes Format auf S. 225.
- Klicken Sie auf Weiter, um mit der Auswahl des zu exportierenden Inhalts zu beginnen und die vorgeschlagenen Daten zu prüfen.

Abbildung 10-25
Assistent "Kategorien exportieren", Schritt 3



	A	B	C	D	E
1	Code level	Code	Category name	Annotation	Descriptor co...
2	1	25	consumer electronics	10/26/09 1:03 PM -...	5
3	2	26	computers	10/26/09 1:03 PM -...	2
4			_computers		
5			_pc		
6	2	27	home audio	10/26/09 1:03 PM -...	3
7			_cd player		
8	3	28	speakers	10/26/09 1:03 PM -...	2
9			_speakers		
10			_speaker system		
11	1	29	tracks	10/26/09 1:04 PM -...	2
12			_* tracks | building track list		
13			_building track list		
14	1	30	headphones	10/26/09 1:04 PM -...	2
15			_headphone cable		
16			_headphones		
17	1	31	listening	10/26/09 1:04 PM -...	3
18			_listening		
19			_freedom to listen		
20			_house to listen		
21	1	32	photo	10/26/09 1:04 PM -...	2
22			_photo * | viewing digital ...		
23			_viewing digital photos		
24	1	33	size	10/26/09 1:04 PM -...	1
25			_size *		

- Prüfen Sie den Inhalt für die exportierte Datei.
- Aktivieren oder deaktivieren Sie zusätzliche zu exportierende Inhaltseinstellungen wie Anmerkungen oder Deskriptornamen..
- Klicken Sie auf Fertigstellen, um die Kategorien zu exportieren.

Verwendung von Text Analysis Packages

Ein Text Analysis Package, auch TAP genannt, dient als Vorlage für die Kategorisierung von Textdaten. Der Einsatz eines TAP ist eine einfache Methode zur Kategorisierung Ihrer Textdaten, die minimalen Benutzereingriff erfordert, da das TAP die vordefinierten Kategoriensets und die linguistischen Ressourcen enthält, die zur schnellen und automatischen Kodierung einer großen Anzahl von Datensätzen nötig sind. Mit Hilfe der linguistischen Ressourcen werden Textdaten analysiert und dem Mining-Verfahren unterzogen, um die wichtigsten Konzepte zu extrahieren. Auf Basis von wichtigen Konzepten und Mustern, die im Text gefunden werden, können die Datensätze dem Kategorienset zugeordnet werden, das Sie im TAP ausgewählt haben. Sie können Ihr eigenes TAP erstellen oder ein TAP aktualisieren.

Ein TAP besteht aus folgenden Elementen:

- **Kategorienset(s)**. Ein Kategorie-Set besteht im Wesentlichen aus vordefinierten Kategorien, Kategoriecodes, Deskriptoren für jede Kategorie und schließlich aus einem Namen für das gesamte Kategorie-Set. Deskriptoren sind linguistische Elemente (Konzepte, Typen, Muster und Regeln), so z. B. der Begriff *billig* oder das Muster *guter Preis*. Deskriptoren werden verwendet, um eine Kategorie zu definieren, sodass das Dokument oder der Datensatz dieser Kategorie zugeordnet wird, wenn der Text mit einem Kategoriedeskriptor übereinstimmt.
- **Linguistische Ressourcen**. Linguistische Ressourcen sind eine Reihe von Bibliotheken und erweiterten Ressourcen, die darauf abgestimmt sind, wichtige Konzepte und Muster zu extrahieren. Diese Extrahierungskonzepte und -muster wiederum werden als Deskriptoren verwendet, die es ermöglichen, Datensätze einer Kategorie im Kategorienset zuzuordnen.

Sie können Ihr eigenes TAP erstellen, ein TAP aktualisieren oder Text Analysis Packages laden.

Nachdem das TAP ausgewählt und ein Kategorie-Set gewählt wurde, kann IBM® SPSS® Modeler Text Analytics Ihre Datensätze extrahieren und kategorisieren.

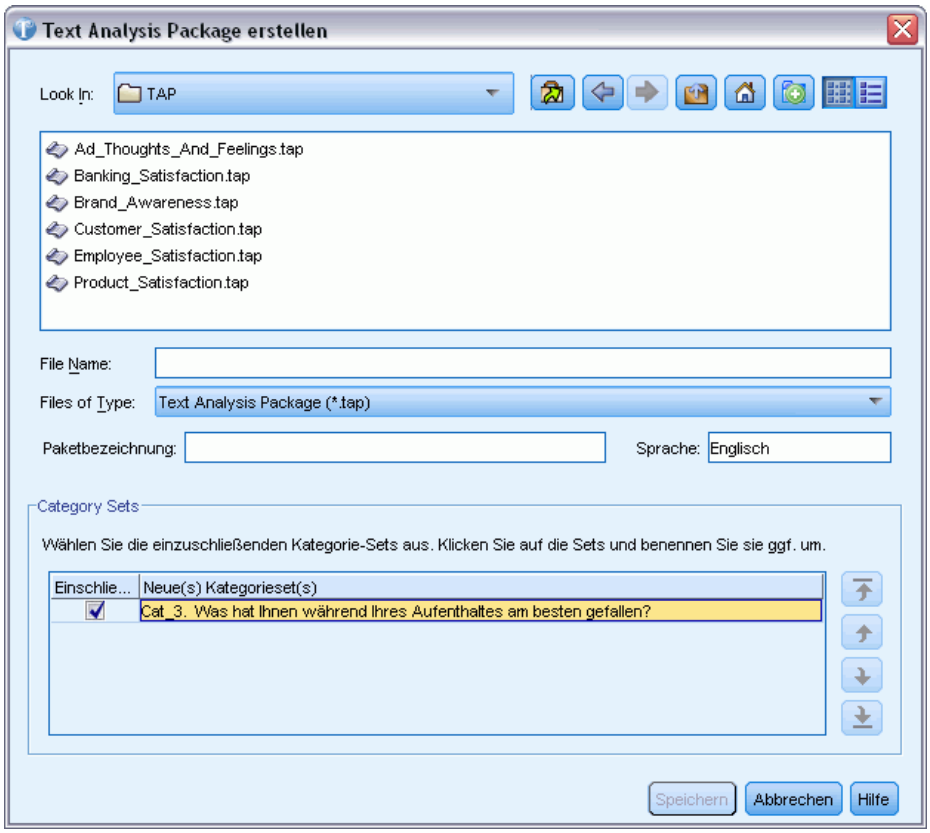
Anmerkung: TAPs können austauschbar zwischen IBM® SPSS® Text Analytics for Surveys und SPSS Modeler Text Analytics verwendet werden.

Erstellung von Text Analysis Packages

Wenn Sie eine sitzung mit mindestens einer Kategorie und einigen Ressourcen haben, können Sie aus dem Inhalt der offenen interaktiven Workbench-Sitzung ein Text Analysis Package (TAP) erstellen. Die Menge der Kategorien und Deskriptoren (Konzepte, Typen, Regeln oder TLA-Musterausgaben) kann zusammen mit allen linguistischen Ressourcen, die im Ressourceneditor geöffnet sind, zur Erstellung eines TAP verwendet werden.

Sie können die Sprache sehen, für die die Ressourcen erstellt wurden. Die Sprache wird auf der Registerkarte “Erweiterte Ressourcen” von Template Editor oder Resource Editor eingestellt.

Abbildung 10-26
Dialogfeld "Text Analysis Package erstellen"

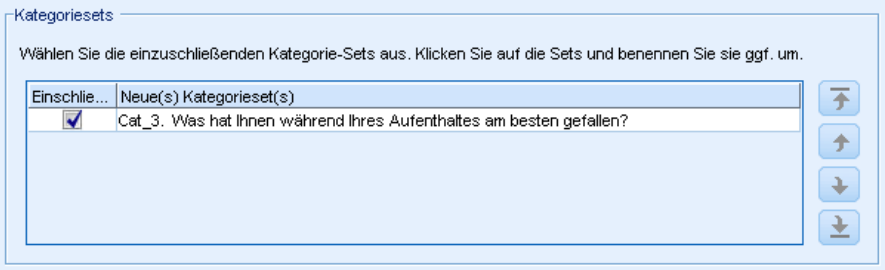


So erstellen Sie ein Text Analysis Package

- Wählen Sie in den Menüs die Optionsfolge Datei > Text Analysis Packages > Paket erstellen. Das Dialogfeld "Paket erstellen" erscheint.
- Wechseln Sie zum Verzeichnis, in dem Sie das TAP speichern möchten. Standardmäßig werden TAPs im Unterverzeichnis \TAP des Produktinstallationsverzeichnisses gespeichert.
- Geben Sie im Feld Dateiname einen Namen für das TAP ein.
- Geben Sie im Feld Paketbezeichnung eine Bezeichnung ein. Wenn Sie einen Dateinamen eingeben, erscheint dieser Name automatisch als Bezeichnung, die Sie aber ändern können.
- Um ein Kategorienset aus dem TAP auszuschließen, deaktivieren Sie das Kontrollkästchen Einschließen. Dadurch wird sichergestellt, dass es Ihrem Paket nicht hinzugefügt wird. Standardmäßig ist ein Kategorienset pro Frage im TAP enthalten. Im TAP muss sich immer mindestens ein Kategorienset befinden.
- Umbenennen von Kategoriensets. Die Spalte Neues Kategorienset enthält standardmäßig generische Namen, die durch Hinzufügen des Präfix Cat_ zum Textvariablennamen generiert werden. Durch einmaliges Klicken in die Zelle können Sie den Namen bearbeiten. Durch Drücken der Eingabetaste oder Klicken an einer anderen Stelle wird die Umbenennung angewendet. Wenn

Sie ein Kategorienset umbenennen, ändert sich der Name nur im TAP; der Variablenname in der offenen sitzung wird nicht geändert.

Abbildung 10-27
Kategoriensets umbenennen



- Falls gewünscht, ändern Sie die Reihenfolge der Kategoriensets mit Hilfe der Pfeiltasten rechts von der Tabelle “Kategorienset”.
- Klicken Sie auf Speichern, um das Text Analysis Package zu erstellen. Das Dialogfeld wird geschlossen.

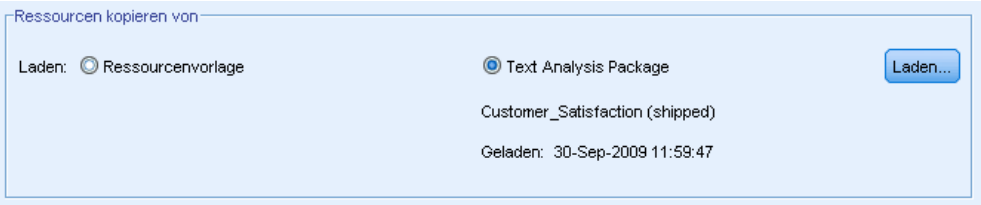
Laden von Text Analysis Packages

Bei der Konfiguration eines Text-Mining-Modellierungsknotens müssen Sie angeben, welche Ressourcen für die Extrahierung verwendet werden. Anstatt eine Ressourcenvorlage auszuwählen, können Sie ein Text Analysis Package (TAP) auswählen, um nicht nur seine Ressourcen, sondern auch ein Kategorienset in den Knoten zu kopieren.

TAPs sind vor allem für die interaktive Erstellung eines Kategorienmodells sinnvoll, da Sie das Kategorienset als Ausgangspunkt für die Kategorisierung verwenden können. Wenn Sie den Stream ausführen, wird die interaktive Workbench-Sitzung gestartet und dieses Kategorienset im Bereich “Kategorien” angezeigt. So können Sie Ihre Dokumente und Datensätze unmittelbar mithilfe dieser Kategorien scoren und diese Kategorien anschließend verfeinern, erstellen und erweitern, bis sie Ihre Anforderungen erfüllen. Für weitere Informationen siehe Thema Methoden und Strategien zur Erstellung von Kategorien auf S. 171.

Ab Version 14 können Sie auch die Sprache sehen, für die die Ressourcen in diesem TAP definiert wurden, wenn Sie auf Laden klicken und das TAP auswählen.

Abbildung 10-28
Optionen auf der Registerkarte “Modell” für das Kopieren von Ressourcen in den Knoten



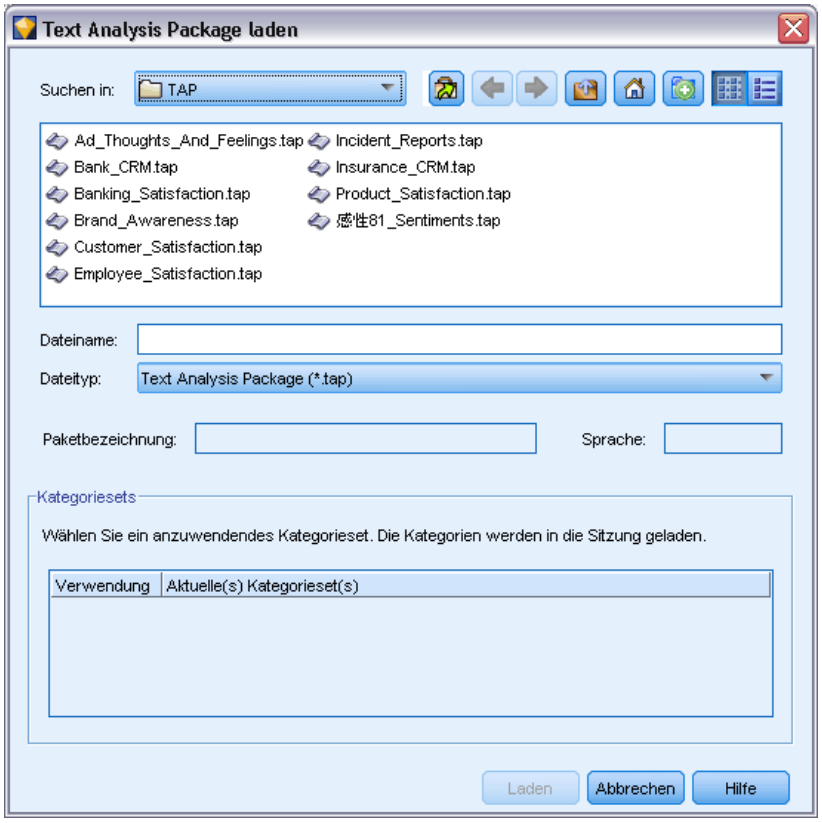
So laden Sie ein Text Analysis Package:

- Bearbeiten Sie den Text-Mining-Modellierungsknoten.

- Wählen Sie auf der Registerkarte “Modell” im Abschnitt Ressourcen kopieren von die Option *Text Analysis Package*.

- Klicken Sie auf Laden. Das Dialogfeld “Text Analysis Package laden” wird angezeigt.

Abbildung 10-29
Dialogfeld “Text Analysis Package laden”



- Wechseln Sie an den Speicherort des TAP mit den Ressourcen und dem Kategorie-Set, die Sie in diesen Knoten kopieren möchten. Standardmäßig werden TAPs im Unterverzeichnis \TAP des Produktinstallationsverzeichnisses gespeichert.

- Geben Sie im Feld Dateiname einen Namen für das TAP ein. Die Bezeichnung wird automatisch angezeigt.

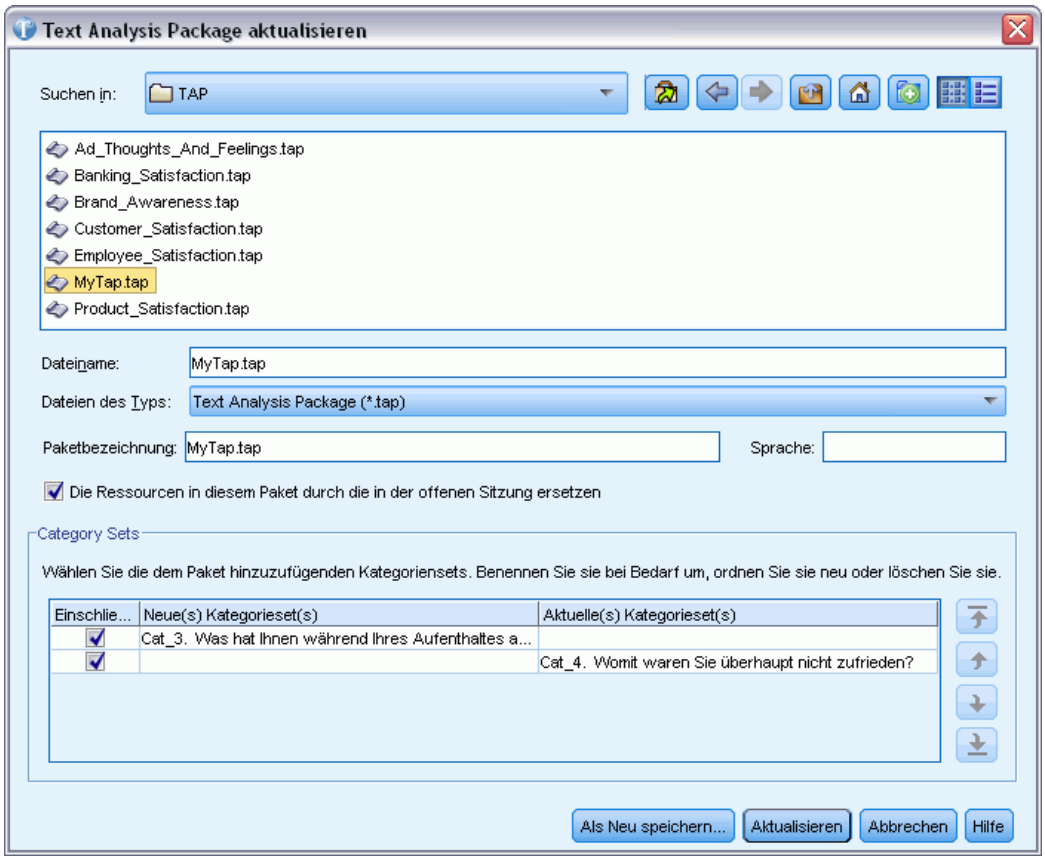
- Wählen Sie das Kategorienset aus, das Sie verwenden möchten. Dieses Kategorienset wird in der interaktiven Workbench-Sitzung angezeigt. Sie können diese Kategorien anschließend manuell oder über die Optionen zum Aufbauen und Erweitern von Kategorien optimieren und verbessern.

- Klicken Sie auf Laden, um den Inhalt des Text Analysis Package in den Knoten zu kopieren. Das Dialogfeld wird geschlossen. Wenn ein TAP geladen wird, wird eine Kopie des TAP in den Knoten kopiert; daher werden Änderungen, die Sie an Ressourcen und Kategorien vornehmen, nur dann in das TAP übernommen, wenn Sie es gesondert aktualisieren und neu laden.

Aktualisierung von Text Analysis Packages

Um Verbesserungen an einem Kategorie-Set oder an linguistischen Ressourcen vorzunehmen oder um ein völlig neues Kategorienset zu erstellen, können Sie ein Text Analysis Package (TAP) aktualisieren, um die spätere Nutzung dieser Verbesserungen zu erleichtern. Dafür müssen Sie die sitzung geöffnet haben, die die Informationen enthält, die Sie in das TAP aufnehmen möchten. Beim Aktualisieren können Sie wählen, ob Sie Kategoriensets anfügen, Ressourcen ersetzen, die Paketbezeichnung ändern oder Kategoriensets umbenennen/neu ordnen möchten.

Abbildung 10-30
Dialogfeld "Text Analysis Package aktualisieren"



So aktualisieren Sie ein Text Analysis Package

- Wählen Sie in den Menüs die Optionsfolge Datei > Text Analysis Packages > Paket aktualisieren. Das Dialogfeld "Text Analysis Package aktualisieren" wird angezeigt.
- Wechseln Sie in das Verzeichnis, das das Text Analysis Package enthält, das Sie aktualisieren möchten.
- Geben Sie im Feld Dateiname einen Namen für das TAP ein.
- Um die linguistischen Ressourcen im TAP durch die in der aktuellen sitzung zu ersetzen, wählen Sie die Option Ressourcen in diesem Paket durch die in der offenen Sitzung ersetzen. Es macht für

gewöhnlich Sinn, die linguistischen Ressourcen zu aktualisieren, da sie dazu verwendet wurden, die wichtigen Konzepte und Muster zu extrahieren, mit denen die Kategoriedefinitionen erstellt wurden. Durch die Verwendung der aktuellen linguistischen Ressourcen können Sie sicherstellen, dass Sie bei der Kategorisierung Ihrer Datensätze die besten Ergebnisse erzielen. Wenn Sie diese Option nicht auswählen, bleiben die linguistischen Ressourcen, die bereits im Paket vorhanden waren, unverändert.

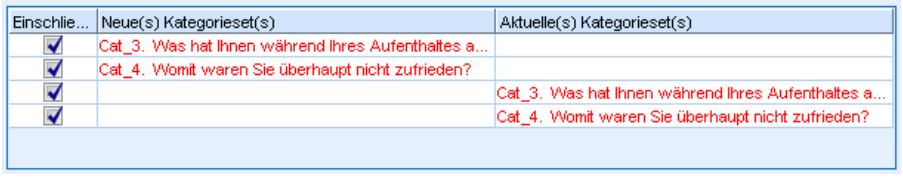
- Um nur die linguistischen Ressourcen zu aktualisieren, stellen Sie sicher, dass die Option Ressourcen in diesem Paket durch die in der offenen Sitzung ersetzen aktiviert ist und Sie nur die aktuellen Kategoriensets auswählen, die bereits im TAP vorhanden waren.

- Um das (die) neue(n) Kategorie-Set aus der offenen sitzung in das TAP zu übernehmen, aktivieren Sie das Kontrollkästchen für jede Kategorie, die hinzugefügt werden soll. Sie können ein Kategorienset, mehrere oder keine Kategoriensets hinzufügen.

- Um Kategoriensets aus dem TAP zu entfernen, deaktivieren Sie die jeweiligen Einschließen-Kontrollkästchen. Unter Umständen möchten Sie ein Kategorienset, das bereits im TAP vorhanden war, entfernen, da Sie ein verbessertes Set hinzufügen. Deaktivieren Sie zu diesem Zweck das Kontrollkästchen Einschließen für das jeweilige Kategorienset in der Spalte "Aktuelles Kategorienset". Im TAP muss sich immer mindestens ein Kategorienset befinden.

- Falls nötig, benennen Sie die Kategoriensets um. Durch einmaliges Klicken in die Zelle können Sie den Namen bearbeiten. Durch Drücken der Eingabetaste oder Klicken an einer anderen Stelle wird die Umbenennung angewendet. Wenn Sie ein Kategorienset umbenennen, ändert sich der Name nur im TAP; der Variablenname in der offenen sitzung wird nicht geändert. Wenn zwei Kategoriensets denselben Namen haben, werden ihre Namen rot angezeigt, bis Sie das Duplikat korrigieren.

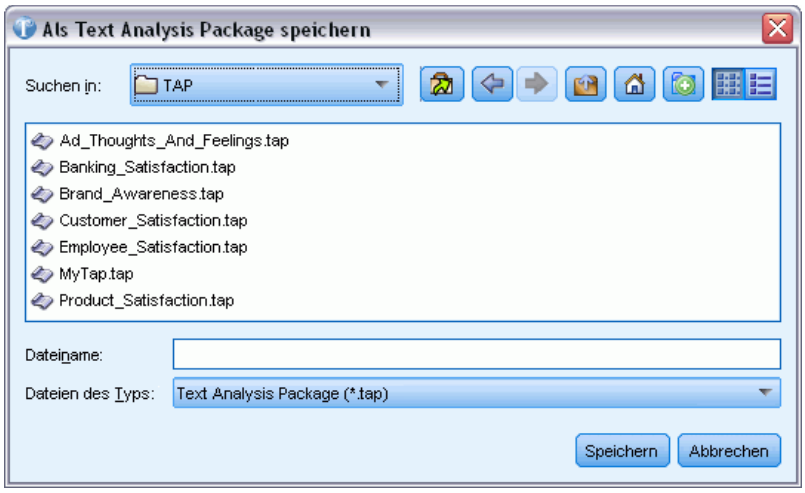
Abbildung 10-31
Doppelt vorhandene Namen



- Um ein neues Paket mit den Inhalten der Sitzung und den Inhalten des ausgewählten TAP zu erstellen, klicken Sie auf Als neu speichern. Das Dialogfeld "Als Text Analysis Package speichern" wird angezeigt. Beachten Sie die folgenden Anweisungen.

- Klicken Sie auf Aktualisieren, um die Änderungen, die Sie am ausgewählten TAP vorgenommen haben, zu speichern.

Abbildung 10-32
Dialogfeld "Als Text Analysis Package speichern"



So speichern Sie ein Text Analysis Package

- Wechseln Sie zum Verzeichnis, in dem Sie die TAP-Datei speichern möchten. Standardmäßig werden TAP-Dateien im Unterverzeichnis "TAP" des Installationsverzeichnisses gespeichert.
- Geben Sie im Feld "Dateiname" einen Namen für die TAP-Datei ein.
- Geben Sie im Feld "Paketbezeichnung" eine Bezeichnung ein. Wenn Sie einen Dateinamen eingeben, wird dieser Name automatisch als Bezeichnung verwendet. Sie können diese Bezeichnung jedoch umbenennen. Sie müssen eine Bezeichnung haben.
- Klicken Sie auf Speichern, um das neue Paket zu erstellen.

Bearbeiten und Verfeinern von Kategorien

Nachdem Sie einige Kategorien erstellt haben, sollten Sie diese untersuchen und Anpassungen vornehmen. Neben der Verfeinerung der linguistischen Ressourcen sollten Sie Ihre Kategorien überprüfen, indem Sie nach Möglichkeiten suchen, die zugehörigen Definitionen zu kombinieren oder zu bereinigen, und einige der kategorisierten Dokumente bzw. Datensätze prüfen. Außerdem können Sie die Dokumente bzw. Datensätze in einer Kategorie überprüfen und Anpassungen vornehmen, sodass die Kategorien so definiert sind, dass Nuancen und Unterschiede erfasst werden.

Sie können die integrierten, automatisierten Kategorieerstellungsverfahren zum Erstellen Ihrer Kategorien verwenden; Sie werden jedoch wahrscheinlich noch einige Optimierungen an diesen Kategorien vornehmen wollen. Nach der Verwendung eines oder mehrerer Verfahren wird eine Reihe neuer Kategorien im Fenster angezeigt. Anschließend können Sie die Daten in einer Kategorie überprüfen und Anpassungen vornehmen, bis Sie mit den Kategoriedefinitionen zufrieden sind. Für weitere Informationen siehe Thema Erläuterung von Kategorien auf S. 178.

Hier sind einige Optionen zur Verfeinerung Ihrer Kategorien, von denen die meisten auf den folgenden Seiten beschrieben werden:

- Hinzufügen von Deskriptoren zu Ihren Kategorien
- Bearbeiten von Kategorien
- Verschieben von Kategorien
- Glätten hierarchischer Kategorien
- Zusammenführen von Kategorien
- Löschen von Kategorien
- Vornehmen von Änderungen an den linguistischen Ressourcen und erneutes Extrahieren
- Visualisieren, wie die Kategorien zusammenarbeiten, und Vornehmen von Anpassungen. Für weitere Informationen siehe Thema Kategoriendiagramme und Grafiken in Kapitel 13 auf S. 262.

Hinzufügen von Deskriptoren zu Kategorien

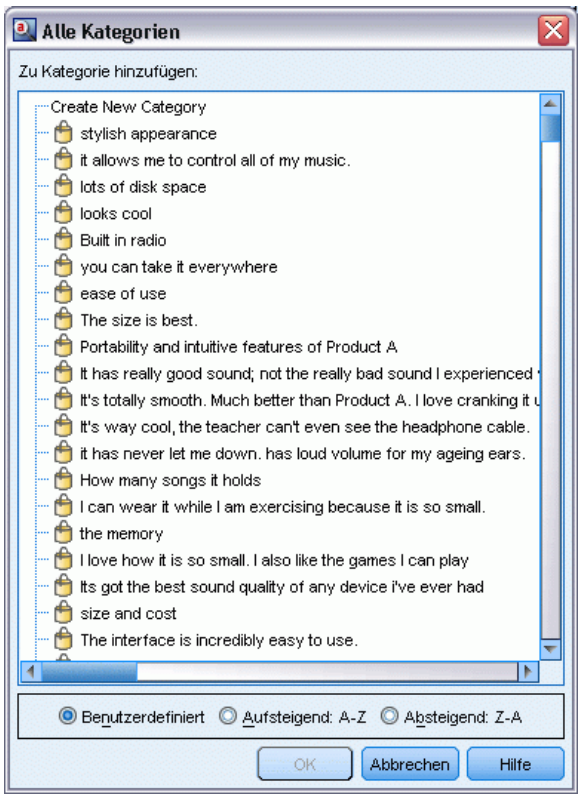
Nach der Verwendung von automatisierten Verfahren liegen Ihnen mit großer Wahrscheinlichkeit noch Extrahierungsergebnisse vor, die in keiner der Kategoriedefinitionen verwendet wurden. Sie sollten die betreffende Liste im Fensterbereich “Extrahierungsergebnisse” durchgehen. Wenn Sie Elemente finden, die Sie in eine Kategorie verschieben möchten, können Sie diese zu einer bestehenden oder zu einer neuen Kategorie hinzufügen.

So fügen Sie ein Konzept bzw. einen Typ zu einer Kategorie hinzu:

▶ Wählen Sie in den Fensterbereichen “Extrahierungsergebnisse” und “Daten” die Elemente, die Sie einer neuen oder einer bestehenden Kategorie hinzufügen möchten.

▶ Wählen Sie in den Menüs Kategorien > Zu Kategorie hinzufügen. Im Dialogfeld “Alle Kategorien” wird das Kategorie-Set angezeigt. Wählen Sie die Kategorie aus, zu der die ausgewählten Elemente hinzugefügt werden sollen. Wenn Sie die Elemente zu einer neuen Kategorie hinzufügen möchten, wählen Sie Neue Kategorie. Im Fensterbereich “Kategorien” wird eine neue Kategorie angezeigt, für die der Name des ersten ausgewählten Elements verwendet wird.

Abbildung 10-33
Dialogfeld "Alle Kategorien"



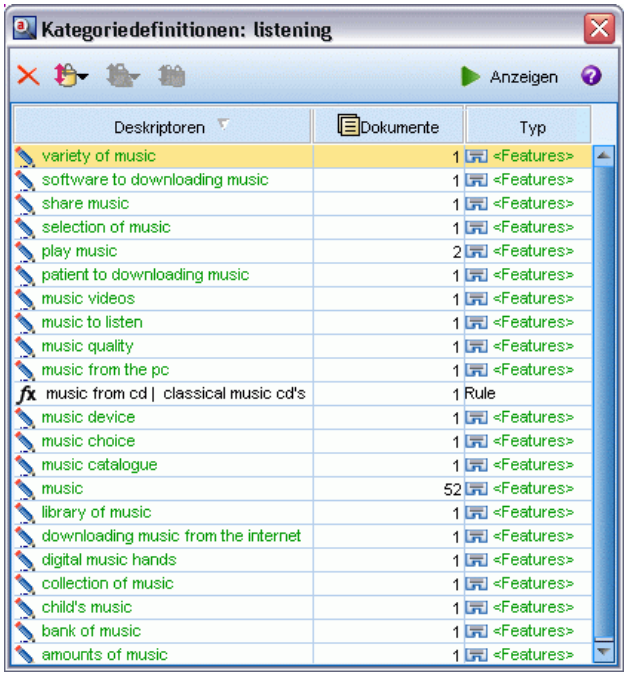
Bearbeiten von Kategoriedeskriptoren

Nachdem Sie einige Kategorien erstellt haben, können Sie die einzelnen Kategorien öffnen, um alle Deskriptoren anzuzeigen, aus denen die Definition der jeweiligen Kategorie besteht. Im Dialogfeld "Kategoriedefinitionen" können Sie eine Reihe von Bearbeitungsschritten an den Kategoriedeskriptoren vornehmen. Außerdem können Sie mit ihnen dort arbeiten, wenn Kategorien im Kategoriebaum angezeigt werden.

So bearbeiten Sie eine Kategorie:

- Wählen Sie die zu bearbeitende Kategorie im Fensterbereich "Kategorien" aus.
- Wählen Sie Ansicht > Kategoriedefinitionen. Das Dialogfeld "Kategoriedefinitionen" wird geöffnet.

Abbildung 10-34
Dialogfeld "Kategoriedefinitionen"



- Wählen Sie den zu bearbeitenden Deskriptor aus und klicken Sie auf die entsprechende Schaltfläche auf der Symbolleiste.

In der folgenden Tabelle sind die einzelnen Symbolleistenschaltflächen beschrieben, mit denen Sie die Kategoriedefinitionen bearbeiten können.

Tabelle 10-13
Symbolleistenschaltflächen und Beschreibungen

Symbole	Beschreibung
	Löscht die ausgewählten Deskriptoren aus der Kategorie .
	Verschiebt die ausgewählten Deskriptoren in eine neue oder bereits vorhandene Kategorie.
	Verschiebt die ausgewählten Deskriptoren in Form einer &-Kategorieregel in eine Kategorie. Für weitere Informationen siehe Thema Verwenden von Kategorieregeln auf S. 207.
	Verschiebt die einzelnen ausgewählten Deskriptoren jeweils als eigene neue Kategorie
 Anzeigen	Aktualisiert die Anzeige des Daten- und des Visualisierungsbereichs in Funktion der ausgewählten Deskriptoren

Verschieben von Kategorien

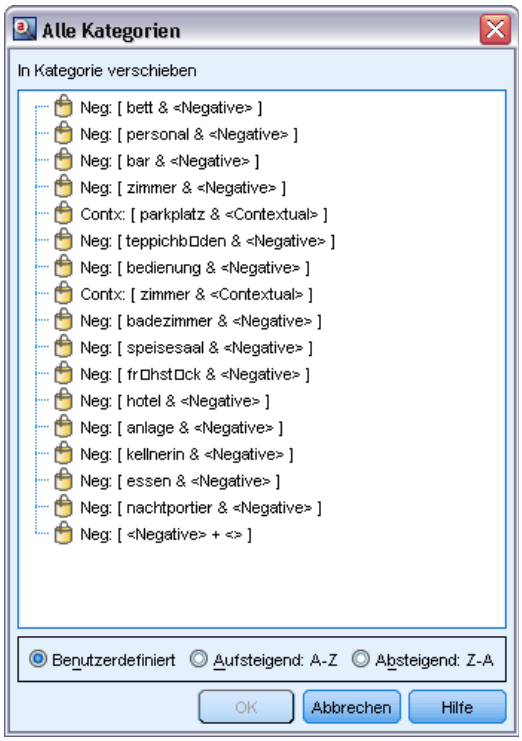
Wenn Sie eine Kategorie in eine andere vorhandene Kategorie einordnen oder Deskriptoren in eine andere Kategorie verschieben möchten, können Sie dies tun.

So verschieben Sie eine Kategorie:

- Wählen Sie im Fensterbereich “Kategorien” die Kategorien aus, die Sie in eine andere Kategorie verschieben möchten.

- Wählen Sie in den Menüs Kategorien > In Kategorie verschieben. Im Menü wird eine Menge von Kategorien angezeigt, wobei sich die zuletzt erstellte Kategorie ganz oben in der Liste befindet. Wählen Sie den Namen der Kategorie aus, in die die ausgewählten Konzepte verschoben werden sollen.
  - Wenn Sie den gesuchten Namen gefunden haben, wählen Sie ihn aus und die ausgewählten Elemente werden zu der betreffenden Kategorie hinzugefügt.
  - Wenn er nicht angezeigt wird, wählen Sie Weitere aus, um das Dialogfeld “Alle Kategorien” anzuzeigen, und wählen Sie die Kategorie in der Liste aus.

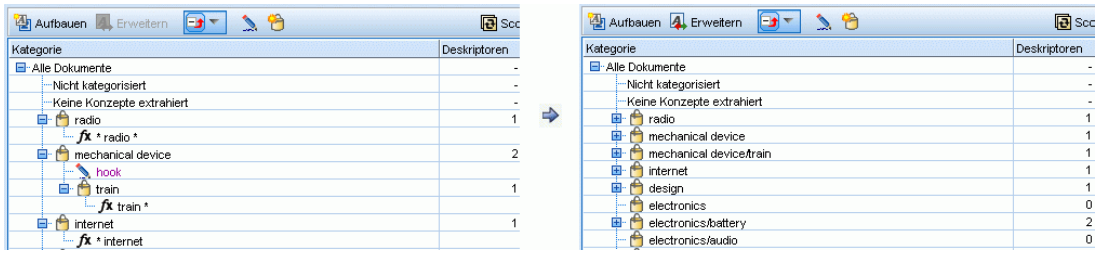
Abbildung 10-35
Dialogfeld “Alle Kategorien”



Glätten von Kategorien

Wenn Sie eine hierarchische Kategoriestruktur mit Kategorien und Unterkategorien haben, können Sie Ihre Struktur glätten. Wenn Sie eine Kategorie glätten, werden alle Deskriptoren in den Unterkategorien dieser Kategorie in die ausgewählte Kategorie verschoben und die nun leeren Unterkategorien werden gelöscht. Auf diese Weise werden alle Dokumente, die zuvor mit den Unterkategorien übereinstimmten, nun der ausgewählten Kategorie zugeordnet.

Abbildung 10-36
Geglättete Kategorien



So glätten Sie eine Kategorie:

- Wählen Sie im Kategoriebereich die Kategorie (oberste Ebene oder Unterkategorie) aus, die Sie glätten möchten.
- Wählen Sie in den Menüs die Optionsfolge Kategorien > Kategorien glätten. Die Unterkategorien werden entfernt und die Deskriptoren werden in die ausgewählte Kategorie zusammengeführt.

Zusammenführen bzw. Kombinieren von Kategorien

Um zwei oder mehr vorhandene Kategorien in einer neuen Kategorie zu kombinieren, können Sie sie zusammenführen. Wenn Sie Kategorien zusammenführen, wird eine Kategorie mit einem generischen Namen erstellt. Alle Konzepte, Ausdrücke und Muster, die in den Kategoriedeskriptoren verwendet werden, werden in diese neue Kategorie verschoben. Sie können die Kategorie später durch Bearbeiten der Kategorieeigenschaften umbenennen.

So führen Sie eine Kategorie bzw. einen Teil einer Kategorie zusammen:

- Wählen Sie im Fensterbereich “Kategorien” die Elemente aus, die zusammengeführt werden sollen.
- Wählen Sie in den Menüs Kategorien > Kategorien zusammenführen. Das Dialogfeld “Kategorieeigenschaften” wird angezeigt und Sie können einen Namen für die neu erstellte Kategorie eingeben. Die ausgewählten Kategorien werden als Unterkategorien in die ausgewählte Kategorie zusammengeführt.

Löschen von Kategorien

Wenn Sie eine Kategorie nicht mehr benötigen, können Sie sie löschen.

So löschen Sie eine Kategorie:

- Wählen Sie im Fensterbereich “Kategorien” die zu löschende(n) Kategorie(n) aus.
- Wählen Sie im Menü Bearbeiten > Löschen.

Kapitel
11

Analyse von Clustern

Konzeptcluster können Sie in der Clusteransicht erstellen und untersuchen (Ansicht > Cluster). Ein **Cluster** ist eine Gruppierung zusammengehöriger Konzepte, die durch Clusteringalgorithmen auf der Grundlage der Häufigkeit ihres Vorkommens im Dokument/Daten-Set sowie der Häufigkeit des gemeinsamen Vorkommens in demselben Dokument, auch als **Kookkurrenz** bezeichnet, generiert wurden. Jedes in einem Cluster enthaltene Konzept tritt mit mindestens einem anderen im Cluster enthaltenen Konzept gemeinsam auf. Cluster zielen darauf ab, Konzepte zu gruppieren, die gemeinsam auftreten. Kategorien hingegen zielen darauf ab, Dokumente oder Datensätze auf der Grundlage dessen zu gruppieren, wie der enthaltene Text den Deskriptoren (Konzepten, Regeln, Mustern) für jede Kategorie entspricht.

Ein guter Cluster enthält Konzepte, die einen starken Zusammenhang besitzen, häufig gemeinsam auftreten und nur wenig Zusammenhang mit in anderen Clustern enthaltenen Konzepten besitzen. Bei der Arbeit mit größeren Daten-Sets kann diese Technik erheblich längere Verarbeitungszeiten nach sich ziehen.

Hinweis: Verwenden Sie die Option Maximal zu verwendende Anzahl von Dokumenten beim Berechnen von Clustern im Dialogfeld "Cluster aufbauen", um Cluster mit nur einer Untermenge aller Dokumente oder Datensätze aufzubauen.

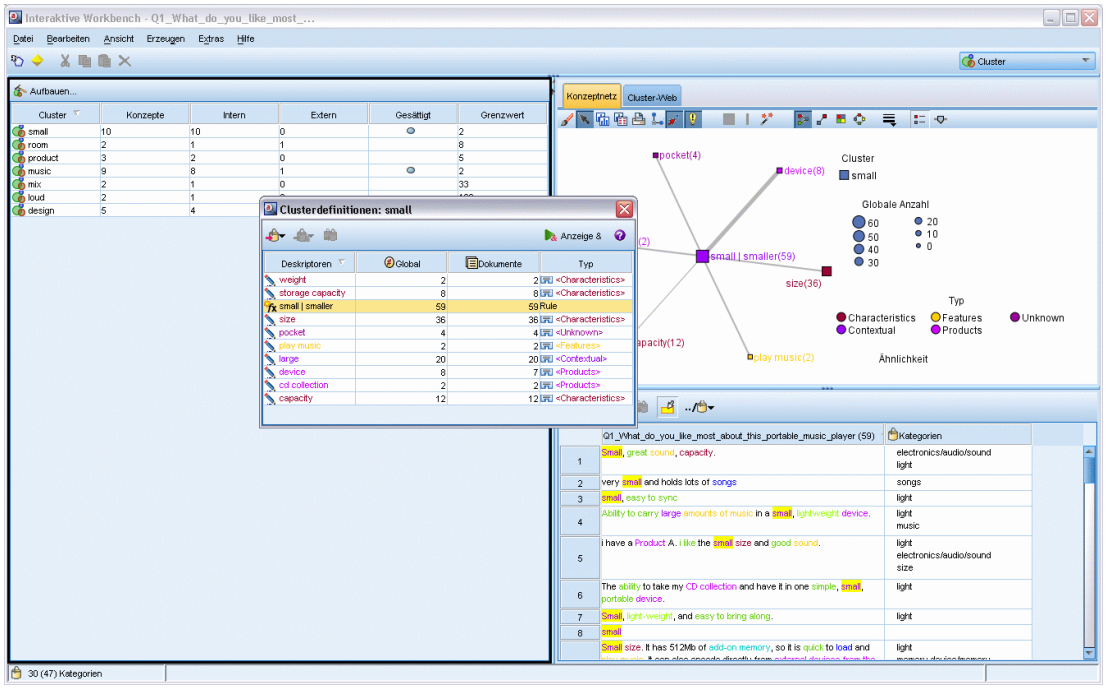
Die Clusterbildung ist ein Prozess, der mit der Analyse eines Satzes von Konzepten und der Suche nach Konzepten beginnt, die häufig gemeinsam in Dokumenten vorkommen. Zwei Konzepte, die gemeinsam in einem Dokument vorkommen, werden als Konzeptpaar betrachtet. Anschließend ermittelt der Clusteringprozess den **Ähnlichkeitswert** der einzelnen Konzeptpaare, indem die Anzahl der Dokumente, in denen das Paar gemeinsam vorkommt, mit der Anzahl der Konzepte verglichen wird, in denen jedes einzelne Konzept vorkommt. Für weitere Informationen siehe Thema Berechnen von Werten für Ähnlichkeitszusammenhänge auf S. 247.

Zuletzt gruppiert der Clusteringprozess ähnliche Konzepte durch Aggregation in Clustern und berücksichtigt dabei deren Zusammenhangswerte sowie die im Dialogfeld "Cluster aufbauen" definierten Einstellungen. Aggregation bedeutet hier, dass so lange Konzepte zu Clustern hinzugefügt oder kleinere Cluster in größere Cluster integriert werden, bis der Cluster gesättigt ist. Ein Cluster ist **gesättigt**, wenn das Hinzufügen weiterer Konzepte oder weiterer kleinerer Cluster dazu führen würde, dass der Cluster die im Dialogfeld "Cluster aufbauen" vorgenommenen Einstellungen (Anzahl der Konzepte, interne Zusammenhänge oder externe Zusammenhänge) überschreitet. Ein Cluster erhält den Namen des im Cluster enthaltenen Konzepts, das die insgesamt höchste Anzahl an Zusammenhängen mit anderen Konzepten innerhalb des Clusters besitzt.

Letztendlich gelangen nicht alle Konzeptpaare zusammen in denselben Cluster, da ein stärkerer Zusammenhang mit einem anderen Cluster vorliegen kann oder weil die Sättigung verhindert, dass die Cluster aufgenommen werden, in denen sie vorkommen. Aus diesem Grund gibt es sowohl interne als auch externe Zusammenhänge.

- **Interne Zusammenhänge** sind Zusammenhänge zwischen Konzeptpaaren innerhalb eines Clusters. Nicht alle Konzepte in einem Cluster stehen miteinander in Zusammenhang. Alle Konzepte stehen jedoch mit mindestens einem anderen, im Cluster enthaltenen Konzept in Zusammenhang.
- **Externe Zusammenhänge** sind Zusammenhänge zwischen Konzeptpaaren, die sich in unterschiedlichen Clustern befinden (ein Konzept befindet sich in einem und das andere Konzept in einem anderen Cluster).

Abbildung 11-1
Clusteransicht



Die Clusteransicht ist in drei Bereiche unterteilt, die jeweils durch Auswahl des entsprechenden Namens in der Clusteransicht aus- oder eingeblendet werden können.

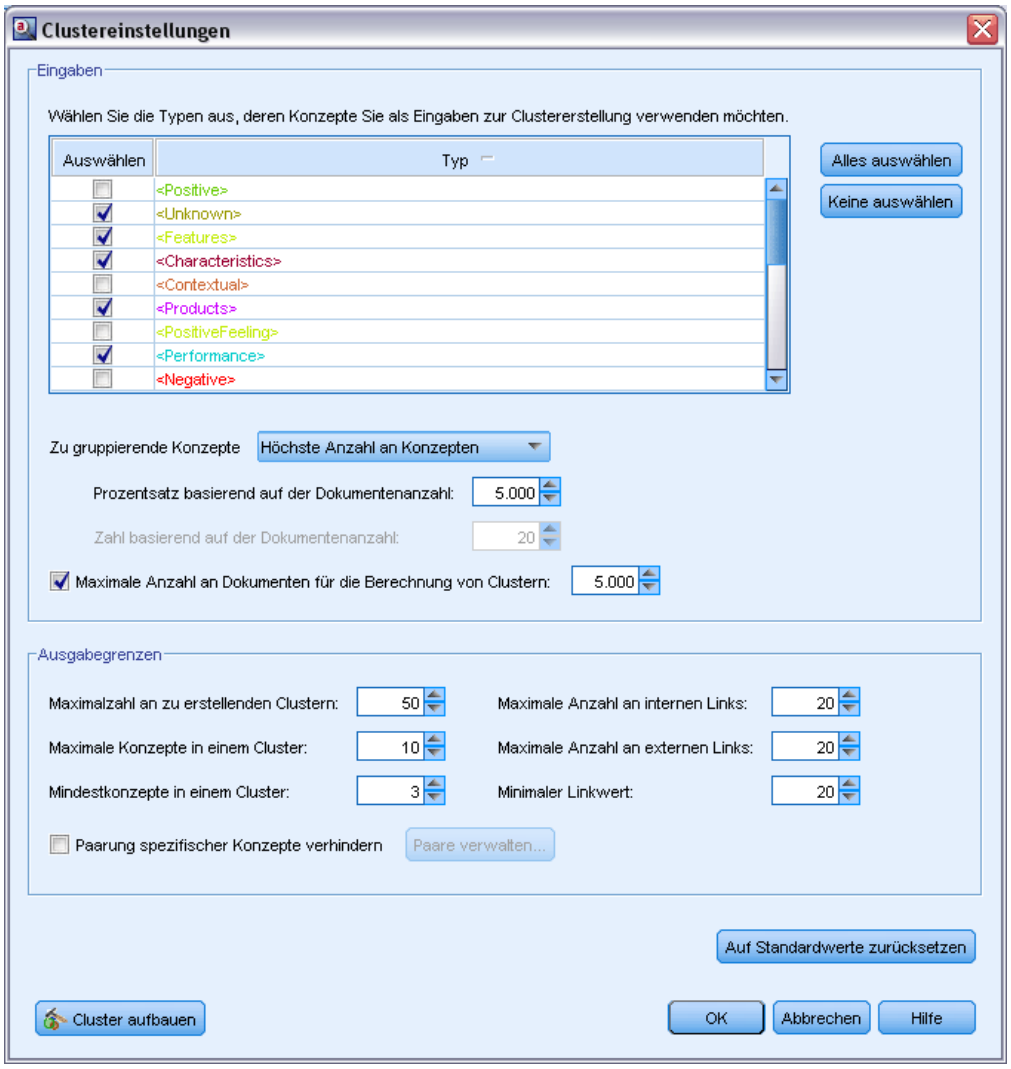
- **Bereich "Cluster".** In diesem Bereich erstellen und verwalten Sie Ihre Cluster. Für weitere Informationen siehe Thema Untersuchen von Clustern auf S. 248.
- **Visualisierungsbereich.** In diesem Bereich können Sie Ihre Cluster und deren Interaktionen visuell untersuchen. Für weitere Informationen siehe Thema Clusterdiagramme in Kapitel 13 auf S. 266.
- **Datenbereich.** Hier können Sie Text, der in den Dokumenten und Datensätzen enthalten ist, die im Dialogfeld "Clusterdefinitionen" ausgewählt wurden, untersuchen und prüfen. Für weitere Informationen siehe Thema Clusterdefinitionen auf S. 249.

Cluster aufbauen

Wenn Sie die Clusteransicht zum ersten Mal aufrufen, werden keine Cluster angezeigt. Sie können Cluster über das Menü (Extras > Cluster aufbauen) erstellen oder indem Sie in der Symbolleiste auf die Schaltfläche Aufbauen... klicken. Daraufhin wird das Dialogfeld Cluster aufbauen geöffnet, in dem Sie die Einstellungen und Grenzwerte für das Erstellen Ihrer Cluster definieren können.

Hinweis: Wenn die Extrahierungsergebnisse nicht mehr mit den Ressourcen übereinstimmen, wird der Bereich, ebenso wie der Bereich "Extrahierungsergebnisse", gelb dargestellt. Sie können dann eine erneute Extrahierung durchführen, um die neusten Extrahierungsergebnisse zu erhalten. Der Bereich wird anschließend nicht mehr gelb angezeigt. Bei jeder Extrahierung wird jedoch der Clusterbereich gelöscht. Anschließend müssen Sie Ihre Cluster neu aufbauen. Cluster werden nicht von einer Sitzung zur nächsten gespeichert.

Abbildung 11-2
Dialogfeld "Cluster aufbauen"



Eingaben

Eingabe-Tabelle. Cluster werden aus Deskriptoren erstellt, die aus bestimmten Typen abgeleitet wurden. Sie können in der Tabelle die Typen auswählen, die in den Erstellprozess eingeschlossen werden sollen. Standardmäßig sind die Typen ausgewählt, die die meisten Datensätze oder Dokumente erfassen.

Konzepte für das Clustering: Wählen Sie das Auswahlverfahren für die Konzepte aus, die Sie für das Clustering verwenden möchten. Indem Sie die Anzahl der Konzepte reduzieren, können Sie den Clusteringprozess beschleunigen. Sie können das Clustering mit einer Anzahl von Top-Konzepten, einem Prozentsatz von Top-Konzepten oder mit allen Konzepten durchführen:

- **Anzahl basierend auf Dokumentenanzahl.** Wenn Sie Größte Anzahl an Konzepten auswählen, geben Sie die Anzahl an Konzepten ein, die für das Clustering berücksichtigt werden sollen. Die Konzepte werden auf der Grundlage der höchsten Dokumentenanzahl ausgewählt. Die Dokumentenanzahl ist die Anzahl der Dokumente oder Datensätze, in denen die Konzepte vorkommen.
- **Prozentsatz basierend auf Dokumentenanzahl**Wenn Sie Oberster Prozentsatz der Konzepte auswählen, geben Sie den Prozentsatz der Konzepte ein, die für das Clustering berücksichtigt werden sollen. Die Konzepte werden auf der Grundlage dieses Prozentsatzes der Konzepte mit der höchsten Dokumentenanzahl ausgewählt.

Maximal zu verwendende Anzahl von Dokumenten beim Berechnen von Clustern. Standardmäßig werden die Zusammenhangswerte anhand des gesamten Satzes an Dokumenten oder Datensätzen berechnet. In einigen Fällen möchten Sie jedoch den Clusteringprozess eventuell beschleunigen, indem Sie die Anzahl der zum Berechnen der Zusammenhänge verwendeten Dokumente oder Datensätze einschränken. Eine Einschränkung der Dokumente kann die Qualität der Cluster verringern. Um diese Option zu verwenden, aktivieren Sie das linke Kontrollkästchen und geben Sie die maximale Anzahl der zu verwendenden Dokumente oder Datensätze ein.

Ausgabegrenzen

Maximal zu erstellende Anzahl an Clustern. Dieser Wert gibt die maximale Anzahl von Clustern vor, die generiert und im Clusterbereich angezeigt werden. Während des Clusteringprozesses werden gesättigte Cluster vor ungesättigten bearbeitet. Aus diesem Grund sind viele der resultierenden Cluster gesättigt. Um mehr ungesättigte Cluster zu erhalten, können Sie für diese Einstellung einen Wert angeben, der über der Anzahl der gesättigten Cluster liegt.

Maximale Anzahl an Konzepten in einem Cluster. Dieser Wert legt die maximale Anzahl an Konzepten fest, die ein Cluster enthalten kann.

Minimale Anzahl an Konzepten in einem Cluster. Dieser Wert legt die minimale Anzahl von Konzepten fest, die in Zusammenhang stehen müssen, damit ein Cluster aufgebaut wird.

Maximale Anzahl an internen Zusammenhängen. Dieser Wert legt die maximale Anzahl interner Zusammenhänge fest, die ein Cluster enthalten kann. Interne Zusammenhänge sind Zusammenhänge zwischen Konzeptpaaren innerhalb eines Clusters.

Maximale Anzahl an externen Zusammenhängen. Dieser Wert bestimmt die maximale Anzahl an Zusammenhängen mit Konzepten außerhalb des Clusters. Externe Zusammenhänge sind Zusammenhänge zwischen Konzeptpaaren, die sich in unterschiedlichen Clustern befinden.

Minimaler Zusammenhangswert. Dieser Wert ist der kleinste Zusammenhangswert, der für ein Konzeptpaar erforderlich ist, damit es für das Clustering berücksichtigt wird. Der Zusammenhangswert wird mithilfe einer Ähnlichkeitsformel berechnet. Für weitere Informationen siehe Thema Berechnen von Werten für Ähnlichkeitszusammenhänge auf S. 247.

Paarung spezifischer Konzepte verhindern. Markieren Sie dieses Kontrollkästchen, um den Vorgang der Gruppierung oder Paarung von zwei Konzepten in der Ausgabe zu verhindern. Klicken Sie zum Erstellen oder Verwalten von Konzeptpaaren auf Paare verwalten. Für weitere Informationen siehe Thema Verwalten von Verknüpfungsausnahmepaaren in Kapitel 10 auf S. 191.

Berechnen von Werten für Ähnlichkeitszusammenhänge

Wenn Sie lediglich wissen, in wie vielen Dokumenten ein Konzeptpaar gemeinsam vorkommt, sagt dies nichts darüber aus, wie ähnlich die beiden Konzepte sind. In diesen Fällen kann der Ähnlichkeitswert hilfreich sein. Der Ähnlichkeitszusammenhangswert wird gemessen, indem die Anzahl der Dokumente mit Kookkurrenz mit den Dokumentenwerten der einzelnen an der Beziehung beteiligten Konzepte verglichen wird. Beim Berechnen der Ähnlichkeit dient die Anzahl der Dokumente (Dokumentenanzahl) als Maßstab, in denen ein Konzept oder ein Konzeptpaar gefunden wird. Ein Konzept oder ein Konzeptpaar gilt als in einem Dokument "gefunden", wenn es *mindestens* einmal im Dokument vorkommt. Sie können wahlweise festlegen, dass die Linienstärke im Konzeptdiagramm dem Wert des Ähnlichkeitszusammenhangs entspricht.

Der Algorithmus zeigt die stärksten Beziehungen auf, was bedeutet, dass die Tendenz, dass die entsprechenden Konzepte in den Textdaten gemeinsam vorkommen, viel größer ist als die Tendenz, dass beide unabhängig voneinander vorkommen. Intern ergibt der Algorithmus einen Ähnlichkeitskoeffizienten zwischen 0 und 1, wobei der Wert 1 bedeutet, dass die beiden Konzepte immer gemeinsam und niemals getrennt vorkommen. Der Ähnlichkeitskoeffizient wird dann mit 100 multipliziert und auf die nächste ganze Zahl gerundet. Für die Berechnung des Ähnlichkeitskoeffizienten wird die in der folgenden Abbildung dargestellte Formel verwendet.

Abbildung 11-3
Formel des Ähnlichkeitskoeffizienten

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

Erläuterung:

- C_I ist die Anzahl der Dokumente oder Datensätze, in denen das Konzept `I` vorkommt.
- C_J ist die Anzahl der Dokumente oder Datensätze, in denen das Konzept `J` vorkommt.
- C_{IJ} ist die Anzahl der Dokumente oder Datensätze, in denen das Konzeptpaar `I` und `J` gemeinsam im Dokumentensatz vorkommt.

Angenommen, Sie haben 5.000 Dokumente. `I` und `J` wären die extrahierten Konzepte und `IJ` die Kookkurrenz eines Konzeptpaars aus `I` und `J`. Die folgende Tabelle enthält zwei Szenarien, die verdeutlichen, wie der Koeffizient und der Zusammenhangswert berechnet werden.

Tabelle 11-1
Beispiel für Konzepthäufigkeiten

Konzept/Paar	Szenario A	Szenario B
Konzept: I	Kommt in 20 Dokumenten vor	Kommt in 30 Dokumenten vor
Konzept: J	Kommt in 20 Dokumenten vor	Kommt in 60 Dokumenten vor
Konzept/Paar: IJ	Kommt in 20 Dokumenten gemeinsam vor	Kommt in 20 Dokumenten gemeinsam vor
Ähnlichkeitskoeffizient	1	0.22222
Ähnlichkeits-zusammenhangswert	100	22

In Szenario A kommen die Konzepte I und J sowie das Paar IJ in 20 Dokumenten vor, woraus sich ein Ähnlichkeitskoeffizient von 1 ergibt, der aussagt, dass die Konzepte immer gemeinsam vorkommen. Der Wert des Ähnlichkeitszusammenhangs für das Paar wäre demnach 100.

In Szenario B kommt das Konzept I in 30 Dokumenten vor und das Konzept J in 60 Dokumenten, das Paar IJ kommt jedoch nur in 20 Dokumenten vor. Der Ähnlichkeitskoeffizient ist demnach 0,22222. Der Wert des Ähnlichkeitszusammenhangs für dieses Paar würde auf 22 abgerundet.

Untersuchen von Clustern

Nachdem Sie Cluster aufgebaut haben, wird im Clusterbereich ein Satz von Ergebnissen angezeigt. Für alle Cluster stehen in der Tabelle folgende Informationen zur Verfügung:

- **Cluster.** Der Name des Clusters. Cluster werden nach dem Konzept benannt, das die höchste Anzahl interner Zusammenhänge besitzt.
- **Konzepte.** Die Anzahl der im Cluster enthaltenen Konzepte. Für weitere Informationen siehe Thema Clusterdefinitionen auf S. 249.
- **Intern.** Die Anzahl der im Cluster enthaltenen internen Zusammenhänge. Interne Zusammenhänge sind Zusammenhänge zwischen Konzeptpaaren innerhalb eines Clusters.
- **Extern.** Die Anzahl der im Cluster enthaltenen externen Zusammenhänge. Externe Zusammenhänge sind Zusammenhänge zwischen Konzeptpaaren, bei denen sich ein Konzept in einem Cluster und das andere Konzept in einem anderen Cluster befindet.
- **Ges.** Wenn ein Symbol vorhanden ist, zeigt dies an, dass dieser Cluster hätte größer sein können, dass dadurch aber eine oder mehrere Einschränkungen überschritten worden wären, weshalb der Clusteringprozess für diesen Cluster beendet wurde und der Cluster als *gesättigt* erachtet wird. Am Ende des Clusteringprozesses werden gesättigte Cluster vor ungesättigten bearbeitet. Aus diesem Grund sind viele der resultierenden Cluster gesättigt. Um mehr ungesättigte Cluster zu erhalten, können Sie für Maximal zu erstellende Anzahl an Clustern einen Wert angeben, der über der Anzahl der gesättigten Cluster liegt oder den

Minimalen Zusammenhangswert erhöhen. Für weitere Informationen siehe Thema Cluster aufbauen auf S. 245.

- **Schwelle.** Für alle im Cluster enthaltenen gemeinsam vorkommenden Konzeptpaare ist dies der niedrigste Ähnlichkeitszusammenhangswert von allen im Cluster vorhandenen. Für weitere Informationen siehe Thema Berechnen von Werten für Ähnlichkeitszusammenhänge auf S. 247. Ein Cluster mit einem hohen Schwellenwert zeigt an, dass die in diesem Cluster enthaltenen Konzepte eine höhere Gesamtähnlichkeit besitzen und enger zusammenhängen als diejenigen, die sich in einem Cluster mit einem niedrigeren Schwellenwert befinden.

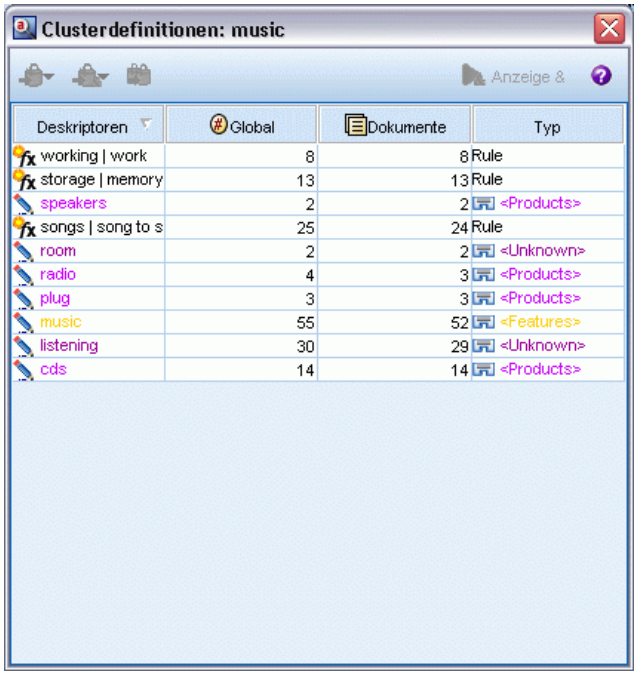
Um einen bestimmten Cluster genauer zu untersuchen, wählen Sie diesen aus, damit rechts im Visualisierungsbereich zwei Diagramme angezeigt werden, die eine Analyse ermöglichen. Für weitere Informationen siehe Thema Clusterdiagramme in Kapitel 13 auf S. 266. Sie können den Inhalt der Tabelle auch kopieren und in eine andere Anwendung einfügen.

Wenn die Extrahierungsergebnisse nicht mehr mit den Ressourcen übereinstimmen, wird der Bereich, ebenso wie der Bereich “Extrahierungsergebnisse”, gelb dargestellt. Sie können dann eine erneute Extrahierung durchführen, um die neusten Extrahierungsergebnisse zu erhalten. Der Bereich wird anschließend nicht mehr gelb angezeigt. Bei jeder Extrahierung wird jedoch der Clusterbereich gelöscht. Anschließend müssen Sie Ihre Cluster neu aufbauen. Cluster werden nicht von einer Sitzung zur nächsten gespeichert.

Clusterdefinitionen

Sie können alle in einem Cluster enthaltenen Konzepte anzeigen, indem Sie den Cluster im Clusterbereich auswählen und das Dialogfeld “Clusterdefinitionen” öffnen (Ansicht > Clusterdefinitionen).

Abbildung 11-4
Dialogfeld “Clusterdefinitionen”



Im Dialogfeld “Clusterdefinitionen” werden alle im ausgewählten Cluster enthaltenen Konzepte angezeigt. Wenn Sie im Dialogfeld “Clusterdefinitionen” eines oder mehrere Konzepte auswählen und auf Anzeigen & klicken, werden im Datenbereich alle Datensätze oder Dokumente angezeigt, in denen *alle ausgewählten Konzepte gemeinsam vorkommen*. Im Datenbereich werden jedoch keine Textdatensätze oder Dokumente angezeigt, wenn Sie einen Cluster lediglich im Clusterbereich auswählen. Allgemeine Informationen zum Datenbereich finden Sie unter Der Fensterbereich “Daten” in Kapitel 10.

Wenn Sie in diesem Dialogfeld Konzepte auswählen, ändert dies auch das Netzdiagramm des Konzepts. Für weitere Informationen siehe Thema Clusterdiagramme in Kapitel 13 auf S. 266. Wenn Sie im Dialogfeld “Clusterdefinitionen” eines oder mehrere Konzepte auswählen, werden für diese Konzepte im Visualisierungsbereich die externen und internen Zusammenhänge angezeigt.

Spaltenbeschreibungen

Es werden Symbole angezeigt, über die Sie die einzelnen Deskriptoren problemlos identifizieren können.

Tabelle 11-2
Symbole für Spalten und Deskriptoren

Spalten	Beschreibung
Deskriptoren	Der Name des Konzepts.
Globalwert	Zeigt, wie oft dieser Deskriptor im gesamten Daten-Set vorkommt. Dies wird auch als globale Häufigkeit bezeichnet.
Dokumente	Zeigt die Anzahl der Dokumente oder Datensätze, in denen dieser Deskriptor vorkommt. Dies wird auch als Dokumentenhäufigkeit bezeichnet.
Typ	Zeigt den oder die Typen, zu denen der Deskriptor gehört. Wenn es sich bei dem Deskriptor um eine Kategorieregel handelt, wird in dieser Spalte kein Name angezeigt.

Symbolleistenaktionen

In diesem Dialogfeld können Sie außerdem eines oder mehrere Konzepte auswählen, die in einer Kategorie verwendet werden sollen. Dies ist auf verschiedene Weisen möglich. Am interessantesten ist es jedoch, Konzepte auszuwählen, die gemeinsam in einem Cluster vorkommen, und diese zu einer Kategorieregel hinzuzufügen. Für weitere Informationen siehe Thema Kookkurrenzregeln in Kapitel 10 auf S. 197. Über die Schaltflächen der Symbolleiste können Sie Konzepte zu Kategorien hinzufügen.

Tabelle 11-3
Schaltflächen der Symbolleiste zum Hinzufügen von Konzepten zu Kategorien

Symbole	Beschreibung
	Fügt die ausgewählten Konzepte zu einer neuen oder einer vorhandenen Kategorie hinzu
	Fügt die ausgewählten Konzepte in Form einer &-Kategorieregel zu einer neuen oder einer vorhandenen Kategorie hinzu. Für weitere Informationen siehe Thema Verwenden von Kategorieregeln in Kapitel 10 auf S. 207.
	Fügt alle ausgewählten Konzepte als eigene neue Kategorie hinzu
Anzeigen &	Aktualisiert die Anzeige des Daten- und des Visualisierungsbereichs in Funktion der ausgewählten Deskriptoren

Hinweis: Sie können Konzepte auch mithilfe der Kontextmenüs zu einem Typ hinzufügen, als Synonyme oder als Ausschlusselemente.

Kapitel
12

Untersuchen von Textlinkanalyse

In der Textlinkanalyse- (TLA-) Ansicht können Sie Textlinkanalyse-Musterergebnisse untersuchen. Die Textlinkanalyse ist ein Verfahren zum Musterabgleich, mit der Sie Musterregeln definieren und mit tatsächlich extrahierten Konzepten und Beziehungen, die in Ihrem Text aufgefunden werden, vergleichen können.

Beispiel: Die bloße Extrahierung von allgemeinen Daten zu einem Unternehmen ist für Sie wenig aussagekräftig. Mit der TLA erkennen Sie ggf. auch die Zusammenhänge zwischen diesem und anderen Unternehmen oder zwischen den Mitarbeitern in einem Unternehmen. Mit der TLA können Sie auch Meinungen zu Produkten oder für einige Sprachen die Beziehungen zwischen Genen extrahieren.

Nachdem Sie einige TLA-Muster extrahiert haben, können Sie diese in den Typ- und Fachausdruckmuster-Bereichen der Textlinkanalyse-Ansicht überprüfen. Für weitere Informationen siehe Thema Typ- und Konzeptmuster auf S. 254. Außerdem können Sie sie in den Daten- und Visualisierungsbereichen in dieser Ansicht untersuchen. Was wahrscheinlich am wichtigsten ist: Sie können sie Kategorien hinzufügen.

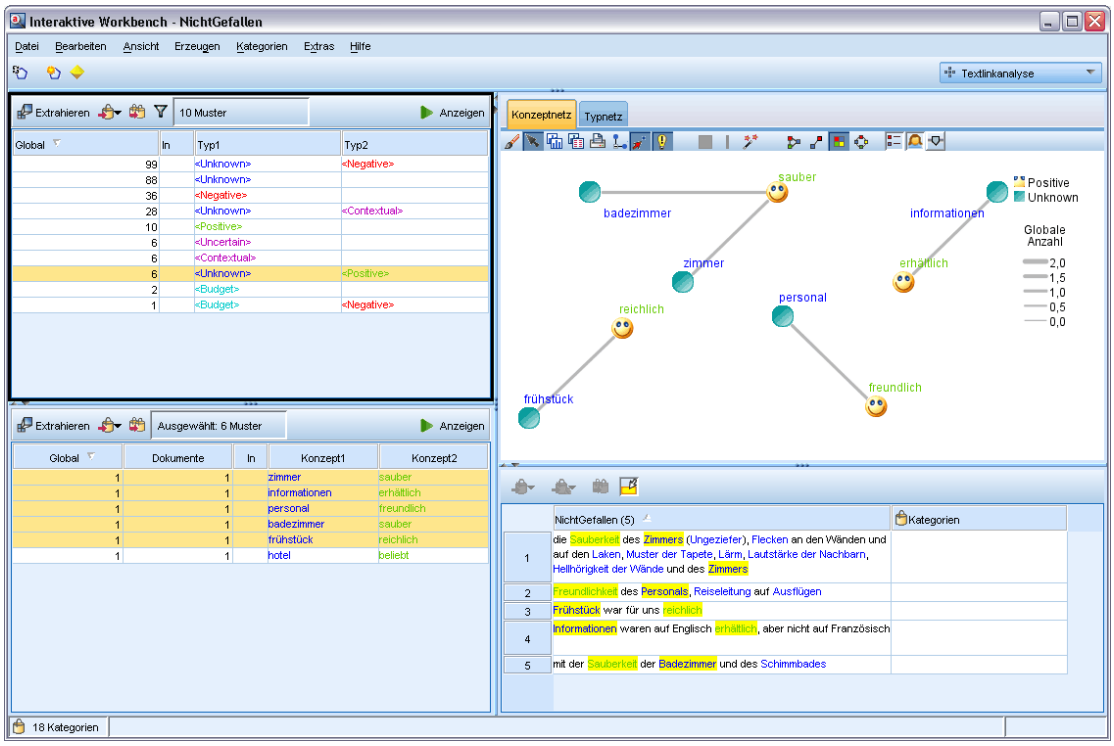
Wenn Sie keine entsprechende Auswahl getroffen haben, können Sie auf Extrahieren klicken und Musterextrahierung für Textlinkanalyse aktivieren im Dialogfeld “Extrahierungseinstellungen” auswählen. Für weitere Informationen siehe Thema TLA-Musterergebnisse extrahieren auf S. 253.

Um TLA-Musterergebnisse extrahieren zu können, müssen in der verwendeten Ressourcenvorlage bzw. in den verwendeten Bibliotheken TLA-Musterregeln definiert sein. Sie können die TLA-Muster in bestimmten Ressourcenvorlagen verwenden, die mit IBM® SPSS® Modeler Text Analytics ausgeliefert werden. Die Art von Beziehungen und Mustern, die Sie extrahieren können, hängt völlig von den TLA-Regeln ab, die in Ihren Ressourcen definiert sind. Sie können Ihre eigenen TLA-Regeln für alle Sprachen *außer* Japanisch definieren. Muster bestehen aus Makros, Wortlisten sowie Wortlücken und bilden eine Boole’sche Abfrage (Regel), die mit dem Eingangstext abgeglichen wird. Für weitere Informationen siehe Thema Textlinkregeln in Kapitel 19 auf S. 346.

Wenn eine TLA-Musterregel mit Text übereinstimmt, kann dieser Text als Muster extrahiert und als Ausgabedaten neu strukturiert werden. Die Ergebnisse sind dann in den Bereichen der Textlinkanalyse-Ansicht sichtbar. Jeder Bereich kann aus- oder eingeblendet werden, indem Sie seinen Namen aus dem Menü “Ansicht” wählen:

- **Typ- und Konzeptmusterbereiche.** In diesen beiden Bereichen können Sie Ihre Muster erstellen und untersuchen. Für weitere Informationen siehe Thema Typ- und Konzeptmuster auf S. 254.
- **Visualisierungsbereich.** In diesem Bereich können Sie die Interaktionen der Konzepte und Typen in Ihren Mustern visuell untersuchen. Für weitere Informationen siehe Thema Textlinkanalyse-Diagramme in Kapitel 13 auf S. 268.
- **Datenbereich.** Sie können Text, der in Dokumenten und Datensätzen enthalten ist, die mit der Auswahl in einem anderen Bereich übereinstimmen, untersuchen und überprüfen. Für weitere Informationen siehe Thema Datenbereich auf S. 259.

Abbildung 12-1
Ansicht "Text Link Analysis"



TLA-Musterergebnisse extrahieren

Der Extrahierungsprozess liefert eine Reihe von Konzepten und Typen und, sofern aktiviert, Textlinkanalyse-(TLA-)Muster. Extrahierte TLA-Muster könnnen Sie in der Textlinkanalysenansicht sehen. Wenn die Extrahierungsergebnisse nicht mit den Ressourcen übereinstimmen, färben sich die Musterbereiche gelb und signalisieren damit, dass eine erneute Extrahierung andere Ergebnisse liefern würde.

Sie müssen die Extrahierung dieser Muster in der Knoteneinstellung für oder im Dialogfeld "Extrahieren" über die Option Musterextrahierung für Textlinkanalyse aktivieren auswählen. Für weitere Informationen siehe Thema Daten extrahieren in Kapitel 9 auf S. 147.

Hinweis: Die für den Extrahierungsprozess benötigte Zeit steht in direkter Beziehung zur Größe Ihrer Datenmenge. Weitere Informationen zu Leistungsstatistiken und Empfehlungen finden Sie in den Installationsanweisungen. Sie haben jederzeit die Möglichkeit, einen vorgeordneten Stichprobenknoten einzufügen oder die Konfiguration Ihres Computers zu optimieren.

Daten extrahieren

- Wählen Sie in den Menüs die Optionsfolge Extras > Extrahieren aus. Alternativ können Sie auf die Symbolleistenschaltfläche Extrahieren klicken.
- Ändern Sie nach Bedarf die Optionen, die Sie verwenden möchten. Beachten Sie, dass die Option Musterextrahierung für Textlinkanalyse aktivieren auf dieser Registerkarte ausgewählt sein muss und

außerdem TLA-Regeln in Ihrer Vorlage vorhanden sein müssen, damit TLA-Musterergebnisse extrahiert werden können. Für weitere Informationen siehe Thema Daten extrahieren in Kapitel 9 auf S. 147.

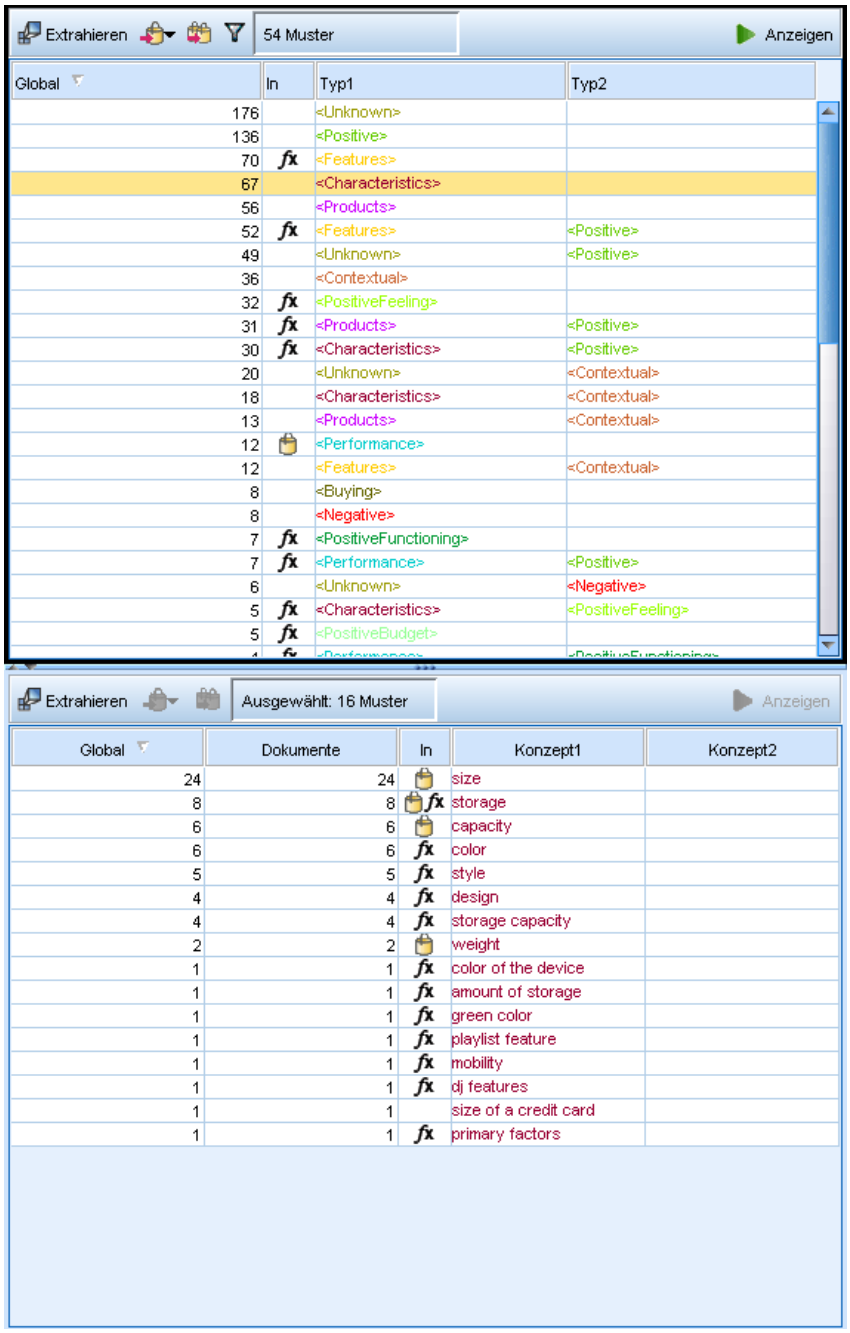
- Klicken Sie auf Extrahieren, um die Extrahierung zu starten.

  Sobald die Extrahierung beginnt, öffnet sich die Statusanzeige. Wenn Sie die Extrahierung abbrechen möchten, klicken Sie auf Abbrechen. Beim Abschluss der Extrahierung wird das Dialogfeld geschlossen und die Ergebnisse werden im Bereich angezeigt. Für weitere Informationen siehe Thema Typ- und Konzeptmuster auf S. 254.

Typ- und Konzeptmuster

Muster bestehen aus zwei Teilen: einer Kombination aus Konzepten und Typen. Muster sind am nützlichsten, wenn man versucht, Meinungen zu einem bestimmten Thema oder Beziehungen zwischen Konzepten herauszufinden. Es könnte sein, dass es Ihnen nicht reicht, den Produktnamen Ihres Mitbewerbers zu extrahieren. In diesem Fall können Sie die extrahierten Muster dahingehend überprüfen, ob Sie Beispiele von Dokumenten oder Datensätzen finden können, die Text mit Aussagen zum Produkt (gut, schlecht, teuer) enthalten.

Abbildung 12-2
Textlinkanalysenansicht: Typ- und Konzeptmusterbereiche



Muster können aus bis zu sechs Typen oder sechs Konzepten bestehen. Aus diesem Grund enthalten die Zeilen in beiden Musterbereichen bis zu sechs Slots bzw. Positionen. Jeder Slot entspricht der jeweiligen Position eines Elements im TLA-Muster gemäß seiner Definition in den linguistischen Ressourcen. Wenn ein Slot in der interaktiven Workbench keinen Wert enthält, wird er nicht in der Tabelle angezeigt. Beispiel: Wenn die längsten Musterergebnisse nicht mehr als

vier Slots enthalten, werden die letzten beiden Slots nicht angezeigt. Für weitere Informationen siehe Thema Textlinkregeln in Kapitel 19 auf S. 346.

Bei der Extrahierung von Musterergebnissen werden die ersten Ergebnisse auf Typebene gruppiert und anschließend in Konzeptmuster unterteilt. Aus diesem Grund gibt es zwei verschiedene Ergebnisbereiche: Typmuster (oben links) und Konzeptmuster (unten links). Wenn Sie alle zurückgegebenen Konzeptmuster sehen möchten, wählen Sie alle Typmuster aus. Im unteren Konzeptmusterbereich werden dann alle Konzeptmuster bis zum maximalen Rangwert (gemäß Definition im Dialogfeld “Filter”) angezeigt.

Typmuster. Dieser Bereich stellt Musterergebnisse dar, die aus mindestens einem verwandten, einer TLA-Musterregel entsprechenden Typ bestehen. Typmuster werden angezeigt als `<Organization>` + `<Location>` + `<Positive>`, was ein positives Feedback zu einem Unternehmen an einem bestimmten Ort bereitstellen könnte. Die Syntax lautet wie folgt:

```
<Typ1> + <Typ2> + <Typ3> + <Typ4> + <Typ5> + <Typ6>
```

Konzeptmuster. In diesem Bereich werden die Musterergebnisse auf Konzeptebene für alle Typmuster dargestellt, die aktuell im darüber liegenden Typmusterbereich ausgewählt sind. Konzeptmuster folgen einer Struktur, wie beispielsweise `Hotel` + `Paris` + `wunderbar`. Die Syntax lautet wie folgt:

```
Konzept1 + Konzept2 + Konzept3 + Konzept4 + Konzept5 + Konzept6
```

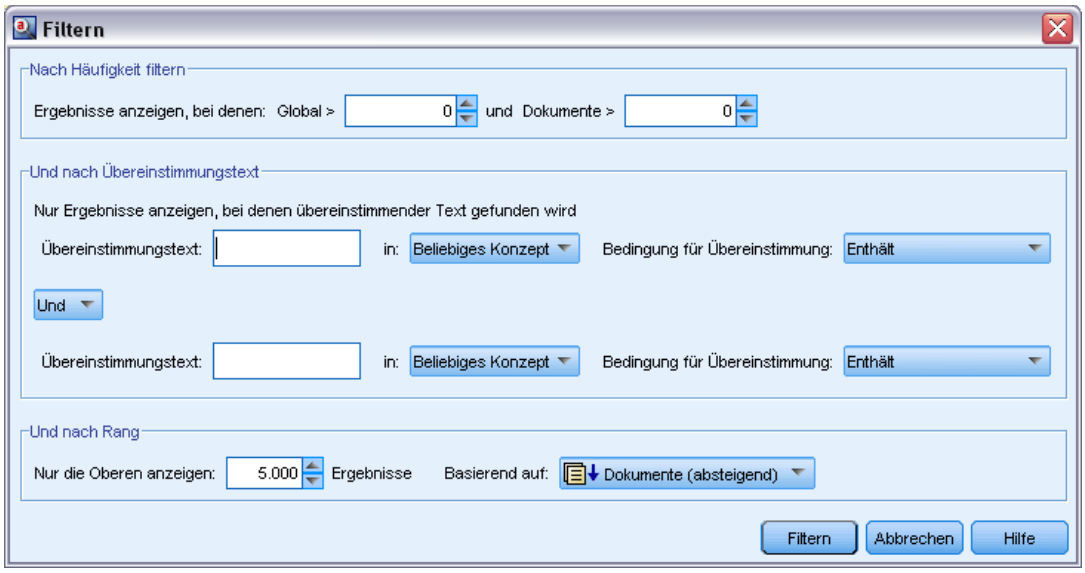
Wenn Musterergebnisse weniger als die sechs maximalen Slots verwenden, werden nur die benötigten Slots (oder Spalten) angezeigt. Leere Slots, die sich zwischen zwei gefüllten Slots befinden, werden verworfen. Folglich kann das Muster `<Typ1>+<>+<Typ2>+<>+<>+<>` durch `<Typ1>+<>+<Typ3>` dargestellt werden. Beispiel für ein Konzeptmuster: `Konzept1+.+Konzept2` (wobei `.` einen Nullwert darstellt).

Wie bei den Extrahierungsergebnissen in der Ansicht “Kategorien und Konzepte” können Sie hier die Ergebnisse überprüfen. Wenn Sie die Typen und Konzepte, aus denen diese Muster bestehen, weiter verfeinern möchten, können Sie das im Bereich “Extrahierungsergebnisse” in der Ansicht “Kategorien und Konzepte” oder direkt im Ressourceneditor vornehmen und Ihre Muster erneut extrahieren. Wenn ein Konzept, ein Typ oder ein Muster in einer Kategoriedefinition als solches oder als Teil einer Regel verwendet wird, wird ein Kategorieregelsymbol in der Spalte In in der Muster- oder Extrahierungsergebnissetabelle angezeigt.

Filtern von TLA-Ergebnissen

Wenn Sie mit sehr großen Datenmengen arbeiten, kann der Extrahierungsprozess Millionen von Ergebnissen liefern. Durch diese Menge ist eine effektive Überprüfung der Ergebnisse für viele Benutzer mühsam. Sie können jedoch diese Ergebnisse filtern, um die interessantesten Ergebnisse näher heranzuholen. Über die Einstellungen im Filterdialogfeld können Sie eingrenzen, welche Muster angezeigt werden sollen. Alle Einstellungen werden gemeinsam verwendet.

Abbildung 12-3
Filterdialogfeld (in der TLA-Ansicht)



Nach Häufigkeit filtern. Mit diesem Filter werden nur Ergebnisse mit einem bestimmten globalen oder Dokumenthäufigkeitswert angezeigt.

- **Die globale Häufigkeit** gibt an, wie oft ein Muster in der gesamten Menge der Dokumente oder Datensätze insgesamt auftritt, und wird in der Spalte Global angezeigt.
- **Die Dokumenthäufigkeit** gibt die Gesamtzahl der Dokumente oder Datensätze an, in denen ein Muster aufgefunden wird, und wird in der Spalte Dokumente angezeigt.

Beispiel: Wenn ein Muster 300-mal in 500 Datensätzen aufgefunden wird, hat dieses Muster eine globale Häufigkeit von 300 und eine Dokumenthäufigkeit von 500.

Und nach Übereinstimmungstext. Sie können auch einen Filter anwenden, durch den nur Ergebnisse angezeigt werden, die mit den hier definierten Regeln übereinstimmen. Geben Sie im Feld Übereinstimmungstext die Zeichen ein, mit denen eine Übereinstimmung vorhanden sein muss, und wählen Sie aus, ob nach diesem Text innerhalb von Konzept- oder Typnamen gesucht werden soll, indem Sie die Slotnummer oder alle Slots auswählen. Wählen Sie anschließend die Bedingung aus, unter der eine Übereinstimmung angewendet werden soll (es ist nicht erforderlich, den Beginn oder das Ende eines Typnamens mit spitzen Klammern anzugeben). Wählen Sie entweder Und oder Oder aus der Dropdown-Liste aus, damit die Regel einen Vergleich mit beiden Anweisungen oder mit nur einer der Anweisungen vornimmt, und definieren Sie die zweite Übereinstimmungstextanweisung analog zur ersten Anweisung.

Tabelle 12-1
Bedingungen für Übereinstimmungstext

Bedingung	Beschreibung
Enthält	Text stimmt überein, wenn die Zeichenfolge irgendwo vorkommt. (Standardauswahl)
Beginnt mit	Text stimmt nur überein, wenn das Konzept oder der Typ mit dem angegebenen Text beginnt.

Bedingung	Beschreibung
Endet mit	Text stimmt nur überein, wenn das Konzept oder der Typ mit dem angegebenen Text endet.
Exakte Übereinstimmung	Die gesamte Zeichenfolge muss mit dem Konzept- oder dem Typnamen übereinstimmen.

Und nach Rang. Sie können Ihre Ergebnisse auch filtern, um nur die oberste Anzahl der Muster nach globaler Häufigkeit (Global) oder Dokumenthäufigkeit (Dokumente) in auf- oder absteigender Reihenfolge anzuzeigen. Mit diesem maximalen Rangwert wird die Gesamtzahl der zur Anzeige zurückgegebenen Muster begrenzt.

Bei Anwendung des Filters werden Typmuster hinzugefügt, bis die maximale Gesamtzahl der Konzeptmuster (maximaler Rang) überstiegen würde. Zunächst wird das Typmuster mit dem höchsten Rang untersucht und anschließend die Summe der entsprechenden Konzeptmuster genommen. Wenn diese Summe den maximalen Rang nicht übersteigt, werden die Muster in der Ansicht angezeigt. Anschließend wird die Anzahl der Konzeptmuster für das nächste Typmuster summiert. Wenn diese Anzahl zuzüglich der Gesamtzahl der Konzeptmuster im vorhergehenden Typmuster unter dem maximalen Rang liegt, werden diese Muster ebenfalls in der Ansicht angezeigt. Dies wird fortgeführt, bis so viele Muster wie nur möglich ohne Übersteigen des maximalen Rangs angezeigt werden.

Anzeige der Ergebnisse im Musterbereich

Hier einige Beispiele, wie die Ergebnisse auf der Grundlage von Filtern in der Symbolleiste des Musterbereichs angezeigt werden könnten.

Abbildung 12-4
Filterergebnisse, Beispiel 1

In diesem Beispiel zeigt die Symbolleiste an, dass die Anzahl der zurückgegebenen Muster aufgrund des im Filter spezifizierten maximalen Rangs begrenzt wurde. Ein lilafarbenes Symbol bedeutet, dass die maximale Anzahl an Mustern erreicht wurde. Für weitere Informationen bewegen Sie die Maus über das Symbol. Näheres erfahren Sie in der obigen Erklärung zum Filter Und nach Rang.

Abbildung 12-5
Filterergebnisse, Beispiel 2

In diesem Beipiel zeigt die Symbolleiste an, dass die Ergebnisse durch einen Übereinstimmungstextfilter begrenzt wurden (siehe Vergrößerungsglassymbol). Bewegen Sie die Maus über das Symbol, um den Übereinstimmungstext zu sehen.

Ergebnisse filtern

- Wählen Sie in den Menüs die Optionsfolge Extras > Filter aus. Das Filterdialogfeld wird geöffnet.
- Wählen und verfeinern Sie die Filter, die Sie verwenden möchten.
- Klicken Sie auf OK, um die Filter anzuwenden und die neuen Ergebnisse anzuzeigen.

Datenbereich

Bei der Extrahierung und Untersuchung von Textlinkanalyse-Mustern sollen manche Daten, mit denen Sie arbeiten, überprüft werden. Beispielsweise sollen die tatsächlichen Datensätze, in denen eine Gruppe von Mustern aufgefunden wurde, angezeigt werden. Sie können Datensätze oder Dokumente im Datenbereich überprüfen, der sich unten rechts befindet. Wird dieser nicht standardmäßig angezeigt, wählen Sie die Befehlsfolge Ansicht > Bereiche > Daten.

Im Datenbereich wird für jedes Dokument oder jeden Datensatz, der in der Ansicht ausgewählt ist, eine Zeile angezeigt, bis eine bestimmte Anzeigegrenze erreicht ist. Standardmäßig ist die Anzeige der im Datenbereich anzuzeigenden Dokumente oder Datensätze begrenzt, damit Sie Ihre Daten schneller sehen können. Sie können diese Einstellung jedoch im Optionsdialogfeld ändern. Für weitere Informationen siehe Thema Optionen: Registerkarte “Sitzung” in Kapitel 8 auf S. 135.

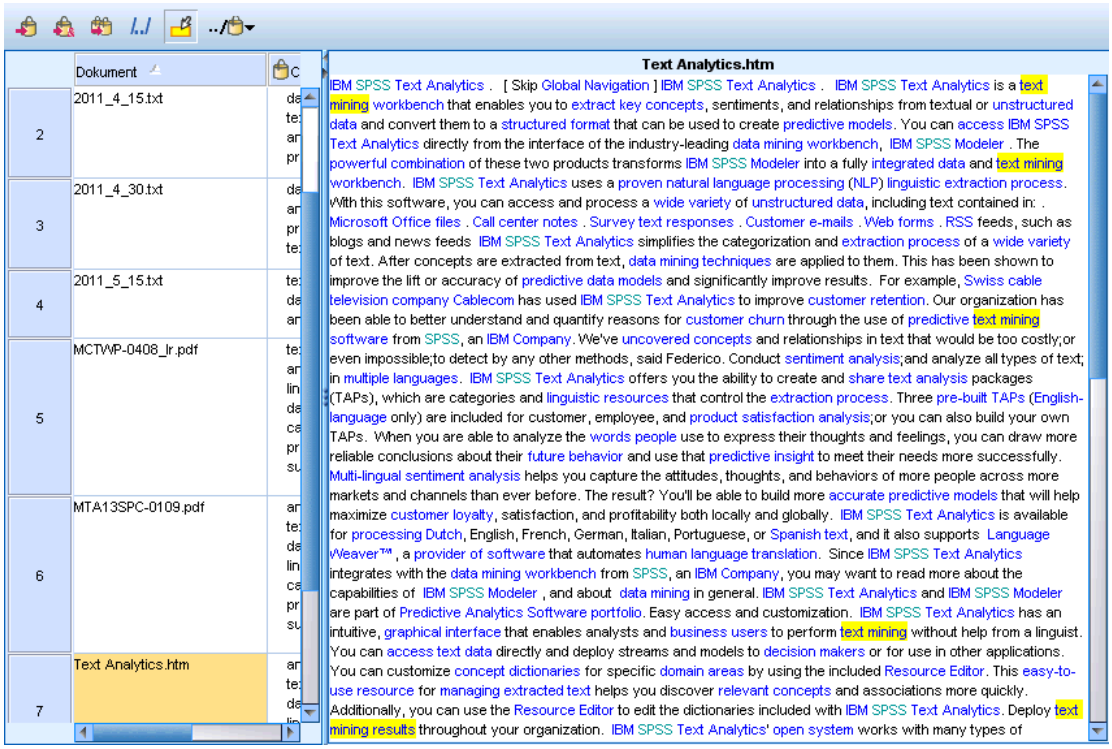
Datenbereich anzeigen und aktualisieren

Die Anzeige des Datenbereichs wird nicht automatisch aktualisiert, da die automatische Datenaktualisierung bei umfangreicheren Datenmengen sehr zeitaufwändig sein könnte. Daher können Sie bei der Auswahl von Typ- oder Konzeptmustern in dieser Ansicht auf Anzeige klicken, um den Inhalt des Datenbereichs zu aktualisieren.

Textdokumente oder Datensätze

Wenn Ihre Textdaten als Datensätze vorliegen und der Text relativ kurz ist, zeigt das Textfeld im Datenbereich die Textdaten vollständig an. Wenn Sie jedoch mit Datensätzen und größeren Datenmengen arbeiten, zeigt die Textfeldspalte einen kurzen Abschnitt des Texts und öffnet einen Textvorschaubereich auf der rechten Seite, in dem ein größerer Teil oder der ganze Text des in der Tabelle markierten Datensatzes angezeigt wird. Wenn Ihre Textdaten als einzelne Dokumente vorliegen, wird im Datenbereich der Dateiname des Dokuments angezeigt. Wenn Sie ein Dokument markieren, wird der Textvorschaubereich geöffnet und der Text des ausgewählten Dokuments angezeigt.

Abbildung 12-6
Datenbereich mit Textvorschaubereich



Farben und Hervorheben

Wenn Sie die Daten anzeigen, werden die in diesen Dokumenten oder Datensätzen gefundenen Konzepte und Deskriptoren farblich hervorgehoben, damit Sie diese leicht im Text identifizieren können. Die Farbkodierung entspricht den Typen, die den Konzepten zugewiesen sind. Alternativ können Sie die Maus über farbkodierte Elemente bewegen, um das Konzept anzuzeigen, unter dem das betreffende Element extrahiert wurde, und den Typ, dem es zugewiesen wurde. Nicht extrahierter Text wird schwarz angezeigt. Bei diesen nicht extrahierten Wörtern handelt es sich meistens um Verbindungselemente (*und* oder *mit*), Pronomen (*mich* oder *sie*) und Verben (*ist*, *haben* oder *nehmen*).

Datenbereichsspalten

Die Textfeldspalte wird immer angezeigt. Sie können jedoch auch andere Spalten anzeigen. Wählen Sie für die Anzeige anderer Spalten die Optionsfolge Ansicht > Datenbereich und anschließend die Spalte aus, die im Datenbereich angezeigt werden soll. Folgende Spalten stehen ggf. zur Anzeige zur Verfügung:

- **"Textfeldname" (Anzahl)/Dokumente.** Fügt eine Spalte für die Textdaten hinzu, aus denen Konzepte und Typ extrahiert wurden. Wenn sich Ihre Daten in Dokumenten befinden, trägt die Spalte den Titel "Dokumente" und nur der Dateiname des Dokuments oder der vollständige Pfad ist sichtbar. Um den Text für diese Dokumente einzusehen, müssen Sie den Fensterbereich "Textvorschau" betrachten. Die Anzahl der Zeilen im Fensterbereich "Daten"

wird in Klammern nach diesem Spaltennamen angezeigt. Es kann vorkommen, dass aufgrund einer Einschränkung im Dialogfeld “Optionen”, die zur Beschleunigung des Ladevorgangs dient, nicht alle Dokumente bzw. Datensätze angezeigt werden. Wenn die maximale Anzahl erreicht wurde, steht nach der Zahl die Angabe - Maximum.Für weitere Informationen siehe Thema Optionen: Registerkarte “Sitzung” in Kapitel 8 auf S. 135.

- **Kategorien.** Führt jede Kategorie auf, der ein Datensatz angehört. Wenn diese Spalte angezeigt wird, kann die Aktualisierung des Datenbereichs ein wenig länger dauern, da jeweils die aktuellsten Informationen angezeigt werden.
- **Relevanzrang.** Führt den Rang jedes Datensatzes einer einzelnen Kategorie auf. Dieser Rang gibt an, wie gut der Datensatz im Verhältnis zu den anderen Datensätzen in der Kategorie zu der Kategorie passt. Wählen Sie eine Kategorie im Fensterbereich “Kategorien” (oben links) aus, um den Rang anzuzeigen. Für weitere Informationen siehe Thema Kategorierelevanz in Kapitel 10 auf S. 181.
- **Kategorienanzahl.** Führt die Anzahl der Kategorien auf, denen ein Datensatz angehört.

Kapitel
13

Visualisierung von Diagrammen

Die Kategorie- und Konzeptansicht, die Clusteransicht und die Textlinkanalysenansicht verfügen jeweils über einen Visualisierungsbereich in der rechten oberen Ecke des Fensters. Sie können diesen Bereich nutzen, um Ihre Daten visuell zu untersuchen. Die folgenden Diagramme und Grafiken stehen zur Verfügung.

- **Kategorie- und Konzeptansicht.** In dieser Ansicht sind drei Diagramme und Grafiken zu finden: *Kategoriebalken*, *Kategorienetzdiagramm* und *Tabelle für Kategorienetzdiagramm*. In dieser Ansicht werden die Diagramme nur aktualisiert, wenn Sie auf Anzeigen klicken. Für weitere Informationen siehe Thema Kategoriendiagramme und Grafiken auf S. 262.
- **Clusteransicht.** In dieser Ansicht sind zwei Netzdiagramme zu finden: *Konzeptnetzdiagramm* und *Clusternetzdiagramm*. Für weitere Informationen siehe Thema Clusterdiagramme auf S. 266.
- **Textlinkanalysenansicht.** In dieser Ansicht sind zwei Netzdiagramme zu finden: *Konzeptnetzdiagramm* und *Typnetzdiagramm*. Für weitere Informationen siehe Thema Textlinkanalyse-Diagramme auf S. 268.

Weitere Informationen über alle zum Bearbeiten von Diagrammen verwendeten allgemeinen Symbolleisten und Paletten finden Sie im Abschnitt über das Bearbeiten von Diagrammen in der Online-Hilfe oder in der Datei *SourceProcessOutputNodes.pdf*, die sich im Ordner *\Documentation\en* auf dem IBM® SPSS® ModelerDVD befindet.

Kategoriendiagramme und Grafiken

Bei der Erstellung von Kategorien ist es wichtig, die Definitionen der Kategorien, die darin enthaltenen Dokumente oder Datensätze und die Überschneidung von Kategorien sorgfältig zu überprüfen. Im Visualisierungsbereich können Kategorien von etlichen Perspektiven betrachtet werden. Der Visualisierungsbereich befindet sich rechts oben im Fenster “Kategorien und Konzepte”. Sollte er noch nicht sichtbar sein, können Sie auf diesen Bereich über das Menü “Ansicht” (Ansicht > Bereiche > Visualisierung) zugreifen.

In dieser Ansicht bietet der Visualisierungsbereich drei Darstellungen der Gemeinsamkeiten bei der Kategorisierung von Dokumenten oder Datensätzen. Mit den Grafiken und Diagrammen in diesem Bereich können Sie Ihre Kategorisierungsergebnisse analysieren und Kategorien oder Berichte feiner abstimmen. Bei der Verfeinerung von Kategorien können Sie in diesem Bereich Ihre Kategoriedefinitionen überprüfen, um Kategorien aufzudecken, die zu ähnlich sind (z. B. über 75 % ihrer Dokumente oder Datensätze gemein haben) oder zu unterschiedlich sind. Wenn zwei Kategorien zu ähnlich sind, kann es sinnvoll sein, die beiden Kategorien zu kombinieren. Alternativ können Sie auch die Kategoriedefinitionen ausarbeiten, indem Sie bestimmte Deskriptoren von der einen oder anderen Kategorie entfernen.

Je nach Auswahl im Bereich “Extrahierungsergebnisse” , im Bereich “Kategorien” oder im Dialogfeld “Kategoriedefinitionen” können Sie die entsprechenden Interaktionen zwischen Dokumenten/Datensätzen und Kategorien auf allen Registerkarten in diesem Bereich anzeigen. Auf jeder Registerkarte werden ähnliche Informationen dargestellt, jedoch auf unterschiedliche Weise oder unterschiedlichen Detailebenen. Um ein Diagramm jedoch für die aktuelle Auswahl zu aktualisieren, klicken Sie in der Symbolleiste des Fensterbereichs oder Dialogfelds, in dem Sie Ihre Auswahl vorgenommen haben, auf Anzeigen.

Der Visualisierungsbereich im Fenster “Kategorien und Konzepte” bietet die folgenden Diagramme:

- **Kategoriebalkendiagramm.** Eine Tabelle und ein Balkendiagramm zeigen gemäß Ihrer Auswahl und den assoziierten Kategorien die Überschneidungen zwischen den Dokumenten/Datensätzen an. Das Balkendiagramm zeigt auch das jeweilige Verhältnis der Dokumente/Datensätze in den Kategorien zu der Gesamtzahl der Dokumente/Datensätze an. Für weitere Informationen siehe Thema Kategoriebalkendiagramm auf S. 263.
- **Kategorienetzdiagramm.** Dieses Diagramm zeigt die Überschneidung der Dokumente/Datensätze der Kategorien an, denen die Dokumente/Datensätze gemäß der Auswahl in den anderen Bereichen zugeordnet sind. Für weitere Informationen siehe Thema Kategorienetzdiagramm auf S. 264.
- **Tabelle für Kategorienetzdiagramm.** Auf dieser Registerkarte werden dieselben Informationen wie auf der Registerkarte “Kategorienetzdiagramm” dargestellt, allerdings als Tabelle. Die drei Spalten in der Tabelle können durch Klicken auf den Spaltentitel sortiert werden. Für weitere Informationen siehe Thema Tabelle für Kategorienetzdiagramm auf S. 265.

Für weitere Informationen siehe Thema Kategorisieren von Textdaten in Kapitel 10 auf S. 167.

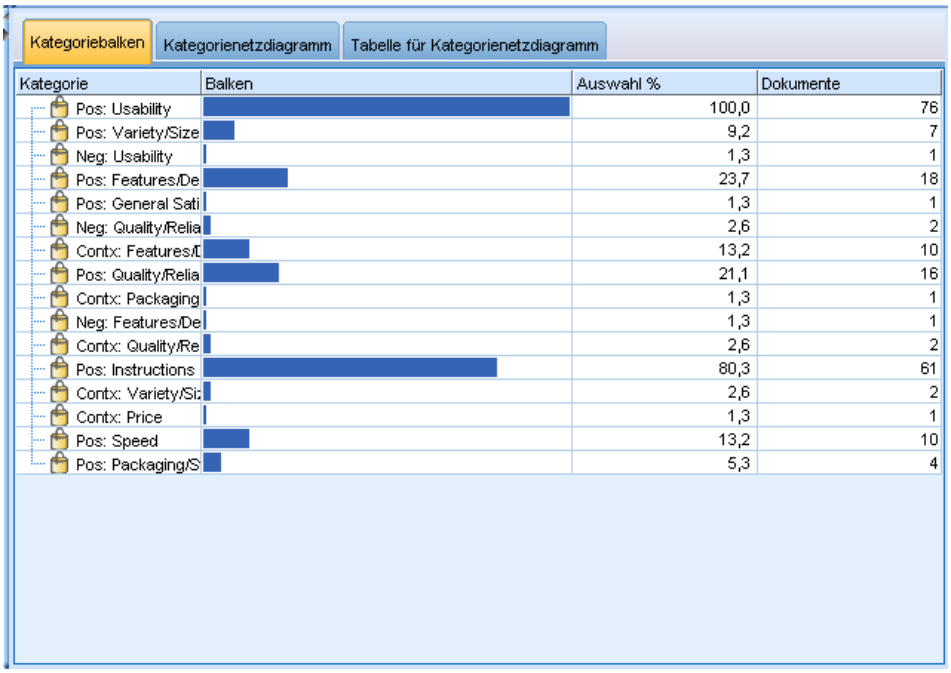
Kategoriebalkendiagramm

Auf dieser Registerkarte werden eine Tabelle und ein Balkendiagramm angezeigt, die gemäß Ihrer Auswahl und den assoziierten Kategorien die Überschneidungen zwischen den Dokumenten/Datensätzen anzeigen. Das Balkendiagramm zeigt auch das jeweilige Verhältnis der Dokumente/Datensätze in den Kategorien zu der Gesamtzahl der Dokumente oder Datensätze an. Das Layout dieses Diagramms kann nicht geändert werden. Sie können jedoch durch Klicken auf die Spaltenüberschriften die Spalten sortieren.

Das Diagramm enthält die folgenden Spalten:

- **Kategorie.** In dieser Spalte wird der Name der Kategorien in Ihrer Auswahl angezeigt. Standardmäßig wird die häufigste Kategorie in Ihrer Auswahl an erster Stelle aufgeführt.
- **Balken.** Diese Spalte zeigt auf visuelle Art das jeweilige Verhältnis der Dokumente oder Datensätze in einer Kategorie zu der Gesamtzahl der Dokumente oder Datensätze an.
- **Auswahl %.** Diese Spalte zeigt einen Prozentsatz auf der Basis des jeweiligen Verhältnisses der Dokumente oder Datensätze in einer Kategorie zu der Gesamtzahl der in der Auswahl enthaltenen Dokumente oder Datensätze an.
- **Dokumente.** In dieser Spalte wird die Anzahl der Dokumente oder Datensätze in einer Auswahl für die jeweilige Kategorie angezeigt.

Abbildung 13-1
Kategoriebalkendiagramm

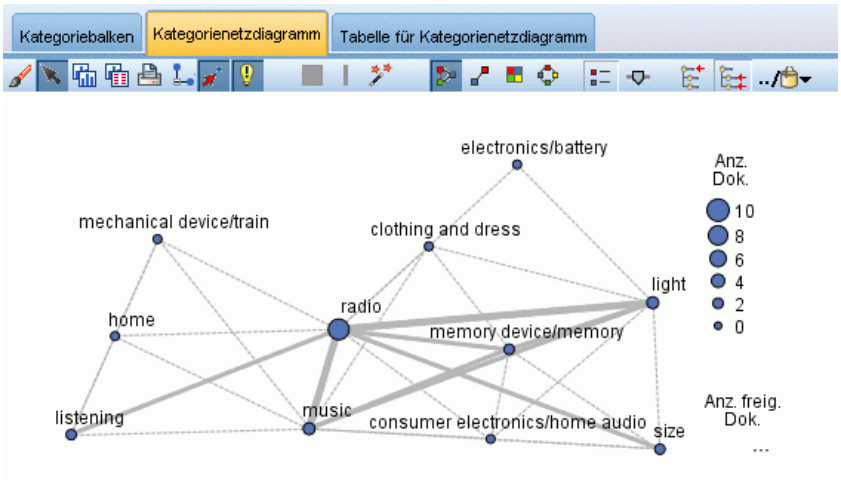


Kategorie	Balken	Auswahl %	Dokumente
Pos: Usability		100,0	76
Pos: Variety/Size		9,2	7
Neg: Usability		1,3	1
Pos: Features/De		23,7	18
Pos: General Sati		1,3	1
Neg: Quality/Relia		2,6	2
Contx: Features/I		13,2	10
Pos: Quality/Relia		21,1	16
Contx: Packaging		1,3	1
Neg: Features/De		1,3	1
Contx: Quality/Re		2,6	2
Pos: Instructions		80,3	61
Contx: Variety/Siz		2,6	2
Contx: Price		1,3	1
Pos: Speed		13,2	10
Pos: Packaging/S		5,3	4

Kategorienetzdiagramm

Auf dieser Registerkarte wird ein Kategorienetzdiagramm angezeigt. Das Netzdiagramm zeigt die Überschneidung der Dokumente/Datensätze der Kategorien an, denen die Dokumente bzw. Datensätze gemäß der Auswahl in den anderen Bereichen zugeordnet sind. Vorhandene Kategoriebezeichnungen werden im Diagramm angezeigt. Sie können über die Symbolleistenschaltflächen in diesem Bereich das Layout des Diagramms auswählen (Netz-, Kreis-, gerichtetes oder Gitterlayout).

Abbildung 13-2
Kategorienetzdiagramm, Gitterlayout



Im Netzdiagramm stellt jeder Knoten eine Kategorie dar. Sie können mit der Maus innerhalb des Bereichs Knoten auswählen und verschieben. Die Größe des Knotens stellt die relative Größe auf der Basis der Anzahl der Dokumente oder Datensätze der Kategorie in Ihrer Auswahl dar. Die Stärke und Farbe der Linie zwischen zwei Kategorien gibt die Anzahl der Dokumente oder Datensätze an, die in beiden Kategorien vorhanden sind. Wenn Sie im Untersuchungsmodus den Mauszeiger auf einen Knoten platzieren, zeigt eine QuickInfo den Namen (oder die Bezeichnung) der Kategorie und die Gesamtanzahl an Dokumenten oder Datensätzen in der Kategorie an.

Hinweis: Standardmäßig ist der Untersuchungsmodus für die Diagramme aktiviert, in denen Sie Knoten verschieben können. Sie können das Layout Ihrer Diagramme im Bearbeitungsmodus bearbeiten, u. a. Farben und Schriftarten, Legenden usw. Für weitere Informationen siehe Thema Verwenden von Diagrammsymbolleisten und Paletten auf S. 271.

Tabelle für Kategorienetzdiagramm

Auf dieser Registerkarte werden dieselben Informationen wie auf der Registerkarte "Kategorienetzdiagramm" dargestellt, allerdings als Tabelle. Die drei Spalten in der Tabelle können durch Klicken auf den Spaltentitel sortiert werden:

- **Anzahl.** In dieser Spalte wird die Anzahl der in beiden Kategorien vorhandenen oder gemeinsamen Dokumente oder Datensätze angezeigt.
- **Kategorie 1.** Diese Spalte zeigt den Namen der ersten Kategorie an, gefolgt von der – in Klammern eingeschlossenen – Gesamtzahl der darin enthaltenen Dokumente oder Datensätze.
- **Kategorie 2.** Diese Spalte zeigt den Namen der zweiten Kategorie an, gefolgt von der – in Klammern eingeschlossenen – Gesamtzahl der darin enthaltenen Dokumente oder Datensätze.

Abbildung 13-3
Tabelle für Kategorienetzdiagramm

Kategoriebalken | Kategorienetzdiagramm | Tabelle für Kategorienetzdiagramm

Anzahl	Kategorie 1	Kategorie 2
61	Pos: Usability(76)	Pos: Instructions(61)
18	Pos: Usability(76)	Pos: Features/Design(18)
16	Pos: Usability(76)	Pos: Quality/Reliability(16)
14	Pos: Features/Design(18)	Pos: Instructions(61)
13	Pos: Instructions(61)	Pos: Quality/Reliability(16)
10	Pos: Speed(10)	Pos: Usability(76)
10	Pos: Speed(10)	Pos: Instructions(61)
10	Pos: Usability(76)	Contx: Features/Design(10)
9	Contx: Features/Design(10)	Pos: Instructions(61)
8	Pos: Features/Design(18)	Pos: Quality/Reliability(16)
7	Pos: Variety/Size/Weight(7)	Pos: Usability(76)
6	Pos: Instructions(61)	Pos: Variety/Size/Weight(7)
4	Pos: Speed(10)	Pos: Quality/Reliability(16)
4	Pos: Usability(76)	Pos: Packaging/Storage(4)
3	Pos: Speed(10)	Pos: Features/Design(18)
2	Contx: Features/Design(10)	Pos: Quality/Reliability(16)
2	Contx: Quality/Reliability(2)	Contx: Features/Design(10)
2	Contx: Quality/Reliability(2)	Pos: Instructions(61)
2	Neg: Quality/Reliability(2)	Pos: Instructions(61)
2	Neg: Quality/Reliability(2)	Pos: Quality/Reliability(16)
2	Pos: Features/Design(18)	Pos: Packaging/Storage(4)
2	Pos: Features/Design(18)	Contx: Features/Design(10)
2	Pos: Instructions(61)	Contx: Variety/Size/Weight(2)

Clusterdiagramme

Nach dem Erstellen der Cluster können Sie diese visuell in den Netzdiagrammen im Visualisierungsbereich untersuchen. Im Visualisierungsbereich sind zwei Clustering-Ansichten möglich: ein Konzeptnetzdiagramm und ein Clusternetzdiagramm. Mit den Netzdiagrammen in diesem Bereich können Sie Clusterbildungsergebnisse analysieren und Konzepte und Regeln aufdecken, die Sie Ihren Kategorien hinzufügen sollten. Der Visualisierungsbereich befindet sich in der rechten oberen Ecke der Clusteransicht. Sollte er noch nicht sichtbar sein, können Sie auf diesen Bereich über das Menü "Ansicht" (Ansicht > Bereiche > Visualisierung) zugreifen. Durch die Auswahl eines Clusters im Clusterbereich können Sie automatisch die entsprechenden Diagramme im Visualisierungsbereich anzeigen.

Anmerkung: Standardmäßig befinden sich die Diagramme im Interaktions-/Auswahlmodus, in dem Sie Knoten verschieben können. Sie können das Layout Ihrer Diagramme im Bearbeitungsmodus bearbeiten, u. a. Farben und Schriftarten, Legenden usw. Für weitere Informationen siehe Thema Verwenden von Diagrammsymbolleisten und Paletten auf S. 271.

In der Clusteransicht sind zwei Netzdiagramme zu finden.

- **Konzeptnetzdiagramm**. Dieses Diagramm stellt alle Konzepte in dem bzw. den ausgewählten Cluster(n) sowie verknüpfte Konzepte außerhalb des Clusters dar. Anhand dieses Diagramms können Sie erkennen, wie die Konzepte innerhalb eines Clusters miteinander verknüpft sind sowie alle externen Links. Für weitere Informationen siehe Thema Konzeptnetzdiagramm auf S. 266.
- **Clusternetzdiagramm**. Dieses Diagramm zeigt den oder die ausgewählten Cluster und alle externen Links zwischen den ausgewählten Clustern als gepunktete Linien an. Für weitere Informationen siehe Thema Clusternetzdiagramm auf S. 267.

Für weitere Informationen siehe Thema Analyse von Clustern in Kapitel 11 auf S. 243.

Konzeptnetzdiagramm

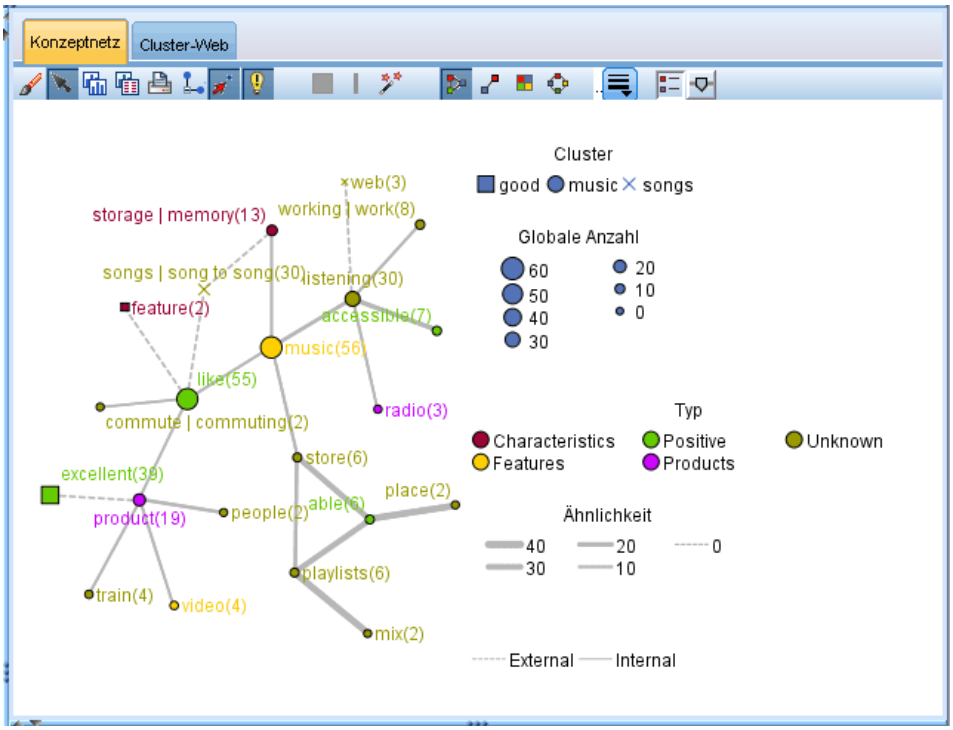
Auf dieser Registerkarte wird ein Netzdiagramm angezeigt mit allen Konzepten, die sich innerhalb des oder der ausgewählten Cluster(s) befinden, sowie verknüpfte Konzepte außerhalb des Clusters. Anhand dieses Diagramms können Sie erkennen, wie die Konzepte innerhalb eines Clusters miteinander verknüpft sind sowie alle externen Links. Jedes Konzept in einem Cluster wird als ein Knoten dargestellt, der gemäß der Typfarbe farbkodiert ist. Für weitere Informationen siehe Thema Erstellen von Typen in Kapitel 17 auf S. 314.

Die internen Links zwischen den Konzepten innerhalb eines Clusters werden als Linien dargestellt, deren Stärke je nach Auswahl in der Diagrammsymbolleiste unmittelbar entweder von der Anzahl der Dokumente, in denen jedes Konzeptpaar gleichzeitig vorkommt, oder dem Ähnlichkeits-Zusammenhangswert abhängt. Die externen Links zwischen den Konzepten eines Clusters und den Konzepten außerhalb des Clusters werden ebenfalls dargestellt.

Wenn Konzepte im Dialogfeld "Clusterdefinitionen" ausgewählt sind, zeigt das Konzeptnetzdiagramm diese Konzepte sowie alle assoziierten internen und externen Links zu diesen Konzepten an. Links zwischen anderen Konzepten, die keine der auswählten Konzepte enthalten, werden nicht im Diagramm angezeigt.

Anmerkung: Standardmäßig befinden sich die Diagramme im Interaktions-/Auswahlmodus, in dem Sie Knoten verschieben können. Sie können das Layout Ihrer Diagramme im Bearbeitungsmodus bearbeiten, u. a. Farben und Schriftarten, Legenden usw. Für weitere Informationen siehe Thema Verwenden von Diagrammsymbolleisten und Paletten auf S. 271.

Abbildung 13-4
Konzeptnetzdiagramm



Clusternetzdiagramm

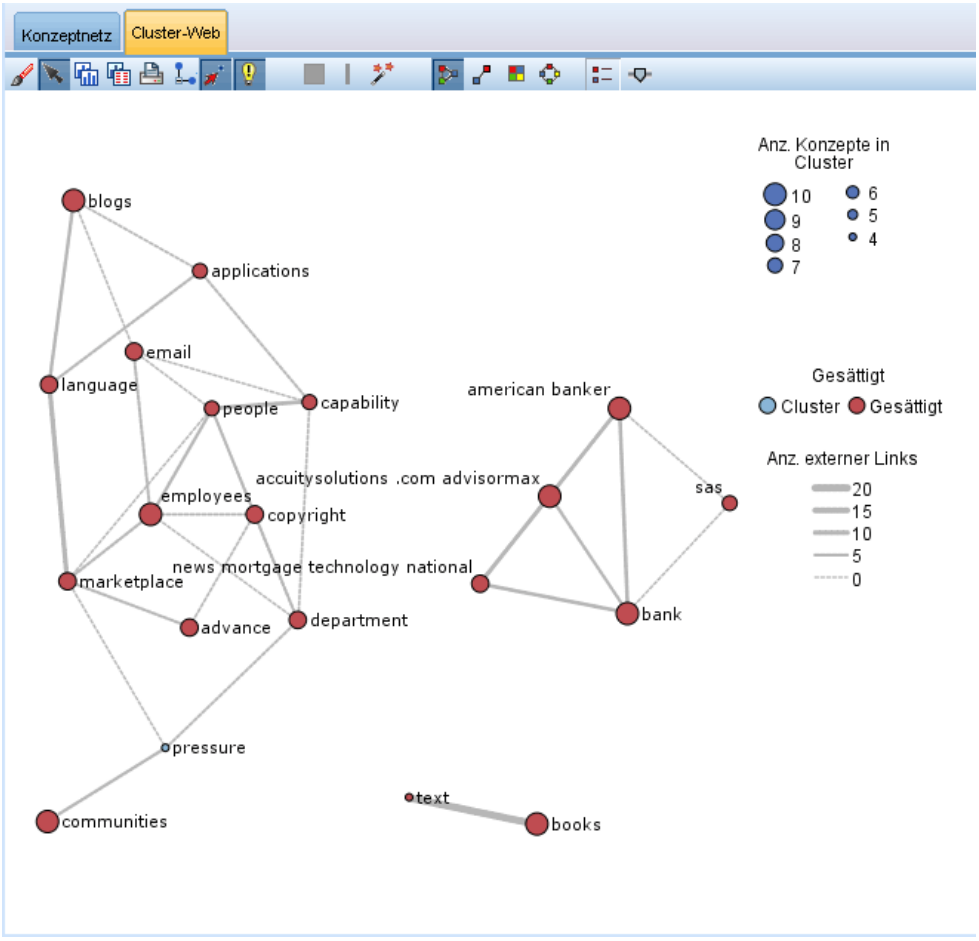
In dieser Registerkarte wird ein Netzdiagramm mit dem bzw. den ausgewählten Cluster(n) angezeigt. Die externen Links zwischen den ausgewählten Clustern sowie Links zwischen anderen Clustern werden als gepunktete Linien dargestellt. In einem Clusternetzdiagramm stellt jeder Knoten einen gesamten Cluster dar und die Stärke der Linien zwischen den Knoten stellt die Anzahl der externen Links zwischen zwei Clustern dar.

Wichtig: Um ein Clusternetzdiagramm anzuzeigen, müssen Sie bereits Cluster mit externen Links erstellt haben. Externe Links sind Verknüpfungen zwischen Konzeptpaaren, die sich in unterschiedlichen Clustern befinden (ein Konzept befindet sich in einem und das andere Konzept in einem anderen Cluster).

Beispiel: Nehmen wir an, wir haben zwei Cluster: `Cluster A` umfasst drei Konzepte: `A1`, `A2` und `A3`. `Cluster B` umfasst zwei Konzepte: `B1` und `B2`. Die folgenden Konzepte sind miteinander verknüpft: `A1-A2`, `A1-A3`, `A2-B1` (extern), `A2-B2` (extern), `A1-B2` (extern) und `B1-B2`. Das bedeutet, dass die Stärke der Linie im Clusternetzdiagramm die drei externen Links darstellen würde.

Anmerkung: Standardmäßig befinden sich die Diagramme im Interaktions-/Auswahlmodus, in dem Sie Knoten verschieben können. Sie können das Layout Ihrer Diagramme im Bearbeitungsmodus bearbeiten, u. a. Farben und Schriftarten, Legenden usw. Für weitere Informationen siehe Thema Verwenden von Diagrammsymbolleisten und Paletten auf S. 271.

Abbildung 13-5
Clusternetzdiagramm



Textlinkanalyse-Diagramme

Nach dem Extrahieren Ihrer Textlinkanalyse-(TLA-)Muster können Sie diese visuell in den Netzdiagrammen im Visualisierungsbereich untersuchen. Der Visualisierungsbereich bietet zwei Ansichten Ihrer TLA-Muster: ein Konzept-(Muster-)Netzdiagramm und ein Typ-(Muster-)Netzdiagramm. Anhand der Netzdiagramme in diesem Bereich können Muster visuell dargestellt werden. Der Visualisierungsbereich befindet sich in der rechten oberen Ecke der Textlinkanalyse. Sollte er noch nicht sichtbar sein, können Sie auf diesen Bereich über das Menü “Ansicht” (Ansicht > Bereiche > Visualisierung) zugreifen. Falls keine Auswahl getroffen wurde, ist der Diagrammbereich leer.

Anmerkung: Standardmäßig befinden sich die Diagramme im Interaktions-/Auswahlmodus, in dem Sie Knoten verschieben können. Sie können das Layout Ihrer Diagramme im Bearbeitungsmodus bearbeiten, u. a. Farben und Schriftarten, Legenden usw. Für weitere Informationen siehe Thema Verwenden von Diagrammsymbolleisten und Paletten auf S. 271.

In der Textlinkanalysenansicht sind zwei Netzdiagramme zu finden.

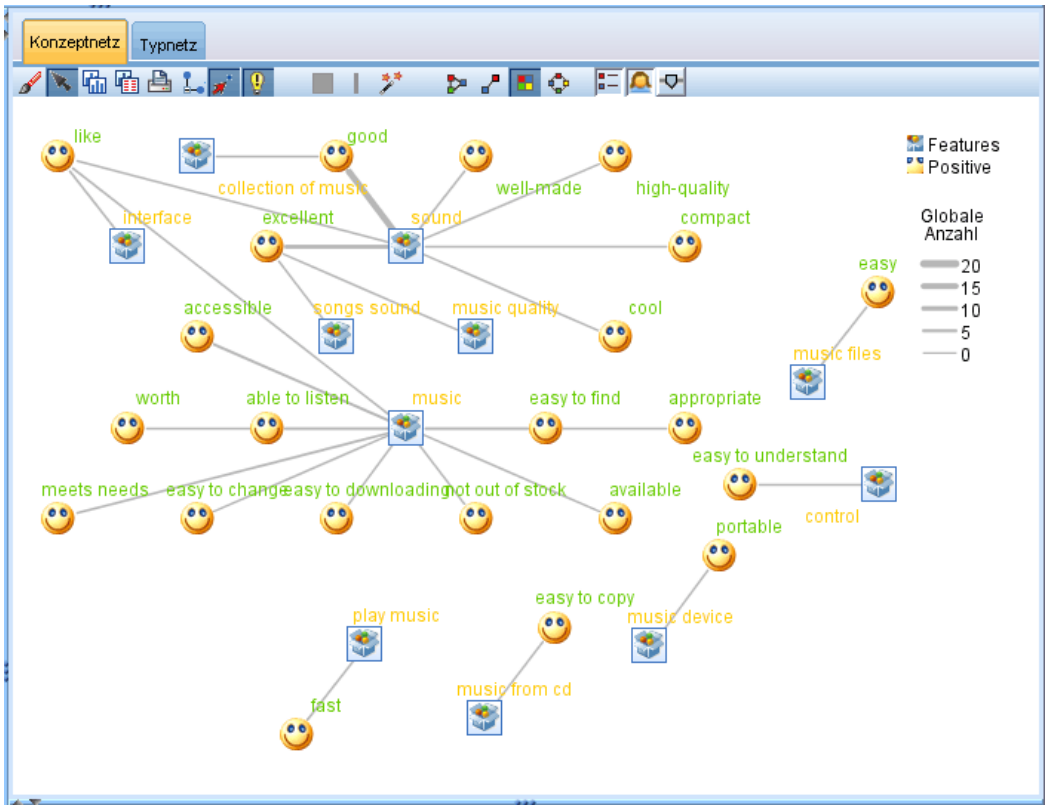
- **Konzeptnetzdiagramm**. In diesem Diagramm werden alle Konzepte in dem oder den ausgewählten Muster(n) angezeigt. Die Zeilenstärke und die Knotengrößen (sofern keine Typsymbole angezeigt werden) in einem Konzeptdiagramm zeigen die Anzahl der globalen Vorkommnisse in der ausgewählten Tabelle an. Für weitere Informationen siehe Thema Konzeptnetzdiagramm auf S. 269.
- **Typnetzdiagramm**. In diesem Diagramm werden alle Typen in dem oder den ausgewählten Muster(n) angezeigt. Die Zeilenstärke und die Knotengrößen (sofern keine Typsymbole angezeigt werden) im Diagramm zeigen die Anzahl der globalen Vorkommnisse in der ausgewählten Tabelle an. Knoten werden entweder durch eine Typfarbe oder durch ein Symbol dargestellt. Für weitere Informationen siehe Thema Typnetzdiagramm auf S. 270.

Für weitere Informationen siehe Thema Untersuchen von Textlinkanalyse in Kapitel 12 auf S. 252.

Konzeptnetzdiagramm

In diesem Netzdiagramm werden alle in der aktuellen Auswahl dargestellten Konzepte angezeigt. Beispiel: Wenn Sie ein Typmuster mit drei übereinstimmenden Konzeptmustern ausgewählt haben, zeigt dieses Diagramm drei Mengen mit verknüpften Konzepten an. Die Zeilenstärke und die Knotengrößen in einem Konzeptdiagramm stellen die globalen Häufigkeitswerte dar. Das Diagramm stellt die gleichen Informationen visuell dar, die auch in den Musterbereichen ausgewählt sind. Die Typen jedes einzelnen Konzepts werden je nach Auswahl über die Diagrammsymbolleiste entweder farblich dargestellt oder durch ein Symbol. Für weitere Informationen siehe Thema Verwenden von Diagrammsymbolleisten und Paletten auf S. 271.

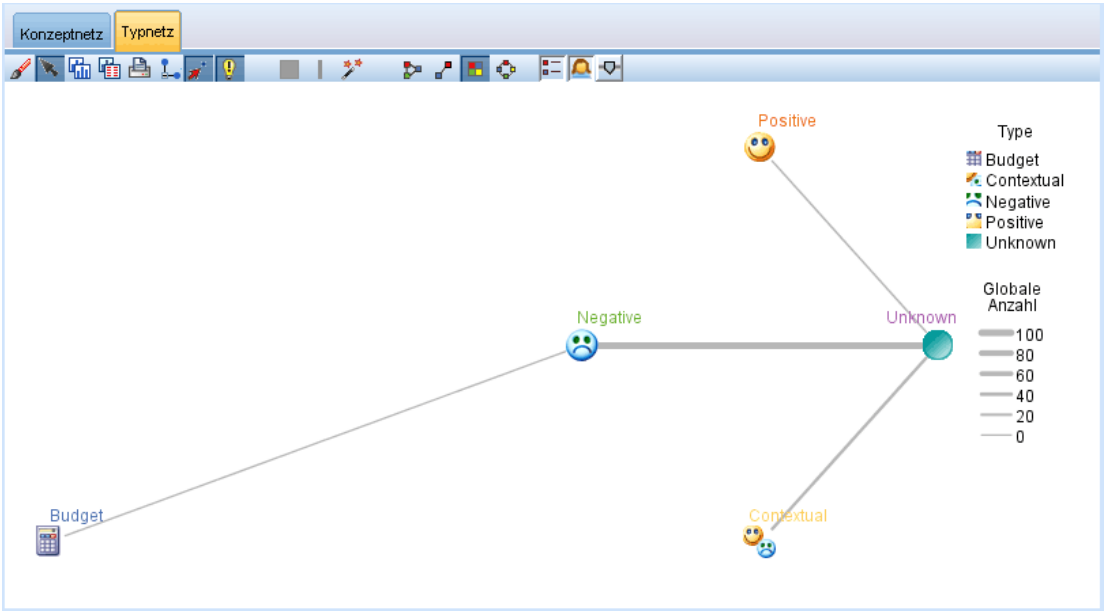
Abbildung 13-6
Konzeptnetzdiagramm



Typnetzdiagramm

In diesem Netzdiagramm werden alle Typmuster für die aktuelle Auswahl dargestellt. Beispiel: Wenn Sie zwei Konzeptmuster ausgewählt haben, zeigt dieses Diagramm pro Typ einen Knoten in den ausgewählten Mustern an sowie die dazwischen liegenden Links, die es im gleichen Muster aufgefunden hat. Die Zeilenstärke und die Knotengrößen stellen die globalen Häufigkeitswerte der Menge dar. Das Diagramm stellt die gleichen Informationen visuell dar, die auch in den Musterbereichen ausgewählt sind. Zusätzlich zu den im Diagramm angezeigten Typnamen werden die Typen je nach Auswahl über die Diagrammsymbolleiste entweder farblich gekennzeichnet oder durch ein Symbol. Für weitere Informationen siehe Thema Verwenden von Diagrammsymbolleisten und Paletten auf S. 271.

Abbildung 13-7
Typnetzdiagramm



Verwenden von Diagrammsymbolleisten und Paletten

Für jedes Diagramm steht eine Symbolleiste zur Verfügung, über die Sie schnell auf häufig verwendete Paletten zugreifen können, mit denen Sie zahlreiche Aktionen mit Ihren Diagrammen durchführen können. Jede Ansicht (Kategorien und Konzepte, Cluster, Textlinkanalyse) hat eine etwas andere Symbolleiste. Sie können zwischen den folgenden Ansichtsmodi wählen: *Untersuchen* oder *Bearbeiten*.

Während Sie im Sondierungsmodus die durch die Visualisierung dargestellten Daten und Werte erforschen, gestattet Ihnen der Bearbeitungsmodus, das Layout und Aussehen der Visualisierung zu ändern. Sie können beispielsweise die Schriftarten und die Farben so ändern, dass sie den Stilvorgaben Ihres Unternehmens entsprechen. In diesen Modus wechseln Sie, indem Sie Ansicht > Bereich “Visualisierung” > Bearbeitungsmodus aus den Menüs wählen (oder auf das entsprechende Symbol in der Symbolleiste klicken).

Im Bearbeitungsmodus stehen mehrere Symbolleisten zur Verfügung, mit denen sich die verschiedenen Aspekte des Visualisierungslayouts beeinflussen lassen. Wenn Sie einige der Symbolleisten nicht benötigen, können Sie sie ausblenden und im Dialogfeld, in dem die Grafik angezeigt wird, mehr Platz schaffen. Sie blenden Symbolleisten aus bzw. deaktivieren sie, indem Sie im Menü “Ansicht” auf den entsprechenden Symbolleisten- oder Palettennamen klicken.

Weitere Informationen über alle zum Bearbeiten von Diagrammen verwendeten allgemeinen Symbolleisten und Paletten finden Sie im Abschnitt über das Bearbeiten von Diagrammen in der Online-Hilfe oder in der Datei *SourceProcessOutputNodes.pdf*, die sich im Ordner *\Documentation\en* auf dem IBM® SPSS® ModelerDVD befindet.

Tabelle 13-1
Schaltflächen der Symbolleiste "Textanalyse"

Schaltfläche/ Liste	**Beschreibung**
	Aktiviert den Bearbeitungsmodus. Wechseln Sie in den Bearbeitungsmodus, wenn Sie das Aussehen des Diagramms ändern möchten, z. B. die Schrift vergrößern, die Farben Ihren Unternehmensrichtlinien anpassen oder Bezeichnungen und Legenden entfernen möchten.
	Aktiviert den Untersuchungsmodus. Standardmäßig ist der Untersuchungsmodus aktiv; das bedeutet, dass Sie Knoten im Diagramm verschieben und ziehen und auch mit der Maus über Diagrammobjekte fahren können, um weitere QuickInfos anzuzeigen.
	Wählen Sie die Art der Netzanzeige für die Diagramme in der Kategorie- und Konzeptansicht und der Textlinkanalysenansicht aus. ■ Kreislayout. Allgemeines Layout, das für alle Diagramme verwendet werden kann. Bei diesem Layout wird das Diagramm unter der Voraussetzung erstellt, dass alle Links ungerichtet sind, und alle Knoten werden gleich behandelt. Knoten werden nur um den Kreisumfang herum platziert. ■ Netzlayout. Allgemeines Layout, das für alle Diagramme verwendet werden kann. Bei diesem Layout wird das Diagramm unter der Voraussetzung erstellt, dass alle Links ungerichtet sind, und alle Knoten werden gleich behandelt. Knoten werden frei innerhalb des Layouts platziert. ■ Gerichtetes Layout. Layout, das nur für gerichtete Diagramme verwendet werden sollte. In diesem Layout werden baumförmige Strukturen von den Stammknoten bis zu den Blattknoten erstellt und nach Farben sortiert. Hierarchische Daten lassen sich mit diesem Layout sehr gut anzeigen. ■ Gitterlayout. Allgemeines Layout, das für alle Diagramme verwendet werden kann. Bei diesem Layout wird das Diagramm unter der Voraussetzung erstellt, dass alle Links ungerichtet sind, und alle Knoten werden gleich behandelt. Knoten werden nur an Gitterpunkten innerhalb dieses Bereichs platziert.
	Linkgrößendarstellung. Wählen Sie, was die Stärke der Linie im Diagramm darstellt. Dies gilt nur für die Clusteransicht. Das Clusternetzdiagramm zeigt nur die Anzahl der externen Links zwischen Clustern an. Sie können zwischen folgenden Optionen wählen: ■ Ähnlichkeit. Die Stärke gibt die Anzahl externer Verknüpfungen zwischen zwei Clustern an. ■ Kookkurrenz. Die Stärke gibt die Anzahl von Dokumenten an, in denen es eine Kookkurrenz von Deskriptoren gibt.
	Eine Umschaltfläche, mit der die Legende angezeigt wird. Wenn diese Schaltfläche nicht gedrückt ist, wird die Legende nicht angezeigt.
	Eine Umschaltfläche, mit der anstelle der Farben der Typen deren Symbole angezeigt werden. Dies gilt nur für die Textlinkanalysenansicht.
	Eine Umschaltfläche, mit der der Links-Schieberegler unter dem Diagramm angezeigt wird. Sie können die Ergebnisse durch Verschieben des Pfeils filtern.
	Zeigt das Diagramm für die höchste Ebene der ausgewählten Kategorien anstelle deren Unterkategorien an.

Schaltfläche/ Liste	Beschreibung
	Zeigt das Diagramm für die niedrigste Ebene der ausgewählten Kategorien an.
	Diese Option steuert, wie die Namen von Unterkategorien in der Ausgabe angezeigt werden. ■ **Vollständiger Kategoriepfad.** Diese Option gibt den Namen der Kategorie und den vollständigen Pfad von übergeordneten Kategorien (falls zutreffend) mit Schrägstrichen zwischen den Namen von Kategorien und Unterkategorien an. ■ **Kurzer Kategoriepfad.** Diese Option gibt nur den Namen der Kategorie aus, verwendet aber Auslassungszeichen, um die Anzahl der übergeordneten Kategorien für die betreffende Kategorie anzuzeigen. ■ **Kategorie der untersten Ebene.** Diese Option gibt nur den Namen der Kategorie aus, ohne dass der vollständige Pfad oder übergeordnete Kategorien angezeigt werden.

Kapitel
14

Sitzungsressourceneditor

IBM® SPSS® Modeler Text Analytics erfasst und extrahiert Schlüsselkonzepte aus Textdaten schnell und genau. Dieser Extrahierungsprozess beruht erheblich darauf, dass durch die linguistischen Ressourcen festgelegt wird, wie Informationen aus Textdaten extrahiert werden. Standardmäßig stammen diese Ressourcen aus Ressourcenvorlagen.

SPSS Modeler Text Analytics wird mit einer Reihe von spezialisierten **Ressourcenvorlagen** ausgeliefert, die in Form von Bibliotheken und erweiterten Ressourcen eine Reihe von linguistischen und nicht linguistischen Ressourcen zur Unterstützung bei der Definition enthalten, wie Ihre Daten gehandhabt und extrahiert werden sollen. Für weitere Informationen siehe Thema Vorlagen und Ressourcen in Kapitel 15 auf S. 280.

Im Knotendialogfeld können Sie eine Kopie der Ressourcen der Vorlage in den Knoten laden. Sobald Sie eine interaktive Workbench-Sitzung gestartet haben, können Sie diese Ressourcen bei Bedarf an die Daten dieses Knotens anpassen. Während einer interaktiven Workbench-Sitzung können Sie in der Resource Editorenansicht mit Ihren Ressourcen arbeiten. Wenn Sie eine interaktive Sitzung starten, erfolgt eine Extrahierung mit den Ressourcen, die im Knotendialogfeld geladen wurden, sofern Sie die Daten und Extrahierungsergebnisse im Knoten nicht zwischengespeichert haben.

Bearbeiten von Ressourcen im Ressourceneditor

Der Resource Editor ermöglicht den Zugriff auf eine Reihe von Ressourcen, die für die Gewinnung von Extrahierungsergebnissen (Konzepten, Typen und Mustern) für interaktive Workbench-Sitzungen verwendet werden. Dieser Editor ist dem Template Editor sehr ähnlich, im Resource Editor bearbeiten Sie jedoch die Ressourcen für diese Sitzung. Wenn Sie die Arbeiten an den Ressourcen sowie andere Arbeiten durchgeführt haben, können Sie den Modellierungsknoten aktualisieren, um diese Arbeiten zu speichern. Dadurch können sie in nachfolgenden interaktiven Workbench-Sitzungen wiederhergestellt werden. Für weitere Informationen siehe Thema Aktualisieren von Modellierungsknoten und Speichern in Kapitel 8 auf S. 140.

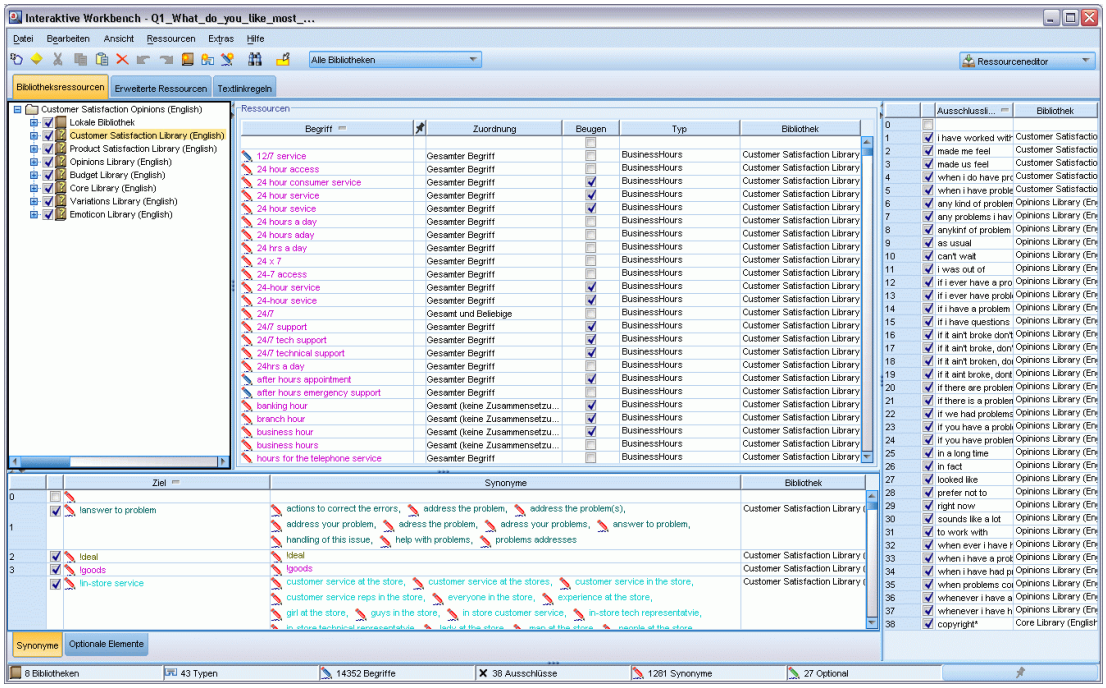
Falls Sie die Vorlagen direkt bearbeiten möchten, die zum Laden von Ressourcen in Knoten verwendet werden, empfiehlt sich die Verwendung des Template Editors. Viele der Aufgaben, die Sie im Resource Editor durchführen können, werden auf dieselbe Weise durchgeführt wie im Template Editor. Dazu zählen die folgenden Aufgaben:

- **Arbeiten mit Bibliotheken**Für weitere Informationen siehe Thema Mit Bibliotheken arbeiten in Kapitel 16 auf S. 297.
- **Erstellen von Typ-Wörterbüchern**Für weitere Informationen siehe Thema Erstellen von Typen in Kapitel 17 auf S. 314.
- **Hinzufügen von Fachausdrücken zu Wörterbüchern**Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken in Kapitel 17 auf S. 316.
- **Erstellen von Synonymen**Für weitere Informationen siehe Thema Definieren von Synonymen in Kapitel 17 auf S. 325.

- **Importieren und Exportieren von Vorlagen**Für weitere Informationen siehe Thema Importieren und Exportieren von Vorlagen in Kapitel 15 auf S. 290.
- **Veröffentlichen von Bibliotheken**Für weitere Informationen siehe Thema Bibliotheken veröffentlichen in Kapitel 16 auf S. 307.

Für niederländischen, englischen, französischen, deutschen, italienischen, portugiesischen und spanischen Text

Abbildung 14-1
Ressourceneditor-Ansicht für andere Sprachen als japanisch

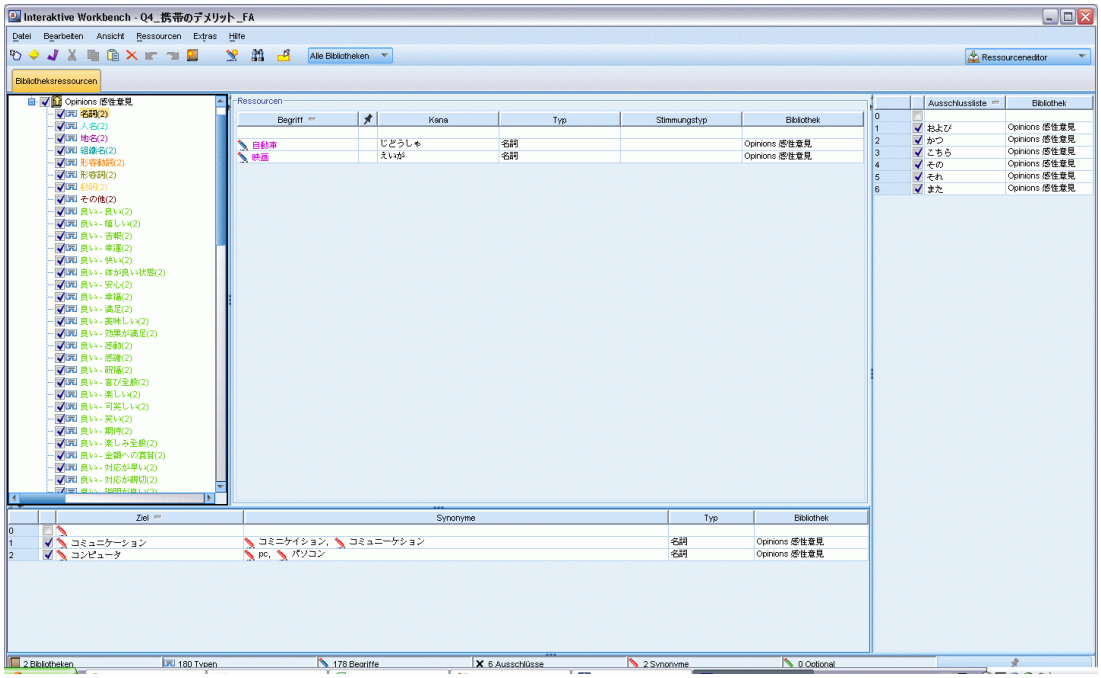


Für japanischen Text

Die Editoroberfläche für die japanische Sprache unterscheidet sich von der Oberfläche für andere Sprachen. Für weitere Informationen siehe Thema Bearbeitungsressourcen für japanischen Text in Anhang A auf S. 383.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Abbildung 14-2
Ressourceneditor-Ansicht für japanischen Text

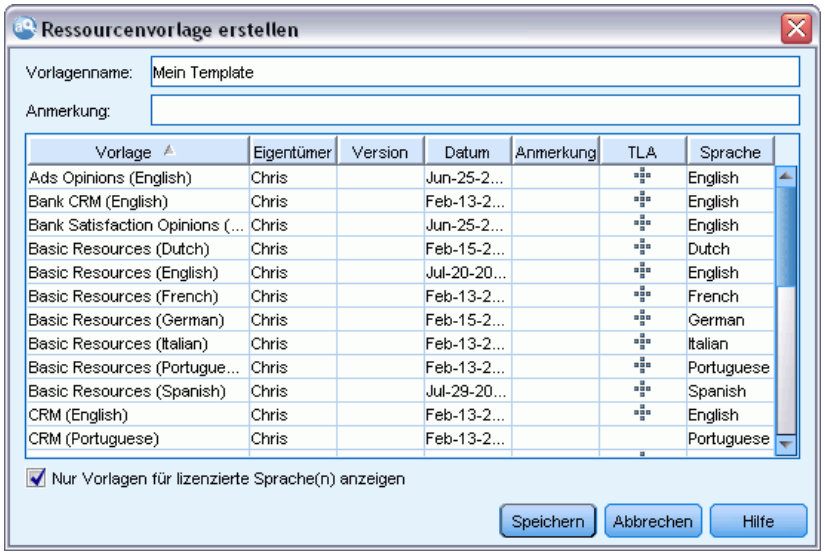


Erstellen und Aktualisieren von Vorlagen

Wenn Sie Änderungen an Ihren Ressourcen vornehmen und sie in Zukunft erneut verwenden möchten, speichern Sie die Ressourcen als Vorlage. Dabei können Sie die Vorlage unter einem bestehenden Vorlagennamen oder unter einem neuen Namen speichern. Wenn Sie künftig diese Vorlage laden, erhalten Sie dieselben Ressourcen. Für weitere Informationen siehe Thema Kopieren von Ressourcen aus TAPs und Vorlagen in Kapitel 3 auf S. 42.

Anmerkung: Sie können Ihre Bibliotheken auch veröffentlichen und mit anderen Benutzern gemeinsam nutzen. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen in Kapitel 16 auf S. 306.

Abbildung 14-3
Dialogfeld "Ressourcenvorlage erstellen"



So erstellen (oder aktualisieren) Sie eine Vorlage:

- Wählen Sie in den Menüs der Resource Editor-Ansicht die Optionsfolge Ressourcen > Ressourcenvorlage erstellen. Das Dialogfeld "Ressourcenvorlage erstellen" wird geöffnet.
- Geben Sie einen neuen Namen in das Feld "Vorlagenname" ein, wenn Sie eine neue Vorlage erstellen möchten. Wählen Sie eine Vorlage in der Tabelle aus, um eine vorhandene Vorlage mit den derzeit geladenen Ressourcen zu überschreiben.
- Klicken Sie auf Speichern, um die Vorlage zu erstellen.

Wichtig: Da die Vorlagen bei der Auswahl im Knoten und nicht bei der Ausführung des Streams geladen werden, müssen Sie die Ressourcenvorlage erneut in alle anderen Knoten laden, in denen sie verwendet wird, um die neuesten Änderungen zu erhalten. Für weitere Informationen siehe Thema Aktualisieren von Knotenressourcen nach dem Laden in Kapitel 15 auf S. 288.

Wechseln von Ressourcenvorlagen

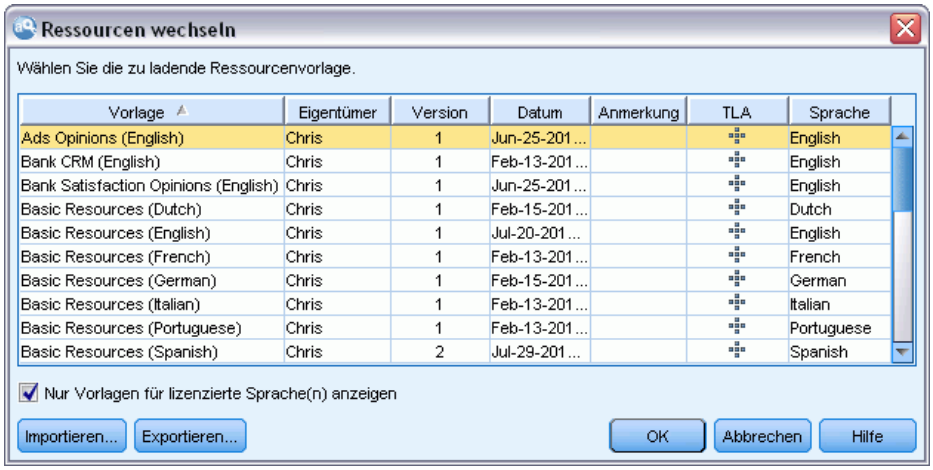
Wenn Sie die aktuell in der Sitzung geladenen Ressourcen durch eine Kopie der Ressourcen einer anderen Vorlage ersetzen möchten, können Sie zu diesen wechseln. Dadurch werden die in dieser Sitzung geladenen Ressourcen überschrieben. Wenn Sie die Ressourcen wechseln, um vordefinierte Text Link Analysis-Musterregeln (TLA-Musterregeln) zu erhalten, stellen Sie sicher, dass Sie eine Vorlage auswählen, bei der die Musterregeln in der TLA-Spalte hervorgehoben sind.

Wichtig: Sie können nicht von einer japanischen Vorlage in eine nicht-japanische Vorlage wechseln und umgekehrt. *Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Das Wechseln der Ressourcen ist vor allem dann sinnvoll, wenn Sie die Arbeit der Sitzung wiederherstellen möchten (Kategorien, Muster und Ressourcen), aber eine aktualisierte Kopie der Ressourcen aus einer Vorlage laden möchten, ohne dabei die übrigen Gegenstände Ihrer Sitzung zu verlieren.Sie können die Vorlage auswählen, deren Inhalt Sie in den Resource Editor kopieren möchten, und dann auf OK klicken. Dadurch werden die Ressourcen in dieser sitzung ersetzt. Stellen Sie sicher, dass Sie am Ende der Sitzung den Modellierungsknoten aktualisieren, wenn Sie diese Änderungen für den nächsten Start der interaktiven Workbench-Sitzung beibehalten möchten.

Anmerkung: Wenn Sie während einer interaktiven Sitzung zum Inhalt einer anderen Vorlage wechseln, bleibt der Name der im Knoten aufgeführten Vorlage derselbe wie bei der zuletzt geladenen und kopierten Vorlage. Um diese Ressourcen bzw. andere Arbeiten der Sitzung nutzen zu können, müssen Sie den Modellierungsknoten vor Beenden der Sitzung aktualisieren und im Knoten die Option Arbeit der Sitzung verwenden auswählen. Für weitere Informationen siehe Thema Aktualisieren von Modellierungsknoten und Speichern in Kapitel 8 auf S. 140.

Abbildung 14-4
Dialogfeld "Ressourcen wechseln"



So wechseln Sie Ressourcen:

- Wählen Sie in den Menüs der Resource Editor-Ansicht die Optionsfolge Ressourcen > Ressourcenvorlage wechseln. Das Dialogfeld "Ressourcenvorlagen wechseln" wird geöffnet.
- Wählen Sie die gewünschte Vorlage aus der Tabelle aus.
- Klicken Sie auf OK, um die derzeit geladenen Ressourcen zu entfernen und stattdessen eine Kopie der Ressourcen der ausgewählten Vorlage zu verwenden. Wenn Sie Änderungen in Ihren Ressourcen vorgenommen haben und Ihre Bibliotheken für künftige Verwendungen speichern möchten, können Sie sie vor dem Wechseln veröffentlichen, aktualisieren und für die gemeinsame Nutzung freigeben. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen in Kapitel 16 auf S. 306.

Teil III:
Vorlagen und Ressourcen

Kapitel
15

Vorlagen und Ressourcen

IBM® SPSS® Modeler Text Analytics erfasst und extrahiert Schlüsselkonzepte aus Textdaten schnell und genau. Dieser Extrahierungsprozess beruht erheblich darauf, dass durch die linguistischen Ressourcen festgelegt wird, wie Informationen aus Textdaten extrahiert werden. Für weitere Informationen siehe Thema Wie Extrahierung funktioniert in Kapitel 1 auf S. 7. Sie können die Feinabstimmung für diese Ressourcen in der Resource Editor-Ansicht vornehmen.

Bei der Installation der Software erhalten Sie außerdem eine Reihe von spezialisierten Ressourcen. Diese im Lieferumfang enthaltenen Ressourcen ermöglichen es Ihnen, von jahrelangen Forschungen zu profitieren und Feinabstimmungen für spezifische Sprachen und Anwendungen vorzunehmen. Da die mitgelieferten Ressourcen ggf. nicht perfekt an den Kontext Ihrer Daten angepasst sind, können Sie diese Ressourcenvorlagen bearbeiten oder sogar benutzerdefinierte Bibliotheken erstellen und verwenden, die spezifisch auf die Daten Ihrer Organisation abgestimmt sind. Diese Ressourcen liegen in verschiedenen Formaten vor und jede davon kann in Ihrem Sitzungs verwendet werden. Ressourcen finden Sie an folgenden Stellen:

- **Ressourcenvorlagen**. Vorlagen setzen sich aus einer Reihe von Bibliotheken, Typen und einigen erweiterten Ressourcen zusammen, die ein spezialisiertes Ressourcen-Set bilden, das auf eine bestimmte Domäne (wie z. B. Meinungen zu einem Produkt) ausgerichtet ist.
- **Text Analysis Packages (TAP)**. Zusätzlich zu den Ressourcen, die in einer Vorlage gespeichert sind, bündeln TAPs auch ein oder mehrere spezielle Kategoriensets, die mithilfe jener Ressourcen erstellt wurden, so dass die Kategorien und die Ressourcen gemeinsam gespeichert werden und wiederverwendbar sind. Für weitere Informationen siehe Thema Verwendung von Text Analysis Packages in Kapitel 10 auf S. 230.
- **Bibliotheken**. Bibliotheken werden als Bausteine für TAPs und Vorlagen verwendet. So können Ressourcen innerhalb Ihrer Sitzung auch einzeln hinzugefügt werden. Jede Bibliothek besteht aus mehreren Wörterbüchern zur Definition und Verwaltung von Typen, Synonymen und Ausschlusslisten. Während Wörterbücher auch einzeln bezogen werden können, werden sie in Vorlagen und TAPs vorab gemeinsam verpackt. Für weitere Informationen siehe Thema Mit Bibliotheken arbeiten in Kapitel 16 auf S. 297.

Anmerkung: Im Laufe der Extrahierung werden auch einige kompilierte interne Ressourcen verwendet. Diese kompilierten Ressourcen enthalten eine große Anzahl von Definitionen, die die Typen in der Core Library ergänzen. Diese kompilierten Ressourcen können nicht bearbeitet werden.

Der Resource Editor ermöglicht den Zugriff auf eine Reihe von Ressourcen, die für die Gewinnung von Extrahierungsergebnissen (Konzepten, Typen und Mustern) verwendet werden. Im Resource Editor können Sie eine Vielzahl von Aufgaben durchführen, darunter:

- **Arbeiten mit Bibliotheken**Für weitere Informationen siehe Thema Mit Bibliotheken arbeiten in Kapitel 16 auf S. 297.
- **Erstellen von Typ-Wörterbüchern**Für weitere Informationen siehe Thema Erstellen von Typen in Kapitel 17 auf S. 314.

- **Hinzufügen von Fachausdrücken zu Wörterbüchern**Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken in Kapitel 17 auf S. 316.
- **Erstellen von Synonymen**Für weitere Informationen siehe Thema Definieren von Synonymen in Kapitel 17 auf S. 325.
- **Aktualisierung der Ressourcen in TAPs**. Für weitere Informationen siehe Thema Aktualisierung von Text Analysis Packages in Kapitel 10 auf S. 235.
- **Erstellen von Vorlagen**. Für weitere Informationen siehe Thema Erstellen und Aktualisieren von Vorlagen in Kapitel 14 auf S. 276.
- **Importieren und Exportieren von Vorlagen**Für weitere Informationen siehe Thema Importieren und Exportieren von Vorlagen auf S. 290.
- **Veröffentlichen von Bibliotheken**Für weitere Informationen siehe Thema Bibliotheken veröffentlichen in Kapitel 16 auf S. 307.

Vorlageneditor im Vergleich zum Ressourceneditor

Es gibt zwei Methoden zum Arbeiten mit und Bearbeiten von Vorlagen, Bibliotheken und deren Ressourcen. Sie können im Template Editor oder Resource Editor mit linguistischen Ressourcen arbeiten.

Template Editor

Im Template Editor können Sie Ressourcenvorlagen ohne eine interaktive Workbench-Sitzung und unabhängig von einem bestimmten Knoten oder Stream erstellen und bearbeiten. Sie können diesen Editor dazu verwenden, Ressourcenvorlagen vor dem Laden in den Textlinkanalyse- und Text-Mining-Modellierungsknoten zu erstellen bzw. zu bearbeiten.

Der Zugriff auf den Template Editor erfolgt über die IBM® SPSS® Modeler-Hauptsymbolleiste bzw. über das Menü Extras > Template Editor.

Resource Editor

Der Resource Editor, auf den in einer interaktiven Workbench-Sitzung zugegriffen werden kann, bietet die Möglichkeit, mit den Ressourcen innerhalb des Kontexts eines bestimmten Knotens und Datensatzes zu arbeiten. Beim Hinzufügen eines Text-Mining-Modellierungsknotens zu einem Stream können Sie eine Kopie des Inhalts einer Ressourcenvorlage oder eines Textanalysepakets (Kategorie-Sets *und* linguistische Ressourcen) laden, um zu steuern, wie Text für Text-Mining extrahiert wird. Beim Starten einer interaktiven Workbench-Sitzung können Sie neben dem Erstellen von Kategorien, Extrahieren von Textlinkanalyse-Mustern und Erstellen von Category-Modellen darüber hinaus auch die Ressourcen für die Daten dieser Sitzung in der integrierten Resource Editorenansicht durch Feinabstimmung optimieren. Für weitere Informationen siehe Thema Bearbeiten von Ressourcen im Ressourceneditor in Kapitel 14 auf S. 274.

Beim Arbeiten mit Ressourcen in einer interaktiven Workbench-Sitzung wirken sich diese Änderungen nur auf diese Sitzung aus. Wenn Sie Ihre Arbeit (Ressourcen, Kategorien, Muster usw.) speichern möchten, um in einer nachfolgenden Sitzung weiterzuarbeiten, müssen Sie den

Modellierungsknoten aktualisieren. Für weitere Informationen siehe Thema Aktualisieren von Modellierungsknoten und Speichern in Kapitel 8 auf S. 140.

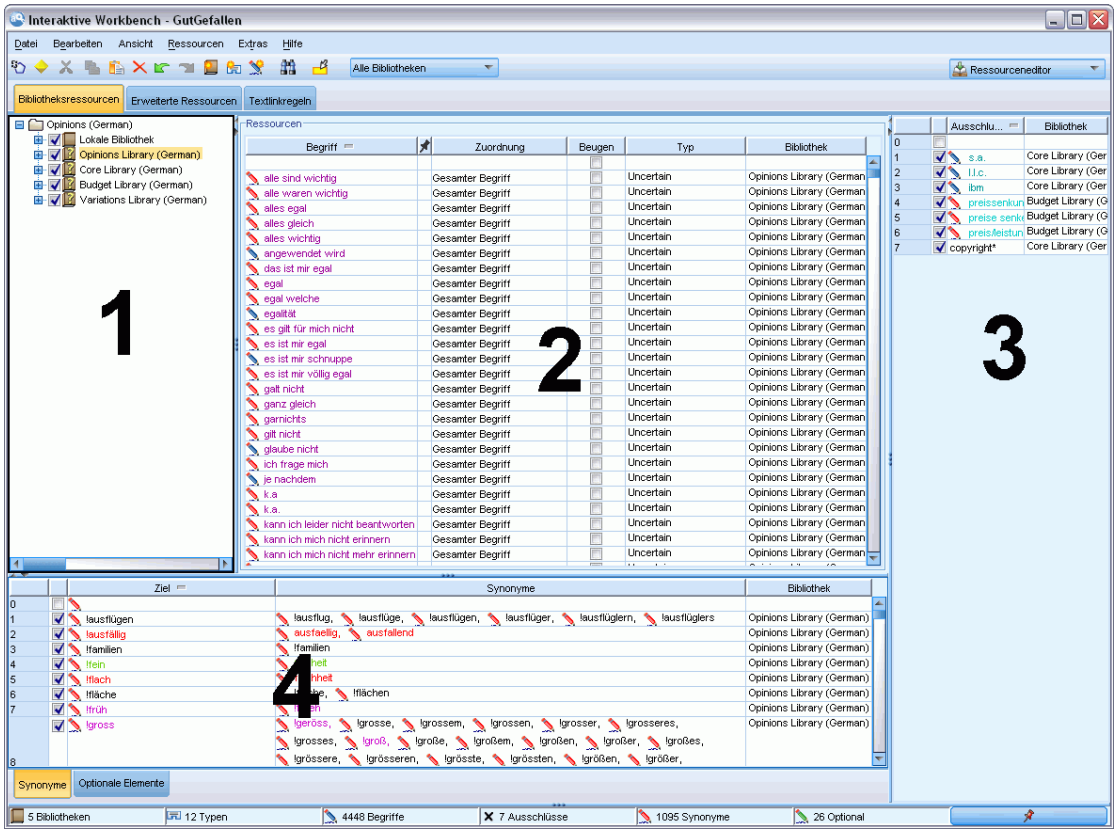
Falls Sie Ihre Änderungen unter der ursprünglichen Vorlage speichern möchten, deren Inhalt in den Modellierungsknoten kopiert wurden, um diese aktualisierte Vorlage in anderen Knoten laden zu können, können Sie aus den Ressourcen eine Vorlage erstellen. Für weitere Informationen siehe Thema Erstellen und Aktualisieren von Vorlagen in Kapitel 14 auf S. 276.

Die Editoroberfläche

Bei den im Template Editor oder im Resource Editor durchgeführten Operationen geht es um die Verwaltung und Feinabstimmung der linguistischen Ressourcen. Diese Ressourcen werden in Form von Vorlagen und Bibliotheken gespeichert. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

Registerkarte "Bibliotheksressourcen"

Abbildung 15-1
Text-Mining-Vorlageneditor



Die Oberfläche gliedert sich in die vier folgenden Bereiche:

1. Fensterbereich "Bibliotheksbaum". Dieser Plan befindet sich links oben und zeigt eine Baumstruktur mit den Bibliotheken. Sie können die Bibliotheken in diesem Baum aktivieren bzw. deaktivieren sowie die Ansichten in den anderen Fensterbereichen filtern, indem Sie eine Bibliothek im Baum auswählen. In diesem Baum können Sie mithilfe der Kontextmenüs eine Vielzahl von Operationen durchführen. Wenn Sie eine Bibliothek im Baum erweitern, wird das darin enthaltene Typen-Set angezeigt. Sie können diese Liste auch über das Menü Ansicht filtern, wenn Sie sich nur auf eine bestimmte Bibliothek konzentrieren möchten.

2. Bereich für Fachausdruckslisten von Typen-Wörterbüchern. Dieser Fensterbereich befindet sich rechts neben dem Bibliotheksbaum und zeigt die Fachausdruckslisten der Typ-Wörterbücher für die im Baum ausgewählten Bibliotheken an. Ein **Typ-Wörterbuch** ist eine Zusammenstellung von Fachausdrücken, die unter einer Bezeichnung (bzw. einem Typ oder einem Namen) gruppiert werden sollen. Beim Lesen der Textdaten vergleicht die Extrahierungsengine die im Text gefundenen Wörter mit den Fachausdrücken in den Typ-Wörterbüchern. Wenn ein extrahiertes Konzept als Fachausdruck in einem Typ-Wörterbuch vorkommt, wird der entsprechende Typname zugewiesen. Sie können das Typ-Wörterbuch als eigenes Wörterbuch mit Fachausdrücken betrachten, die etwas gemeinsam haben. Der Typ `<Location>` (Ort) in der Core Library beispielsweise enthält Konzepte wie `new orleans`, `great britain` und `new york`. Diese Fachausdrücke stehen jeweils für geografische Orte. Eine Bibliothek kann ein oder mehrere Typ-Wörterbücher enthalten. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

3. Fensterbereich "Ausschlusswörterbuch". Dieser Fensterbereich befindet sich auf der rechten Seite und zeigt die Sammlung der Fachausdrücke, die aus den endgültigen Extrahierungsergebnissen ausgeschlossen werden. Die Fachausdrücke in diesem Ausschlusswörterbuch werden nicht im Fensterbereich "Extrahierungsergebnisse" angezeigt. Ausgeschlossene Fachausdrücke werden in der Bibliothek Ihrer Wahl gespeichert. Jedoch zeigt der Bereich "Ausschlussbibliothek" alle ausgeschlossenen Fachausdrücke für alle Bibiotheken an, die im Bibliotheksbaum sichtbar sind.Für weitere Informationen siehe Thema Ausschlusswörterbücher in Kapitel 17 auf S. 329.

4. Fensterbereich "Substitutionswörterbuch". Dieser Fensterbereich befindet sich links unten und zeigt Synonyme und optionale Elemente in einer jeweils eigenen Registerkarte an. Synonyme und optionale Elemente helfen bei der Gruppierung ähnlicher Fachbegriffe unter einem Haupt- oder Zielkonzept in den endgültigen Extrahierungsergebnissen. Das Wörterbuch kann bekannte Synonyme, benutzerdefinierte Synonyme und Elemente sowie häufig vorkommende Rechtschreibfehler zusammen mit der korrekten Schreibung enthalten. Synonyme und optionale Elemente können in einer Bibliothek Ihrer Wahl gespeichert werden. Jedoch zeigt der Bereich "Substitutionswörterbuch" den Inhalt für alle Bibliotheken an, die im Bibliotheksbaum sichtbar sind. Während dieser Bereich alle Synonyme und optionale Elemente aus allen Bibliotheken anzeigt, werden die Substitutionen für alle Bibliotheken in der Baumstruktur zusammen in diesem Fensterbereich angezeigt. Eine Bibliothek kann nur ein einziges Substitutionswörterbuch enthalten. Für weitere Informationen siehe Thema Substitutions-/Synonymwörterbücher in Kapitel 17 auf S. 323. Bitte beachten Sie, dass die Registerkarte „Optionale Elemente" nicht für die japanische Textsprache gilt.

Anmerkungen:

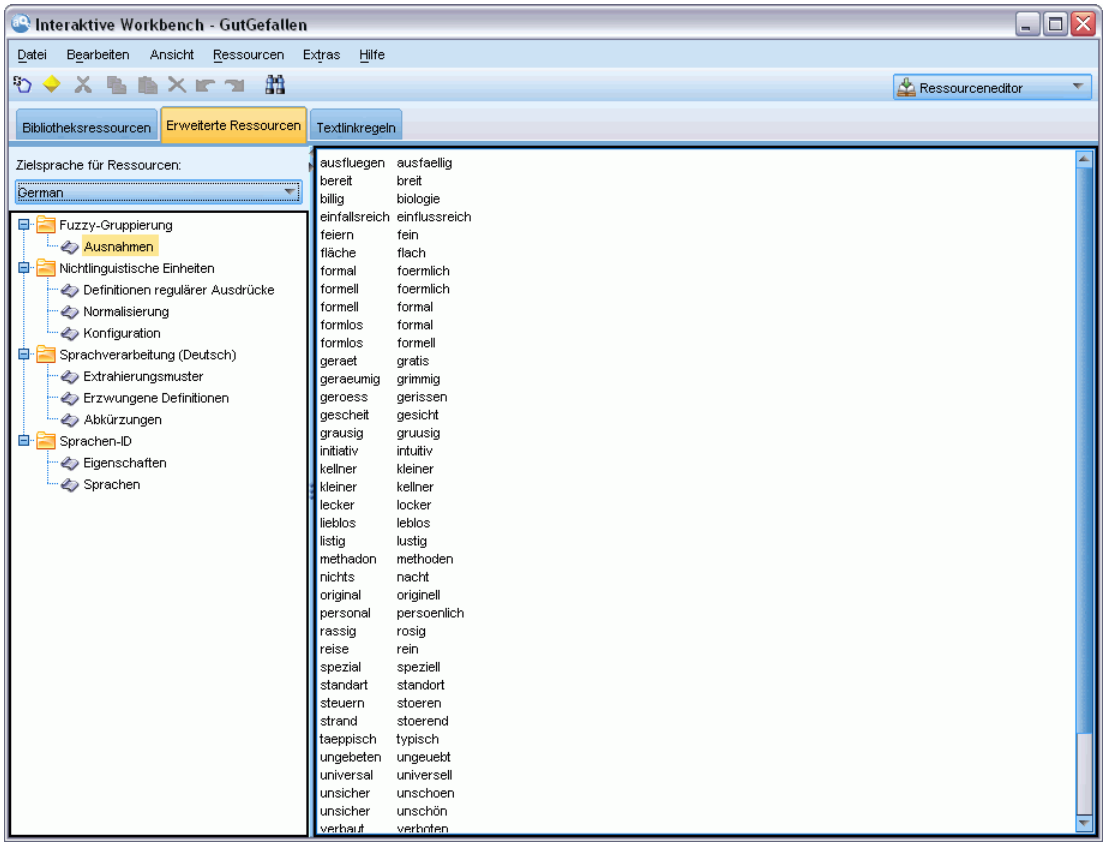
- Wenn Sie einen Filtervorgang durchführen möchten, sodass nur die Informationen angezeigt werden, die zu einer einzigen Bibliothek gehören, können Sie die Bibliotheksansicht mithilfe der Dropdown-Liste in der Symbolleiste ändern. Sie enthält auf der obersten Ebene den Eintrag Alle Bibliotheken sowie einen zusätzlichen Eintrag für jede einzelne Bibliothek. Für weitere Informationen siehe Thema Bibliotheken anzeigen in Kapitel 16 auf S. 301.
- Die Editoroberfläche für die japanische Sprache unterscheidet sich von der Oberfläche für andere Sprachen. Für weitere Informationen siehe Thema Bearbeitungsressourcen für japanischen Text in Anhang A auf S. 383.*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Registerkarte "Erweiterte Ressourcen"

Die erweiterten Ressourcen stehen in der zweiten Registerkarte der Editoransicht zur Verfügung. Sie können die erweiterten Ressourcen in dieser Registerkarte überprüfen und bearbeiten. Für weitere Informationen siehe Thema Informationen zu erweiterten Ressourcen in Kapitel 18 auf S. 332.

Wichtig: Diese Registerkarte ist nicht für Ressourcen verfügbar, die für japanischen Text eingestellt sind.

Abbildung 15-2
Text-Mining-Vorlageneditor - Registerkarte "Erweiterte Ressourcen"

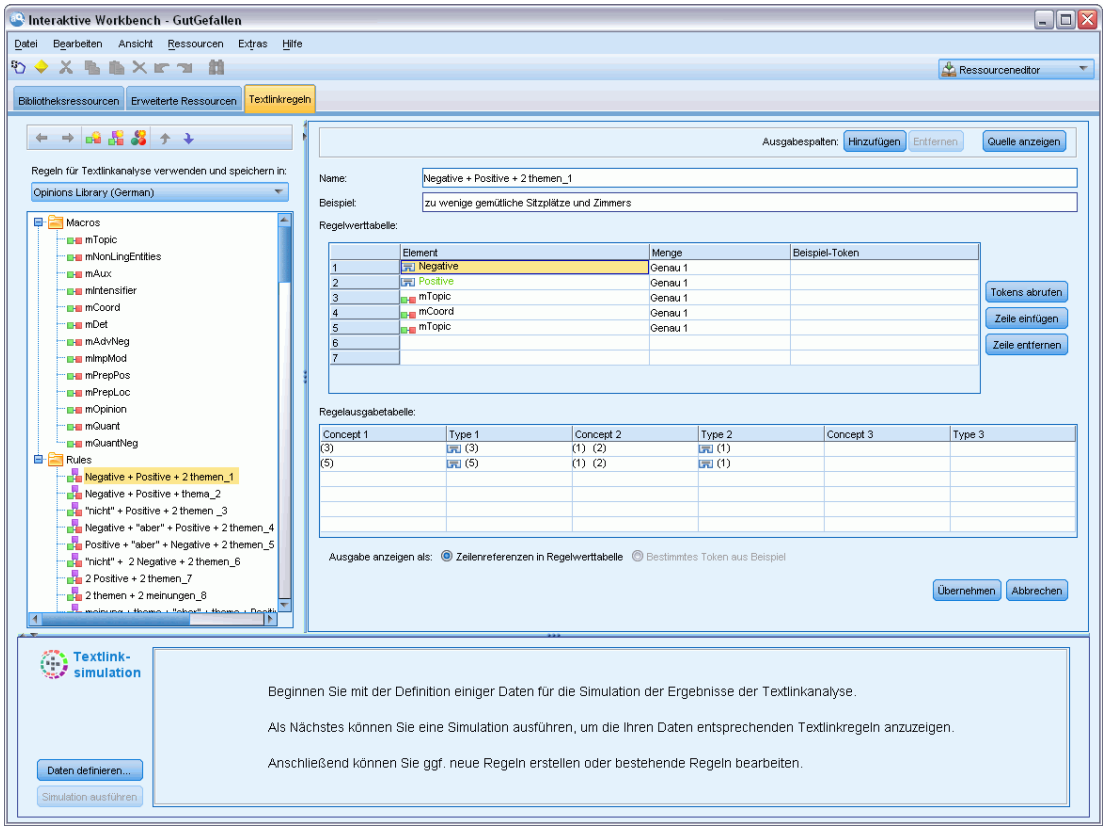


Registerkarte "Textlinkanalyseregeln"

Ab Version 14 sind die Textlinkanalyseregeln in ihrer eigenen Registerkarte der Editoransicht editierbar. Sie können im Regeleditor arbeiten, Ihre eigenen Regeln erstellen und sogar Simulationen ausführen, um zu sehen, wie sich Ihre Regeln auf die TLA-Ergebnisse auswirken. Für weitere Informationen siehe Thema Textlinkregeln in Kapitel 19 auf S. 346.

Wichtig: Diese Registerkarte ist nicht für Ressourcen verfügbar, die für japanischen Text eingestellt sind.

Abbildung 15-3
Text-Mining-Vorlageneditor - Registerkarte "Textlinkregeln"

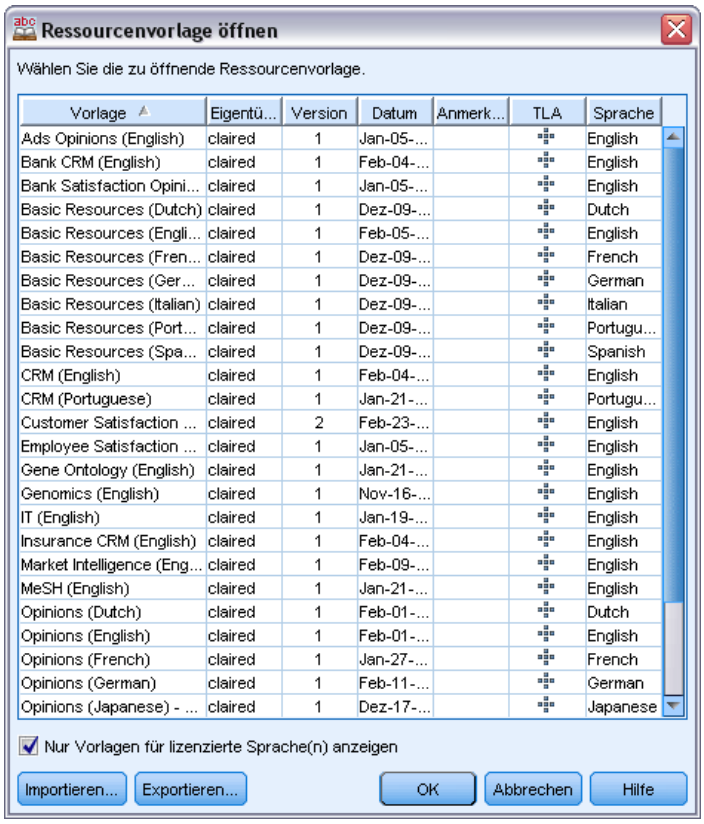


Öffnen von Vorlagen

Beim Starten des Template Editors werden Sie zum Öffnen einer Vorlage aufgefordert. Vorlagen können auch über das Menü "Datei" geöffnet werden. Wenn Sie eine Vorlage mit einigen Textlinkanalyseregeln (TLA) benötigen, stellen Sie sicher, dass Sie eine Vorlage auswählen, bei der ein Symbol in der Spalte "TLA" angezeigt wird. Die Sprache, für die eine Vorlage erstellt wurde, wird in der Spalte "Sprache" angezeigt.

Wenn Sie eine Vorlage importieren möchten, die nicht in der Tabelle angezeigt wird, oder wenn Sie eine Vorlage exportieren möchten, können Sie die Schaltflächen im Dialogfeld “Projekt öffnen” verwenden. Für weitere Informationen siehe Thema Importieren und Exportieren von Vorlagen auf S. 290.

Abbildung 15-4
Dialogfeld “Ressourcenvorlage öffnen”



So öffnen Sie eine Vorlage:

- Wählen Sie aus den Menüs im Template Editor die Optionsfolge Datei > Ressourcenvorlage öffnen aus. Das Dialogfeld “Ressourcenvorlage öffnen” wird geöffnet.
- Wählen Sie die gewünschte Vorlage aus der Tabelle aus.
- Klicken Sie auf OK, um diese Vorlage zu öffnen. Falls im Editor im Moment eine andere Vorlage geöffnet ist, wird diese Vorlage durch Klicken auf “OK” verworfen und die Vorlage angezeigt, die Sie hier ausgewählt haben. Wenn Sie Änderungen in Ihren Ressourcen vorgenommen haben und Ihre Bibliotheken für künftige Verwendungen speichern möchten, können Sie sie vor dem Öffnen einer weiteren veröffentlichen, aktualisieren und für die gemeinsame Nutzung freigeben. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen in Kapitel 16 auf S. 306.

Speichern von Vorlagen

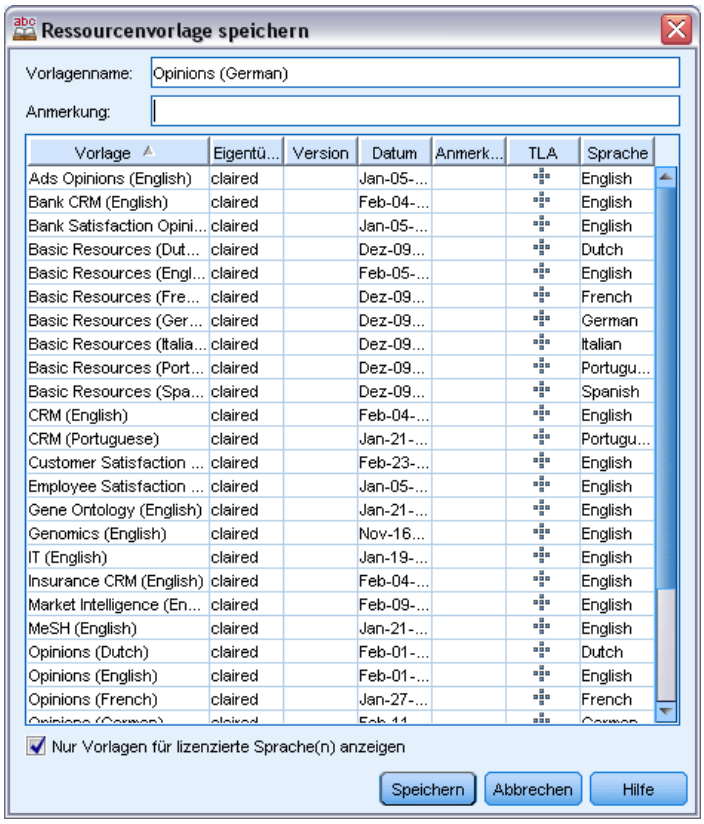
Im Template Editor können Sie die Änderungen speichern, die Sie an einer Vorlage vorgenommen haben. Sie können die Vorlage unter einem bestehenden Vorlagennamen oder unter einem neuen Namen speichern.

Wenn Sie Änderungen an einer Vorlage vornehmen, die Sie zuvor bereits in einen Knoten geladen haben, müssen Sie den Inhalt der Vorlage in den Knoten laden, um die aktuellen Änderungen zu erhalten. Für weitere Informationen siehe Thema Kopieren von Ressourcen aus TAPs und Vorlagen in Kapitel 3 auf S. 42.

Alternativ müssen Sie bei Verwendung der Option Gespeicherte interaktive Arbeit verwenden in der Registerkarte “Modell” des Text-Mining-Knotens (Sie verwenden also Ressourcen aus einer vorangehenden interaktiven Workbench-Sitzung) innerhalb der interaktiven Workbench-Sitzung zu den Ressourcen dieser Vorlage wechseln. Für weitere Informationen siehe Thema Wechseln von Ressourcenvorlagen in Kapitel 14 auf S. 277.

Anmerkung: Sie können Ihre Bibliotheken auch veröffentlichen und mit anderen Benutzern gemeinsam nutzen. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen in Kapitel 16 auf S. 306.

Abbildung 15-5
Dialogfeld “Ressourcenvorlage speichern”



So speichern Sie eine Vorlage:

- Wählen Sie aus den Menüs im Template Editor die Optionsfolge Datei > Ressourcenvorlage speichern aus. Das Dialogfeld “Ressourcenvorlage speichern” wird geöffnet.
- Geben Sie einen neuen Namen in das Feld “Vorlagenname” ein, um diese Vorlage als neue Vorlage zu speichern. Wählen Sie eine Vorlage in der Tabelle aus, um eine vorhandene Vorlage mit den derzeit geladenen Ressourcen zu überschreiben.
- Geben Sie bei Bedarf eine Beschreibung ein, um in der Tabelle einen Kommentar oder eine Anmerkung anzuzeigen.
- Klicken Sie auf Speichern, um die Vorlage zu speichern.

Wichtig: Da Ressourcen aus Vorlagen oder TAPs in den Knoten geladen/kopiert werden, müssen Sie die Ressourcen aktualisieren, indem Sie sie erneut laden, nachdem Sie eine Vorlage geändert haben, damit Sie in einem bestehenden Stream von diesen Änderungen profitieren können. Für weitere Informationen siehe Thema Aktualisieren von Knotenressourcen nach dem Laden auf S. 288.

Aktualisieren von Knotenressourcen nach dem Laden

Standardmäßig wird beim Hinzufügen eines Knotens zu einem Stream ein Set von Ressourcen aus einer Standardvorlage geladen und in Ihren Knoten eingebettet. Und wenn Sie beim Laden Vorlagen ändern oder ein TAP verwenden, überschreibt eine Kopie davon die vorhandenen Ressourcen. Da Vorlagen und TAPs nicht direkt mit dem Knoten verknüpft sind, werden Änderungen an einer Vorlage oder einem TAP nicht automatisch in einem bestehenden Knoten verfügbar. Um von diesen Änderungen zu profitieren, müssen Sie die Ressourcen in diesem Knoten aktualisieren. Es sind zwei Möglichkeiten zum Aktualisieren der Ressourcen verfügbar.

Methode 1: Erneutes Laden der Ressourcen auf der Registerkarte “Modell”

Wenn Sie die Ressourcen im Knoten mit einer neuen oder aktualisierten Vorlage oder einem TAP aktualisieren möchten, können Sie diese Vorlage auf der Registerkarte “Modell” des betreffenden Knotens erneut laden. Durch das erneute Laden wird die Kopie der Ressourcen im Knoten mit einer aktuelleren Kopie ersetzt. Um Ihnen die Arbeit zu erleichtern, werden die Uhrzeit und das Datum der Aktualisierung auf der Registerkarte “Modell” zusammen mit dem Namen der Ursprungsvorlage angezeigt. Für weitere Informationen siehe Thema Kopieren von Ressourcen aus TAPs und Vorlagen in Kapitel 3 auf S. 42.

Wenn Sie jedoch mit interaktiven Sitzungsdaten in einem Text-Mining-Modellierungsknoten arbeiten und auf der Registerkarte “Modell” die Option Arbeit der Sitzung verwenden ausgewählt haben, werden die gespeicherte Arbeit und die Ressourcen der Sitzung verwendet, während die Schaltfläche Laden deaktiviert ist. Die Schaltfläche ist deaktiviert, da Sie zu einem bestimmten Zeitpunkt während der interaktiven Workbench-Sitzung die Option Modellierungsknoten aktualisieren gewählt und die Kategorien, Ressourcen und andere Arbeit der Sitzung beibehalten haben. Wenn Sie in diesem Fall diese Ressourcen ändern oder aktualisieren möchten, können Sie versuchen, die nächste Methode zu verwenden, bei der die Ressourcen im Resource Editor gewechselt werden.

Methode 2: Wechseln von Ressourcen im Resource Editor

Wenn Sie während einer interaktiven Sitzung andere Ressourcen verwenden möchten, können Sie diese Ressourcen mit dem Dialogfeld “Ressourcen wechseln” austauschen. Dies ist insbesondere dann nützlich, wenn Sie vorhandene Arbeit an Kategorien wiederverwenden, jedoch die Ressourcen ersetzen möchten. In diesem Fall können Sie die Option Arbeit der Sitzung verwenden auf der Registerkarte “Modell” eines Text-Mining-Modellierungsknotens auswählen. Dadurch steht die Möglichkeit zum erneuten Laden einer Vorlage über das Knotendialogfeld nicht mehr zur Verfügung und die Einstellungen und Änderungen, die während Ihrer Sitzung vorgenommen wurden, bleiben stattdessen erhalten. Danach können Sie die interaktive Workbench-Sitzung durch Ausführen des Streams und Wechseln der Ressourcen im Resource Editor starten. Für weitere Informationen siehe Thema Wechseln von Ressourcenvorlagen in Kapitel 14 auf S. 277.

Um die Arbeit der Sitzung inklusive Ressourcen für nachfolgende Sitzungen beizubehalten, müssen Sie den Modellierungsknoten innerhalb der interaktiven Workbench-Sitzung aktualisieren, sodass die Ressourcen (und weitere Daten) unter dem Knoten gespeichert werden. Für weitere Informationen siehe Thema Aktualisieren von Modellierungsknoten und Speichern in Kapitel 8 auf S. 140.

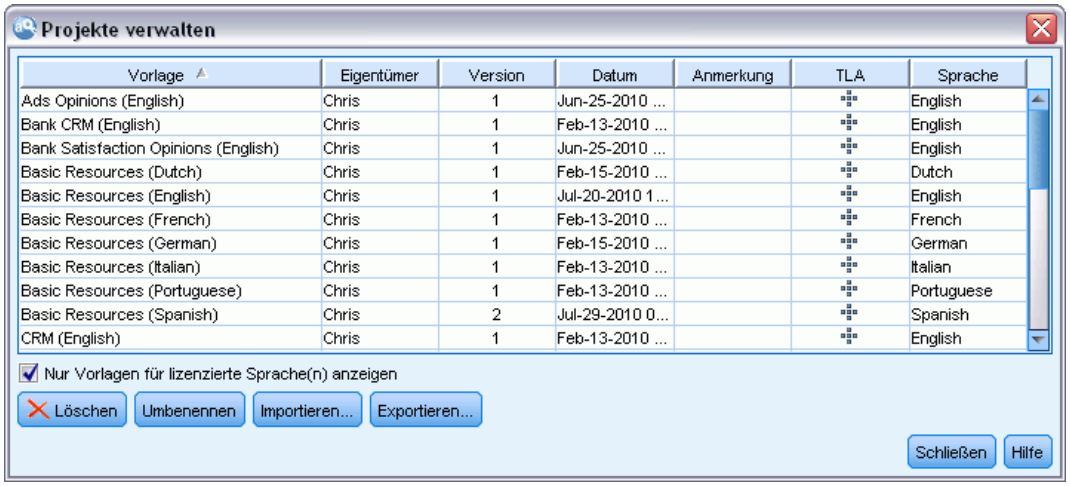
Anmerkung: Wenn Sie während einer interaktiven Sitzung zum Inhalt einer anderen Vorlage wechseln, bleibt der Name der im Knoten aufgeführten Vorlage derselbe wie bei der zuletzt geladenen und kopierten Vorlage. Aktualisieren Sie den Modellierungsknoten vor dem Beenden der Sitzung, um diese Ressourcen bzw. andere Arbeit der Sitzung nutzen zu können.

Vorlagen verwalten

Gelegentlich sollten Sie einige grundlegende Verwaltungsmaßnahmen an Ihren Vorlagen durchführen, z. B. das Umbenennen der Vorlagen, Importieren und Exportieren von Vorlagen oder Löschen veralteter Vorlagen. Diese Aufgaben werden im Dialogfeld “Vorlagen verwalten” durchgeführt. Durch Importieren und Exportieren von Vorlagen können Sie Vorlagen mit anderen Benutzern gemeinsam nutzen. Für weitere Informationen siehe Thema Importieren und Exportieren von Vorlagen auf S. 290.

Anmerkung: Mit diesem Produkt installierte (oder mitgelieferte) Vorlagen können nicht umbenannt oder gelöscht werden. Wenn Sie eine Vorlage umbenennen möchten, können Sie die installierte Vorlage öffnen und eine neue mit einem Namen Ihrer Wahl erstellen. Ihre benutzerdefinierten Vorlagen können Sie löschen. Wenn Sie jedoch versuchen, eine mitgelieferte Vorlage zu lsöchen, wird sie auf die installierte Version zurückgesetzt.

Abbildung 15-6
Dialogfeld "Vorlagen verwalten"



So benennen Sie eine Vorlage um:

- Wählen Sie in den Menüs die Optionsfolge Ressourcen > Ressourcenvorlagen verwalten aus. Das Dialogfeld "Vorlagen verwalten" wird geöffnet.
- Wählen Sie die Vorlage aus, die Sie umbenennen möchten, und klicken Sie auf Umbenennen. Das Namensfeld wird in der Tabelle editierbar.
- Geben Sie einen neuen Namen ein und drücken Sie die Eingabetaste. Es wird ein Bestätigungs-Dialogfeld geöffnet.
- Wenn Sie die Namensänderung anwenden möchten, klicken Sie auf Ja. Wenn Sie sie verwerfen möchten, klicken Sie auf Nein.

So löschen Sie eine Vorlage:

- Wählen Sie in den Menüs die Optionsfolge Ressourcen > Ressourcenvorlagen verwalten aus. Das Dialogfeld "Vorlagen verwalten" wird geöffnet.
- Wählen Sie im Dialogfeld "Vorlagen verwalten" die zu löschende Vorlage aus.
- Klicken Sie auf Löschen. Es wird ein Bestätigungs-Dialogfeld geöffnet.
- Klicken Sie auf Ja, um zu löschen, oder auf Nein, um den Anforderungsvorgang abzubrechen. Wenn Sie auf Ja klicken, wird die Vorlage gelöscht.

Importieren und Exportieren von Vorlagen

Sie können Vorlagen mit anderen Benutzern oder Computern gemeinsam nutzen, indem Sie sie importieren und exportieren. Vorlagen werden in einer internen Datenbank gespeichert, können jedoch als **.lrt*-Dateien auf die Festplatte exportiert werden.

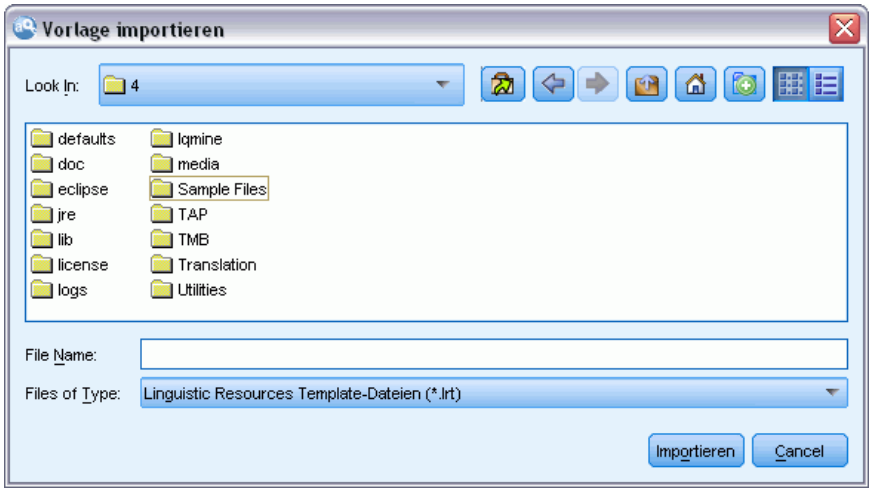
Da verschiedene Situationen gegeben sein können, in denen Sie Vorlagen importieren bzw. exportieren möchten, stehen diese Funktionen auf verschiedenen Dialogfeldern zur Verfügung.

- Dialogfeld “Projekt öffnen” im Template Editor
- Dialogfeld “Ressourcen laden” im Textlinkanalyse- und Text-Mining-Modellierungsknoten
- Dialogfeld “Vorlagen verwalten” im Template Editor und Resource Editor

So importieren Sie eine Vorlage:

▶ Klicken Sie im Dialogfeld auf Importieren. Das Dialogfeld “Vorlage importieren” wird geöffnet.

Abbildung 15-7
Dialogfeld “Vorlage importieren”

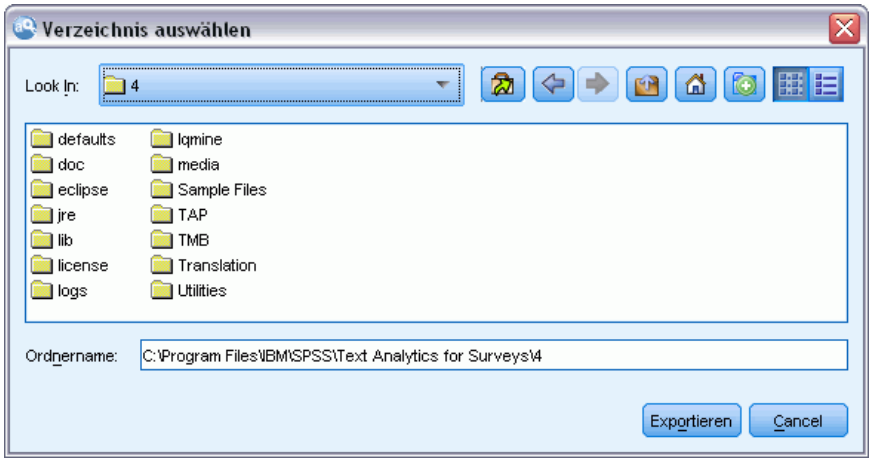


▶ Wählen Sie die zu importierende Ressourcenvorlagendatei (*.*lrt*) aus und klicken Sie auf Importieren. Sie können die zu importierende Vorlage unter einem neuen Namen speichern oder die vorhandene Vorlage überschreiben. Das Dialogfeld wird geschlossen und die Vorlage wird nun in der Tabelle angezeigt.

So exportieren Sie eine Vorlage:

▶ Wählen Sie im Dialogfeld die zu exportierende Vorlage aus und klicken Sie auf Exportieren. Das Dialogfeld “Verzeichnis auswählen” wird geöffnet.

Abbildung 15-8
Dialogfeld "Verzeichnis auswählen"



- Wählen Sie das Verzeichnis aus, in das exportiert werden soll, und klicken Sie auf Exportieren. Das Dialogfeld wird geschlossen und die Vorlage wird mit der Dateierweiterung (*.*lrt*) exportiert.

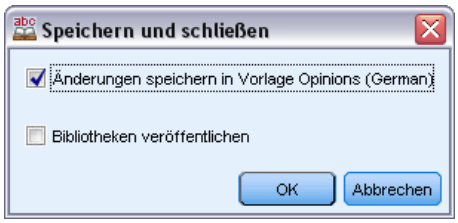
Beenden des Template Editors

Wenn Sie die Arbeit im Template Editor beendet haben, können Sie Ihre Arbeit speichern und den Editor beenden.

So beenden Sie den Template Editor:

- Wählen Sie in den Menüs die Optionsfolge Datei > Schließen. Das Dialogfeld "Speichern und schließen" wird geöffnet.

Abbildung 15-9
Dialogfeld "Speichern und schließen"



- Wählen Sie die Option Änderungen an Projekt speichern aus, um die geöffnete Vorlage vor dem Schließen des Editors zu speichern.

- Wählen Sie die Option Bibliotheken veröffentlichen aus, wenn Sie Bibliotheken in der geöffneten Vorlage veröffentlichen möchten, bevor Sie den Editor schließen. Bei Auswahl dieser Option werden Sie dazu aufgefordert, die zu veröffentlichenden Bibliotheken auszuwählen. Für weitere Informationen siehe Thema Bibliotheken veröffentlichen in Kapitel 16 auf S. 307.

Ressourcen sichern

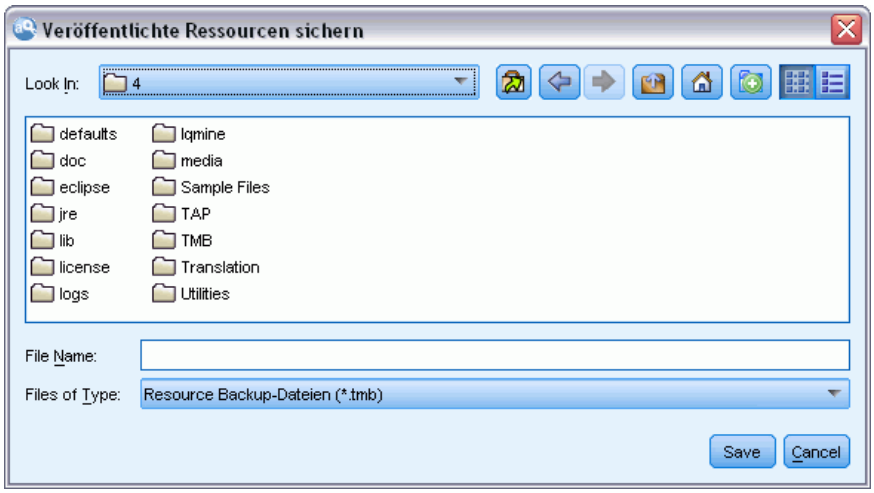
Sicherheitshalber sollten Sie Ihre Ressourcen gelegentlich sichern.

Wichtig: Bei der Wiederherstellung wird der gesamte Inhalt Ihrer Ressourcen gelöscht und nur der Inhalt der Sicherungsdatei ist in dem Produkt verfügbar. Dies gilt auch für geöffnete Arbeiten.

So sichern Sie die Ressourcen:

- Wählen Sie in den Menüs die Optionsfolge Ressourcen > Sicherungstools > Ressourcen sichern. Das Dialogfeld "Sicherung" wird geöffnet.

Abbildung 15-10
Dialogfeld "Ressourcen sichern"

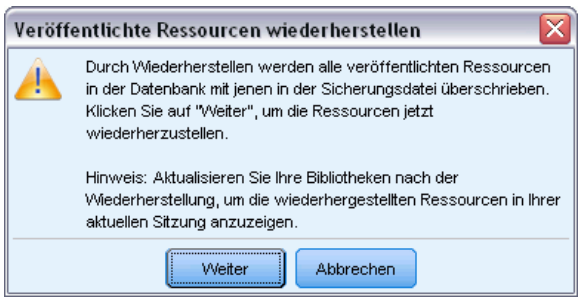


- Geben Sie einen Namen für Ihre Sicherungsdatei ein und klicken Sie auf Speichern. Das Dialogfeld wird geschlossen und die Sicherungsdatei erstellt.

So stellen Sie die Ressourcen wieder her:

- Wählen Sie in den Menüs die Optionsfolge Ressourcen > Sicherungstools > Ressourcen wiederherstellen. Ein Warnhinweis informiert Sie darüber, dass bei einer Wiederherstellung der Inhalt Ihrer Datenbank überschrieben wird.

Abbildung 15-11
Warnhinweis bei Überschreibung



▶ Klicken Sie auf Ja, um fortzufahren. Das Dialogfeld wird geöffnet.

Abbildung 15-12
Dialogfeld "Ressourcen wiederherstellen"

▶ Wählen Sie die Sicherungsdatei aus, die Sie wiederherstellen möchten, und klicken Sie auf Öffnen. Das Dialogfeld wird geschlossen und die Ressourcen werden in der Anwendung wiederhergestellt.

Wichtig: Bei der Wiederherstellung wird der gesamte Inhalt Ihrer Ressourcen gelöscht, und nur der Inhalt der Sicherungsdatei ist in dem Produkt verfügbar. Dies gilt auch für geöffnete Arbeiten.

Ressourcendateien importieren

Wenn Sie Änderungen direkt in Ressourcendateien außerhalb dieses Produkts vorgenommen haben, können Sie sie in eine ausgewählte Bibliothek importieren, indem Sie diese Bibliothek auswählen und mit dem Import fortfahren. Wenn Sie ein Verzeichnis importieren, können Sie auch alle unterstützten Dateien in eine bestimmte geöffnete Bibliothek importieren. Es können nur **.txt*-Dateien importiert werden.

Wichtig: Für Dateien in japanischer Sprache müssen die .txt-Dateien, die Sie importieren möchten, in UTF8 kodiert sein. Außerdem können Sie für Japanisch keine Ausschlusslisten importieren. *Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Jede importierte Datei darf nur einen Eintrag pro Zeile enthalten und wenn die Inhalte folgendermaßen struktiert sind:

- Eine Liste mit Wörtern oder Phrasen (eine pro Zeile). Die Datei wird als Fachausdrucksliste für ein Typ-Wörterbuch importiert. Dabei übernimmt das Typ-Wörterbuch den Namen der Datei ohne die Erweiterung. Bei japanischem Text muss der Dateiname einem bekannten japanischen Typ entsprechen, damit er importiert wird. Wenn der Dateiname einem Stimmungstypnamen anstelle eines japanischen Basistyps entspricht, werden die

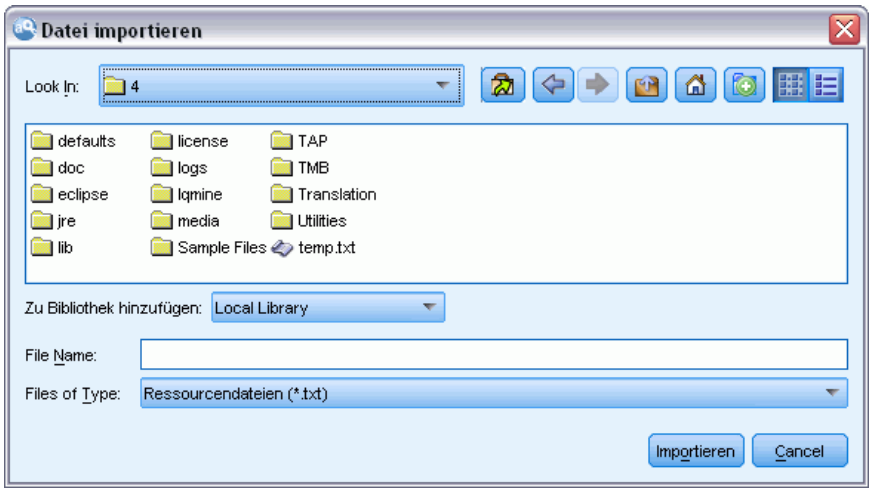
Fachausdrücke in der Datei dem Stimmungstyp sowie dem Standard-Basistyp 名詞 zugeordnet..

- Als Liste von Einträgen wie *Fachausdruck1*<TAB>*Fachausdruck2*. In diesem Fall wird die Datei als Liste von Synonymen importiert. Dabei ist *Fachausdruck1* das Set des zugrunde liegenden Begriffs und *Fachausdruck2* ist der Zielausdruck. Bei japanischem Text wird dem Zielausdruck der Standardwert 名詞 zugewiesen.

So importieren Sie eine einzelne Ressourcendatei:

▶ Wählen Sie in den Menüs die Optionsfolge Ressourcen > Dateien importieren > Einzeldatei importieren. Das Dialogfeld “Datei importieren” wird geöffnet.

Abbildung 15-13
Dialogfeld “Datei importieren”

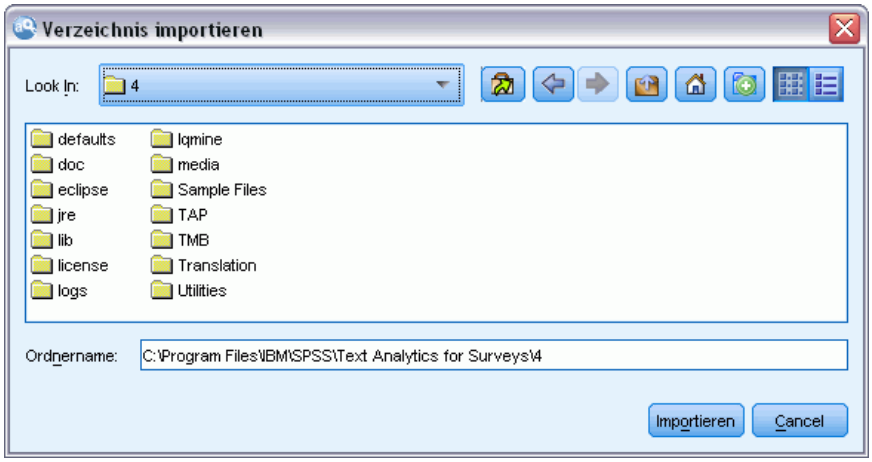


▶ Wählen Sie die zu importierende Datei aus und klicken Sie auf Importieren. Der Inhalt der Datei wird in ein internes Format umgewandelt und zu Ihrer Bibliothek hinzugefügt.

So importieren Sie alle Dateien eines Verzeichnisses:

▶ Wählen Sie in den Menüs die Optionsfolge Ressourcen > Dateien importieren > Ganzes Verzeichnis importieren. Das Dialogfeld “Verzeichnis importieren” wird geöffnet.

Abbildung 15-14
Dialogfeld "Verzeichnis importieren"



- Wählen Sie die Bibliothek, in die alle Ressourcendateien importiert werden sollen, aus der Liste Importieren aus. Wenn Sie die Option Standard auswählen, wird eine neue Bibliothek mit dem Namen des Verzeichnisses erstellt.
- Wählen Sie das Verzeichnis aus, aus dem die Dateien importiert werden sollen. Es werden keine Unterverzeichnisse gelesen.
- Klicken Sie auf Importieren. Das Dialogfeld wird geschlossen und der Inhalt der importierten Ressourcendateien wird im Editor als Verzeichnisse und erweiterte Ressourcendateien aufgeführt.

Kapitel
16

Mit Bibliotheken arbeiten

Die Ressourcen, die von der Extrahierungsengine zum Extrahieren und Gruppieren von Fachausdrücken verwendet werden, enthalten immer eine oder mehrere Bibliotheken. Die Bibliotheken werden im Bibliotheksbaum im oberen linken Bereich des Template Editors und Resource Editors aufgeführt. Sie bestehen aus zwei Arten von Wörterbüchern: Typ, Substitution und Ausschluss. Für weitere Informationen siehe Thema Informationen zu Bibliothekwörterbüchern in Kapitel 17 auf S. 312.

Die ausgewählte Ressourcenvorlage oder die Ressourcen aus dem TAP enthalten verschiedene Bibliotheken, die Ihnen die unverzügliche Extrahierung von Konzepten aus Ihren Textdaten ermöglichen. Sie können jedoch auch eigene Bibliotheken erstellen und zur Wiederverwendung auch veröffentlichen. Für weitere Informationen siehe Thema Bibliotheken veröffentlichen auf S. 307.

Angenommen, Sie arbeiten z. B. häufig mit Textdaten zur Automobilindustrie. Nach der Analyse Ihrer Daten möchten Sie benutzerdefinierte Ressourcen zur Verarbeitung von branchenspezifischem Wortschatz oder von Fachsprache erstellen. Mit dem Template Editor können Sie eine neue Vorlage und darin eine Bibliothek zum Extrahieren und Gruppieren von Begriffen aus der Automobilindustrie erstellen. Da Sie die Informationen in dieser Bibliothek erneut benötigen werden, veröffentlichen Sie Ihre Bibliothek in einem zentralen Archiv, das Sie in dem Dialogfeld **Bibliotheken verwalten** aufrufen können, sodass sie unabhängig in verschiedenen Streamsitzungen wiederverwendet werden kann.

Angenommen, Sie möchten außerdem Fachausdrücke aus verschiedenen Unterbranchen, z. B. zu elektronischen Geräten, Motoren, Kühlsystemen, oder sogar Fachausdrücke eines bestimmten Herstellers oder Markts gruppieren. Sie können für jede Gruppe eine Bibliothek erstellen und sie dann so veröffentlichen, dass sie mit mehreren Textdatenmengen verwendet werden kann. So können Sie Bibliotheken hinzufügen, die dem Kontext Ihrer Textdaten am besten entsprechen.

Anmerkung: Zusätzliche Ressourcen können in der Registerkarte “Erweiterte Ressourcen” konfiguriert und verwaltet werden. Einige beziehen sich auf alle Bibliotheken und verwalten nicht linguistische Elemente, Ausnahmen bei der Fuzzy-Gruppierung usw. Zudem können Sie die bibliotheksspezifischen Musterregeln für Textlinkanalysen in der Registerkarte “Textlinkregeln” bearbeiten. Für weitere Informationen siehe Thema Informationen zu erweiterten Ressourcen in Kapitel 18 auf S. 332.

Mitgelieferte Bibliotheken

Es werden standardmäßig mehrere Bibliotheken mit IBM® SPSS® Modeler Text Analytics installiert. Mit diesen vorformatierten Bibliotheken können Sie auf Tausende vordefinierter Fachausdrücke und Synonyme sowie auf viele verschiedene Typen zugreifen. Diese mitgelieferten Bibliotheken sind auf verschiedene Domänen abgestimmt und in verschiedenen Sprachen verfügbar.

Anmerkung: Spezifische Informationen zu japanischen Textressourcen und Ausnahmen finden Sie unter Ausnahmen bei japanischem Text auf S. 377*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Es gibt eine Vielzahl von Bibliotheken, am häufigsten werden jedoch die folgenden verwendet:

- **Lokale Bibliothek.** Zum Speichern von benutzerdefinierten Wörterbüchern. Dies ist eine leere Bibliothek, die standardmäßig allen Ressourcen hinzugefügt wird. Sie enthält außerdem ein leeres Typ-Wörterbuch. Diese ist besonders bei Änderungen oder Feineinstellungen nützlich, die direkt über die Kategorien- und Konzepte-Ansicht, die Cluster-Ansicht und die Textlinkanalyse-Ansicht vorgenommen werden (z. B. Hinzufügen eines Worts zu einem Typ). In diesem Fall werden die Änderungen und Feineinstellungen automatisch in der ersten Bibliothek im Bibliotheksbaum im Resource Editor gespeichert; standardmäßig ist dies die *Lokale Bibliothek*. Diese Bibliothek können Sie nicht veröffentlichen, weil sie für die Sitzungs-Daten bestimmt ist. Um ihren Inhalt zu veröffentlichen, müssen Sie die Bibliothek zunächst umbenennen.
- **Core Library.** Wird in den meisten Fällen verwendet, da sie die fünf integrierten Grundtypen, d. h. Personen, Orte, Organisationen, Produkte und Unbekannt, umfasst. Obwohl möglicherweise nur wenige Fachausdrücke in einem ihrer Typ-Wörterbücher aufgeführt werden, sind die in der Core Library dargestellten Typen tatsächlich Ergänzungen zu den robusten Typen in den internen, kompilierten Ressourcen, die im Lieferumfang Ihres Text-Mining-Produkts enthalten sind. Diese internen, kompilierten Ressourcen enthalten Tausende von Fachausdrücken für jeden Typ. Daher kann ein Fachausdruck dennoch extrahiert und mit einem Kerntyp versehen werden, auch wenn Sie ihn nicht in der Fachausdruckliste des Typ-Wörterbuchs sehen. Dies erklärt, wie Namen, z. B. *George*, extrahiert werden und den Typ `<Person>` erhalten können, wenn nur *John*> in dem Typ-Wörterbuch `<Person>` in der Core Library aufgeführt wird. Wenn Sie die Core Library nicht einschließen, werden diese Typen ggf. trotzdem in den Extrahierungsergebnissen aufgeführt, da die kompilierten Ressourcen, die diese Typen enthalten, weiterhin von der Extrahierungsengine verwendet werden.
- **Opinions Library.** Am häufigsten zum Extrahieren von Meinungen und Stimmungen aus Textdaten verwendet. Diese Vorlage beinhaltet Tausende von Wörtern für Einstellungen, Vermerke und Präferenzen, die—wenn sie zusammen mit anderen Fachausdrücken verwendet werden—eine Meinung über ein Thema ausdrücken. Diese Bibliothek enthält eine Reihe von integrierten Typen, Synonymen und Ausschlüssen. Sie enthält außerdem eine große Menge an Musterregeln für die Textlinkanalyse. Damit die Textlinkanalyseregeln in dieser Bibliothek sowie die erzeugten Musterergebnisse genutzt werden können, muss diese Bibliothek in der Registerkarte “Textlinkregeln” angegeben werden. Für weitere Informationen siehe Thema Textlinkregeln in Kapitel 19 auf S. 346.
- **Budget Library.** Wird zum Extrahieren von Begriffen zum Thema “Kosten” verwendet. Diese Bibliothek enthält zahlreiche Wörter und Ausdrücke, die Adjektive, Vermerke und Entscheidungen zu den Themen “Preis” oder “Qualität” darstellen.
- **Variations Library.** Wird verwendet, um Fälle einzuschließen, in denen bestimmte Sprachvarianten zur richtigen Gruppierung Synonymdefinitionen erfordern. Diese Bibliothek enthält nur Synonymdefinitionen.

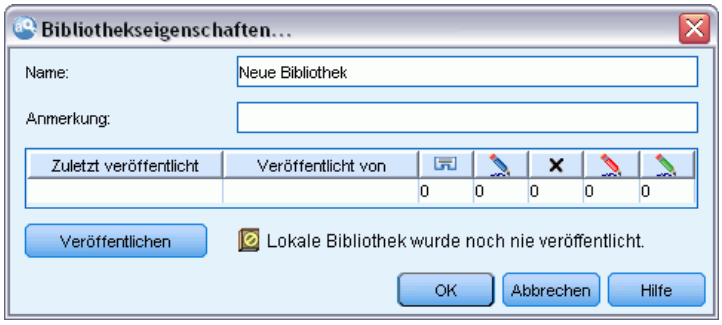
Obwohl einige der mitgelieferten Bibliotheken außerhalb der Vorlagen dem Inhalt einiger Vorlagen ähneln, sind die Vorlagen speziell auf bestimmte Anwendungen abgestimmt und enthalten zusätzliche erweiterte Ressourcen. Es wird empfohlen, dass Sie es mit einer speziellen Vorlage für die Art der verwendeten Textdaten versuchen und Ihre Änderungen an diesen Ressourcen ausführen anstatt einer allgemeineren Vorlage nur individuelle Bibliotheken hinzuzufügen.

Es sind außerdem kompilierte Ressourcen im Lieferumfang von SPSS Modeler Text Analytics enthalten. Sie werden bei jedem Extrahierungsprozess verwendet und enthalten eine große Anzahl ergänzender Definitionen zu den integrierten Wörterbüchern in den Standardbibliotheken. Da diese Ressourcen kompiliert sind, können sie nicht angezeigt oder bearbeitet werden. Sie können jedoch die Aufnahme eines Fachausdrucks, der einen Typ durch diese kompilierten Ressourcen erhalten hat, in ein anderes Wörterbuch erzwingen. Für weitere Informationen siehe Thema Erzwingen von Fachausdrücken in Kapitel 17 auf S. 320.

Bibliotheken erstellen

Sie können beliebig viele Bibliotheken erstellen. Nach dem Erstellen einer neuen Bibliothek können Sie Typ-Wörterbücher in dieser Bibliothek erstellen und Fachausdrücke, Synonyme und Ausschlüsse eingeben.

Abbildung 16-1
Dialogfeld "Bibliotheken-Eigenschaften"



So erstellen Sie eine Bibliothek:

- Wählen Sie in den Menüs die Optionsfolge Ressourcen > Neue Bibliothek. Das Dialogfeld Bibliothekseigenschaften wird geöffnet.
- Geben Sie im Textfeld "Name" einen Namen für die Bibliothek ein.
- Bei Bedarf geben Sie einen Kommentar im Textfeld "Anmerkungen" ein.
- Klicken Sie auf Veröffentlichen, wenn Sie diese Bibliothek nun veröffentlichen möchten, bevor Sie Einträge in die Bibliothek vornehmen. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen auf S. 306. Sie können sie auch später zu einem beliebigen Zeitpunkt veröffentlichen.
- Klicken Sie auf OK, um die Bibliothek zu erstellen. Das Dialogfeld wird geschlossen und die Bibliothek wird in der Baumansicht aufgeführt. Wenn Sie alle Bibliotheken in dem Baum anzeigen, wird ein leeres Typ-Wörterbuch angezeigt, das automatisch in die Bibliothek

aufgenommen worden ist. In dieses können Sie sofort Fachausdrücke aufnehmen. Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken in Kapitel 17 auf S. 316.

Öffentliche Bibliotheken hinzufügen

Um eine Bibliothek aus anderen Sitzungsdaten wiederzuverwenden, fügen Sie sie Ihren aktuellen Ressourcen hinzu, vorausgesetzt, es handelt sich um eine Public Library. Eine **Public Library** ist eine Bibliothek, die veröffentlicht worden ist. Für weitere Informationen siehe Thema Bibliotheken veröffentlichen auf S. 307.

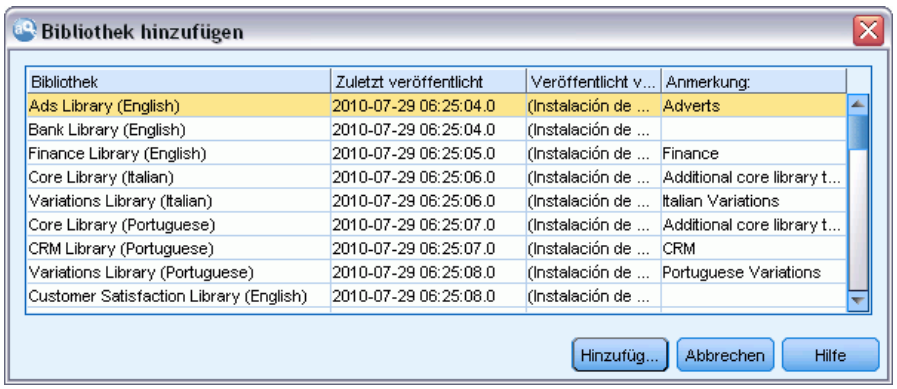
Wichtig: Es ist nicht möglich, nicht japanischen Ressourcen eine japanische Bibliothek hinzuzufügen oder umgekehrt. *Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Wenn Sie eine Public Library hinzufügen, wird eine **lokale** Kopie in Ihre Sitzungsdaten eingebettet. Sie können Änderungen an dieser Bibliothek vornehmen, müssen jedoch die öffentliche Version der Bibliothek erneut veröffentlichen, um die Änderungen mit anderen gemeinsam nutzen zu können.

Beim Hinzufügen einer öffentlichen Bibliothek wird möglicherweise das Dialogfeld “Konflikte lösen” angezeigt, wenn Konflikte zwischen den Fachausdrücken und Typen in einer Bibliothek und den anderen lokalen Bibliotheken entdeckt werden. Lösen Sie diese Konflikte auf oder akzeptieren Sie die vorgeschlagenen Lösungen, um den Vorgang abzuschließen. Für weitere Informationen siehe Thema Konflikte auflösen auf S. 309.

Anmerkung: Wenn Sie Ihre Bibliotheken jedesmal aktualisieren, wenn Sie eine interaktive Workbench-Sitzung starten, oder veröffentlichen, wenn Sie eine Sitzung schließen, ist die Gefahr geringer, dass Bibliotheken asynchron werden. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen auf S. 306.

Abbildung 16-2
Dialogfeld “Bibliothek hinzufügen”



So fügen Sie eine Bibliothek hinzu:

- Wählen Sie in den Menüs die Optionsfolge Ressourcen > Bibliothek hinzufügen. Das Dialogfeld “Bibliotheken hinzufügen” wird geöffnet.
- Wählen Sie die Bibliotheken aus der Liste aus.

- Klicken Sie auf Hinzufügen. Wenn Konflikte zwischen den neu hinzugefügten Bibliotheken und bereits vorhandenen Bibliotheken auftreten, werden Sie aufgefordert, die Konfliktauflösungen vor Abschluss des Vorgangs zu bestätigen oder zu ändern. Für weitere Informationen siehe Thema Konflikte auflösen auf S. 309.

Fachausdrücke und Typen suchen

Sie können mit der Suchfunktion in den verschiedenen Bereichen im Editor suchen. Im Editor können Sie die Optionsfolge Bearbeiten > Suchen in den Menüs wählen, woraufhin die Symbolleiste “Suchen” angezeigt wird. Mit dieser Symbolleiste können Sie jeweils ein Vorkommnis suchen. Durch erneutes Klicken auf Suchen können Sie nachfolgende Vorkommnisse Ihres Suchausdrucks suchen.

Bei der Suche durchsucht der Editor nur die in der Dropdown-Liste der Symbolleiste “Suchen” aufgeführten Bibliotheken. Wenn Alle Bibliotheken ausgewählt ist, durchsucht das Programm alle Bibliotheken im Editor.

Eine Suche beginnt in dem fokussierten Bereich. Sie wird in Schleifen durch alle Abschnitte fortgesetzt, bis sie zur aktiven Zelle zurückkehrt. Mit den Richtungspfeilen können Sie die Suchrichtung umkehren. Sie können auch wählen, ob die Suche Groß-/Kleinschreibung unterscheiden soll.

So finden Sie Zeichenketten in der Ansicht:

- Wählen Sie im Menü Bearbeiten > Suchen. Die Symbolleiste “Suchen” wird angezeigt.
- Geben Sie die zu suchende Zeichenfolge ein.
- Klicken Sie auf die Schaltfläche Suchen, um die Suche zu starten. Das nächste Vorkommnis des Fachausdrucks oder Typs wird hervorgehoben.
- Klicken Sie erneut auf die Schaltfläche, um zu den nachfolgenden Vorkommnissen zu navigieren.

Bibliotheken anzeigen

Sie können den Inhalt einer bestimmten Bibliothek oder aller Bibliotheken anzeigen. Dies kann hilfreich sein, wenn Sie mit vielen Bibliotheken arbeiten oder wenn Sie den Inhalt einer bestimmten Bibliothek vor der Veröffentlichung überprüfen möchten. Eine Änderung der Ansicht beeinflusst nur die Anzeige in der Registerkarte “Bibliotheksressourcen”, deaktiviert aber keine Bibliotheken bei der Extrahierung. Für weitere Informationen siehe Thema Lokale Bibliotheken deaktivieren auf S. 303.

Die Standardansicht ist Alle Bibliotheken, die alle Bibliotheken im Baum und ihren Inhalt in anderen Bereichen anzeigt. Sie können diese Auswahl mit der Dropdown-Liste auf der Symbolleiste oder durch eine Menüauswahl (Ansicht > Bibliotheken) ändern. Wenn eine einzelne Bibliothek angezeigt wird, werden alle Elemente in anderen Bibliotheken aus der Ansicht entfernt, aber trotzdem bei der Extrahierung gelesen.

So ändern Sie die Bibliothekenansicht:

- Wählen Sie in den Menüs der Registerkarte "Bibliotheksressourcen" die Optionsfolge Ansicht > Bibliotheken. Ein Menü mit allen lokalen Bibliotheken wird geöffnet.
- Wählen Sie die anzuzeigende Bibliothek aus oder wählen Sie die Option Alle Bibliotheken, um den Inhalt aller Bibliotheken anzuzeigen. Der Inhalt der Ansicht wird entsprechend Ihrer Auswahl gefiltert.

Lokale Bibliotheken verwalten

Lokale Bibliotheken sind im Gegensatz zu öffentlichen Bibliotheken die Bibliotheken innerhalb Ihrer interaktiven Workbench-Sitzung oder innerhalb einer Vorlage. Für weitere Informationen siehe Thema Public Libraries verwalten auf S. 303. Es gibt auch einige grundlegende Maßnahmen zur Verwaltung lokaler Bibliotheken, die Sie eventuell durchführen möchte. Dazu gehören: Umbenennen, Deaktivieren bzw. Löschen von lokalen Bibliotheken.

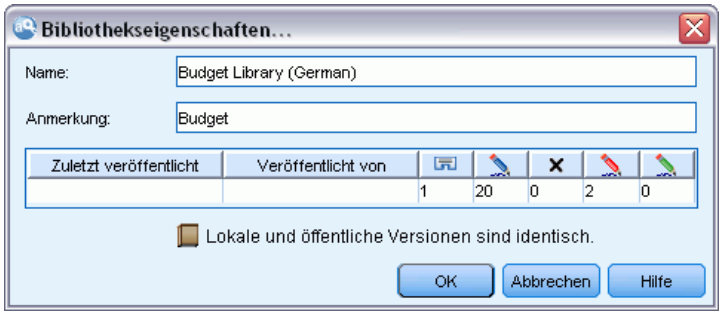
Lokale Bibliotheken umbenennen

Sie haben die Möglichkeit, lokale Bibliotheken umzubenennen. Wenn Sie eine lokale Bibliothek umbenennen, trennen Sie ihre Verknüpfung mit der öffentlichen Version, sofern eine öffentliche Version existiert. In diesem Fall können nachfolgende Änderungen nicht mehr in der öffentlichen Version gemeinsam genutzt werden. Sie können diese lokale Bibliothek unter einem neuen Namen erneut veröffentlichen. Dies bedeutet außerdem, dass Sie in der ursprünglichen öffentlichen Version keine Änderungen aktualisieren können, die Sie in der lokalen Version vornehmen.

Anmerkung: Eine Public Library kann nicht umbenannt werden.

- Wählen Sie in den Menüs die Optionsfolge Bearbeiten > Bibliotheken-Eigenschaften. Das Dialogfeld "Bibliotheken-Eigenschaften" wird geöffnet.

Abbildung 16-3
Dialogfeld "Bibliotheken-Eigenschaften"



So benennen Sie eine lokale Bibliothek um:

- Wählen Sie in der Baumansicht die Bibliothek aus, die Sie umbenennen möchten.
- Geben Sie im Textfeld "Name" einen neuen Namen für die Bibliothek ein.

- Klicken Sie auf OK, um den neuen Namen der Bibliothek zu akzeptieren. Das Dialogfeld wird geschlossen und der Name der Bibliothek wird in der Baumansicht aktualisiert.

Lokale Bibliotheken deaktivieren

Um eine Bibliothek vorübergehend aus dem Extrahierungsprozess auszuschließen, heben Sie die Auswahl des Kontrollkästchens links von dem Namen der Bibliothek in der Baumansicht auf. Dies signalisiert, dass die Bibliothek beibehalten, aber der Inhalt bei der Konfliktprüfung und der Extrahierung ignoriert werden soll.

So deaktivieren Sie eine Bibliothek:

- Wählen Sie im Bibliotheksbaum die Bibliothek aus, die Sie deaktivieren möchten.
- Drücken Sie die Leertaste. Das Kontrollkästchen links neben dem Namen wird deaktiviert.

Lokale Bibliotheken löschen

Sie können eine Bibliothek entfernen, ohne die öffentliche Version der Bibliothek zu löschen, und umgekehrt. Beim Löschen einer lokalen Bibliothek werden die Bibliothek und ihr gesamter Inhalt nur aus der Sitzung gelöscht. Wenn Sie eine lokale Version einer Bibliothek löschen, wird diese Bibliothek nicht aus anderen Sitzungen entfernt und auch die öffentliche Version wird nicht entfernt. Für weitere Informationen siehe Thema Public Libraries verwalten auf S. 303.

So löschen Sie eine lokale Bibliothek:

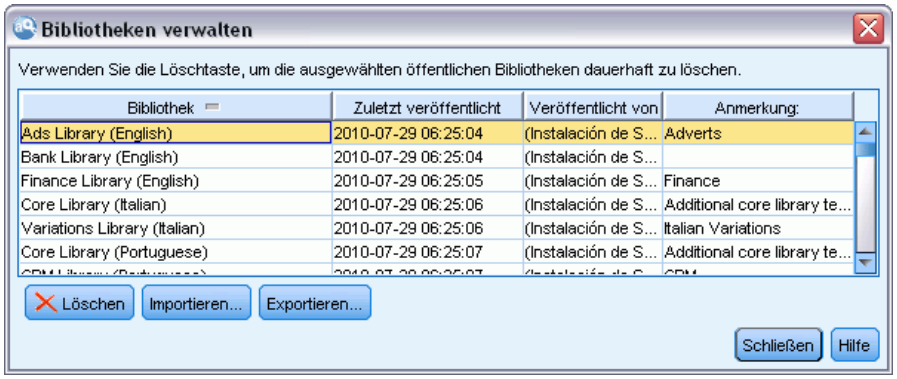
- Wählen Sie in der Baumansicht die zu löschende Bibliothek aus.
- Wählen Sie in den Menüs die Optionsfolge Bearbeiten > Löschen, um die Bibliothek zu löschen. Die Bibliothek wird entfernt.
- Wenn diese Bibliothek zuvor nicht veröffentlicht worden ist, werden Sie aufgefordert zu entscheiden, ob Sie die Bibliothek löschen oder beibehalten möchten. Klicken Sie auf Löschen, um fortzufahren, oder auf Beibehalten, um die Bibliothek beizubehalten.

  Anmerkung: Es muss immer eine Bibliothek bestehen bleiben.

Public Libraries verwalten

Um lokale Bibliotheken wiederzuverwenden, können Sie sie veröffentlichen und dann über die Dialogbox “Bibliotheken verwalten” (Ressourcen > Bibliotheken verwalten) anzeigen und so mit ihnen arbeiten. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen auf S. 306. Zu den grundlegenden Maßnahmen zur Verwaltung von Public Libraries gehören das Importieren, Exportieren oder Löschen einer Public Library. Eine Public Library kann nicht umbenannt werden.

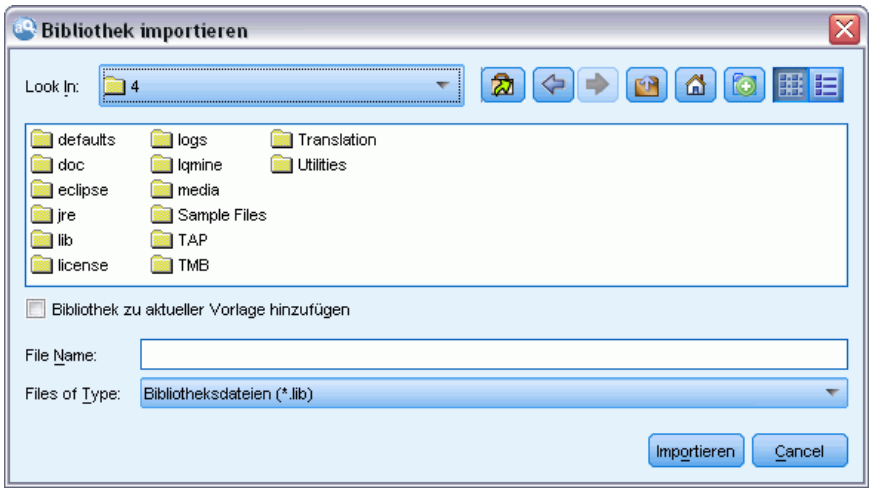
Abbildung 16-4
Dialogfeld "Bibliotheken verwalten"



Public Libraries importieren

- Klicken Sie im Dialogfeld "Bibliotheken verwalten" auf Importieren.... Das Dialogfeld "Bibliothek importieren" wird geöffnet.

Abbildung 16-5
Dialogfeld "Bibliothek importieren"



- Wählen Sie die zu importierende Bibliothekendatei (**.lib*) und wählen Sie Bibliothek zu aktuellem Projekt hinzufügen, wenn Sie diese Bibliothek außerdem lokal hinzufügen möchten.

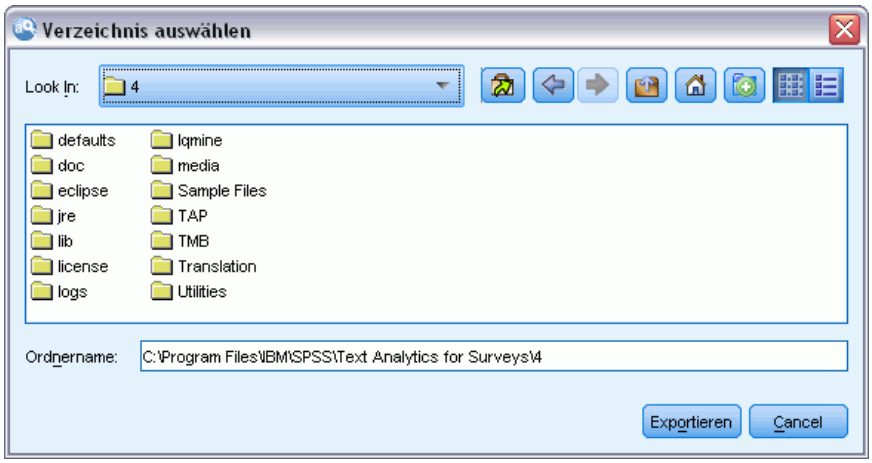
- Klicken Sie auf Importieren. Das Dialogfeld wird geschlossen. Wenn bereits eine Public Library mit demselben Namen vorhanden ist, werden Sie aufgefordert, die zu importierende Bibliothek umzubenennen oder die aktuelle Public Library zu überschreiben.

Public Libraries exportieren

Sie können Public Libraries in das *.lib*-Format exportieren, sodass Sie sie mit anderen Benutzern gemeinsam nutzen können.

- Wählen Sie im Dialogfeld "Bibliotheken verwalten" die Bibliothek, die in die Liste exportiert werden soll.
- Klicken Sie auf Exportieren. Das Dialogfeld "Verzeichnis auswählen" wird geöffnet.

Abbildung 16-6
Dialogfeld "Verzeichnis auswählen"



- Wählen Sie das Verzeichnis aus, in das exportiert werden soll, und klicken Sie auf Exportieren. Das Dialogfeld wird geschlossen und die Bibliothekendatei (**.lib*) wird exportiert.

Public Libraries löschen

Sie können eine lokale Bibliothek entfernen, ohne die öffentliche Version der Bibliothek zu löschen, und umgekehrt. Wenn die Bibliothek jedoch aus diesem Dialogfeld entfernt wird, kann sie nicht mehr zu Sitzungsressourcen hinzugefügt werden, bis wieder eine lokale Version veröffentlicht wird.

Wenn eine Bibliothek gelöscht wird, die zusammen mit dem Produkt installiert wurde, wird die ursprünglich installierte Version wiederhergestellt.

- Wählen Sie im Dialogfeld "Bibliotheken verwalten" die zu löschende Bibliothek aus. Sortieren Sie die Liste, indem Sie auf den entsprechenden Titel klicken.
- Klicken Sie zum Löschen der Bibliothek auf Löschen. IBM® SPSS® Modeler Text Analytics bestätigt, ob die lokale Version der Bibliothek mit der öffentlichen Bibliothek übereinstimmt. In dem Fall wird die Bibliothek ohne Warnhinweis entfernt. Wenn sich die Versionen der Bibliothek unterscheiden, wird ein Warnhinweis geöffnet, der Sie auffordert zu entscheiden, ob Sie die öffentliche Version beibehalten oder entfernen möchten.

Bibliotheken gemeinsam nutzen

Mithilfe von Bibliotheken können Sie Ressourcen leicht in mehreren interaktiven Workbench-Sitzungen gemeinsam nutzen. Bibliotheken können in zwei Status oder Versionen vorliegen. Bibliotheken, die im Editor bearbeitet werden können und Teil einer interaktiven Workbench-Sitzung sind, werden als **lokale Bibliotheken** bezeichnet. Während der Arbeit an einer interaktiven Workbench-Sitzung können Sie zahlreiche Änderungen, z. B. in der *Vegetables*-Bibliothek, vornehmen. Wenn Ihre Änderungen auch für andere Daten nützlich sein können, können Sie diese Ressourcen durch Erstellen einer **Public Library**-Version der *Vegetables*-Bibliothek verfügbar machen. Wie der Name andeutet, ist eine Public Library für alle anderen Ressourcen in allen interaktiven Workbench-Sitzungen verfügbar.

Die Public Libraries werden im Dialogfeld “Bibliotheken verwalten” angezeigt. Sobald diese öffentliche Version der Bibliothek vorhanden ist, können Sie sie zu den Ressourcen in anderen Kontexten hinzufügen, sodass diese benutzerdefinierten linguistischen Ressourcen gemeinsam genutzt werden können.

Die mitgelieferten Bibliotheken sind anfänglich Public Libraries. Sie können die Ressourcen in diesen Bibliotheken bearbeiten und dann eine neue öffentliche Version erstellen. Die neuen Versionen sind dann in anderen interaktiven Workbench-Sitzungen verfügbar.

Wenn Sie die Arbeit mit Ihren Bibliotheken fortsetzen und Änderungen vornehmen, werden die Versionen der Bibliothek asynchron. In einigen Fällen kann eine lokale Version aktueller als die öffentliche Version sein und in anderen Fällen kann die öffentliche Version aktueller als die lokale Version sein. Es ist auch möglich, dass sowohl die öffentliche als auch die lokale Version Änderungen enthalten, die nicht in der anderen enthalten sind, wenn die öffentliche Version aus einer anderen interaktiven Workbench-Sitzung aktualisiert wurde. Wenn Ihre Bibliotheken asynchron werden, können Sie sie neu synchronisieren. Zum Synchronisieren der Versionen einer Bibliothek werden Local Libraries erneut veröffentlicht und/oder aktualisiert.

Wenn Sie eine interaktive Workbench-Sitzung starten oder schließen, werden Sie aufgefordert, alle Bibliotheken zu synchronisieren, die aktualisiert oder erneut veröffentlicht werden sollten. Außerdem können Sie problemlos den Synchronisierungsstatus Ihrer lokalen Bibliothek an dem Symbol ablesen, das in der Baumansicht neben dem Bibliotheksnamen angezeigt wird, oder indem Sie das Dialogfeld “Bibliothekseigenschaften” anzeigen. Dies ist auch zu einem beliebigen Zeitpunkt durch Menüauswahl möglich. Die folgende Tabelle beschreibt die fünf möglichen Status und die ihnen zugeordneten Symbole.

Tabelle 16-1
Status der Synchronisierung von Local Libraries

Symbol	Statusbeschreibung von Local Libraries
	Unveröffentlicht—Die Local Library ist nie veröffentlicht worden.
	Synchronisiert—Die lokale und die öffentliche Version der Bibliothek stimmen überein. Dies gilt auch für die *Local Library*, die nicht veröffentlicht werden kann, da sie nur sitzungsspezifische Ressourcen enthalten soll.
	Veraltet—Die öffentliche Version der Bibliothek ist aktueller als die lokale Version. Sie können die lokale Version mit den Änderungen aktualisieren.

Symbol	Statusbeschreibung von Local Libraries
	Neuer—Die lokale Version der Bibliothek ist aktueller als die öffentliche Version. Sie können die lokale Version erneut als öffentliche Version veröffentlichen.
	Asynchron—Sowohl die Local Library als auch die Public Library enthalten Änderungen, die in der anderen nicht enthalten sind. Sie müssen die Local Library entweder aktualisieren oder veröffentlichen. Bei einer Aktualisierung verlieren Sie die Änderungen, die Sie seit der letzten Aktualisierung oder Veröffentlichung vorgenommen haben. Bei einer Veröffentlichung überschreiben Sie die Änderungen in der öffentlichen Version.

Anmerkung: Wenn Sie Ihre Bibliotheken jedesmal aktualisieren, wenn Sie eine interaktive Workbench-Sitzung starten, oder veröffentlichen, wenn Sie eine Sitzung schließen, ist die Gefahr geringer, dass Bibliotheken asynchron werden.

Sie können eine Bibliothek zu einem beliebigen Zeitpunkt erneut veröffentlichen, zu dem die Änderungen in der Bibliothek anderen Streams, die diese Bibliothek ebenfalls enthalten, zugute kommen. Wenn die Änderungen anderen Streams zugute kommen, können Sie die lokalen Versionen in diesen Streams aktualisieren. So können Sie Streams für jeden Kontext oder jede Domäne erstellen, die zu Ihren Daten passen, indem Sie neue Bibliotheken erstellen und/oder Ihren Ressourcen eine beliebige Anzahl von Public Libraries hinzufügen.

Wenn eine öffentliche Version einer Bibliothek gemeinsam genutzt wird, ist die Wahrscheinlichkeit größer, dass Unterschiede zwischen der lokalen und der öffentlichen Version entstehen. Wenn Sie eine interaktive Workbench-Sitzung starten oder eine interaktive Workbench-Sitzung schließen und aus dieser Bibliothek veröffentlichen bzw. eine Vorlage aus dem Template Editor schließen , wird eine Meldung angezeigt, damit Sie Bibliotheken veröffentlichen und/oder aktualisieren können, die nicht mehr synchron mit denjenigen im Dialogfeld “Bibliotheken verwalten” sind. Wenn die öffentliche Version der Bibliothek aktueller als die lokale Version ist, wird ein Dialogfeld geöffnet, das Sie auffordert zu entscheiden, ob Sie aktualisieren möchten. Wählen Sie, ob Sie die lokale Version unverändert beibehalten möchten, statt sie mit der öffentlichen Version zu aktualisieren, oder ob Sie die Aktualisierungen in der Local Library zusammenführen möchten.

Bibliotheken veröffentlichen

Wenn eine bestimmte Bibliothek noch nie veröffentlicht worden ist, gehört zur Veröffentlichung die Erstellung einer öffentlichen Kopie Ihrer Local Library in der Datenbank. Bei der erneuten Veröffentlichung einer Bibliothek wird der Inhalt der öffentlichen Version durch den Inhalt der Local Library ersetzt. Nach der erneuten Veröffentlichung können Sie diese Bibliothek in allen anderen Stream-Sitzungen aktualisieren, sodass ihre lokalen Versionen synchron mit der öffentlichen Version sind. Obwohl eine Bibliothek veröffentlicht werden kann, wird immer eine lokale Version in der Sitzung gespeichert.

Wichtig: Wenn Sie Änderungen an Ihrer Local Library vornehmen und in der Zwischenzeit auch die öffentliche Version der Bibliothek geändert wurde, gilt Ihre Bibliothek als asynchron. Sie sollten zunächst die lokale Version mit den öffentlichen Änderungen aktualisieren, danach die gewünschten Änderungen vornehmen und dann Ihre lokale Version erneut veröffentlichen, damit beide Versionen übereinstimmen. Wenn Sie zuerst Änderungen vornehmen und die Bibliothek veröffentlichen, überschreiben Sie die Änderungen in der öffentlichen Version.

Abbildung 16-7
Dialogfeld "Bibliotheken veröffentlichen"

So veröffentlichen Sie Local Libraries in der Datenbank:

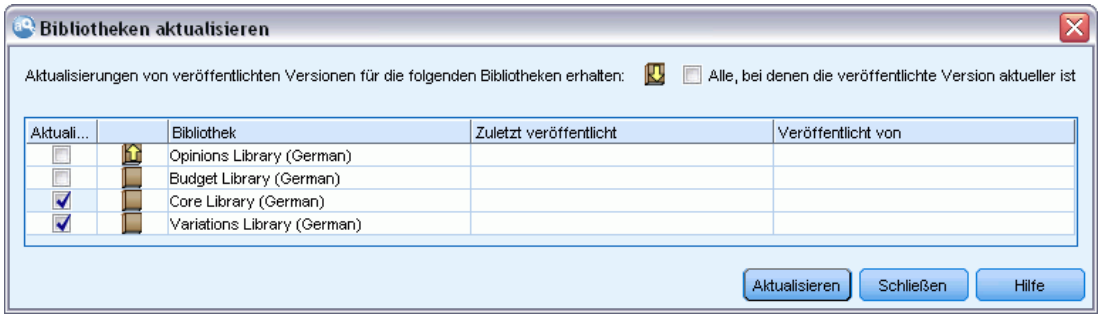
- Wählen Sie in den Menüs die Optionsfolge Ressourcen > Bibliotheken veröffentlichen. Das Dialogfeld "Bibliotheken veröffentlichen" wird geöffnet, wobei alle Bibliotheken, die veröffentlicht werden sollten, standardmäßig ausgewählt sind.
- Wählen Sie das Kontrollkästchen links von jeder Bibliothek, die Sie veröffentlichen oder erneut veröffentlichen möchten.
- Klicken Sie auf Veröffentlichen, um die Bibliotheken in der Datenbank "Bibliotheken verwalten" zu veröffentlichen.

Bibliotheken aktualisieren

Wenn Sie eine interaktive Workbench-Sitzung starten oder schließen, können Sie alle Bibliotheken aktualisieren oder veröffentlichen, die nicht mehr synchron mit den öffentlichen Versionen sind. Wenn die öffentliche Version der Bibliothek aktueller als die lokale Version ist, wird ein Dialogfeld geöffnet, das Sie auffordert zu entscheiden, ob Sie die Bibliothek aktualisieren möchten. Wählen Sie, ob Sie die lokale Version beibehalten möchten, statt sie mit der öffentlichen Version zu aktualisieren oder die lokale Version durch die öffentliche zu ersetzen. Wenn eine öffentliche Version einer Bibliothek aktueller als Ihre lokale Version ist, können Sie die lokale Version aktualisieren, um ihren Inhalt mit dem der öffentlichen Version zu synchronisieren. Aktualisieren bedeutet, die Änderungen in der öffentlichen Version in die lokale Version zu integrieren.

Anmerkung: Wenn Sie Ihre Bibliotheken jedesmal aktualisieren, wenn Sie eine interaktive Workbench-Sitzung starten, oder veröffentlichen, wenn Sie eine Sitzung schließen, ist die Gefahr geringer, dass Bibliotheken asynchron werden. Für weitere Informationen siehe Thema Bibliotheken gemeinsam nutzen auf S. 306.

Abbildung 16-8
Dialogfeld "Bibliotheken aktualisieren"



So aktualisieren Sie Local Libraries:

- Wählen Sie in den Menüs die Optionsfolge Ressourcen > Bibliotheken aktualisieren. Das Dialogfeld "Bibliotheken aktualisieren" wird geöffnet, wobei alle Bibliotheken, die veröffentlicht werden sollten, standardmäßig ausgewählt sind.
- Wählen Sie das Kontrollkästchen links von jeder Bibliothek, die Sie veröffentlichen oder erneut veröffentlichen möchten.
- Klicken Sie zum Aktualisieren der Local Library auf Aktualisieren.

Konflikte auflösen

Konflikte zwischen Local und Public Libraries

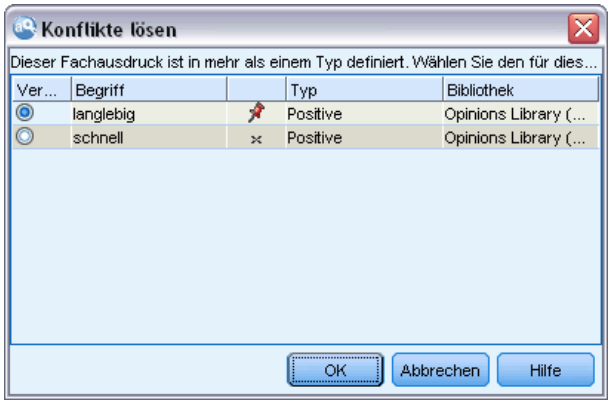
Wenn Sie eine Stream-Sitzung starten, führt IBM® SPSS® Modeler Text Analytics einen Vergleich der Local Libraries mit den im Dialogfeld "Bibliotheken verwalten" aufgeführten Bibliotheken durch. Falls Local Libraries in Ihrem Sitzungs nicht mit den veröffentlichten Versionen synchron sind, wird das Dialogfeld "Warnung zur Synchronisierung der Bibliothek" geöffnet. Wählen Sie eine der folgenden Optionen, um die Versionen der Bibliotheken auszuwählen, die Sie hier verwenden möchten:

- **Alle Bibliotheken, die in der Datei als lokal eingestuft werden.** Mit dieser Option werden alle Local Libraries unverändert beibehalten. Sie können Sie später erneut veröffentlichen oder aktualisieren.
- **Alle auf diesem Computer veröffentlichten Bibliotheken.** Diese Option ersetzt die angezeigten Local Libraries durch die Versionen in der Datenbank.
- **Alle aktuelleren Bibliotheken.** Diese Option ersetzt die älteren Local Libraries durch die aktuelleren öffentlichen Versionen in der Datenbank.
- **Andere.** Mit dieser Option können Sie die gewünschten Versionen manuell in der Tabelle auswählen.

Konflikte durch erzwungene Fachausdrücke

Durch Hinzufügen einer Public Library oder Aktualisierung einer Local Library können Konflikte oder doppelte Einträge unter den Fachausdrücken und Typen in diesem Wörterbuch und den Fachausdrücken und Typen in den anderen Wörterbüchern in Ihren Ressourcen aufgedeckt werden. In diesem Fall werden Sie aufgefordert, die vorgeschlagenen Konfliktauflösungen vor Abschluss des Vorgangs in dem Dialogfeld "Erzwungene Fachausdrücke bearbeiten" zu verifizieren oder zu ändern. Für weitere Informationen siehe Thema Erzwingen von Fachausdrücken in Kapitel 17 auf S. 320.

Abbildung 16-9
Dialogfeld "Erzwungene Fachausdrücke bearbeiten"



Das Dialogfeld "Erzwungene Fachausdrücke bearbeiten" enthält alle Paare der in Konflikt stehenden Fachausdrücke oder Typen. Die Konfliktpaare werden durch abwechselnde Hintergrundfarben optisch unterschieden. Diese Farben können im Dialogfeld "Optionen" geändert werden. Für weitere Informationen siehe Thema Optionen: Registerkarte "Anzeige" in Kapitel 8 auf S. 136. Das Dialogfeld "Erzwungene Fachausdrücke bearbeiten" enthält zwei Registerkarten:

- **Doppelte Einträge.** Diese Registerkarte enthält die in den Bibliotheken ermittelten doppelten Einträge. Wenn ein Reißzwecken-Symbol hinter einem Fachausdruck angezeigt wird, ist das Vorkommnis dieses Fachausdrucks erzwungen worden. Wenn ein schwarzes X-Symbol angezeigt wird, wird das Vorkommnis dieses Fachausdrucks bei der Extrahierung ignoriert, weil er an anderer Stelle erzwungen worden ist.
- **Benutzerdefiniert.** Diese Registerkarte enthält eine Liste der Fachausdrücke, die manuell im Fachausdrucksbereich des Typ-Wörterbuchs und nicht durch Konflikte erzwungen worden sind.

Anmerkung: Das Dialogfeld "Erzwungene Fachausdrücke bearbeiten" wird geöffnet, nachdem Sie Einträge zu einer Bibliothek hinzugefügt oder die Bibliothek aktualisiert haben. Wenn Sie den Vorgang in diesem Dialogfeld abbrechen, brechen Sie nicht die Aktualisierung der Bibliothek oder das Hinzufügen ab.

So lösen Sie Konflikte auf:

▶ Wählen Sie im Dialogfeld "Erzwungene Fachausdrücke bearbeiten" die Optionsschaltfläche in der Spalte "Verwenden" für den zu erzwingenden Fachausdruck.

- Nach Abschluss der Auswahl klicken Sie auf OK, um die erzwungenen Fachausdrücke anzuwenden und das Dialogfeld zu schließen. Klicken Sie auf Abbrechen, um den Änderungsvorgang in diesem Dialogfeld abzubrechen.

Kapitel
17

Informationen zu Bibliothekwörterbüchern

Die zum Extrahieren von Textdaten verwendeten Ressourcen werden in Form von Vorlagen und Bibliotheken gespeichert. Eine Bibliothek kann aus drei Wörterbüchern bestehen.

- Das **Typ-Wörterbuch** enthält eine Zusammenstellung von Fachausdrücken, die unter einer Bezeichnung oder einem Typnamen gruppiert sind. Beim Lesen der Textdaten vergleicht die Extrahierungsengine die im Text aufgefundenen Wörter mit den Ausdrücken, die in den Typ-Wörterbüchern definiert sind. Bei der Extrahierung werden die gebeugten Formen der Fachausdrücke und Synonyme eines Typs unter einem Zielausdruck zusammengefasst, der als Konzept bezeichnet wird. Extrahierte Konzepte werden dem Typ-Wörterbuch zugewiesen, in dem sie als Ausdrücke erscheinen. Sie können Ihre Typ-Wörterbücher in den oben links und in der Mitte gelegenen Bereichen des Editors verwalten — in der Bibliotheksstruktur und im Ausdrucksbereich. Für weitere Informationen siehe Thema Typ-Wörterbücher auf S. 312.
- Das **Substitutionswörterbuch** enthält eine Zusammenstellung von Wörtern, die als Synonyme oder optionale Elemente definiert sind und zur Gruppierung ähnlicher Ausdrücke unter einem Zielausdruck verwendet werden, dem sogenannten “Konzept” in den Extrahierungsendergebnissen. Sie können Ihre Substitutionswörterbücher im unten links gelegenen Bereich des Editors über die Registerkarte “Synonyme” und die Registerkarte “Optional” verwalten. Für weitere Informationen siehe Thema Substitutions-/Synonymwörterbücher auf S. 323.
- Das **Ausschlusswörterbuch** enthält eine Sammlung von Fachausdrücken und Typen, die aus den endgültigen Extrahierungsergebnissen entfernt werden. Sie können Ihre Ausschlusswörterbücher im ganz rechts gelegenen Bereich des Editors verwalten. Für weitere Informationen siehe Thema Ausschlusswörterbücher auf S. 329.

Für weitere Informationen siehe Thema Mit Bibliotheken arbeiten in Kapitel 16 auf S. 297.

Typ-Wörterbücher

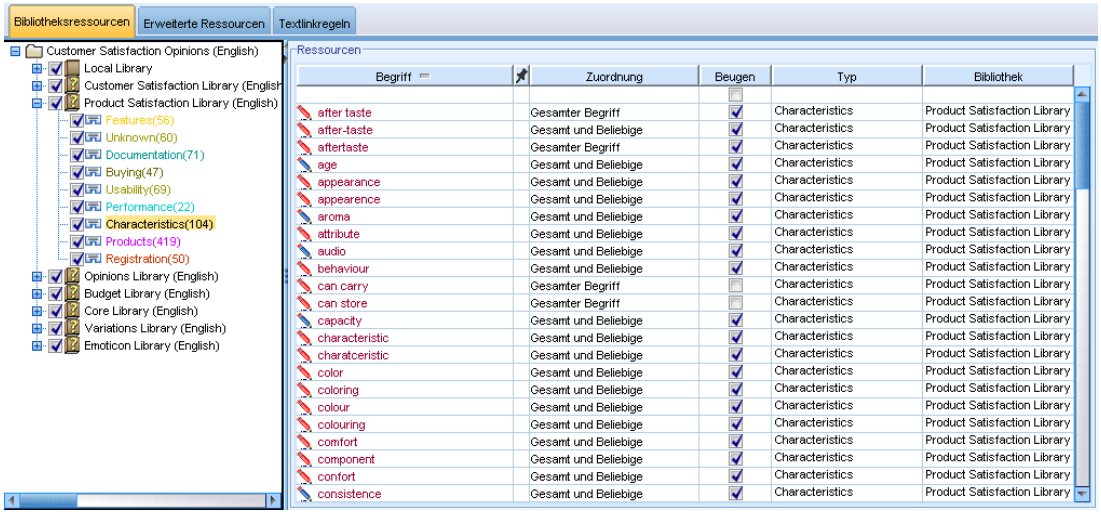
Ein **Typ-Wörterbuch** besteht aus dem Typnamen bzw. der Überschrift und einer Liste von Fachausdrücken. Typ-Wörterbücher werden im oberen linken und im mittleren Bereich der Registerkarte “Bibliotheksressourcen” im Editor verwaltet. Sie können diese Ansicht mit der Optionsfolge Ansicht > Ressourceneditor in den Menüs aufrufen, wenn Sie sich in einer interaktiven Workbench-Sitzung befinden. Andernfalls können Sie Wörterbücher für eine bestimmte Vorlage im Template Editor bearbeiten.

Wenn die Extrahierungsengine Ihre Textdaten liest, vergleicht sie die im Text gefundenen Wörter mit den in Ihren Typ-Wörterbüchern definierten Fachausdrücken. Fachausdrücke sind Wörter oder Ausdrücke in den Typ-Wörterbüchern in Ihren linguistischen Ressourcen.

Wenn ein Wort mit einem Fachausdruck übereinstimmt, wird es dem Typnamen für diesen Fachausdruck zugeordnet. Wenn die Ressourcen während des Extrahierungsprozesses gelesen werden, werden die im Text gefundenen Fachausdrücke einigen Verarbeitungsschritten

unterzogen, bevor sie Konzepte im Bereich “Extrahierungsergebnisse” werden. Wenn mehrere Fachausdrücke, die zum selben Typen-Wörterbuch gehören, von der Extrahierungsengine als synonym eingestuft werden, werden sie unter dem am häufigsten auftretenden Fachausdruck zusammengefasst und im Bereich “Extrahierungsergebnisse” als **Konzept** zusammengefasst. So könnten beispielsweise die Fachbegriffe `Frage` und `Abfrage` letzten Endes unter dem Konzeptnamen `Frage` zusammengefasst werden.

Abbildung 17-1
Bibliotheksbaum und Fachausdrucksbereich



Die Liste der Typ-Wörterbücher wird links im Bereich des Wörterbuchbaums angezeigt. Der Inhalt der einzelnen Typ-Wörterbücher wird im mittleren Bereich angezeigt. Typ-Wörterbücher enthalten mehr als nur eine Liste mit Fachausdrücken. Die festgelegte Abgleichsoption bestimmt, wie die Übereinstimmung der in Ihren Textdaten enthaltenen Wörter und Wortfolgen mit den in den Typ-Wörterbüchern definierten Fachausdrücken ermittelt wird. Eine **Abgleichsoption** legt fest, wie ein Fachausdruck hinsichtlich eines in den Textdaten enthaltenen möglichen Worts oder einer möglichen Wortfolge verankert ist. Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken auf S. 316.

Anmerkung: Für japanischen Text gelten nicht alle Optionen, z. B. die Abgleichoption und flektierte Formen. Für weitere Informationen siehe Thema Bearbeiten von japanischen Typeigenschaften in Anhang A auf S. 391.*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Darüber hinaus können Sie die Fachbegriffe in Ihrem Wörterbuch ausweiten, indem Sie angeben, ob Sie im Wörterbuch automatisch gebeugte Formen der Fachbegriffe generieren und hinzufügen möchten. Indem Sie gebeugte Formen erzeugen, werden im Typ-Wörterbuch automatisch Pluralformen für im Singular angegebene Fachausdrücke, Singularformen von Fachausdrücken und Adjektive hinzugefügt, die im Plural angegeben sind. Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken auf S. 316.

Anmerkung: In den meisten Sprachen gilt: Konzepte, die in keinem Typen-Wörterbuch gefunden, aber aus dem Text extrahiert werden, erhalten automatisch den Typ `<Unbekannt>`. Bei japanischem Text jedoch erhalten sie automatisch den Typ <名詞>*Anmerkung*: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Integrierte Typen

IBM® SPSS® Modeler Text Analytics wird mit einem Satz linguistischer Ressourcen geliefert, die als Bibliotheken und kompilierte Ressourcen vorliegen. Die mitgelieferten Bibliotheken enthalten einen Satz integrierter Wörterbücher: `<Location>`, `<Organization>`, `<Person>` und `<Product>`

Anmerkung: Das Set der integrierten Standardtypen ist für japanischen Text unterschiedlich. Weitere Informationen über die Typen, die mit japanischen Ressourcen geliefert werden, finden Sie unterVerfügbare Typen für japanischen Text. *Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Diese Typwörterbücher werden von der Extraktionsengine benutzt, um den extrahierten Begriffen Typen zuzuweisen, z. B. den Typ `<Location>` zum Begriff `Paris`. Obwohl in den integrierten Wörterbüchern eine Vielzahl von Fachausdrücken definiert ist, decken diese nicht alle Möglichkeiten ab. Sie können daher Fachausdrücke hinzufügen oder eigene Wörterbücher anlegen. Eine Beschreibung der Inhalte eines der mitgelieferten Typ-Wörterbücher finden Sie in den Anmerkungen, die das Dialogfeld “Typeigenschaften” aufführt. Wählen Sie den Typ im Baum aus und wählen Sie die Optionsfolge Bearbeiten > Eigenschaften aus dem Kontextmenü.

Anmerkung: Neben den mitgelieferten Bibliotheken enthalten die kompilierten Ressourcen (die ebenfalls von der Extrahierungsengine verwendet werden) eine Vielzahl von Definitionen, die die integrierten Wörterbücher ergänzen. Ihre Inhalte sind im Produkt allerdings nicht einsehbar. Sie können jedoch erzwingen, dass ein Fachausdruck, dessen Typ durch die kompilierten Wörterbücher bestimmt wurde, in ein anderes Wörterbuch übertragen wird. Für weitere Informationen siehe Thema Erzwingen von Fachausdrücken auf S. 320.

Erstellen von Typen

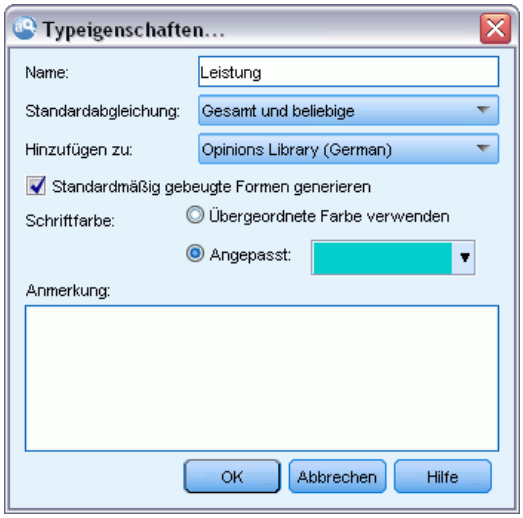
Sie können Typ-Wörterbücher erstellen, mit denen ähnliche Fachausdrücke gruppiert werden können. Wenn während des Extrahierungsprozesses in diesen Wörterbüchern vorhandene Fachausdrücke erkannt werden, wird diesen der entsprechende Typname zugewiesen und sie werden unter einem Konzeptnamen extrahiert. Wenn Sie eine Bibliothek erstellen, enthält diese automatisch eine leere Typbibliothek. Sie können daher sofort mit der Eingabe von Fachausdrücken beginnen.

Wichtig!: Für japanische Ressourcen können Sie keine Typen erstellen. Weitere Informationen über die Typ-Wörterbücher für japanische Ressourcen finden Sie unterJapanischer Bibliotheksbaum, Typen und Fachausdrucksbereich. *Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Wenn Sie Text über Nahrungsmittel analysieren und Fachausdrücke zum Thema Gemüse gruppieren möchten, können Sie Ihr eigenes Typ-Wörterbuch `<Gemüse>` erstellen. Dort können Sie dann Begriffe wie `Karotte`, `Brokkoli` und `Spinat` hinzufügen, sofern Sie der Meinung sind, dass es sich um wichtige Fachausdrücke handelt, die im Text vorkommen. Wenn dann während der Extrahierung einer dieser Begriffe gefunden wird, wird dieser als Konzept extrahiert und dem Typ `<Gemüse>` zugeordnet.

Es ist nicht erforderlich, dass Sie alle Formen eines Worts oder eines Ausdrucks definieren, Sie können die gebeugten Formen der Begriffe stattdessen generieren lassen. Wenn Sie diese Option wählen, erkennt die Extrahierungsengine automatisch die Singular- bzw. Pluralformen der Fachausdrücke sowie andere zum Begriff gehörende Formen. Dies Option ist besonders hilfreich, wenn Ihr Typ überwiegend Substantive enthält, da es eher unwahrscheinlich ist, dass Sie gebeugte Formen von Verben oder Adjektiven erzeugen möchten.

Abbildung 17-2
Dialogfeld "Typ-Eigenschaften"



Name. Der Name des Typ-Wörterbuchs, das sie gerade erstellen. Es wird empfohlen, keine Leerzeichen in Typnamen zu verwenden, insbesondere wenn zwei oder mehr Typnamen mit demselben Wort beginnen.

Anmerkung: Es gibt einige Einschränkungen hinsichtlich Typnamen und der Verwendung von Symbolen. Beispielsweise dürfen keine Symbole wie "@" oder "!" im Namen verwendet werden.

Standardabgleich. Das Attribut "Standardabgleich" legt fest, wie die Extrahierungsengine die Übereinstimmung dieses Fachausdrucks mit den Textdaten ermittelt. Jedes Mal, wenn Sie einen Fachausdruck zu diesem Typ-Wörterbuch hinzufügen, wird diesem automatisch dieses Abgleichsattribut zugewiesen. Die Abgleichsseinstellung können Sie in der Fachausdrucksliste jederzeit manuell ändern. Zu den Optionen gehören: Gesamter Begriff, Start, Ende, Beliebig, Start oder Ende, Gesamt und Start, Gesamt und Ende, Gesamt und (Start oder Ende) und Gesamt (keine Zusammensetzungen). Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken auf S. 316. Diese Option gilt nicht für japanische Ressourcen.

Hinzufügen zu. Dieses Feld gibt die Bibliothek an, in der Sie Ihr neues Typ-Wörterbuch erstellen.

Gebeugte Formen standardmäßig generieren. Diese Option weist die Extrahierungsengine an, die grammatikalische Morphologie zu verwenden, um ähnliche Formen der Fachausdrücke, die Sie zu diesem Wörterbuch hinzufügen, zu erfassen und zusammenzufassen, wie Singular- oder Pluralformen des Worts. Diese Option ist besonders hilfreich, wenn Ihr Typ überwiegend Substantive enthält. Wenn Sie diese Option auswählen, besitzen alle neu zu diesem Typ hinzugefügten Fachausdrücke automatisch diese Option, die Sie in der Liste allerdings manuell ändern können. Diese Option gilt nicht für japanische Ressourcen.

Schriftfarbe. Über dieses Feld können Sie ein Unterscheidungsmerkmal für diesen Typ in Bezug auf die anderen auf der Benutzeroberfläche dargestellten Typen festlegen. Wenn Sie die Option Übergeordnete Farbe verwenden auswählen, wird für dieses Typ-Wörterbuch die Standardtypfarbe verwendet. Die Standardfarbe wird im Dialogfeld "Optionen" festgelegt. Für weitere Informationen siehe Thema Optionen: Registerkarte "Anzeige" in Kapitel 8 auf S. 136. Wenn Sie Benutzerdefiniert auswählen, wählen Sie aus der Dropdown-Liste eine Farbe aus.

Anmerkung. Dieses Feld ist optional und wird für Kommentare und Beschreibungen verwendet.

So erstellen Sie ein Typ-Wörterbuch:

- Wählen Sie eine Bibliothek aus, in der Sie ein neues Typ-Wörterbuch erstellen möchten.
- Wählen Sie im Menü Extras > Neuer Typ aus. Das Dialogfeld "Typ-Eigenschaften" wird geöffnet.
- Geben Sie den Namen für Ihr Typ-Wörterbuch in das Textfeld Name ein und wählen Sie die gewünschten Optionen aus.
- Klicken Sie auf OK, um das Typ-Wörterbuch zu erstellen. Der neue Typ wird im Bibliotheksbaum und im mittleren Bereich angezeigt. Sie können sofort beginnen, Fachausdrücke hinzuzufügen. Weitere Informationen finden Sie unter Hinzufügen von Fachausdrücken.

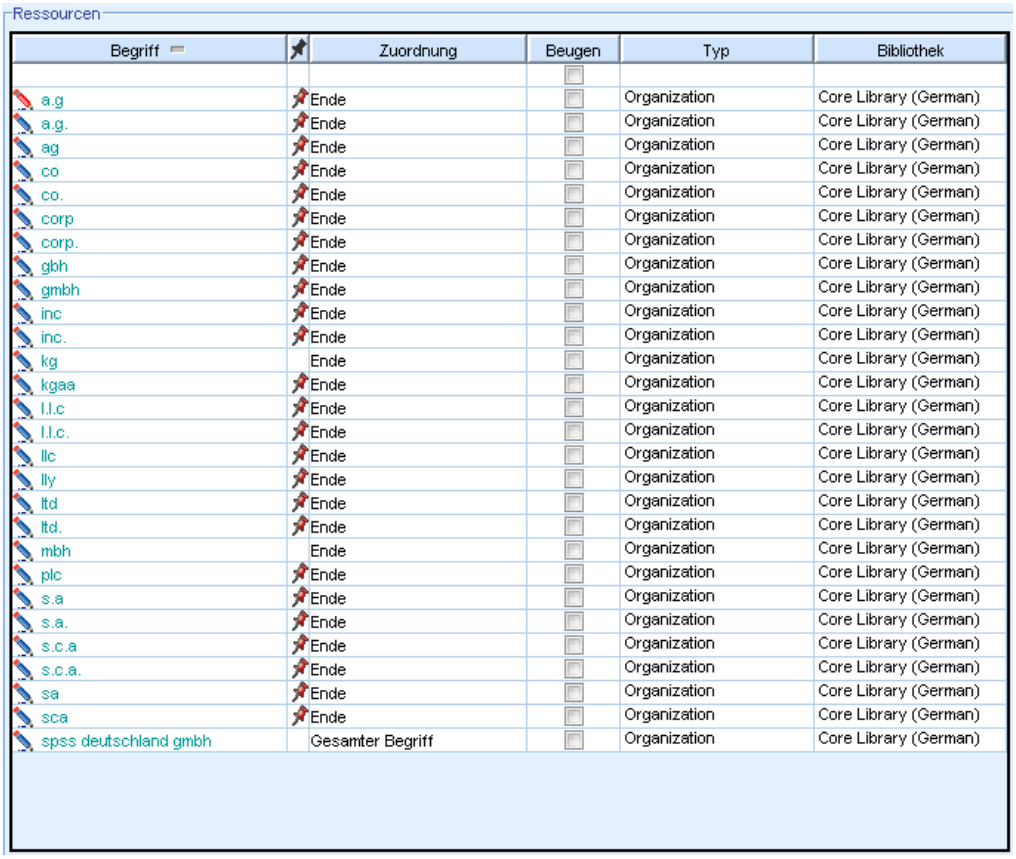
  Hinweis: Diese Anweisungen zeigen Ihnen, wie Sie in der Resource Editor-Ansicht oder im Template Editor Änderungen vornehmen können. Beachten Sie, dass Sie solche Feinabstimmungen auch direkt im Bereich "Extrahierungsergebnisse", im Bereich "Daten", im Bereich "Kategorien" oder im Dialogfeld "Clusterdefinitionen" in den anderen Ansichten vornehmen können. Für weitere Informationen siehe Thema Extrahierungsergebnisse verfeinern in Kapitel 9 auf S. 158.

Hinzufügen von Fachausdrücken

Im Bereich des Bibliotheksbaums werden die Bibliotheken angezeigt. Diese können erweitert werden, um die darin enthaltenen Typ-Wörterbücher anzuzeigen. Im mittleren Bereich zeigt eine Liste je nach Auswahl im Baum die in der ausgewählten Bibliothek oder im ausgewählten Typ-Wörterbuch enthaltenen Fachausdrücke an.

Wichtig: Fachausdrücke werden für japanische Ressourcen abweichend definiert. Für weitere Informationen siehe Thema Japanischer Bibliotheksbaum, Typen und Fachausdrucksbereich in Anhang A auf S. 385.*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Abbildung 17-3
Ausdrucksbereich


Ressourcen

Begriff	Zuordnung	Beugen	Typ	Bibliothek
a.g	Ende		Organization	Core Library (German)
a.g.	Ende		Organization	Core Library (German)
ag	Ende		Organization	Core Library (German)
co	Ende		Organization	Core Library (German)
co.	Ende		Organization	Core Library (German)
corp	Ende		Organization	Core Library (German)
corp.	Ende		Organization	Core Library (German)
gbh	Ende		Organization	Core Library (German)
gmbh	Ende		Organization	Core Library (German)
inc	Ende		Organization	Core Library (German)
inc.	Ende		Organization	Core Library (German)
kg	Ende		Organization	Core Library (German)
kgaa	Ende		Organization	Core Library (German)
l.l.c	Ende		Organization	Core Library (German)
l.l.c.	Ende		Organization	Core Library (German)
llc	Ende		Organization	Core Library (German)
lly	Ende		Organization	Core Library (German)
ltd	Ende		Organization	Core Library (German)
ltd.	Ende		Organization	Core Library (German)
mbh	Ende		Organization	Core Library (German)
plc	Ende		Organization	Core Library (German)
s.a	Ende		Organization	Core Library (German)
s.a.	Ende		Organization	Core Library (German)
s.c.a	Ende		Organization	Core Library (German)
s.c.a.	Ende		Organization	Core Library (German)
sa	Ende		Organization	Core Library (German)
sca	Ende		Organization	Core Library (German)
spss deutschland gmbh	Gesamter Begriff		Organization	Core Library (German)

Im Resource Editor können Sie Fachausdrücke zu einem Typ-Wörterbuch direkt im Fachausdrucksbereich oder über das Dialogfeld “Neue Fachausdrücke hinzufügen” hinzufügen. Bei den hinzugefügten Fachausdrücken kann es sich um einzelne Wörter oder um Wortfolgen handeln. Am Anfang der Liste befindet sich stets eine leere Zeile, in die Sie ein neues Wort eingeben können.

Hinweis: Diese Anweisungen zeigen Ihnen, wie Sie in der Resource Editor-Ansicht oder im Template Editor Änderungen vornehmen können. Beachten Sie, dass Sie solche Feinabstimmungen auch direkt im Bereich “Extrahierungsergebnisse”, im Bereich “Daten”, im Bereich “Kategorien” oder im Dialogfeld “Clusterdefinitionen” in den anderen Ansichten vornehmen können. Für weitere Informationen siehe Thema Extrahierungsergebnisse verfeinern in Kapitel 9 auf S. 158.

Ausdrucksspalte

Geben Sie in diese Spalte ein Wort oder eine Wortfolge in die Zelle ein. In welcher Farbe der Fachausdruck angezeigt wird, hängt von der Farbe des Typs ab, in dem der Ausdruck gespeichert wurde. Die Typfarben können Sie im Dialogfeld “Typeigenschaften” ändern. Für weitere Informationen siehe Thema Erstellen von Typen auf S. 314.

Erzwingen-Spalte

Wenn Sie in dieser Spalte in dieser Zelle ein Reißzweckensymbol hinzufügen, weisen Sie die Extrahierungsengine an, das Vorkommen desselben Ausdrucks in allen anderen Bibliotheken zu ignorieren. Für weitere Informationen siehe Thema Erzwingen von Fachausdrücken auf S. 320.

Übereinstimmungsspalte

Wählen Sie in dieser Spalte eine Übereinstimmungsoption aus, um festzulegen, wie die Extrahierungsengine die Übereinstimmung dieses Fachausdrucks mit den Textdaten ermittelt. Beispiele finden Sie in der Tabelle. Sie können den Standardwert ändern, indem Sie die Typeigenschaften bearbeiten. Für weitere Informationen siehe Thema Erstellen von Typen auf S. 314. Wählen Sie im Menü Bearbeiten > Abgleich ändern. Folgende Beispiele sind grundlegende Abgleiche, da Kombinationen aus ihnen ebenfalls möglich sind:

- Start. Dieser Typ wird zugewiesen, wenn das im Wörterbuch gefundene Konzept mit dem ersten Wort eines aus dem Text extrahierten Fachausdrucks übereinstimmt. Wenn Sie beispielsweise `Apfel` eingeben, gilt `Apfel vom Bioladen` als Übereinstimmung.
- Ende. Dieser Typ wird zugewiesen, wenn das im Wörterbuch gefundene Konzept mit dem letzten Wort eines aus dem Text extrahierten Fachausdrucks übereinstimmt. Wenn Sie beispielsweise `Apfel` eingeben, gilt `grüner Apfel` als Übereinstimmung.
- Beliebig. Dieser Typ wird zugewiesen, wenn das im Wörterbuch gefundene Konzept mit irgendeinem Wort eines aus dem Text extrahierten Fachausdrucks übereinstimmt. Wenn Sie z. B. `Apfel` eingeben, sorgt die Option Beliebig dafür, dass `Apfel vom Bioladen` , `grüner Apfel` und `grüner Apfel vom Bioladen` demselben Typ zugeordnet werden.
- Gesamter Begriff. Wenn das gesamte aus dem Text extrahierte Konzept exakt mit dem in der Bibliothek vorhandenen Ausdruck übereinstimmt, wird dieser Typ zugewiesen. Das Hinzufügen eines Fachausdrucks als Gesamter Begriff, Start und Ende, Gesamt und Ende, Gesamt und Beliebig oder Gesamt (keine Zusammensetzungen) erzwingt die Extrahierung eines Fachausdrucks.

  Da der Typ `<Person>` außerdem nur zweiteilige Namen wie *edith piaf* oder *mohandas gandhi* extrahiert, sollten Sie gegebenenfalls die Vornamen in dieses Typ-Wörterbuch aufnehmen, wenn Sie versuchen, einen Vornamen zu extrahieren, wenn kein Nachname angegeben ist. Wenn Sie beispielsweise alle Instanzen von *edith* als Name erfassen möchten, sollten Sie `edith` dem Typ `<Person>` über Gesamter Begriff oder Gesamt und Start hinzufügen.
- Gesamt (keine Zusammensetzungen). Wenn das gesamte aus dem Text extrahierte Konzept exakt mit dem Fachausdruck im Wörterbuch übereinstimmt, wird dieser Typ zugewiesen und die Extrahierung wird beendet, um die Extrahierung aufgrund einer Übereinstimmung des Fachausdrucks mit einem längeren zusammengesetzten Ausdruck zu verhindern. Wenn Sie beispielsweise `Apfel` eingeben, wird der Begriff durch die Option Gesamt (keine Zusammensetzung) als Übereinstimmung für `Apfel` gehandhabt, nicht jedoch für den zusammengesetzten Ausdruck `Grüner Apfel`, sofern dies nicht an anderer Stelle erzwungen wird.

Die folgende Tabelle geht von einem Fachausdruck `Apfel` in einem Typ-Wörterbuch aus. Je nach Abgleichsoption zeigt diese Tabelle an, welche Konzepte extrahiert und mit Typen versehen würden, wenn sie im Text gefunden würden.

Tabelle 17-1
Abgleichsbeispiele

Abgleichsoptionen für den Fachausdruck:  **`Apfel`**	**Extrahierte Konzepte**			
	`Apfel`	***`Apfel` vom Bioladen***	***grüner* `Apfel`**	***grüner `Apfel` vom Bioladen***
Gesamter Begriff	✔			
Start		✔		
Ende			✔	
Start oder Ende		✔	✔	
Gesamt und Start	✔	✔		
Gesamt und Ende	✔		✔	
Gesamt und (Start oder Ende)	✔	✔	✔	
Beliebig		✔	✔	✔
Gesamt und Beliebig	✔	✔	✔	✔
Gesamt (keine Zusammensetzungen)	✔	*nie extrahiert*	*nie extrahiert*	*nie extrahiert*

Flektierungsspalte

Wählen Sie in dieser Spalte aus, ob die Extrahierungsengine während der Extrahierung gebeute Formen dieses Fachausdrucks erzeugen soll, so dass sie alle zusammengefasst werden. Der Standardwert für diese Spalte ist in den Typeigenschaften definiert. Sie können diese Option direkt in der Spalte, aber für jeden einzelnen Fall ändern. Wählen Sie im Menü Bearbeiten > Beugung ändern.

Typspalte

Wählen Sie in dieser Spalte ein Typ-Wörterbuch aus der Dropdown-Liste aus. Die Liste der Typen wird anhand der von Ihnen im Bibliotheksbaum getroffenen Auswahl gefiltert. Als erster Typ wird in der Liste immer der im Bibliotheksbaum ausgewählte Standardtyp angezeigt. Wählen Sie im Menü Bearbeiten > Typ ändern.

Bibliotheksspalte

In dieser Spalte wird die Bibliothek angezeigt, in der Ihr Fachausdruck gespeichert wird. Sie können einen Begriff mit der Maus auf einen anderen Typ im Bibliotheksbaum ziehen, um seine Bibliothek zu ändern.

So fügen Sie einen einzelnen Fachausdruck zu einem Typ-Wörterbuch hinzu:

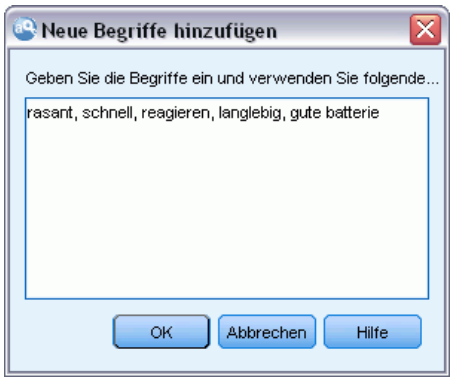
- Wählen Sie im Bibliotheksbaum das Typ-Wörterbuch aus, zu dem Sie den Fachausdruck hinzufügen möchten.

- Geben Sie in der mittleren Liste der Fachausdrücke Ihren Fachausdruck in die erste verfügbare leere Zelle ein und wählen Sie die gewünschten Optionen für diesen Fachausdruck aus.

So fügen Sie mehrere Fachausdrücke zu einem Typ-Wörterbuch hinzu:

- Wählen Sie im Bibliotheksbaumbereich das Typ-Wörterbuch aus, dem Sie die Fachausdrücke hinzufügen möchten.
- Wählen Sie im Menü Extras > Neue Begriffe aus. Das Dialogfeld "Neue Fachausdrücke hinzufügen" wird geöffnet.

Abbildung 17-4
Dialogfeld "Neue Fachausdrücke hinzufügen"



- Geben Sie die Begriffe ein, die Sie zum ausgewählten Typ-Wörterbuch hinzufügen möchten. Sie können die Begriffe eingeben oder mehrere Begriffe kopieren und einfügen. Wenn Sie mehrere Fachausdrücke eingeben, müssen Sie diese mit dem Trennzeichen, das im Dialogfeld "Optionen" definiert ist, trennen oder jeden Begriff in einer neuen Zeile eingeben. Für weitere Informationen siehe Thema Festlegen von Optionen in Kapitel 8 auf S. 135.
- Klicken Sie auf OK, um die Begriffe zum Wörterbuch hinzuzufügen. Als Abgleichsoption wird automatisch die Standardoption für dieses Typ-Wörterbuch festgelegt. Das Dialogfeld wird geschlossen und die neuen Begriffe erscheinen im Wörterbuch.

Erzwingen von Fachausdrücken

Wenn Sie möchten, dass ein Fachausdruck einem bestimmten Typ zugewiesen wird, können Sie ihn zum entsprechenden Typ-Wörterbuch hinzufügen. Wenn jedoch mehrere Fachausdrücke mit demselben Namen vorliegen, muss die Extrahierungsengine wissen, welcher Typ verwendet werden soll. Sie werden daher aufgefordert, den zu verwendenden Typ auszuwählen. Dies wird als **Erzwingen** der Zuordnung eines Fachausdrucks zu einem Typ bezeichnet. Diese Option ist vor allem beim Überschreiben der Typenzuordnung aus einem kompilierten (internen, nicht bearbeitbaren) Wörterbuch hilfreich. Generell empfehlen wir, doppelte Fachbegriffe von vornherein zu vermeiden.

Das Erzwingen *entfernt* die anderen Stellen, an denen der Fachausdruck vorkommt, nicht. Diese werden stattdessen von der Extrahierungsengine ignoriert. Sie können später ändern, welche Fundstelle verwendet wird, indem Sie das Erzwingen eines Fachausdrucks festlegen oder aufheben. Es kann außerdem erforderlich sein, dass Sie erzwingen, dass ein Fachausdruck

in ein Typ-Wörterbuch eingefügt wird, wenn Sie eine öffentliche Bibliothek hinzufügen oder aktualisieren.

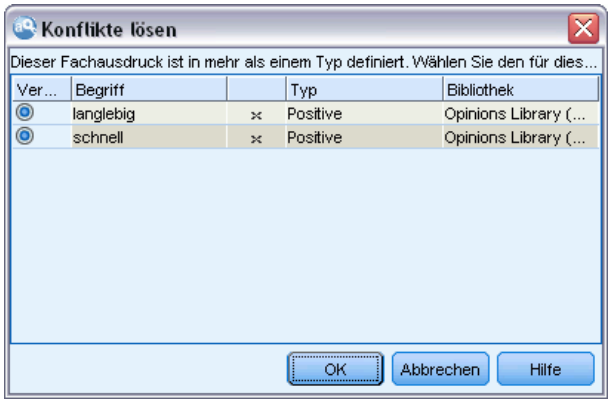
Abbildung 17-5
Symbole für den Erzwingungsstatus

Begriff		Zuordnung	Beugen	Typ	Bibliothek
$ $		Gesamter Begriff	☐	Budget	Budget Library (English)
amount of money	📌	Gesamter Begriff	☐	Budget	Budget Library (English)
arepriced		Gesamter Begriff	☐	Budget	Budget Library (English)
award	✕	Gesamt und (Start oder Ende)	☑	Budget	Budget Library (English)
balance		Gesamt und (Start oder Ende)	☑	Budget	Budget Library (English)
bank charge	✕	Gesamter Begriff	☑	Budget	Budget Library (English)
bill		Gesamt und Beliebige	☑	Budget	Budget Library (English)

In der Erzwingungsspalte (2. Spalte im Fachausdrucksbereich) sehen Sie, welche Ausdrücke erzwungen oder ignoriert werden. Wenn ein Reißzweckensymbol angezeigt wird, bedeutet dies, dass dieses Vorkommen des Ausdrucks erzwungen wurde. Wenn ein schwarzes X-Symbol angezeigt wird, bedeutet dies, dass dieses Vorkommen des Ausdrucks während des Extrahierens ignoriert wird, da er anderweitig erzwungen wurde. Wenn Sie einen Fachausdruck erzwingen, wird er außerdem in der Farbe des Typs angezeigt, für den er erzwungen wurde. Dies bedeutet, dass ein Ausdruck, der sowohl in `Typ 1` als auch in `Typ 2` vorkommt und den Sie als `Typ 1` erzwungen haben, im Fenster immer in der für `Typ 1` definierten Schriftfarbe angezeigt wird.

Um den Status zu ändern, können Sie auf das Symbol doppelklicken. Wenn der Fachausdruck an einer anderen Stelle vorkommt, wird das Dialogfeld “Konflikte auflösen” geöffnet, in dem Sie auswählen können, welche Fundstelle verwendet wird.

Abbildung 17-6
Dialogfeld “Konflikte auflösen”



Umbenennen von Typen

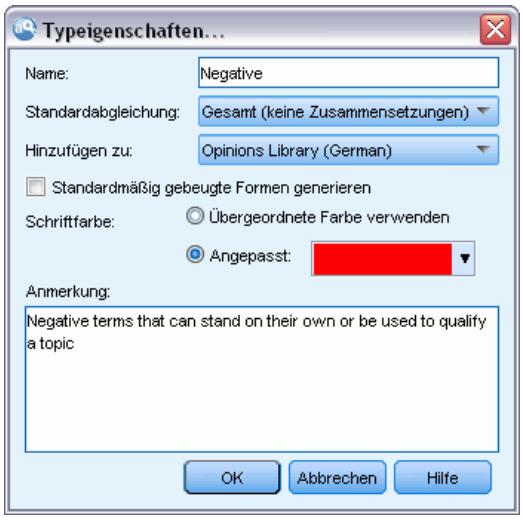
Sie können ein Typ-Wörterbuch umbenennen oder andere Wörterbucheinstellungen ändern, indem Sie die Typeigenschaften bearbeiten.

Wichtig: Es wird empfohlen, keine Leerzeichen in Typnamen zu verwenden, insbesondere wenn zwei oder mehr Typnamen mit demselben Wort beginnen. Es ist außerdem empfehlenswert, die Typen in der Core Library und Opinions Library nicht umzubenennen oder ihre Standard-Übereinstimmungsattribute zu ändern.

So benennen Sie einen Typ um:

- Wählen Sie im Bereich des Bibliotheksbaums das Typ-Wörterbuch aus, das Sie umbenennen möchten.
- Klicken Sie mit der rechten Maustaste und wählen Sie im Kontextmenü Typeigenschaften aus. Das Dialogfeld "Typ-Eigenschaften" wird geöffnet.

Abbildung 17-7
Dialogfeld "Typ-Eigenschaften"



- Geben Sie den neuen Namen für Ihr Typ-Wörterbuch in das Textfeld "Name" ein.
- Klicken Sie auf OK, um den neuen Namen zu übernehmen. Der neue Name wird im Bibliotheksbaum angezeigt.

Verschieben von Typen

Sie können ein Typ-Wörterbuch mit der Maus an eine andere Stelle innerhalb einer Bibliothek oder in eine andere Bibliothek im Baum ziehen.

So ordnen Sie einen Typ in einer Bibliothek neu an:

- Wählen Sie im Bereich des Bibliotheksbaums das Typ-Wörterbuch aus, das Sie verschieben möchten.
- Wählen Sie im Menü Bearbeiten > Nach oben verschieben aus, um das Typ-Wörterbuch im Bibliotheksbaum um eine Position nach oben zu verschieben, bzw. Bearbeiten > Nach unten verschieben, um es um eine Position nach unten zu verschieben.

So verschieben Sie einen Typ in eine andere Bibliothek:

- Wählen Sie im Bereich des Bibliotheksbaums das Typ-Wörterbuch aus, das Sie verschieben möchten.

- ▶ Klicken Sie mit der rechten Maustaste und wählen Sie im Kontextmenü Typeigenschaften aus. Das Dialogfeld "Typ-Eigenschaften" wird geöffnet. (Sie können den Typ auch mit der Maus in eine andere Bibliothek ziehen.)
- ▶ Wählen Sie im Feld "Hinzufügen zu" die Bibliothek aus, in die Sie das Typ-Wörterbuch verschieben möchten.
- ▶ Klicken Sie auf OK. Das Dialogfeld wird geschlossen und der Typ befindet sich in der von Ihnen ausgewählten Bibliothek.

Deaktivieren und Löschen von Typen

Wenn Sie ein Typ-Wörterbuch vorübergehend entfernen möchten, können Sie im Bibliotheksbaum das Kontrollkästchen links neben dem Namen des Wörterbuchs deaktivieren. Dies bewirkt, dass das Wörterbuch in Ihrer Bibliothek bleibt, dass die Inhalte bei der Prüfung von Konflikten und während des Extrahierungsvorgangs aber ignoriert werden.

Sie können Typ-Wörterbücher auch dauerhaft aus einer Bibliothek entfernen.

So deaktivieren Sie ein Typ-Wörterbuch:

- ▶ Wählen Sie im Bereich des Bibliotheksbaums das Typ-Wörterbuch aus, das Sie deaktivieren möchten.
- ▶ Drücken Sie die Leertaste. Das Kontrollkästchen links neben dem Typnamen wird deaktiviert.

So löschen Sie ein Typ-Wörterbuch:

- ▶ Wählen Sie im Bereich des Bibliotheksbaums das Typ-Wörterbuch aus, das Sie löschen möchten.
- ▶ Wählen Sie im Menü Bearbeiten > Löschen aus, um das Typ-Wörterbuch zu löschen.

Substitutions-/Synonymwörterbücher

Ein **Substitutionswörterbuch** ist eine Sammlung von Begriffen, mit deren Hilfe ähnliche Begriffe unter einem Zielbegriff gruppiert werden. Substitutionswörterbücher werden im unteren Bereich der Registerkarte "Bibliotheksressourcen" verwaltet. Sie können diese Ansicht mit der Optionsfolge Ansicht > Ressourceneditor in den Menüs aufrufen, wenn Sie sich in einer interaktiven Workbench-Sitzung befinden. Andernfalls können Sie Wörterbücher für eine bestimmte Vorlage im Template Editor bearbeiten.

In diesen Wörterbüchern können Sie zwei Formen von Substitutionen definieren: **Synonyme** und **optionale Elemente**. Sie können auf die Registerkarten in diesem Bereich klicken, um zwischen ihnen zu wechseln.

Nachdem Sie eine Extrahierung Ihrer Textdaten durchgeführt haben, finden Sie möglicherweise mehrere Konzepte, bei denen es sich um Synonyme oder um gebeugte Formen anderer Konzepte handelt. Indem optionale Elemente und Synonyme identifiziert werden, können Sie erzwingen, dass die Extrahierungsengine diese auf einen einzigen Zielbegriff abbildet.

Die Substituierung der Verwendung von Synonymen und optionalen Elementen verringert die Anzahl an Konzepten im Bereich “Extrahierungsergebnisse”, indem sie in sinnvollere, aussagekräftigere Konzepte mit einer höheren Häufigkeit von Dokument-Anzahlen zusammengefasst werden.

Anmerkung: Für japanische Ressourcen gelten optionale Elemente nicht und sind daher nicht verfügbar. Zusätzlich werden Synonyme für japanischen Text etwas abweichend behandelt. Für weitere Informationen siehe Thema Verwenden des Synonymwörterbuchs für japanischen Text in Anhang A auf S. 392.*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Synonyme

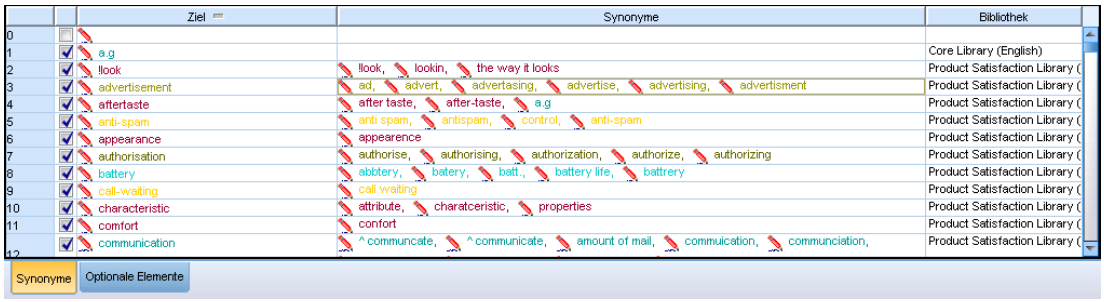
Synonyme verknüpfen zwei oder mehr Wörter mit derselben Bedeutung. Mithilfe von Synonymen können Sie außerdem Fachausdrücke mit ihren Abkürzungen gruppieren oder auch falsch geschriebene Wörter mit dem Fachausdruck in der richtigen Schreibweise. Sie können diese Synonyme auf der Registerkarte “Synonyme” definieren.

Eine Synonymdefinition setzt sich aus zwei Teilen zusammen. Der erste Teil ist ein Zielausdruck, d. h. der Ausdruck, unter dem die Extrahierungsengine alle Synonyme zusammenfassen soll. Wenn dieser Zielausdruck nicht als Synonym eines anderen Zielausdrucks verwendet wird oder er ausgeschlossen wird, ist es wahrscheinlich, dass dieser Ausdruck das Konzept wird, das im Bereich “Extrahierungsergebnisse” angezeigt wird. Der zweite Teil besteht aus einer Liste von Synonymen, die unter dem Zielbegriff zusammengefasst werden.

Wenn Sie beispielsweise `Automobil` durch `Fahrzeug` ersetzen möchten, dann ist `Automobil` das Synonym und `Fahrzeug` der Zielausdruck.

Sie können jedes beliebige Wort in die Spalte Synonym setzen, wenn das Wort jedoch bei der Extrahierung nicht gefunden wird und der Fachausdruck eine Ableichsoption mit `Gesamt` hatte, kann keine Substitution vorgenommen werden. Der Zielausdruck muss jedoch nicht extrahiert werden, damit die Synonyme unter diesem Begriff zusammengefasst werden.

Abbildung 17-8
Substitutionswörterbuch, Registerkarte “Synonyme”



Optionale Elemente

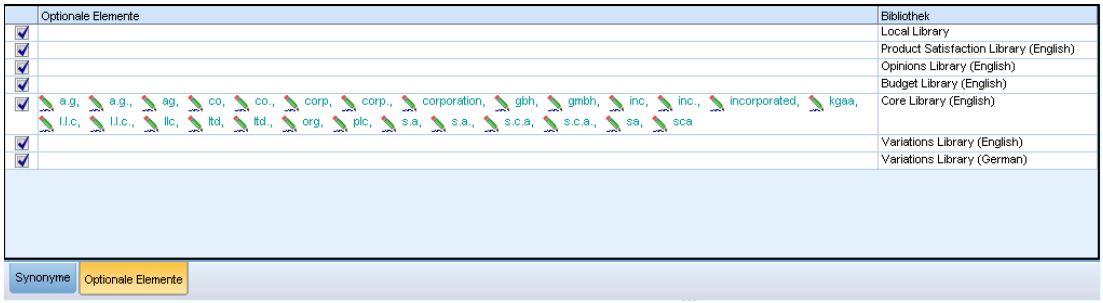
Optionale Elemente kennzeichnen optionale Wörter in einem zusammengesetzten Begriff, die während der Extrahierung ignoriert werden können, um ähnliche Begriffe auch dann zusammenzuhalten, wenn sie im Text leicht unterschiedlich auftreten. Optionale Elemente sind einzelne Wörter, die, wenn sie aus einem zusammengesetzten Ausdruck entfernt werden, eine

Übereinstimmung mit einem anderen Ausdruck darstellen können. Diese einzelnen Wörter können an einer beliebigen Stelle innerhalb des zusammengesetzten Ausdrucks vorkommen – am Anfang, in der Mitte oder am Ende. Sie können optionale Elemente auf der Registerkarte "Optional" definieren.

Um z. B. die Ausdrücke `ibm` und `ibm corp` zu gruppieren, müssen Sie festlegen, dass `corp` in diesem Fall als optionales Element behandelt werden soll. Wenn Sie andererseits festlegen, dass `access` ein optionales Element ist und wenn bei der Extrahierung sowohl `internet access speed` als auch `internet speed` gefunden werden, dann werden diese unter dem Ausdruck gruppiert, der am häufigsten vorkommt.

Anmerkung: Für japanische Textressourcen steht die Registerkarte "Optionale Elemente" nicht zur Verfügung, da optionale Elemente nicht gelten. Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Abbildung 17-9
Substitutionswörterbuch, Registerkarte "Optional"



Definieren von Synonymen

Auf der Registerkarte "Synonyme" können Sie in die Leerzeile am Anfang der Tabelle eine Synonymdefinition eingeben. Beginnen Sie, indem Sie den Zielausdruck und seine Synonyme definieren. Sie können auch die Bibliothek auswählen, in der diese Definition gespeichert werden soll. Während der Extrahierung werden alle Fundstellen der Synonyme für die endgültige Extrahierung unter dem Zielausdruck gruppiert. Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken auf S. 316.

Wenn Ihre Textdaten viele Fachausdrücke aus der Telekommunikation enthalten, liegen beispielsweise folgende Fachausdrücke vor: `cellular phone`, `wireless phone` und `mobile phone`. In diesem Beispiel sollten Sie `cellular` und `mobile` als Synonyme für `wireless` definieren. Wenn Sie diese Synonyme definieren, werden alle extrahierten Fundstellen von `cellular phone` und `mobile phone` als derselbe Ausdruck wie `wireless phone` behandelt und zusammen in der Liste der Fachausdrücke ausgegeben.

Wenn Sie Ihre Typ-Wörterbücher erstellen, können Sie einen Ausdruck eingeben, für den Ihnen drei oder vier Synonyme einfallen. In diesem Fall könnten Sie alle Ausdrücke und anschließend Ihren Zielausdruck in das Substitutionswörterbuch eingeben und dann die Synonyme durch Ziehen übertragen.

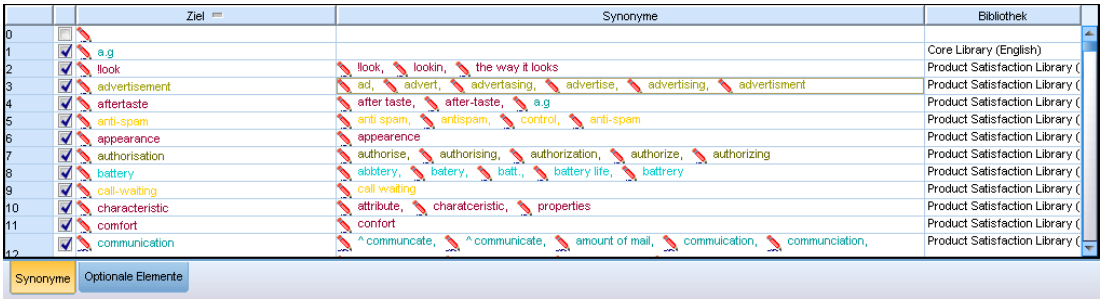
Anmerkung: Synonyme werden für japanischen Text etwas abweichend behandelt. Für weitere Informationen siehe Thema Verwenden des Synonymwörterbuchs für japanischen Text in Anhang A auf S. 392.*Anmerkung*: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Die Substitution von Synonymen wird auch auf gebeugte Formen (wie Pluralformen) der Synonyme angewendet. Je nach Kontext sollten Sie Einschränkungen für die Substitution der Fachausdrücke festlegen. Sie können bestimmte Zeichen verwenden, um Einschränkungen dafür festzulegen, wie weit die Substitution durchgeführt wird:

- **Ausrufezeichen (!).** Wenn das Ausrufezeichen direkt vor dem Synonym steht `!Synonym`, bedeutet das, dass keine gebeugten Formen des Synonyms durch den Zielausdruck ersetzt werden. Ein Ausrufezeichen direkt vor dem Zielausdruck (`!Zielausdruck`) bedeutet hingegen, dass kein weiterer Bestandteil dieses zusammengesetzten Fachausdrucks oder Varianten davon substituiert werden sollen.
- **Sternchen (*).** Ein direkt nach dem Synonym stehendes Sternchen (*) (`Synonym*`) bedeutet, dass dieses Wort durch den Zielausdruck ersetzt werden soll. Wenn Sie beispielsweise den Quellausdruck `manage*` als Synonym und `management` als Zielausdruck definiert haben, dann wird `associate managers` durch den Zielausdruck `associate management` ersetzt. Sie können nach dem Wort auch ein Leerzeichen und ein Sternchen (*) einfügen (`Synonym *`), wie z. B. `internet *`. Wenn Sie den Zielausdruck `internet` sowie die Synonyme `internet * *` und `web *` definiert haben, dann werden `internet access card` und `web portal` entsprechend durch `internet` ersetzt. Ein Wort oder eine Zeichenfolge kann in diesem Wörterbuch nicht mit dem Sternchen (*) als Platzhalterzeichen beginnen.
- **Winkelzeichen (^).** Ein vor einem Synonym stehendes Winkelzeichen und ein Leerzeichen (`^ Synonym>`) bedeuten, dass die Synonymgruppierung nur dann durchgeführt wird, wenn der Fachausdruck mit dem Synonym beginnt. Wenn Sie beispielsweise `^ wage` als das Synonym und `income` als den Zielausdruck definiert haben und beide Ausdrücke extrahiert werden, dann werden beide unter dem Ausdruck `income` gruppiert. Werden dagegen `minimum wage` und `income` extrahiert, erfolgt keine Gruppierung, weil `minimum wage` nicht mit `wage` beginnt. Zwischen diesem Symbol und dem Synonym muss ein Leerzeichen eingefügt werden.
- **Dollarzeichen ($).** Ein nach einem Synonym stehendes Leerzeichen und ein Dollarzeichen (`Synonym $`) bedeuten, dass die Synonymgruppierung nur dann durchgeführt wird, wenn der Fachausdruck mit dem Synonym endet. Wenn Sie beispielsweise `cash $` als das Synonym und `money` als den Zielausdruck definiert haben und beide Ausdrücke extrahiert werden, dann werden beide unter dem Ausdruck `money` gruppiert. Werden dagegen `cash cow` und `money` extrahiert, erfolgt keine Gruppierung, weil `cash cow` nicht auf `cash` endet. Zwischen diesem Symbol und dem Synonym muss ein Leerzeichen eingefügt werden.
- **Winkelzeichen (^) und Dollarzeichen ($).** Wenn Winkel- und Dollarzeichen zusammen verwendet werden
(`^ Synonym $`), stimmt ein Ausdruck nur bei exakter Übereinstimmung mit dem Synonym überein. Dies bedeutet, dass im extrahierten Ausdruck vor oder nach dem Synonym keine Wörter stehen dürfen, damit die Synonymgruppierung stattfindet. Beispiel: Sie definieren `^ van $` als Synonym und `truck` als Zielausdruck, sodass nur `van` mit `truck` gruppiert wird, während `marie van guerin` unverändert bleibt. Wenn Sie ein Synonym mit dem

Winkel- und dem Dollarzeichen definieren und wenn das entsprechende Wort an einer Stelle des Quelltexts auftritt, wird das Synonym außerdem automatisch extrahiert.

Abbildung 17-10
Substitutionswörterbuch, Registerkarte "Synonyme" mit Beispiel



Anmerkung: Diese Sonderzeichen und Platzhalter werden für japanischen Text nicht unterstützt. Für weitere Informationen siehe Thema Verwenden des Synonymwörterbuchs für japanischen Text in Anhang A auf S. 392.*Anmerkung*: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

So fügen Sie einen Synonymeintrag hinzu:

- Klicken Sie im Substitutionsbereich oben links auf die Registerkarte Synonyme.
- Geben Sie in der leeren Zeile am Anfang der Tabelle Ihren Zielausdruck in die Zielspalte ein. Der von Ihnen eingegebene Zielausdruck wird farbig angezeigt. Die Farbe stellt den Typ dar, als der der Ausdruck erscheint oder in den er erzwungen wird, sofern dies der Fall ist. Wenn der Ausdruck schwarz dargestellt wird, bedeutet dies, dass er in keinem der Typ-Wörterbücher vorkommt.
- Klicken Sie auf die zweite Zelle rechts neben dem Zielausdruck und geben Sie die Synonyme ein. Trennen Sie die einzelnen Einträge mithilfe des globalen Trennzeichens, das im Dialogfeld "Optionen" definiert ist. Für weitere Informationen siehe Thema Festlegen von Optionen in Kapitel 8 auf S. 135. Die von Ihnen eingegebenen Ausdrücke werden farbig angezeigt. Die Farbe stellt den Typ dar, in der der Ausdruck angezeigt wird. Wenn der Ausdruck schwarz dargestellt wird, bedeutet dies, dass er in keinem der Typ-Wörterbücher vorkommt.
- Klicken Sie auf die letzte Zelle, um die Bibliothek auszuwählen, in der die Synonymdefinition gespeichert werden soll.

Hinweis: Diese Anweisungen zeigen Ihnen, wie Sie in der Resource Editor-Ansicht oder im Template Editor Änderungen vornehmen können. Beachten Sie, dass Sie solche Feinabstimmungen auch direkt im Bereich "Extrahierungsergebnisse", im Bereich "Daten", im Bereich "Kategorien" oder im Dialogfeld "Clusterdefinitionen" in den anderen Ansichten vornehmen können. Für weitere Informationen siehe Thema Extrahierungsergebnisse verfeinern in Kapitel 9 auf S. 158.

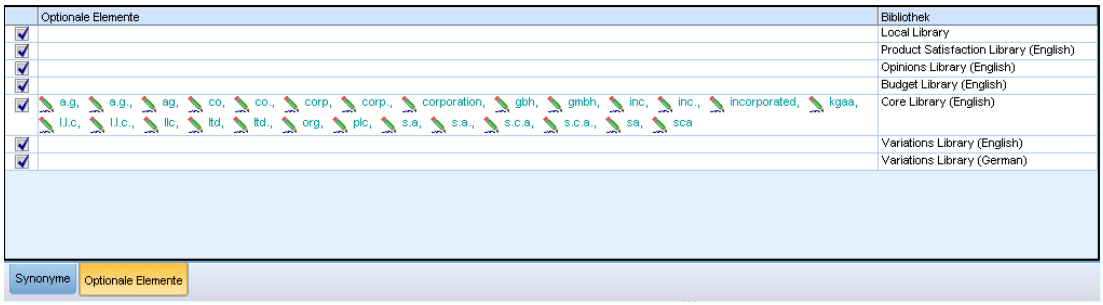
Definieren optionaler Elemente

Auf der Registerkarte “Optional” können Sie für beliebige Bibliotheken optionale Elemente definieren. Diese Einträge werden für alle Bibliotheken miteinander gruppiert. Sobald eine Bibliothek zum Bibliotheksbaum hinzugefügt wird, wird auf der Registerkarte “Optional” eine leere Zeile für optionale Elemente hinzugefügt.

Alle Einträge werden automatisch in Kleinschrift umgewandelt. Die Extrahierungsengine ordnet Wörter, die im Text in Groß- oder Kleinbuchstaben vorkommen, zu.

Anmerkung: Für japanische Ressourcen gelten optionale Elemente nicht und sind daher nicht verfügbar.

Abbildung 17-11
Substitutionswörterbuch, Registerkarte “Optional”



Anmerkung: Ausdrücke werden mithilfe des Trennzeichens getrennt, das im Dialogfeld “Optionen” definiert ist. Für weitere Informationen siehe Thema Festlegen von Optionen in Kapitel 8 auf S. 135. Wenn das von Ihnen eingegebene optionale Element das Trennzeichen als Teil des Ausdrucks enthält, müssen Sie diesem einen umgekehrten Schrägstrich (\) voranstellen.

So fügen Sie einen Eintrag hinzu:

- Klicken Sie im Editor im Substitutionsbereich oben links auf die Registerkarte “Optional”.
- Klicken Sie in der Spalte “Optionale Elemente” auf die Zelle der Bibliothek, zu der Sie diesen Eintrag hinzufügen möchten.
- Geben Sie das optionale Element ein. Trennen Sie die einzelnen Einträge mithilfe des globalen Trennzeichens, das im Dialogfeld “Optionen” definiert ist. Für weitere Informationen siehe Thema Festlegen von Optionen in Kapitel 8 auf S. 135.

Deaktivieren und Löschen von Substitutionen

Sie können einen Eintrag vorübergehend entfernen, indem Sie ihn in Ihrem Wörterbuch deaktivieren. Deaktivierte Einträge werden bei der Extrahierung ignoriert.

Sie können zudem veraltete Einträge aus Ihrem Substitutionswörterbuch löschen.

So deaktivieren Sie einen Eintrag:

- Wählen Sie in Ihrem Wörterbuch den Eintrag aus, den Sie deaktivieren möchten.

- Drücken Sie die Leertaste. Das Kontrollkästchen links neben dem Eintrag wird deaktiviert.

  Anmerkung: Sie können auch links neben dem Eintrag auf das Kontrollkästchen klicken, um es zu deaktivieren.

So löschen Sie einen Synonymeintrag:

- Wählen Sie in Ihrem Wörterbuch den Eintrag aus, den Sie löschen möchten.
- Wählen Sie in den Menüs die Optionen Bearbeiten > Löschen oder drücken Sie die Taste Entfernen auf Ihrer Tastatur. Der Eintrag wird aus dem Wörterbuch entfernt.

So löschen Sie einen Eintrag für ein optionales Element:

- Doppelklicken Sie in Ihrem Wörterbuch auf den Eintrag, den Sie löschen möchten.
- Löschen Sie diesen Ausdruck manuell.
- Drücken Sie die Eingabetaste, damit die Änderung wirksam wird.

Ausschlusswörterbücher

Ein **Ausschlusswörterbuch** ist eine Liste von Wörtern, Ausdrücken oder Teilzeichenfolgen. Ausdrücke, die mit einem Eintrag im Ausschlusswörterbuch übereinstimmen, werden ignoriert oder aus der Extrahierung ausgeschlossen. Ausschlusswörterbücher werden im rechten Bereich des Editors verwaltet. In der Regel handelt es sich bei den zu dieser Liste hinzugefügten Ausdrücken um Füllwörter oder Wortfolgen, die im Text verwendet werden, aber keine wichtigen Informationen enthalten und daher die Extrahierungsergebnisse stören können. Wenn Sie solche Wörter zum Ausschlusswörterbuch hinzufügen, stellen Sie sicher, dass sie nie extrahiert werden.

Ausschlusswörterbücher werden im oberen rechten Bereich der Registerkarte "Bibliotheksressourcen" im Editor verwaltet. Sie können diese Ansicht mit der Optionsfolge Ansicht > Ressourceneditor in den Menüs aufrufen, wenn Sie sich in einer interaktiven Workbench-Sitzung befinden. Andernfalls können Sie Wörterbücher für eine bestimmte Vorlage im Template Editor bearbeiten.

Abbildung 17-12
Ausschlusswörterbuchbereich

		Ausschlussliste	Bibliothek
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	☑	i was out of	Opinions Library (English)
7	☑	if i ever have a problem	Opinions Library (English)
8	☑	if i ever have problems	Opinions Library (English)
9	☑	if i have a problem	Opinions Library (English)
10	☑	if i have questions	Opinions Library (English)
11	☑	if there are problems	Opinions Library (English)
12	☑	if there is a problem	Opinions Library (English)

In das Ausschlusswörterbuch können Wörter, Wortfolgen oder Teile von Zeichenfolgen in die Leerzeile am Anfang der Tabelle eingegeben werden. Sie können Zeichenfolgen zu Ihrem Ausschlusswörterbuch als eines oder mehrere Wörter hinzufügen oder auch als Wortteile, indem Sie das Sternchen als Platzhalter verwenden. Die im Ausschlusswörterbuch erfassten Einträge werden verwendet, um Konzepte von der Extrahierung auszuschließen. Falls ein Eintrag auch an anderer Stelle deklariert ist, beispielsweise in einem Typ-Wörterbuch, wird er in den anderen Wörterbüchern durchgestrichen dargestellt, was anzeigt, dass er zur Zeit ausgeschlossen ist. Diese Zeichenfolge muss nicht in den Textdaten vorkommen oder als Teil eines Typ-Wörterbuchs deklariert sein, um angewendet zu werden.

Anmerkung: Wenn Sie ein Konzept zum Ausschlusswörterbuch hinzufügen, das in einem Synonymeintrag als Zielausdruck verwendet wird, dann werden auch das Ziel und alle Synonyme ausgeschlossen. Für weitere Informationen siehe Thema Definieren von Synonymen auf S. 325.

Verwenden von Platzhaltern (*)

Für alle Sprachen außer Japanisch können Sie das Sternchen als Platzhalter verwenden, wenn der Ausschlusseintrag als Teilzeichenfolge behandelt werden soll. Alle Ausdrücke, die von der Extrahierungsengine aufgefunden werden und ein Wort enthalten, das mit einer im Ausschlusswörterbuch angegebenen Zeichenfolge beginnt oder endet, werden von der endgültigen Extrahierung ausgeschlossen. Es gibt jedoch zwei Fälle, in denen ein Platzhalterzeichen nicht zulässig ist:

- Bindestrich (-), vor dem ein Sternchen (*) als Platzhalterzeichen steht (*-)
- Apostroph (‘), vor dem ein Sternchen (*) als Platzhalterzeichen steht (**’s*)

Tabelle 17-2
Beispiele für Ausschlusseinträge

Aufnahme (Cox Regression)	**Beispiel**	**Ergebnisse**
Wort	*next*	Es werden keine Konzepte (bzw. die darin enthaltenen Ausdrücke) extrahiert, die das Wort *next* enthalten.
Wortfolge	*for example*	Es werden keine Konzepte (bzw. die darin enthaltenen Ausdrücke) extrahiert, die den Ausdruck *for example* enthalten.
Teilweise	*copyright**	Schließt alle Konzepte (bzw. die darin enthaltenen Ausdrücke) aus, die mit Variationen des Worts *copyright* übereinstimmen bzw. diese enthalten, wie `copyrighted`, `copyrighting`, `copyrights` oder `copyright 2010`.
Teilweise	**ware*	Schließt alle Konzepte (bzw. die darin enthaltenen Ausdrücke) aus, die mit Variationen des Worts *ware* übereinstimmen bzw. diese enthalten, wie `freeware`, `shareware`, `software`, `hardware`, `beware` oder `silverware`.

So fügen Sie Einträge hinzu

▶ Geben Sie in die Leerzeilen am Anfang der Tabelle einen Ausdruck ein. Der von Ihnen eingegebene Ausdruck wird farbig angezeigt. Die Farbe stellt den Typ dar, in der der Ausdruck angezeigt wird. Wenn der Ausdruck schwarz dargestellt wird, bedeutet dies, dass er in keinem der Typ-Wörterbücher vorkommt.

So deaktivieren Sie Einträge

Sie können einen Eintrag vorübergehend entfernen, indem Sie ihn in Ihrem Ausschlusswörterbuch deaktivieren. Deaktivierte Einträge werden bei der Extrahierung ignoriert.

- Wählen Sie in Ihrem Ausschlusswörterbuch den Eintrag aus, den Sie deaktivieren möchten.
- Drücken Sie die Leertaste. Das Kontrollkästchen links neben dem Eintrag wird deaktiviert.

  Anmerkung: Sie können auch links neben dem Eintrag auf das Kontrollkästchen klicken, um es zu deaktivieren.

So löschen Sie Einträge

Sie können nicht mehr benötigte Einträge aus Ihrem Ausschlusswörterbuch löschen.

- Wählen Sie in Ihrem Ausschlusswörterbuch den Eintrag aus, den Sie löschen möchten.
- Wählen Sie im Menü Bearbeiten > Löschen. Der Eintrag wird aus dem Wörterbuch entfernt.

Kapitel
18

Informationen zu erweiterten Ressourcen

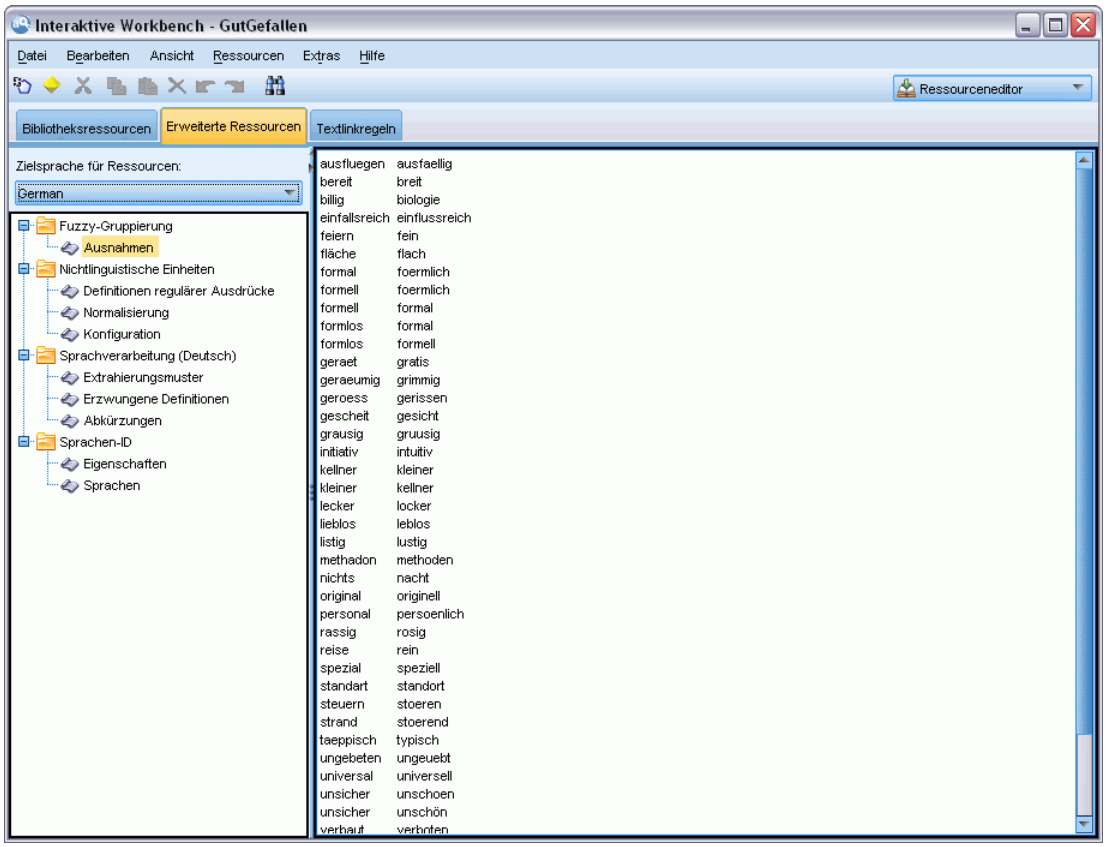
Ergänzend zu den Typ-, Ausschluss- und Substitutionswörterbüchern können Sie auch mit einer Vielzahl erweiterter Ressourceneinstellungen arbeiten, z. B. Einstellungen für unscharfe Gruppierung und nicht-linguistische Typdefinitionen. Sie können in der Registerkarte “Erweiterte Ressourcen” im Template Editor oder Resource Editor mit diesen Ressourcen arbeiten.

Wichtig: Diese Registerkarte ist nicht für Ressourcen verfügbar, die für japanischen Text eingestellt sind.

Wenn Sie die Registerkarte “Erweiterte Ressourcen” aufrufen, können Sie folgende Informationen bearbeiten:

- **Zielsprache für Ressourcen.** Dient zur Auswahl der Sprache, für die die Ressourcen erstellt und angepasst werden. Für weitere Informationen siehe Thema Zielsprache für Ressourcen auf S. 335.
- **Unscharfe Gruppierung (Ausnahmen).** Hiermit werden Wortpaare aus dem Algorithmus für die unscharfe Gruppierung ausgeschlossen (Rechtschreibfehlerkorrektur). Für weitere Informationen siehe Thema Unscharfe Gruppierung auf S. 336.
- **Nicht linguistische Entitäten.** Hiermit wird aktiviert bzw. deaktiviert, welche linguistischen Elemente extrahiert werden und welche regulären Ausdrücke und Normalisierungsregeln bei ihrer Extrahierung angewendet werden. Für weitere Informationen siehe Thema Nicht linguistische Elemente auf S. 337.
- **Sprachbehandlung.** Darüber wird festgelegt, wie Sätze strukturiert werden (Extrahierungsmuster und erzwungene Definitionen) und wie Abkürzungen für die ausgewählte Sprache verwendet werden. Für weitere Informationen siehe Thema Sprachbehandlung auf S. 343.
- **Language Identifier.** Hierüber wird der automatische Language Identifier konfiguriert, der aufgerufen wird, wenn die Sprache auf Alle festgelegt ist. Für weitere Informationen siehe Thema Language Identifier auf S. 345.

Abbildung 18-1
Text-Mining-Vorlageneditor - Registerkarte "Erweiterte Ressourcen"



Anmerkung: Mit der Symbolleiste zum Suchen und Ersetzen können Sie Informationen schnell auffinden oder in einem Abschnitt identische Änderungen durchführen. Für weitere Informationen siehe Thema Ersetzen auf S. 334.

So bearbeiten Sie erweiterte Ressourcen:

- Finden Sie den Ressourcenabschnitt, den Sie bearbeiten möchten, und wählen Sie ihn aus. Der Inhalt wird im rechten Fensterbereich angezeigt.
- Über die Menübefehle oder die Schaltflächen der Symbolleiste können Sie Inhalte ausschneiden, kopieren oder einfügen.
- Bearbeiten Sie die Dateien, die Sie ändern möchten, mithilfe der Formatierungsregeln in diesem Abschnitt. Ihre Änderungen werden direkt gespeichert. Mithilfe der in der Symbolleiste angezeigten Pfeile zum Rückgängigmachen bzw. Wiederholen können Sie Ihre Änderungen rückgängig machen.

Suchen

In manchen Fällen ist es erforderlich, Informationen in einem bestimmten Abschnitt schnell aufzufinden. Wenn Sie z. B. eine Text Link Analysis durchführen, haben Sie eventuell Hunderte von Makros und Musterdefinitionen. Mit der Suchfunktion können Sie eine spezifische Regel schnell auffinden. Für die Suche nach Informationen in einem Abschnitt können Sie die Symbolleiste "Suchen" verwenden.

Abbildung 18-2
Symbolleiste "Suchen"

So verwenden Sie die Suchfunktion:

- Suchen Sie den Ressourcenabschnitt, den Sie durchsuchen möchten, und wählen Sie ihn aus. Der Inhalt wird im rechten Bereich des Editors angezeigt.
- Wählen Sie im Menü Bearbeiten > Suchen. Oben rechts im Dialogfeld "Erweiterte Ressourcen bearbeiten" wird die Symbolleiste "Suchen" angezeigt.
- Geben Sie die Wortfolge, nach der Sie suchen möchten, in das Textfeld ein. Mit den Schaltflächen der Symbolleiste können Sie festlegen, ob zwischen Groß- und Kleinschreibung unterschieden wird, ob eine teilweise Übereinstimmung zulässig ist und in welche Richtung die Suche durchgeführt wird.
- Klicken Sie auf Suchen, um die Suche zu starten. Wenn eine Übereinstimmung gefunden wird, wird der Text im Fenster markiert.
- Klicken Sie erneut auf Suchen, um nach der nächsten Übereinstimmung zu suchen.

Anmerkung: Bei der Arbeit auf der Registerkarte "Textlinkregeln" steht die Option "Suchen" nur zur Verfügung, wenn der Quellcode angezeigt wird.

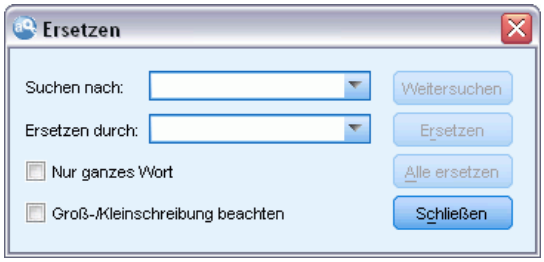
Ersetzen

Manchmal ist es erforderlich, erweiterte Ressourcen umfangreich zu aktualisieren. Mit der Funktion "Ersetzen" können Sie Ihren Inhalt einheitlich aktualisieren.

So verwenden Sie die Funktion "Ersetzen":

- Suchen Sie den Ressourcenabschnitt, in dem Sie suchen und ersetzen möchten, und wählen Sie ihn aus. Der Inhalt wird im rechten Bereich des Editors angezeigt.
- Wählen Sie im Menü Bearbeiten > Ersetzen. Das Dialogfeld "Ersetzen" wird geöffnet.

Abbildung 18-3
Dialogfeld "Ersetzen"



- Geben Sie in das Textfeld Suchen nach das Wort ein, nach dem Sie suchen möchten.
- Geben Sie in das Textfeld Ersetzen durch die Zeichenfolge ein, die Sie anstelle des gefundenen Texts verwenden möchten.
- Wählen Sie die Option Nur ganzes Wort aus, wenn Sie nur vollständige Wörter suchen und ersetzen möchten.
- Wählen Sie Groß-/Kleinschreibung beachten, wenn Sie nur Wörter suchen oder ersetzen möchten, deren Schreibweise exakt übereinstimmt.
- Klicken Sie auf Weitersuchen, um nach einer Übereinstimmung zu suchen. Wenn eine Übereinstimmung gefunden wird, wird der Text im Fenster markiert. Wenn Sie eine gefundene Übereinstimmung nicht ersetzen möchten, klicken Sie so oft erneut auf Weitersuchen, bis Sie einen Treffer erhalten, den Sie ersetzen möchten.
- Klicken Sie auf Ersetzen, um den ausgewählten Treffer zu ersetzen.
- Klicken Sie auf Ersetzen, um alle im Abschnitt gefundenen Treffer zu ersetzen. Anschließend wird eine Meldung mit der Anzahl der durchgeführten Ersetzungen angezeigt.
- Wenn Sie Ihre Ersetzungen durchgeführt haben, klicken Sie auf Schließen. Das Dialogfeld wird geschlossen.

  Anmerkung: Wenn Sie beim Ersetzen einen Fehler gemacht haben, können Sie diesen rückgängig machen, indem Sie das Dialogfeld schließen und im Menü Bearbeiten > Rückgängig auswählen. Dies müssen Sie für jede Änderung, die Sie rückgängig machen möchten, wiederholen.

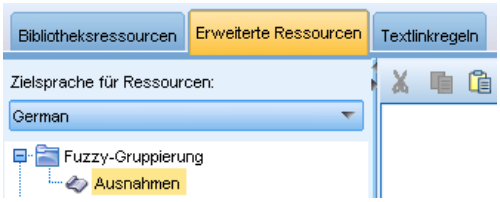
Zielsprache für Ressourcen

Ressourcen werden für eine bestimmte Textsprache erzeugt. Die Sprache, an die diese Ressourcen angepasst werden, wird auf der Registerkarte "Erweiterte Ressourcen" definiert. Sie können, falls gewünscht, eine andere Sprache wählen, indem Sie die jeweilige Sprache in der Kombo-Box Zielsprache für Ressourcen auswählen. Außerdem erscheint die hier aufgelistete Sprache als die Sprache für Textanalysepakete, die Sie mit diesen Ressourcen erstellen.

Wichtig: Sie müssen die Sprache in Ihren Ressourcen nur sehr selten ändern. Eine Änderung der Sprache kann zu Problemen führen, wenn Ihre Ressourcen nicht mehr mit der Extrahierungssprache übereinstimmen. Auch wenn die Sprache nur selten geändert wird, können Sie dies tun, wenn Sie die Sprachoption `ALLE` während der Extrahierung verwenden wollten,

weil Sie Text in mehr als einer Sprache erwarteten. Durch Änderung der Sprache können Sie beispielsweise auf die Sprachverwendungsressourcen für Extrahierungsmuster und Abkürzungen zugreifen und Definitionen für die gewünschte Sekundärsprache erzwingen. Beachten Sie jedoch, vor der Veröffentlichung oder Sicherung der vorgenommenen Ressourcenänderungen oder vor dem Ausführen einer weiteren Extrahierung die Sprache wieder auf die Primärsprache umzustellen, die Sie extrahieren möchten.

Abbildung 18-4
Einstellen der Zielsprache



Unscharfe Gruppierung

Wenn Sie im Text-Mining-Knoten unter Extrahierungseinstellungen Rechtschreibung korrigieren für eine minimale Anzahl an Stammzeichen von auswählen, haben Sie den Algorithmus für die unscharfe Gruppierung aktiviert.

Mit der unscharfen Gruppierung können Sie Wörter einfacher gruppieren, die häufig falsch oder ähnlich geschrieben werden, indem Sie vorübergehend alle Vokale (außer den ersten) und doppelt/dreifach auftretende Konsonanten aus den extrahierten Wörtern entfernen und anschließend einen Vergleich dieser Wörter durchführen. Die Funktion für die unscharfe Gruppierung wird während des Extrahierungsprozesses auf die extrahierten Ausdrücke angewendet und die Ergebnisse werden dann verglichen, um eventuelle Übereinstimmungen zu ermitteln. Wenn dies der Fall ist, werden die ursprünglichen Begriffe in der endgültigen Extrahierungsliste gruppiert. Die Gruppierung erfolgt unter dem Ausdruck, der am häufigsten in den Daten vorkommt.

Anmerkung: Wenn die zwei miteinander verglichenen Begriffe unterschiedlichen Typen zugeordnet werden (ausschließlich des Typs `<Unknown>`), wird die unscharfe Gruppierung auf dieses Paar nicht angewendet. Die Begriffe müssen also zum selben Typ oder zum Typ `<Unknown>` gehören, damit diese Methode angewendet wird.

Wenn Sie diese Funktion aktiviert haben und feststellen, dass zwei Wörter mit einer ähnlichen Schreibweise fälschlicherweise gruppiert wurden, können Sie diese Wörter aus der unscharfen Gruppierung ausschließen. Hierzu geben Sie die fälschlicherweise als übereinstimmend eingestuften Wortpaare auf der Registerkarte “Erweiterte Ressourcen” im Abschnitt “Ausnahmen” ein. Für weitere Informationen siehe Thema Informationen zu erweiterten Ressourcen auf S. 332.

Das folgende Beispiel veranschaulicht, wie eine unscharfe Gruppierung durchgeführt wird. Wenn die unscharfe Gruppierung aktiviert ist, werden die folgenden Wörter als identisch eingestuft:

```
color  -> colr                mountain -> montn
colour -> colr                montana  -> montn

modeling  -> modlng           furniture -> furntr
modelling -> modlng           furnature -> furntr
```

In diesem Beispiel würden Sie sicherlich verhindern wollen, das `mountain` und `montana` miteinander gruppiert werden. Hierzu können Sie diese wie folgt im Abschnitt “Ausnahmen” angeben:

```
mountain      montana
```

Wichtig: In manchen Fällen werden trotz definierter Ausnahmen für unscharfe Gruppierung zwei Wörter gepaart, weil bestimmte Synonymregeln angewendet werden. Versuchen Sie in diesem Fall, Synonyme mit einem Ausrufezeichen (`!`) als Platzhalter einzugeben, um zu verhindern, dass Wörter in der Ausgabe als synonym betrachtet werden. Für weitere Informationen siehe Thema Definieren von Synonymen in Kapitel 17 auf S. 325.

Formatierungsregeln für Ausnahmen bei der unscharfen Gruppierung

- Definieren Sie nur je ein Ausschlusspaar pro Zeile.
- Geben Sie einzelne Wörter oder zusammengesetzte Wörter an.
- Geben Sie die Wörter nur in Kleinbuchstaben ein. Wörter in Großbuchstaben werden ignoriert.
- Trennen Sie die Wörter in einem Paar durch ein `Tabulator`-Zeichen.

Nicht linguistische Elemente

Bei der Arbeit mit bestimmten Datenarten sind Datumsangaben, Sozialversicherungsnummern, Prozentsätze und andere nicht linguistische Elemente von Interesse. Diese Elemente werden explizit in der Konfigurationsdatei deklariert, in der Sie auch die Elemente aktivieren bzw. deaktivieren können. Für weitere Informationen siehe Thema Konfiguration auf S. 341. Um die Ausgabe der Extrahierungsengine zu optimieren, wird die Eingabe aus der nicht linguistischen Verarbeitung so normalisiert, dass ähnliche Elemente gemäß den vordefinierten Formaten gruppiert werden. Für weitere Informationen siehe Thema Normalisierung auf S. 341.

Anmerkung: Sie können die Extrahierung nicht linguistischer Elemente in den Extrahierungseinstellungen aktivieren und deaktivieren.

Verfügbare nicht linguistische Entitäten.

Die in der folgenden Tabelle aufgeführten nicht linguistischen Elemente können extrahiert werden. Der Typname wird in Klammern angegeben.

Adressen (`<Address>`)	Organisationen (`<Organization>`)
Aminosäuren (`<Aminoacid>`)	Prozente (`<Percent>`)
Währungen (`<Currency>`)	Produkte (`<Product>`)
Datumsangaben (`<Date>`)	Proteine (`<Gene>`)
Verspätung (`<Delay>`)	Telefonnummern (`<PhoneNumber>`)
Stellen (`<Digit>`)	Uhrzeiten (`<Time>`)
E-Mail-Adressen (`<email>`)	Sozialversicherungsnummer (USA) (`<SocialSecurityNumber>`)
HTTP/URL-Adressen (`<url>`)	Gewichte und Maßangaben (`<Weights-Measures>`)
IP-Adresse (`<IP>`)	

Bereinigen von Text für die Verarbeitung

Bevor die Extrahierung nicht-linguistischer Einheiten beginnt, wird der Eingabetext bereinigt. Bei diesem Schritt werden die folgenden temporären Änderungen durchgeführt, damit nichtlinguistische Einheiten identifiziert und als solche extrahiert werden können:

- Jede Folge von zwei oder mehr Leerzeichen wird durch ein einzelnes Leerzeichen ersetzt.
- Tabulatorzeichen werden durch Leerzeichen ersetzt.
- Einzelne Zeilenende-Zeichen oder Sequenzzeichen werden durch ein Leerzeichen ersetzt, während mehrere Zeilenende-Zeichen als Absatzende markiert werden. Das Zeilenende kann durch Zeilenschaltungen (Carriage Return, CR) und/oder Zeilenvorschub (Line Feed, LF) gekennzeichnet werden.
- HTML- und XML-Tags werden temporär unterdrückt und ignoriert.

Definitonen regulärer Ausdrücke

Beim Extrahieren nicht linguistischer Elemente sollen ggf. die Definitionen für reguläre Ausdrücke bearbeitet oder ergänzt werden, mit denen diese regulären Ausdrücke erkannt werden. Dies erfolgt im Abschnitt Definitionen regulärer Ausdrücke auf der Registerkarte “Erweiterte Ressourcen”. Für weitere Informationen siehe Thema Informationen zu erweiterten Ressourcen auf S. 332.

Die Datei ist in mehrere Abschnitte gegliedert. Der erste Abschnitt trägt die Bezeichnung `[macros]`. Neben diesem Abschnitt kann jeweils ein zusätzlicher Abschnitt für jedes nicht linguistische Element vorliegen. Sie können weitere Abschnitte in diese Datei aufnehmen. Die Regeln in den einzelnen Abschnitten sind jeweils nummeriert (*regexp1*, *regexp2* usw.). Die Regeln müssen sequenziell von 1 bis –*n* nummeriert werden. Eine Unterbrechung der Nummerierung führt dazu, dass diese Datei überhaupt nicht verarbeitet wird.

In bestimmten Fällen ist ein Element sprachabhängig. Ein Element gilt dann als sprachabhängig, wenn es für den Sprachparameter in der Konfigurationsdatei einen Wert ungleich 0 besitzt. Für weitere Informationen siehe Thema Konfiguration auf S. 341. Ist ein Element sprachabhängig, muss die Sprache dem Abschnittsnamen vorangestellt werden, z. B. `[english/PhoneNumber]`. Wenn Sie dem Element `PhoneNumber` den Wert 2 für die Sprache zuweisen, enthält dieser Abschnitt bestimmte Regeln, die nur für englische Telefonnummern gelten.

Wichtig: Wenn Sie an dieser oder einer anderen Datei im Editor Änderungen vornehmen und wenn die Extrahierungsengine anschließend nicht mehr wie gewünscht arbeitet, dann verwenden Sie die in der Symbolleiste vorhandene Option Auf Original zurücksetzen, um die Datei in den Zustand zurückzuversetzen, den sie bei der Installation hatte. Für diese Datei sollten Sie mit regulären Ausdrücken vertraut sein. Falls Sie weitere Hilfe in diesem Bereich benötigen, wenden Sie sich an IBM Corp..

Sonderzeichen. [] {} () \ * + ? | ^ $

Alle Zeichen entsprechen sich selbst, mit Ausnahme der folgenden Sonderzeichen, die für einen bestimmten Zweck in Ausdrücken benutzt werden: `.[{()\*+?|^$` Um diese Zeichen als solche zu verwenden, muss ihnen in der Definition ein umgekehrter Schrägstrich (\) vorangestellt werden.

Beispiel: Wenn Sie versuchen, Internet-Adressen zu extrahieren, ist der Punkt sehr wichtig für die Entität, daher müssen Sie ihn folgendermaßen mit einem umgekehrten Schrägstrich versehen:

```
www\.[a-z]+\.[a-z]+
```

Wiederholungsoperatoren und -Quantifikatoren ? + * {}

Damit flexiblere Definitionen möglich sind, können Sie mehrere Platzhalterzeichen verwenden, die für reguläre Ausdrücke Standard sind. Dabei handelt es sich um `*` `?` `+`

- Ein *Sternchen* * gibt an, dass *null oder mehr* Instanzen der vorangehenden Zeichenfolge vorhanden sind.
  Beispiel: `ab*c` findet "*ac*", "*abc*", "*abbbc*" usw. als Übereinstimmung.
- Ein *Pluszeichen* + gibt an, dass *eine oder mehrere* Instanzen der vorangehenden Zeichenfolge vorhanden sind.
  Beispiel: `ab+c` findet "*abc*", "*abbc*", "*abbbc*", aber nicht "*ac*" als Übereinstimmung.
- Ein *Fragezeichen ?* gibt an, dass *null oder eine* Instanz der vorangehenden Zeichenfolge vorhanden ist.
  Beispiel: `modell?ing` findet sowohl "*modelling*" als auch "*modeling*".
- *Beschränken von Wiederholungen mit Klammern {}* gibt die Grenzen der Wiederholung an. Zum Beispiel
  - `[0-9]{n}` entspricht einer Ziffer, die exakt *n* Mal wiederholt wird.
    Beispiel: `[0-9]{4}` findet "*1998*", aber weder "*33*" noch "*19983*".
  - `[0-9]{n}` entspricht einer Ziffer, die *n Mal oder öfter* wiederholt wird.
    Beispiel: `[0-9]{3,}` findet "*199*" oder "*1998*", aber nicht "*19*".
  - `[0-9]{n,m}` entspricht einer Ziffer, die zwischen *n und m Mal (inklusive)* wiederholt wird.
    Beispiel: `[0-9]{3,5}` findet "*199*", "*1998*" oder "*19983*", aber nicht "*19*" und nicht "*199835*".

Optionale Leerzeichen und Bindestriche

In manchen Fällen ist es erforderlich, ein optionales Leerzeichen in eine Definition einzufügen. Wenn Sie beispielsweise Währungen extrahieren möchten wie "*uruguayan pesos*", "*uruguayan peso*", "*uruguay pesos*", "*uruguay peso*", "*pesos*" oder "*peso*", müssen Sie sich damit auseinandersetzen, dass zwei Wörter vorhanden sein können, die durch ein Leerzeichen getrennt sind. In diesem Fall könnte die Definition als `(uruguayan |uruguay )?pesos?` geschrieben werden. Da auf *uruguayan* bzw. *uruguay* ein Leerzeichen folgt, wenn es mit *pesos*/*peso* verwendet wird, muss das optionale Leerzeichen innerhalb der optionalen Sequenz `(uruguayan |uruguay )` definiert werden. Wenn das Leerzeichen nicht in der optionalen Sequenz wie `(uruguayan|uruguay)? pesos?` vorhanden wäre, würde es nicht "*pesos*" oder "*peso*" finden, da das Leerzeichen erforderlich wäre.

Wenn Sie eine Reihe von Einträgen einschließlich eines Bindestrichs (-) in einer Liste suchen, muss der Bindestrich als Letztes definiert werden. Beispiel: Wenn Sie ein Komma (,) oder einen Bindestrich (-) suchen, verwenden Sie `[,-]`, aber auf keinen Fall `[-,]`.

Reihenfolge von Zeichenfolgen in Listen und Makros

Sie sollten stets die längste Sequenz vor einer kürzeren definieren. Andernfalls wird die längste Sequenz nie gelesen, da die Übereinstimmung schon mit der kürzeren gefunden wird. Beispiel: Wenn Sie die Zeichenfolgen "*billion*" oder "*bill*" suchen, muss "*billion*" vor "*bill*" definiert werden, also `(billion|bill)` und nicht `(bill|billion)`. Dies gilt auch für Makros, da Makros Listen von Zeichenfolgen sind.

Reihenfolge von Regeln im Definitionsabschnitt

Definieren Sie nur je eine Regel pro Zeile. Die Regeln in den einzelnen Abschnitten sind jeweils nummeriert (*regexp1*, *regexp2* usw.). Die Regeln müssen sequenziell von 1 bis *–n* nummeriert werden. Eine Unterbrechung der Nummerierung führt dazu, dass diese Datei überhaupt nicht verarbeitet wird. Um einen Eintrag zu deaktivieren, geben Sie am Anfang jeder Zeile, in der der reguläre Ausdruck definiert ist, eine Raute (#) ein. Soll ein Eintrag aktiviert werden, entfernen Sie das Nummernzeichen (#) am Anfang der zugehörigen Zeile.

In jedem Abschnitt müssen die spezifischsten Regeln vor den allgemeinsten Regeln definiert werden, um eine korrekte Verarbeitung sicherzustellen. Wenn Sie beispielsweise ein Datum in der Form "*Monat Jahr*" und in der Form "*Monat*" suchen, muss die Regel "*Monat Jahr*" vor der Regel "*Monat*" definiert werden. Die Definition sollte wie folgt aussehen:

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

und nicht:

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

Verwenden von Makros in Regeln

Wenn eine bestimmte Sequenz in mehreren Regeln vorkommt, können Sie ein Makro verwenden. Wenn Sie dann die Definition dieser Sequenz ändern müssen, müssen Sie sie nur einmal ändern und nicht in allen Regeln, die darauf verweisen. Hier ein Beispiel: Angenommen, es liegt folgendes Makro vor:

```
MONTH=((january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?)
```

Bei jedem Verweis auf den Namen des Makros muss dieser von `$()` umgeben sein, z. B.:
`regexp1=$(MONTH)`

Alle Makros müssen im Abschnitt `[macros]` definiert werden.

Normalisierung

Beim Extrahieren nicht linguistischer Elemente werden die aufgefundenen Elemente so normalisiert, dass ähnliche Elemente gemäß den vordefinierten Formaten gruppiert werden. Währungssymbole und die zugehörigen Währungskürzel werden beispielsweise als gleich behandelt. Die Normalisierungseinträge werden im Abschnitt Normalisierung auf der Registerkarte "Erweiterte Ressourcen" gespeichert. Für weitere Informationen siehe Thema Informationen zu erweiterten Ressourcen auf S. 332. Die Datei ist in mehrere Abschnitte gegliedert.

Wichtig: Diese Datei richtet sich nur an fortgeschrittene Benutzer. Änderungen an dieser Datei fallen eher selten an. Falls Sie weitere Hilfe in diesem Bereich benötigen, wenden Sie sich an IBM Corp..

Formatierungsregeln für die Normalisierung

- Fügen Sie nur je einen Normalisierungseintrag pro Zeile hinzu.
- Beachten Sie genauestens die Abschnitte in dieser Datei. Es ist nicht möglich, neue Abschnitte hinzuzufügen.
- Um einen Eintrag zu deaktivieren, geben Sie am Anfang der entsprechenden Zeile eine Raute (#) ein. Soll ein Eintrag aktiviert werden, entfernen Sie das Nummernzeichen (#) am Anfang der zugehörigen Zeile.

Konfiguration

Sie können die nicht linguistischen Elementtypen, die Sie extrahieren möchten, in der Konfigurationsdatei der nicht linguistischen Elemente aktivieren bzw. deaktivieren. Wenn Sie die nicht benötigten Elemente deaktivieren, wird die Verarbeitung beschleunigt. Dies erfolgt im Abschnitt Konfiguration auf der Registerkarte "Erweiterte Ressourcen". Für weitere Informationen siehe Thema Informationen zu erweiterten Ressourcen auf S. 332. Wenn die Extrahierung nicht linguistischer Elemente aktiviert ist, liest die Extrahierungsengine diese Konfigurationsdatei beim Extrahieren, um zu ermitteln, welche Typen nicht linguistischer Elemente extrahiert werden sollen.

Für diese Datei gilt die folgende Syntax:

```
#Name<TAB>Sprache<TAB>Code
```

Tabelle 18-1
Syntax für Konfigurationsdatei

Spaltenüberschrift	Beschreibung
`#Name`	Wortlaut, mit dem die anderen beiden Dateien, die für die Extrahierung nicht linguistischer Elemente erforderlich sind, auf diese nicht linguistischen Elemente verweisen. Bei den hier angegebenen Namen wird zwischen Groß- und Kleinschreibung unterschieden.
`Sprache`	Die Sprache der Dokumente. Im Idealfall sollten Sie die jeweilige Sprache angeben; alternativ steht die Option Alle zur Verfügung. Mögliche Optionen: `0` = Alle, was verwendet wird, wenn ein regexp nicht sprachspezifisch ist und in mehreren Vorlagen mit verschiedenen Sprachen benutzt werden kann, z. B. IP/URL/E-Mail-Adressen; `1` = Französisch; `2` = Englisch; `4` = Deutsch; `5` = Spanisch; `6` = Niederländisch; 8 = Portugiesisch; `10` = Italienisch.

Spaltenüberschrift	Beschreibung
Code	Code für die Wortklasse. Die meisten Elemente erhalten den Wert “s”, bis auf wenige Ausnahmefälle. Mögliche Werte: s = Stoppwort; a = Adjektiv; n = Nomen. Wenn diese Option aktiviert ist, werden zuerst die nicht linguistischen Elemente extrahiert und dann wird die Rolle dieser Elemente mithilfe der Extraktionsmuster in einem größeren Zusammenhang ermittelt. Prozentsätze erhalten beispielsweise den Wert “a” Angenommen, 30 % wird als nicht linguistisches Element extrahiert. Dieses Element wird als Adjektiv erkannt. Wenn der Text dann die Wortgruppe “30% salary increase” enthält, erfüllt das nicht linguistische Element “30%” das Wortklasse-Muster “an” (Adjektiv-Nomen-Nomen).

Reihenfolge bei der Definition von Elementen

Die Reihenfolge, in der Sie die Elemente in dieser Datei deklarieren, ist von Bedeutung und wirkt sich auf die Extrahierung aus. Die Einträge werden in der angegebenen Reihenfolge angewendet. Wenn Sie die Reihenfolge ändern, ändern sich auch die Ergebnisse. Die spezifischsten nicht linguistischen Elemente müssen vor den allgemeineren definiert werden.

Beispielsweise wird das nicht linguistische Element “Aminosäure” definiert durch:

```
regexp1=($(AA)-?$(NUM))
```

Dabei entspricht $(AA) den Werten “(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)”, bei denen es sich um bestimmte Folgen aus drei Buchstaben handelt, die bestimmten Aminosäuren entsprechen.

Andererseits wird das nicht linguistische Element “Gen” allgemeiner und wird definiert durch:

```
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?
```

Wenn im Abschnitt “Konfiguration” das Element “Gen” vor “Aminosäure” definiert wird, wird für “Aminosäure” nie eine Übereinstimmung gefunden, da regexp3 von “Gen” immer die erste Übereinstimmung erzielt.

Formatierungsregeln für die Konfiguration

- Trennen Sie die Einträge in einer Spalte durch ein Tabulator-Zeichen.
- Löschen Sie keine Zeilen.
- Halten Sie sich an die Syntax, die in der vorangehenden Tabelle angegeben ist.
- Um einen Eintrag zu deaktivieren, geben Sie am Anfang der entsprechenden Zeile eine Raute (#) ein. Soll ein Element aktiviert werden, entfernen Sie die Raute (#) am Anfang der zugehörigen Zeile.

Sprachbehandlung

Alle modernen Sprachen besitzen eigene Systeme, um Ideen auszudrücken, Sätze zu strukturieren und Abkürzungen zu verwenden. Im Abschnitt Sprachbehandlung können Sie Extraktionsmuster bearbeiten, Definitionen für diese Muster erzwingen und Abkürzungen für von Ihnen in der Dropdown-Liste “Sprache” ausgewählte Sprachen deklarieren.

- Extraktionsmuster
- Erwungene Definitionen
- Abkürzungen

Extraktionsmuster

Beim Extrahieren von Daten aus Ihren Dokumenten wendet die Extrahierungsengine eine Reihe von Wortklasse-Mustern (Part of Speech) auf einen “Stapel” von Wörtern im Text an, um so Kandidaten (Wörter und zusammengesetzte Ausdrücke) für die Extrahierung zu erkennen. Sie können die Extraktionsmuster hinzufügen und bearbeiten.

Die Wortarten (Part of Speech) bestehen aus grammatischen Elementen, z. B. Nomen, Adjektive, Partizip Präteritum, Determinatoren, Präpositionen, Koordinatoren, Vornamen, Initialen und Partikel. Eine Reihe dieser Elemente bildet ein Wortklasse-Extraktionsmuster. In IBM Corp.-Text-Mining-Produkten ist jede Wortart (Part of Speech) mit einem einzelnen Buchstaben gekennzeichnet, so dass Sie die Muster leichter definieren können. Ein Adjektiv ist beispielsweise am Kleinbuchstaben *a* erkennbar. Die unterstützten Codes werden standardmäßig am Anfang jedes Abschnitts für Standardextraktionsmuster aufgeführt, zusammen mit einer Reihe von Mustern und Beispielen für die Muster, mit denen die verwendeten Codes erläutert werden.

Formatierungsregeln für Extraktionsmuster

- Ein Muster pro Zeile.
- Um ein Muster zu deaktivieren, geben Sie am Anfang der entsprechenden Zeile eine Raute (#) ein.

Die Reihenfolge, in der Sie die Extraktionsmuster aufführen, ist von großer Bedeutung, weil eine gegebene Wortfolge nur einmal in der Extraktionsengine gelesen und dann dem ersten Extraktionsmuster zugewiesen wird, für das die Engine eine Übereinstimmung erkennt.

Erzwungene Definitionen

Beim Extrahieren von Daten aus Ihren Dokumenten scannt die Extrahierungsengine den Text und erkennt dabei die Wortart (Part of Speech) für jedes gefundene Wort. In einigen Fällen kann ein Wort verschiedene POS-Rollen annehmen, je nach Kontext. Soll ein Wort in jedem Fall eine bestimmte Part-of-Speech-Rolle annehmen oder ganz von der Verarbeitung ausgeschlossen werden, legen Sie hierzu die entsprechenden Einstellungen auf der Registerkarte “Erweiterte Ressourcen” im Bereich Erzwungene Definition fest. Für weitere Informationen siehe Thema Informationen zu erweiterten Ressourcen auf S. 332.

Um eine Wortklassen-Rolle für ein bestimmtes Wort zu erzwingen, müssen Sie in diesem Bereich eine Zeile mit der folgenden Syntax hinzufügen:

```
Fachausdruck:Code
```

Tabelle 18-2
Beschreibung der Syntax

Aufnahme (Cox Regression)	Beschreibung
`Fachaus-druck`	Der Name eines Fachausrucks
`Code`	Ein einstelliger Code für die Wortklassen-Regel. Sie können bis zu sechs verschiedene Wortklassen-Codes pro Uniterm angeben. Mit dem Code `s` (Kleinbuchstabe "s") können Sie außerdem die Extrahierung eines bestimmten Wortes in zusammengesetzte Worte bzw. Wortfolgen unterbinden, z. B. `additional:s`.

Formatierungsregeln für erzwungene Definitionen

- Eine Zeile pro Wort.
- Doppelpunkte sind in Fachausdrücken nicht zulässig.
- Mit dem Kleinbuchstaben `s` als Code für die Wortart geben Sie an, dass ein Wort überhaupt nicht extrahiert werden soll.
- Geben Sie bis zu sechs Wortklassen-Codes pro Zeile ein. Unterstützte Wortart-(Part-of-Speech-)Codes werden im Abschnitt "Extraktionsmuster" angezeigt. Für weitere Informationen siehe Thema Extraktionsmuster auf S. 343.
- Das Sternchen (*) am Ende einer Zeichenfolge dient als Platzhalter für teilweise Übereinstimmungen. Wenn Sie beispielsweise `add*:s` eingeben, werden Wörter wie `add`, `additional`, `additionally`, `addendum`, und `additive` weder als Fachausdruck noch als Bestandteil eines zusammengesetzten Worts extrahiert. Ist eine Wortübereinstimmung jedoch explizit als Fachausdruck in einem kompilierten Wörterbuch oder in den erzwungenen Definitionen deklariert, wird diese dennoch extrahiert. Wenn Sie beispielsweise sowohl `add*:s` als auch `addendum:n` eingeben, wird `addendum` auf jeden Fall aus dem Text extrahiert.

Abkürzungen

Im Allgemeinen behandelt die Extrahierungsengine bei der Verarbeitung von Text jeden Punkt als ein Zeichen dafür, dass der Satz beendet ist. In der Regel ist das auch richtig, doch gilt diese Behandlung von Punktzeichen nicht, wenn Abkürzungen im Text enthalten sind.

Wenn Sie Ausdrücke aus Ihrem Text extrahieren und feststellen, dass bestimmte Abkürzungen falsch behandelt wurden, sollten Sie die jeweilige Abkürzung in diesem Bereich explizit deklarieren.

Hinweis: Wenn die Abkürzung bereits als Synonymdefinition erscheint oder im Typ-Wörterbuch als Ausdruck definiert ist, ist es nicht erforderlich, die Abkürzung hier einzutragen.

Formatierungsregeln für Abkürzungen

- Definieren Sie nur je eine Abkürzung pro Zeile.

Language Identifier

Im Idealfall sollte die jeweilige Sprache der analysierten Textdaten angegeben werden; die Option Alle bietet jedoch die Möglichkeit, auch Texte mit verschiedenen oder gar unbekannten Sprachen zu verarbeiten. Die Sprachoption Alle greift auf Language Identifier zurück, eine Engine für die automatische Spracherkennung. Language Identifier scannt die Dokumente und ermittelt dabei die Dokumente, die in einer unterstützten Sprache verfasst sind. Bei der Extrahierung werden anschließend die besten internen Wörterbücher auf diese Dateien angewendet. Die Option Alle wird durch die Parameter im Abschnitt “Eigenschaften” bestimmt.

Eigenschaften

Der Language Identifier wird anhand der in diesem Abschnitt angegebenen Parameter definiert. Die folgende Tabelle beschreibt die Parameter, die Sie im Abschnitt Language Identifier – Eigenschaften auf der Registerkarte “Erweiterte Ressourcen” einstellen können. Für weitere Informationen siehe Thema Informationen zu erweiterten Ressourcen auf S. 332.

Tabelle 18-3
Parameterbeschreibungen

Parameter (Probability Plots)	Beschreibung
NUM_CHARS	Legt die Anzahl der Zeichen fest, die von der Extrahierungsengine gelesen werden sollen, um zu ermitteln, in welcher Sprache der Text verfasst ist. Je höher der Wert, mit desto größerer Genauigkeit wird die Sprache ermittelt. Beim Wert 0 wird der gesamte Text im Dokument gelesen.
USE_FIRST_SUPPORTED _LANGUAGE	Gibt an, ob die Extrahierungsengine die erste unterstützte Sprache verwenden soll, die Language Identifier erkennt. Beim Wert 1 wird die erste unterstützte Sprache herangezogen. Beim Wert 0 dagegen wird die Ausweichsprache herangezogen.
FALLBACK_LANGUAGE	Gibt die Sprache an, die verwendet werden soll, falls die von Language Identifier zurückgegebene Sprache nicht unterstützt wird. Zulässige Werte: english, french, german, spanish, dutch, italian bzw. ignore. Beim Wert ignore werden Dokumente, die keine unterstützte Sprache enthalten, schlichtweg ignoriert.

Sprachen

Der Language Identifier unterstützt eine Vielzahl unterschiedlicher Sprachen. Sie können die Liste der Sprachen im Abschnitt Language Identifier – Sprachen auf der Registerkarte “Erweiterte Ressourcen” anpassen.

Je mehr Sprachen vorhanden sind, desto wahrscheinlicher können diese falsch positive Ergebnisse hervorrufen und die Verarbeitungsgeschwindigkeit senken. Entfernen Sie also gegebenenfalls die weniger gebräuchlichen Sprachen aus dieser Liste. Sie können zu dieser Datei jedoch keine weiteren Sprachen hinzufügen. Sie sollten die wahrscheinlichsten Sprachen an den Anfang der Liste stellen, damit der Language Identifier eine Übereinstimmung mit Ihren Dokumenten schneller ermitteln kann.

Kapitel
19

Textlinkregeln

Die Textlinkanalyse (TLA) ist eine Technologie zum Musterabgleich, mit der anhand eines Regelsets die in Ihrem Text gefundenen Beziehungen extrahiert werden können. Wenn Textlinkanalyse für Extrahierung aktiviert ist, werden die Textdaten mit diesen Regeln abgeglichen. Wird eine Übereinstimmung gefunden, wird das Textlinkanalyse-Muster extrahiert und präsentiert. Diese Regeln werden in der Registerkarte “Textlinkregeln” definiert.

Informationen zu extrahieren, die einfache Ideen zu einer Organisation darstellen, kann für Sie z. B. wenig interessant sein. Indem Sie die TLA verwenden, können Sie allerdings Informationen über die Zusammenhänge zwischen verschiedenen Organisationen oder den zu diesen Organisationen gehörenden Menschen erhalten. TLA kann auch verwendet werden, um Meinungen über Themen zu extrahieren, z. B. ihre Meinung über ein bestimmtes Produkt oder Erlebnis.

Um TLA nutzen zu können, müssen Ressourcen vorhanden sein, die Textlink- (TLA-) Regeln enthalten. Wenn Sie Vorlagen auswählen, wird über ein Symbol in der Spalte “TLA” angezeigt, ob die Vorlagen TLA-Regeln besitzen.

Abbildung 19-1
Spalte “TLA” in Vorlagendialogfeldern

Vorlage	Eigentü...	Version	Datum	Anmerk...	TLA	Spr...
Bank CRM (English)	claired	1	Sep-30-...			English
Insurance CRM (English)	claired	1	Sep-30-...			English
Ads Opinions (English)	claired	1	Sep-17-...			English
Bank Satisfaction Opinio...	claired	1	Sep-17-...			English
Security Intelligence (En...	claired	1	Sep-30-...			English

Textlinkanalysemuster werden während der Musterabgleichsphase des Extrahierungsprozesses in den Textdaten gefunden. Während dieser Phase werden Regeln mit den Textdaten verglichen und wenn eine Übereinstimmung vorliegt, werden diese Informationen als Muster extrahiert. Gelegentlich möchten Sie vielleicht die Textlinkanalyse optimieren oder die Übereinstimmungskriterien verändern. In diesen Fällen können Sie die Regeln ausarbeiten, um sie an Ihre jeweiligen Anforderungen anzupassen. Dies erfolgt über die Registerkarte “Textlinkregeln”.

Anmerkung: Unterstützung von Variablen wurde in Version 13 abgeschafft. Verwenden Sie stattdessen Makros. Für weitere Informationen siehe Thema Arbeiten mit Makros auf S. 356.

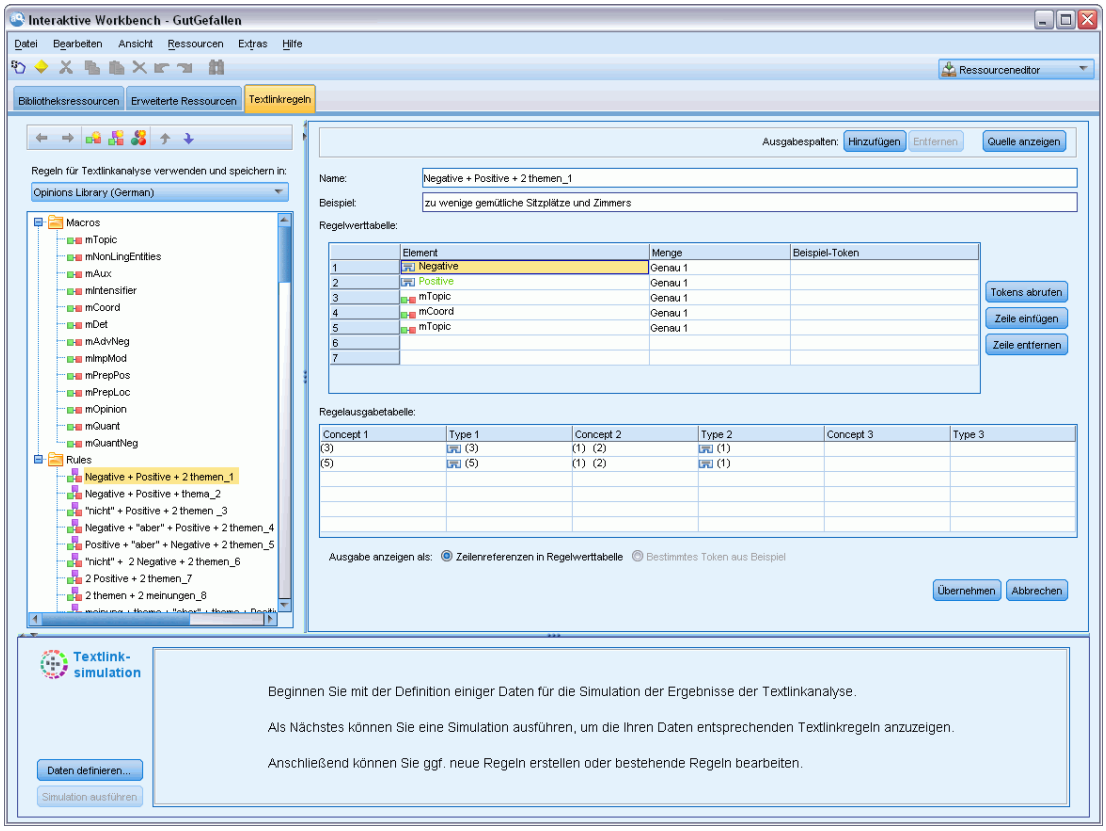
Bearbeiten von Textlinkregeln

Sie können Regeln direkt auf der Registerkarte “Textlinkregeln” in der Ansicht Template Editor oder Resource Editor erstellen und bearbeiten. Um besser sehen zu können, wie Regeln mit Text übereinstimmen können, können Sie auf dieser Registerkarte eine Simulation durchführen. Während der Simulation wird nur bei den Beispielsimulationsdaten eine Extrahierung durchgeführt und die Textlinkregeln werden angewendet, um auf eventuell übereinstimmende Muster zu prüfen. Alle mit dem Text übereinstimmenden Regeln werden im Simulationsbereich

angezeigt. Basierend auf den Übereinstimmungen können Sie Regeln und Makros auswählen, um die Übereinstimmungskriterien für Text zu ändern.

Im Gegensatz zu den anderen erweiterten Ressourcen sind TLA-Regeln bibliothekenspezifisch. Sie können also jeweils nur die TLA-Regeln aus einer Bibliothek verwenden. Wechseln Sie in der Ansicht Template Editor oder Resource Editor auf die Registerkarte Textlinkregeln. Auf dieser Registerkarte können Sie die Bibliothek in Ihrer Vorlage angeben, die die TLA-Regeln enthält, die Sie verwenden oder bearbeiten möchten. Aus diesem Grund wird dringend empfohlen, dass Sie alle Ihre Regeln in einer Bibliothek speichern, es sei denn, dies wird aus einem spezifischen Grund nicht gewünscht.

Abbildung 19-2
Registerkarte "Textlinkanalyseregeln"



Wichtig: Diese Registerkarte ist nicht für japanische Ressourcen verfügbar.

Erste Schritte

Es gibt zahlreiche Möglichkeiten, mit der Arbeit im Editor auf der Registerkarte "Textlinkregeln" zu beginnen:

- Simulieren Sie zunächst einen Beispieltext und erstellen oder bearbeiten Sie übereinstimmende Regeln basierend auf den Kriterien, wie die aktuellen Regeln Muster aus den Simulationsdaten extrahieren.

- Erstellen Sie eine vollständig neue Regel oder bearbeiten Sie eine bestehende Regel.
- Arbeit direkt in der Quellenansicht.

Wann Regeln erstellt oder bearbeitet werden sollten

Obwohl die in jeder Vorlage enthaltenen Textlinkanalyseregeln oft für die Extrahierung vieler einfacher oder komplexer Beziehungen aus Ihrem Text ausreichen, möchten Sie vielleicht gelegentlich Änderungen an diesen Regeln vornehmen oder eigene Regeln erstellen. Beispiel:

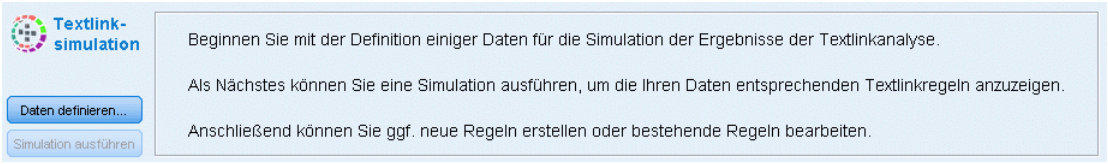
- Um eine Idee oder Beziehung zu erfassen, die nicht mit den bestehenden Regeln extrahiert wurde, indem Sie eine neue Regel oder ein neues Makro erstellen.
- Um das Standardverhalten eines Typs zu ändern, den Sie den Ressourcen hinzugefügt haben. Dazu müssen Sie üblicherweise ein Makro wie `mTopic` oder `mNonLingEntities` bearbeiten. Für weitere Informationen siehe Thema Spezielle Makros: mTopic, mNonLingEntities, SEP auf S. 360.
- Um vorhandenen Textlinkanalyseregeln und Makros neue Typen hinzuzufügen. Wenn Sie beispielsweise der Meinung sind, dass der Typ `<Organization>` zu breit gefasst ist, können Sie neue Typen für Organisationen in unterschiedlichen Sektoren wie `<Pharmazie>`, `<Autoindustrie>`, `<Finanzen>` usw. erstellen. In diesem Fall müssen Sie die Textlinkanalyseregeln ändern und/oder ein Makro erstellen, damit diese neuen Typen berücksichtigt und entsprechend verarbeitet werden.
- Um einer vorhandenen Textlinkanalyseregel Typen hinzuzufügen. Nehmen wir z. B. an, eine Ihrer Regeln erfasst den folgenden Text `Max Mustermann ruft Else Mustermann an`, jedoch möchten Sie, dass diese Regel nicht nur Anrufe, sondern auch E-Mail-Kommunikation erfasst. Sie könnten der Regel einen nichtlinguistischen Elementtyp für E-Mail hinzufügen, damit auch Text wie der folgende erfasst wird: `maxmustermann@ibm.com sendete E-Mail an martinamustermann@ibm.com`.
- Um eine bestehende Regel geringfügig zu ändern, anstatt eine neue zu erstellen. Nehmen wir z. B. an, eine Ihrer Regeln entspricht dem Text `xyz ist sehr gut`, jedoch möchten Sie, dass diese Regel auch `xyz ist extrem gut` erfasst.

Simulation von Textlinkanalyseergebnissen

Zur einfacheren Definition neuer Textlinkregeln oder zum besseren Verständnis dessen, wie bestimmte Sätze während der Textlinkanalyse abgeglichen werden, ist es oft nützlich, eine Simulation eines Mustertexts durchzuführen. Während der Simulation wird nur bei den Beispielsimulationsdaten mithilfe der aktuellen linguistischen Ressourcen und der aktuellen Extrahierungseinstellungen eine Extrahierung durchgeführt. Das Ziel besteht darin, die simulierten Ergebnisse abzurufen und sie zu verwenden, um Ihre Regeln zu verbessern, neue Regeln zu erstellen oder um besser zu verstehen, wie es zu Übereinstimmungen kommt. Für jeden Teil des Texts (Satz, Wort oder Satzteil je nach Kontext) zeigt eine Simulationsausgabe die gesammelten Tokens und TLA-Regeln, die ein Muster in diesem Text aufgedeckt haben. Ein **Token** wird als beliebiges Word oder beliebige Wortfolge definiert, das/die während des Extrahierungsprozesses identifiziert wurde.

Im Gegensatz zu den anderen erweiterten Ressourcen sind TLA-Regeln bibliothekenspezifisch. Sie können also jeweils nur die TLA-Regeln aus einer Bibliothek verwenden. Wechseln Sie in der Ansicht Template Editor oder Resource Editor auf die Registerkarte Textlinkregeln. Auf dieser Registerkarte können Sie die Bibliothek in Ihrer Vorlage angeben, die die TLA-Regeln enthält, die Sie verwenden oder bearbeiten möchten. Aus diesem Grund wird dringend empfohlen, dass Sie alle Ihre Regeln in einer Bibliothek speichern, es sei denn, dies wird aus einem spezifischen Grund nicht gewünscht.

Abbildung 19-3
Bereich "Textlinksimulation" vor der Definition von Daten



Wichtig: Es wird dringend empfohlen, dass Sie bei Verwenden einer Datendatei sicherstellen, dass der enthaltene Text kurz ist, um die Verarbeitungszeit zu minimieren. Das Ziel einer Simulation ist aufzuzeigen, wie ein Text interpretiert wird, und zu verstehen, welche Regeln diesem Text entsprechen. Diese Informationen unterstützen Sie dabei, Ihre eigenen Regeln zu schreiben und zu bearbeiten. Verwenden Sie den Textlinkanalyseknoten oder führen Sie einen Stream mit interaktiver Sitzung mit aktivierter TLA-Extrahierung aus, um Ergebnisse für ein vollständigeres Datenset zu erhalten. Diese Simulation dient nur zu Testzwecken und zum Verfassen von Regeln.

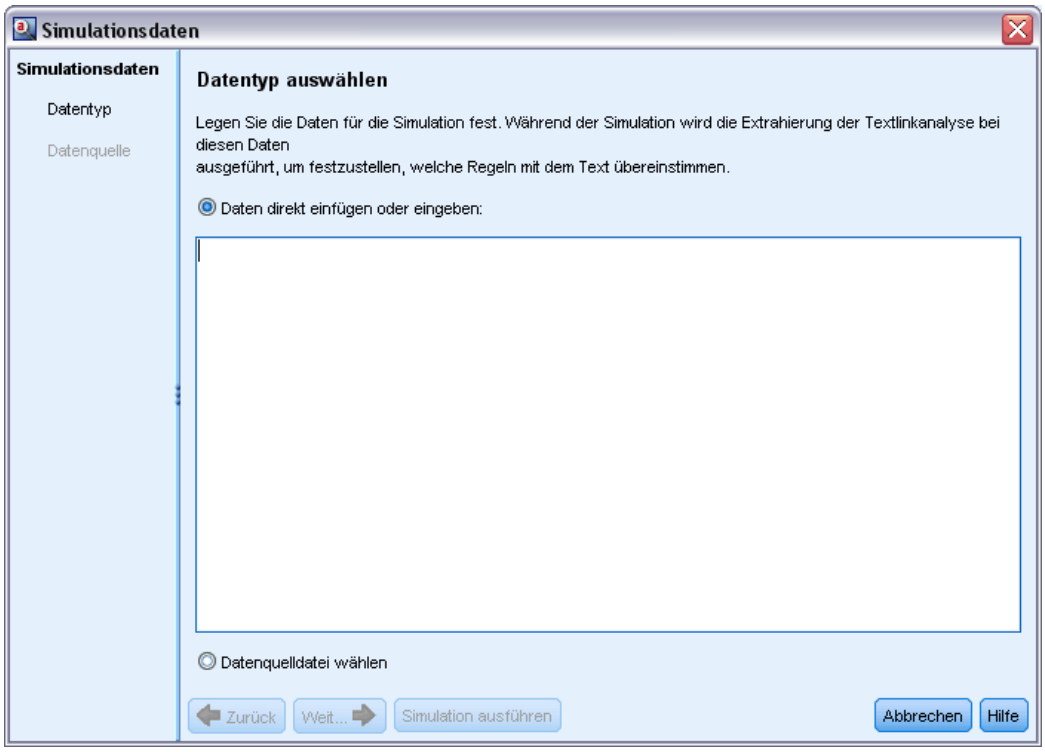
Definition von Daten zur Simulation

Um besser sehen zu können, wie Regeln mit Text übereinstimmen können, können Sie eine Simulation mit Beispieldaten durchführen. Der erste Schritt besteht in der Definition der Daten.

Definition von Daten

- Klicken Sie auf Daten definieren im Simulationsbereich unten auf der Registerkarte Textlinkregeln. Wählen Sie alternativ, sofern zuvor keine Daten definiert wurden, die Optionsfolge Extras > Simulation ausführen aus den Menüs. Der Simulationsdaten-Assistent wird geöffnet.

Abbildung 19-4
Simulations-Assistent

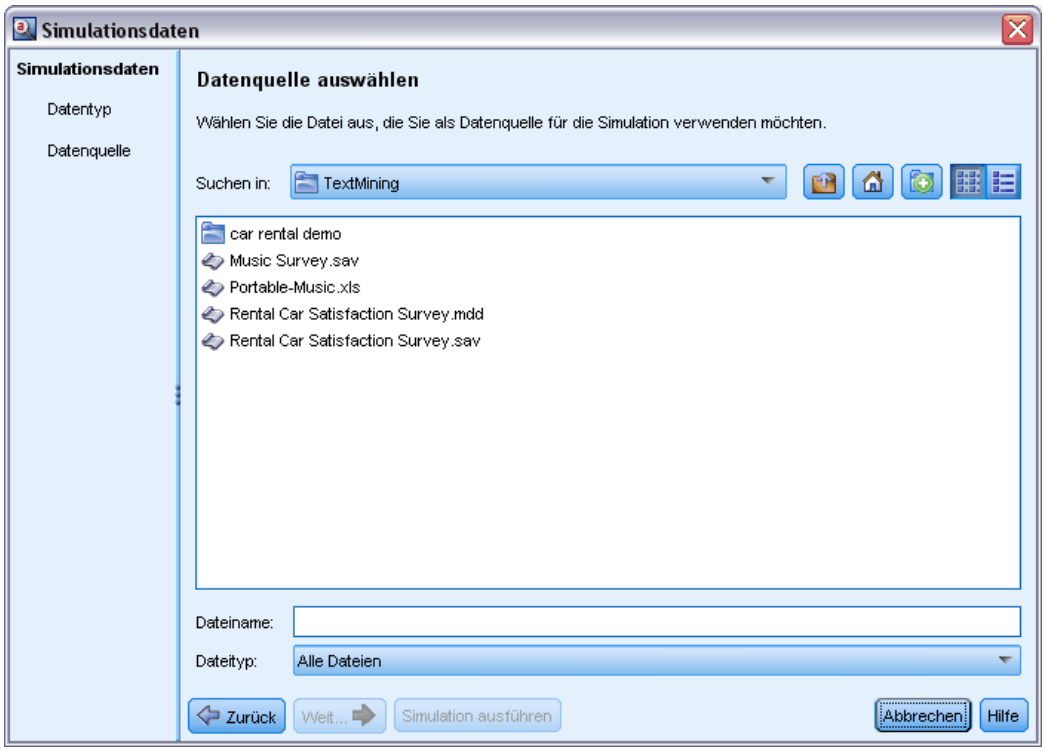


- Legen Sie den Datentyp fest, indem Sie eine der folgenden Optionen auswählen:
  - Text direkt einfügen oder eingeben. Es wird ein Textfeld angezeigt, in das Sie Text aus der Zwischenablage einfügen oder den zu bearbeitenden Text manuell eingeben können. Sie können einen Satz pro Zeile eingeben oder den Satz mithilfe von Interpunktion (z. B. Punkte oder Kommas) aufteilen. Nachdem Sie Ihren Text eingegeben haben, können Sie die Simulation starten, indem Sie auf Simulation ausführen klicken.
  - Bestimmung einer Dateidatenquelle. Diese Option gibt an, dass Sie eine Datei bearbeiten möchten, die Text enthält. Klicken Sie auf Weiter, um zum Schritt im Assistenten fortzufahren, in dem Sie die zu bearbeitende Datei definieren können. Nachdem Sie die Datei ausgewählt haben, können Sie die Simulation starten, indem Sie auf Simulation ausführen klicken. Es werden folgende Dateitypen unterstützt: *.rtf*, *.doc*, *.docx*, *.docm*, *.xls*, *.xlsx*, *.xlsm*, *.htm*, *.html*, *.txt* und Dateien ohne Dateierweiterung. Die von Ihnen gewählte Datendatei wird während der Simulation in der vorliegenden Form gelesen. Wenn Sie beispielsweise eine Microsoft Excel-Datei auswählen, können Sie weder ein bestimmtes Arbeitsblatt noch eine bestimmte Spalte auswählen. Stattdessen wird das vollständige Arbeitsbuch gelesen, ähnlich der Verwendung eines Microsoft Excel-Quellenknotens in IBM® SPSS® Modeler. Die gesamte Datei wird genau so behandelt, als ob Sie einen Dateilistenknoten mit einem Text-Mining-Knoten verbunden hätten.

Wichtig: Es wird dringend empfohlen, dass Sie bei Verwenden einer Datendatei sicherstellen, dass der enthaltene Text kurz ist, um die Verarbeitungszeit zu minimieren. Das Ziel einer Simulation ist aufzuzeigen, wie ein Text interpretiert wird, und zu verstehen, welche Regeln diesem Text entsprechen. Diese Informationen unterstützen Sie dabei, Ihre eigenen Regeln zu schreiben und

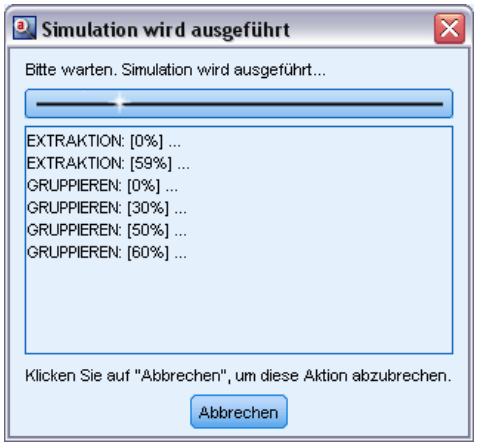
zu bearbeiten. Verwenden Sie den Textlinkanalyseknoten oder führen Sie einen Stream mit interaktiver Sitzung mit aktivierter TLA-Extrahierung aus, um Ergebnisse für ein vollständigeres Datenset zu erhalten. Diese Simulation dient nur zu Testzwecken und zum Verfassen von Regeln.

Abbildung 19-5
Simulations-Assistent – Datenquelle auswählen



- Klicken Sie auf Simulation ausführen, um die Simulation zu starten. Es wird ein Dialogfeld über den Fortschritt angezeigt. Wenn Sie in einer interaktiven Sitzung arbeiten, sind die Extrahierungseinstellungen für die Simulation diejenigen, die derzeit in der interaktiven Sitzung ausgewählt sind (siehe Extras > Extrahierungseinstellungen in der Ansicht “Konzepte und Kategorien”). Wenn Sie im Template Editor arbeiten, werden als Extrahierungseinstellungen für die Simulation die Standard-Extrahierungseinstellungen verwendet, die identisch sind mit denen, die in der Registerkarte “Experten” eines Textlinkanalyseknotens angezeigt werden. Für weitere Informationen siehe Thema Informationen zu den Simulationsergebnissen auf S. 352.

Abbildung 19-6
Dialogfeld "Simulationsfortschritt"



Informationen zu den Simulationsergebnissen

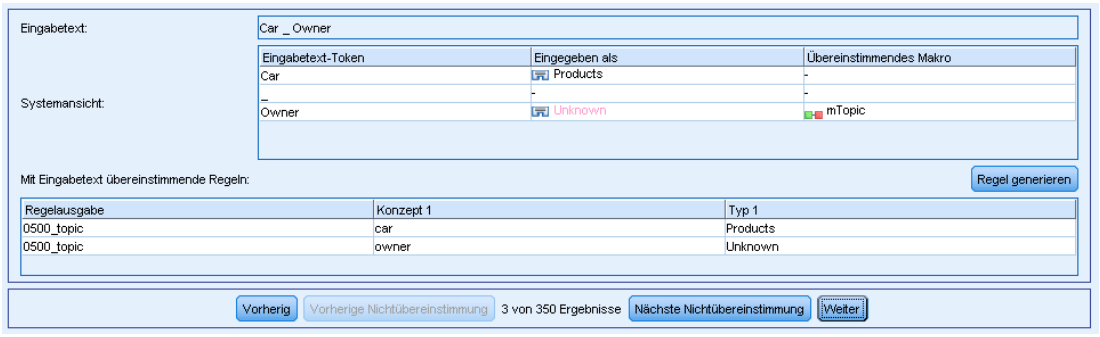
Um besser sehen zu können, wie Regeln mit Text übereinstimmen können, können Sie eine Simulation mit Beispieldaten durchführen und die Ergebnisse überprüfen. Dort können Sie Ihre Regeln besser an Ihre Daten anpassen. Nach Abschluss des Extrahierungs- und Simulationsprozesses werden die Ergebnisse der Simulation angezeigt.

Für jeden während der Extrahierung identifizierten "Satz" erhalten Sie zahlreiche Informationen, darunter den exakten "Satz", die Aufschlüsselung der in diesem Eingabesatz gefundenen Tokens und schließlich alle Regeln, bei denen eine Textübereinstimmung in diesem Satz gefunden wurde. Unter einem "**Satz**" verstehen wir entweder ein Wort, einen Satz oder einen Satzteil, je nach dem, wie der Extraktor den Text in lesbare Abschnitte unterteilt hat.

Ein **Token** wird als beliebiges Word oder beliebige Wortfolge definiert, das/die während des Extrahierungsprozesses identifiziert wurde. Zum Beispiel können in dem Satz *Mein Onkel lebt in New York* die folgenden Tokens während der Extrahierung gefunden werden: *mein*, *Onkel*, *lebt*, *in* und *New York*. Weiterhin könnte *Onkel* als Konzept des Typs `<Unknown>` extrahiert werden und *New York* als Konzept des Typs `<Location>`. Alle Konzepte sind Tokens, doch nicht alle Tokens sind Konzepte. Tokens können auch Makros, Zeichenfolgen und Wortlücken sein. Nur einem Typ zugeordnete Wörter oder Wortfolgen können Konzepte sein.

Wenn Sie in der interaktiven Sitzung oder dem Ressourceneditor arbeiten, befinden Sie sich auf der Konzeptebene. TLA-Regeln sind genauer und einzelne Tokens in einem Satz können in der Definition einer Regel verwendet werden, selbst wenn diese nie extrahiert und typisiert werden. Die Möglichkeit, Tokens zu verwenden, die keine Konzepte sind, bietet zusätzliche Flexibilität für Regeln beim Erfassen komplexer Beziehungen in Ihrem Text.

Abbildung 19-7
Ergebnisse einer Beispielsimulation, die eine Übereinstimmung mit einer Regel anzeigen



Wenn Sie mehr als einen Satz in Ihren Simulationsdaten haben, können Sie sich vorwärts und rückwärts durch die Ergebnisse bewegen, indem Sie auf Weiter und Zurück klicken.

Wenn ein Satz mit keiner TLA-Regel in der ausgewählten Bibliothek übereinstimmt (siehe Bibliotheksname über dem Baum in dieser Registerkarte), werden die Ergebnisse als ohne Übereinstimmung betrachtet und die Schaltflächen Nächste Nicht-Übereinstimmung und Vorherige Nicht-Übereinstimmung werden aktiviert. So wissen Sie, dass Text vorhanden ist, für den keine Regel eine Übereinstimmung gefunden hat, und Sie können schnell zu diesen Textstellen wechseln.

Nachdem Sie neue Regeln erstellt, diese bearbeitet oder Ihre Ressourcen oder Extrahierungseinstellungen geändert haben, möchten Sie vielleicht eine erneute Simulation durchführen. Um eine Simulation erneut auszuführen, klicken Sie auf Simulation ausführen im Simulationsbereich und es werden die gleichen Eingabedaten noch einmal verwendet.

Es werden folgende Felder und Tabellen in den Simulationsergebnissen angezeigt:

Eingabetext. Der eigentliche "Satz", der durch den Extrahierungsprozess aus den Simulationsdaten identifiziert wurde, die Sie im Assistenten definiert haben. Unter einem Satz verstehen wir entweder ein Wort, einen Satz oder einen Satzteil, je nach dem, wie der Extraktor den Text in lesbare Abschnitte unterteilt hat.

Systemansicht. Eine Sammlung von Tokens, die durch den Extrahierungsprozess identifiziert wurden.

- **Eingabetext-Token.** Jedes im Eingabetext gefundene Token. Tokens wurden bereits weiter oben in diesem Thema definiert.
- **Typzuordnung.** Wenn ein Token als Konzept identifiziert und einem Typ zugeordnet wurde, wird der zugehörige Typname (z. B. `<Unbekannt>`, `<Person>`, `<Ort>`) in dieser Spalte angezeigt.
- **Makroübereinstimmung.** Wenn ein Token mit einem bestehenden Makro übereinstimmt, wird der zugehörige Makroname in dieser Spalte angezeigt.

Mit Eingabetext übereinstimmende Regeln. Diese Tabelle zeigt Ihnen alle TLA-Regeln an, mit denen der Eingabetext übereinstimmte. Für jede übereinstimmende Regel sehen Sie den Namen der Regel in der Spalte Regelausgabe und die zugehörigen Ausgabewerte für diese Regel (Paare "Konzept + Typ"). Sie können auf den übereinstimmenden Regelnamen doppelklicken, um die Regel im Editorbereich über dem Simulationsbereich zu öffnen.

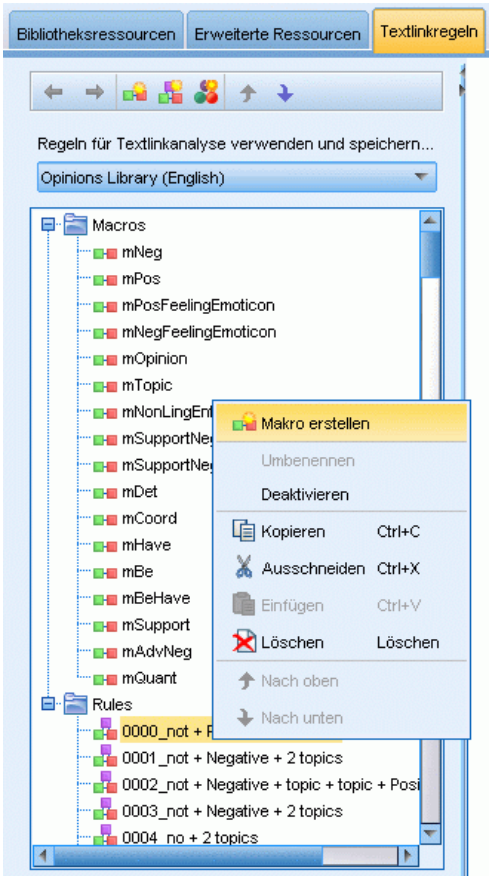
Schaltfläche **Regel generieren**. Wenn Sie auf diese Schaltfläche im Simulationsbereich klicken, wird eine neue Regel im Regeleditorbereich über dem Simulationsbereich geöffnet. Diese Regel wird den Eingabetext als Beispiel verwenden. Ebenso wird jedes Token, das während der Simulation einem Typ zugeordnet oder mit einem Makro abgeglichen wurde, automatisch in die Spalte “Element” in der Regelwert-Tabelle eingefügt. Wenn ein Token einem Typ zugeordnet *und* mit einem Makro abgeglichen wurde, wird der Makrowert zur Vereinfachung der Regel verwendet. Beispielsweise könnte der Satz “*Ich mag Pizza*” während der Simulation dem Typ `<Unbekannt>` zugeordnet und mit dem Makro `mTopic` abgeglichen worden sein, wenn Sie die Basic Resources English verwendet haben. In diesem Fall wird `mTopic` als das Element in der generierten Regel verwendet. Für weitere Informationen siehe Thema Arbeiten mit Textlinkregeln auf S. 361.

Navigation durch Regeln und Makros im Baum

Wenn die Textlinkanalyse während der Extrahierung durchgeführt wird, werden die Textlinkregeln verwendet, die in der ausgewählten Bibliothek in der RegisterkarteTextlinkregeln gespeichert sind.

Im Gegensatz zu den anderen erweiterten Ressourcen sind TLA-Regeln bibliothekenspezifisch. Sie können also jeweils nur die TLA-Regeln aus einer Bibliothek verwenden. Wechseln Sie in der Ansicht Template Editor oder Resource Editor auf die Registerkarte Textlinkregeln. Auf dieser Registerkarte können Sie die Bibliothek in Ihrer Vorlage angeben, die die TLA-Regeln enthält, die Sie verwenden oder bearbeiten möchten. Aus diesem Grund wird dringend empfohlen, dass Sie alle Ihre Regeln in einer Bibliothek speichern, es sei denn, dies wird aus einem wichtigen oder speziellen Grund nicht gewünscht.

Abbildung 19-8
Registerkarte "Textlinkregeln": Regel- und Makrobaum



Sie können auf der Registerkarte "Textlinkregeln" festlegen, in welcher Bibliothek Sie arbeiten möchten, indem Sie diese Bibliothek in der Dropdown-Liste Textlinkanalyseregeln verwenden und speichern in: auf dieser Registerkarte auswählen. Wenn die Textlinkanalyse während der Extrahierung durchgeführt wird, werden die Textlinkregeln verwendet, die in der ausgewählten Bibliothek in der RegisterkarteTextlinkregeln gespeichert sind. Wenn Sie daher Textlinkregeln (TLA-Regeln) in mehr als einer Bibliothek definiert haben, wird nur die erste Bibliothek, in der TLA-Regeln gefunden wurden, für die Textlinkanalyse verwendet. Aus diesem Grund wird dringend empfohlen, dass Sie alle Ihre Regeln in einer Bibliothek speichern, es sei denn, dies wird aus einem spezifischen Grund nicht gewünscht.

Wenn Sie ein Makro oder eine Regel im Baum auswählen, wird der Inhalt rechts im Editorbereich angezeigt. Wenn Sie mit der rechten Maustaste auf ein beliebiges Element im Baum klicken, öffnet sich ein Kontextmenü, über das Sie weitere Aufgaben durchführen können, zum Beispiel:

- Ein neues Makro im Baum öffnen und es rechts im Editor öffnen.
- Eine neue Regel im Baum öffnen und sie rechts im Editor öffnen.
- Ein neues Regelset im Baum öffnen.
- Elemente zur einfacheren Bearbeitung ausschneiden, kopieren und einfügen.

- Makros, Regeln und Regelsets löschen, um sie aus den Ressourcen zu entfernen.
- Makros, Regeln und Regelsets deaktivieren, um anzuzeigen, dass sie während der Verarbeitung ignoriert werden sollten.
- Regeln nach oben oder unten verschieben, um die Verarbeitungsreihenfolge zu ändern.

Warnungen im Baum

Warnungen werden mit einem gelben Dreieck in der Baumstruktur angezeigt und sollen Sie auf potenzielle Probleme hinweisen. Halten Sie den Mauszeiger über das fehlerhafte Makro bzw. die fehlerhafte Regel, um ein Popup-Fenster mit einer Erklärung anzuzeigen. In den meisten Fällen wird ein Text der folgenden Art angezeigt: Warnung: Es wurde kein Beispiel angegeben. Geben Sie ein Beispiel ein. Sie müssen also ein Beispiel eingeben.

Wenn ein Beispiel fehlt oder das Beispiel nicht der Regel entspricht, können Sie die Funktion "Token abrufen" nicht verwenden. Daher sollten Sie nur ein einziges Beispiel pro Regel eingeben.

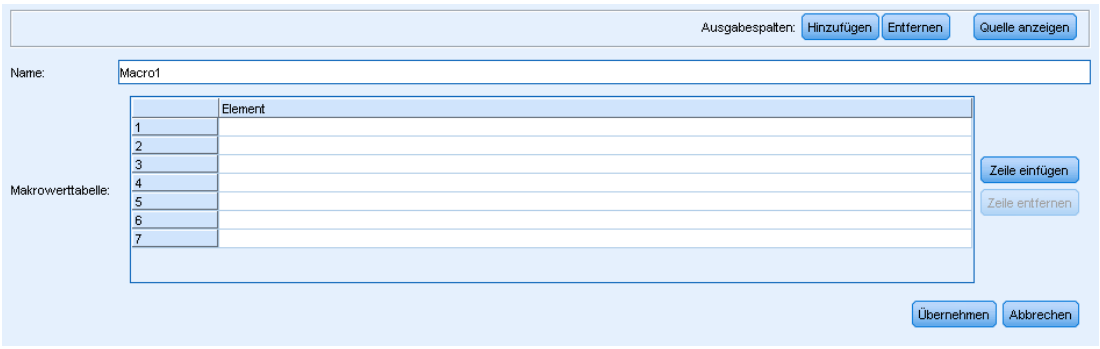
Wenn die Regel gelb markiert ist, bedeutet dies, dass ein Typ oder Makro dem TLA-Editor nicht bekannt ist. Die Meldung lautet in etwa: Warnung: Unbekannter Typ oder Makro.. Damit werden Sie darauf hingewiesen, dass ein Element, das durch $something in der Quellenansicht definiert würde, beispielsweise $myType, kein bestehender Typ in Ihrer Bibliothek und auch kein Makro ist.

Zur Aktualisierung der Syntaxprüfung müssen Sie zu einer anderen Regel bzw. einem anderen Makro wechseln. Es muss nichts neu kompiliert werden. Wenn also beispielsweise für Regel A eine Warnung angezeigt wird, da das Beispiel fehlt, müssen Sie ein Beispiel hinzufügen, entweder auf eine obere oder eine untere Regel klicken und anschließend zu Regel A zurückkehren, um zu überprüfen, ob sie nun korrekt ist.

Arbeiten mit Makros

Makros vereinfachen das Aussehen von Textlinkanalyseregeln, weil Sie hiermit Typen, andere Makros, Zeichenfolgen und Wortfolgen mit dem Operator `ODER` (`|`) verknüpfen können. Der Vorteil von Makros liegt darin, dass Sie nicht nur Makros in mehreren Textlinkanalyseregeln wiederverwenden können, um sie zu vereinfachen, sondern Sie auch Aktualisierungen in einem Makro vornehmen können, ohne sie in all Ihren Textlinkanalyseregeln vornehmen zu müssen. Die meisten mitgelieferten TLA-Regeln enthalten vordefinierte Makros. Makros erscheinen im oberen Bereich des Baums ganz links auf der Registerkarte "Textlinkregeln".

Abbildung 19-9
Registerkarte "Textlinkregeln": Makroeditor



Es werden folgende Felder und Tabellen in den Simulationsergebnissen angezeigt:

Name. Ein eindeutiger Name, der dieses Makro identifiziert. Es wird empfohlen, dass Sie vor den Makronamen den kleingeschriebenen Buchstaben "m" setzen, damit Sie in Ihren Regeln Makros schnell identifizieren können. Wenn Sie manuell auf Makros in Ihren Regeln verweisen (durch Bearbeiten in der Zeile oder in der Quellenansicht), müssen Sie das Zeichen $ voranstellen, damit der Extrahierungsprozess auf diesen besonderen Namen achtet. Wenn Sie den Makronamen jedoch ziehen und ablegen oder ihn über die Kontextmenüs hinzufügen, wird das Produkt ihn automatisch als Makro erkennen und es wird kein $ angefügt.

Tabelle **Makrowert**.

- Eine bestimmte Anzahl von Zeilen, in denen alle möglichen Werte angezeigt werden, die dieses Makro darstellen kann. Bei diesen Werten wird zwischen Groß- und Kleinschreibung unterschieden.
- Diese Werte können einen oder eine Kombination aus Typen, Zeichenfolgen, Wortlücken oder Makros enthalten. Für weitere Informationen siehe Thema Unterstützte Elemente für Regeln und Makros auf S. 369.
- Um in einem Makro einen Wert für ein Element einzugeben, doppelklicken Sie auf die Zeile, in der Sie arbeiten möchten. Es erscheint ein bearbeitbares Textfeld, in das Sie eine Typreferenz, eine Makroreferenz, eine Zeichenfolge oder eine Wortlücke eingeben können. Klicken Sie alternativ mit der rechten Maustaste in die Zelle, um ein Kontextmenü zu öffnen, das übliche Makros, Typennamen und nichtlinguistische Typennamen auflistet. Um auf einen Typ oder ein Makro zu verweisen, müssen Sie dem Makro- oder Typnamen ein '`$`'-Zeichen voranstellen, wie beispielsweise in `$mTopic` für das Makro `mTopic`. Um Argumente zu kombinieren, müssen Sie Klammern `( )` verwenden, um die Argumente zu einer Gruppe zusammenzufassen, und das Zeichen `|` verwenden, um den Boole'schen Operator `ODER` zu kennzeichnen.
- Sie können Zeilen in der Tabelle "Makrowert" mithilfe der Schaltflächen im rechten Bereich hinzufügen oder entfernen.
- Geben Sie jedes Element in seiner eigenen Zeile ein. Wenn Sie zum Beispiel ein Makro erstellen möchten, das eine von drei Zeichenfolgen wie `bin ODER war ODER ist` darstellt, würden Sie jede Zeichenfolge in einer separaten Zeile in der Ansicht eingeben und Ihre Makrotabelle würde aus drei Zeilen bestehen.

Erstellen und Bearbeiten von Makros

Sie können neue Makros erstellen oder bestehende bearbeiten. Folgen Sie den Anweisungen und Beschreibungen für den Makroeditor. Für weitere Informationen siehe Thema Arbeiten mit Makros auf S. 356.

Erstellen neuer Makros

- Wählen Sie im Menü Extras > Neues Makro aus. Klicken Sie alternativ auf das Symbol “Neues Makro” in der Baumsymbolleiste, um ein neues Makro im Editor zu öffnen.
- Geben Sie einen eindeutigen Namen ein und definieren Sie die Makrowertelemente.
- Klicken Sie, wenn Sie fertig sind, auf Übernehmen, um nach möglichen Fehlern zu suchen.

Bearbeiten von Makros

- Klicken Sie auf den Makronamen im Baum. Das Makro öffnet sich rechts im Editorbereich.
- Nehmen Sie die gewünschten Änderungen vor.
- Klicken Sie, wenn Sie fertig sind, auf Übernehmen, um nach möglichen Fehlern zu suchen.

Deaktivieren und Löschen von Makros

Deaktivieren von Makros

Wenn Sie möchten, dass ein Makro während der Verarbeitung ignoriert wird, können Sie es deaktivieren. Dies kann in Regeln, die noch auf dieses deaktivierte Makro verweisen, zu Warn- oder Fehlermeldungen führen. Lassen Sie beim Löschen und Deaktivieren von Makros Vorsicht walten.

- Klicken Sie auf den Makronamen im Baum. Das Makro öffnet sich rechts im Editorbereich.
- Klicken Sie mit der rechten Maustaste auf den Namen.
- Wählen Sie in den Kontextmenüs Deaktivieren. Das Makrosymbol wird grau und das Makro selbst kann nicht mehr bearbeitet werden.

Löschen von Makros

Wenn Sie ein Makro nicht mehr benötigen, können Sie es löschen. Dies kann in Regeln, die noch auf dieses Makro verweisen, zu Fehlermeldungen führen. Lassen Sie beim Löschen und Deaktivieren von Makros Vorsicht walten.

- Klicken Sie auf den Makronamen im Baum. Das Makro öffnet sich rechts im Editorbereich.
- Klicken Sie mit der rechten Maustaste auf den Namen.
- Wählen Sie in den Kontextmenüs Löschen. Das Makro wird aus der Liste entfernt.

Fehlersuche, Speichern und Abbrechen

Makroänderungen übernehmen

Wenn Sie auf eine Stelle außerhalb des Makroeditors oder auf Übernehmen klicken, wird das Makro automatisch nach Fehlern durchsucht. Wird ein Fehler gefunden, müssen Sie diesen beheben, bevor Sie in einen anderen Bereich der Anwendung wechseln.

Wenn jedoch weniger schwerwiegende Fehler gefunden werden, erscheint nur eine Warnmeldung. Wenn Ihr Makro zum Beispiel unvollständige Definitionen oder Definitionen ohne Verweis auf Typen oder andere Makros enthält, wird eine Warnmeldung angezeigt. Wenn Sie auf Übernehmen klicken, erscheint bei allen nicht behobenen Warnungen ein Warnsymbol links neben dem Makronamen im Regel- und Makrobaum im linken Bereich.

Durch das Übernehmen eines Makros wird Ihr Makro nicht permanent gespeichert. Durch das Übernehmen wird im Validierungsprozess nach Fehlern und Warnungen gesucht.

Speichern von Ressourcen in einer interaktiven Workbench-Sitzung

- So speichern Sie die Änderungen an Ihren Ressourcen während einer interaktiven Workbench-Sitzung, damit Sie sie aufrufen können, wenn Sie das nächste Mal Ihren Stream ausführen:

  Aktualisieren Sie Ihren Modellierungsknoten, um sicherzustellen, dass Sie das nächste Mal, wenn Sie Ihren Stream ausführen, auf die gleichen Ressouren zurückgreifen können. Für weitere Informationen siehe Thema Aktualisieren von Modellierungsknoten und Speichern in Kapitel 8 auf S. 140. Speichern Sie anschließend Ihren Stream. Speichern Sie Ihren Stream im Hauptbereich von IBM® SPSS® Modeler nach der Aktualisierung des Modellierungsknotens.

- So speichern Sie die Änderungen an Ihren Ressourcen während einer interaktiven Workbench-Sitzung, damit Sie sie in anderen Streams verwenden können:
  - Aktualisieren Sie die verwendete Vorlage oder erstellen Sie eine neue. Für weitere Informationen siehe Thema Erstellen und Aktualisieren von Vorlagen in Kapitel 14 auf S. 276. Dadurch werden die Änderungen für den aktuellen Knoten nicht übernommen (siehe vorheriger Schritt).
  - Alternativ können Sie auch das verwendete TAP aktualisieren. Für weitere Informationen siehe Thema Aktualisierung von Text Analysis Packages in Kapitel 10 auf S. 235.

Speichern von Ressourcen in Template Editor

- Veröffentlichen Sie zunächst die Bibliothek. Für weitere Informationen siehe Thema Bibliotheken veröffentlichen in Kapitel 16 auf S. 307.

- Speichern Sie anschließend die Vorlage über die Menüfolge Datei > Ressourcenvorlage speichern.

Makroänderungen abbrechen

- Wenn Sie die Änderungen verwerfen möchten, klicken Sie auf Abbrechen.

Spezielle Makros: mTopic, mNonLingEntities, SEP

Die Vorlage “Meinungen” (und ähnliche Vorlagen) sowie die Vorlagen “Basic Resources” werden mit zwei Spezialmakros geliefert: `mTopic` und `mNonLingEntities`.

mTopic

Standardmäßig fasst das Makro `mTopic` alle Typen, die in der Vorlage mitgeliefert werden und wahrscheinlich mit einer Meinung verbunden werden, in Gruppen zusammen. Das gilt beispielsweise für folgende *Kern*bibliothekstypen: `<Person>`, `<Organization>`, `<Location>` usw., solange der Typ kein Meinungstyp (z. B. `<Negative>` oder `<Positive>`) oder ein Typ ist, der in den erweiterten Ressourcen als nicht-linguistische Einheit definiert wurde.

Immer wenn Sie einen neuen Typ in einer Meinungsvorlage (oder einer ähnlichen Vorlage) definieren, behandelt das Produkt diesen Typ auf dieselbe Weise wie die anderen Typen, die im Makro `mTopic` definiert sind, es sei denn, dieser Typ ist in einem anderen Makro oder im Abschnitt für nichtlinguistische Einheiten in der Registerkarte “Erweiterte Ressourcen” angegeben.

Angenommen, Sie haben über eine Vorlage “Meinungen” neue Typen in den Ressourcen erstellt: `<Gemüse>` und `<Obst>`. Ohne dass Änderungen nötig sind, werden Ihre neuen Typen als `mTopic`-Typen behandelt, so dass Sie automatisch die positiven, negativen, neutralen und kontextbezogenen Meinungen über Ihre neuen Typen erfassen können. Bei der Extrahierung würde der Satz “*Ich mag Brokkoli, aber Grapefruit mag ich nicht.*” folgende zwei Ausgabemuster erzeugen:

```
brokkoli <Gemüse> + mag <Positive>
grapefruit <Obst> + mag nicht <Negative>
```

Wenn Sie jedoch diese Typen anders als die anderen Typen in mTopic verarbeiten möchten, können Sie entweder dem bestehenden Makro den Typnamen wie `mPos` hinzufügen, durch den alle positiven Meinungstypen gruppiert werden, oder ein neues Makro erstellen, auf das Sie später in einer oder mehreren Regeln verweisen können.

Wichtig: Wenn Sie einen neuen Typ wie `<Gemüse>` erstellen, wird dieser Typ als ein Typ in mTopic aufgenommen, jedoch ist dieser Typname nicht explizit in der Makrodefinition sichtbar.

mNonLingEntities

Wenn Sie neue nicht linguistische Elemente in den Abschnitt Nicht linguistische Entitäten der Registerkarte “Erweiterte Ressourcen” einfügen, werden sie ebenfalls automatisch als `mNonLingEntities` verarbeitet, wenn nicht anders angegeben. Für weitere Informationen siehe Thema Nicht linguistische Elemente in Kapitel 18 auf S. 337.

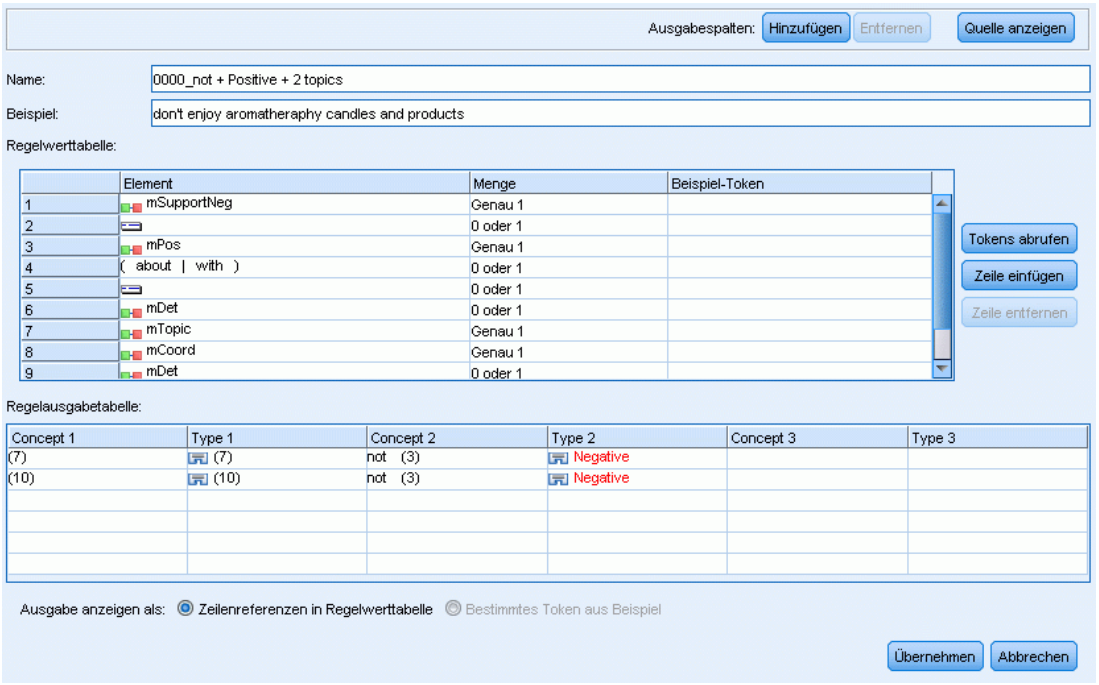
SEP

Sie können auch das vordefinierte Makro `SEP` verwenden, das dem globalen Separator entspricht, der auf dem lokalen Rechner definiert ist, in der Regel ein Komma (,).

Arbeiten mit Textlinkregeln

Eine Textlinkanalyseregel ist eine Boole'sche Abfrage, die für den Abgleich eines Satzes eingesetzt wird. Textlinkanalyseregeln enthalten mindestens eines der folgenden Argumente: Typen, Makros, Zeichenfolgen oder Wortlücken. Sie müssen mindestens eine Textlinkanalyseregel definiert haben, um TLA-Ergebnisse extrahieren zu können.

Abbildung 19-10
Registerkarte "Textlinkregeln": Regeleditor



Feld **Name**. Ein eindeutiger Name für die Textlinkregel.

Feld **Beispiel**. Sie können optional einen Beispielsatz oder eine Wortfolge einschließen, die von dieser Regel erfasst werden würde. Es wird empfohlen, Beispiele zu verwenden. In diesem Editor können Sie Tokens aus diesem Beispieltext erzeugen, um zu sehen, wie dieser mit der Regel übereinstimmt und wie er ausgegeben wird. Ein **Token** wird als beliebiges Word oder beliebige Wortfolge definiert, das/die während des Extrahierungsprozesses identifiziert wurde. Zum Beispiel können in dem Satz *Mein Onkel lebt in New York* die folgenden Tokens während der Extrahierung gefunden werden: *mein*, *Onkel*, *lebt*, *in* und *New York*. Weiterhin könnte *Onkel* als Konzept des Typs `<Unknown>` extrahiert werden und *New York* als Konzept des Typs `<Location>`. Alle Konzepte sind Tokens, doch nicht alle Tokens sind Konzepte. Tokens können auch Makros, Zeichenfolgen und Wortlücken sein. Nur einem Typ zugeordnete Wörter oder Wortfolgen können Konzepte sein.

Regelwert-Tabelle. Diese Tabelle enthält die Elemente der Regel, die zum Abgleich einer Regel mit einem Satz verwendet werden. Sie können Zeilen in der Tabelle mithilfe der Schaltflächen im rechten Bereich hinzufügen oder entfernen. Die Tabelle besteht aus drei Spalten:

- Spalte **Element**. Geben Sie Werte als einen oder als Kombination von Typen, Zeichenfolgen, Wortlücken (`<Beliebiges Token>`) oder Makros ein. Für weitere Informationen siehe Thema Unterstützte Elemente für Regeln und Makros auf S. 369. Doppelklicken Sie auf die Elementzelle, um die Informationen direkt einzugeben. Klicken Sie alternativ mit der rechten Maustaste in die Zelle, um ein Kontextmenü zu öffnen, das übliche Makros, Typennamen und nichtlinguistische Typennamen auflistet. Beachten Sie, dass Sie bei der manuellen Eingabe von Daten in die Zelle dem Makro- oder Typnamen ein '`$`'-Zeichen voranstellen müssen, wie beispielsweise in `$mTopic` für das Makro `mTopic`. Durch die Reihenfolge, in der Sie Ihre Elementzeilen erstellen, wird die Art bestimmt, in der die Regel mit dem Text abgeglichen wird. Um Argumente zu kombinieren, müssen Sie Klammern `( )` verwenden, um die Argumente zu einer Gruppe zusammenzufassen, und das Zeichen `|` verwenden, um den Boole'schen Operator `ODER` zu kennzeichnen. Beachten Sie, dass bei den Werten zwischen Groß- und Kleinschreibung unterschieden wird.
- Spalte **Menge**. Gibt an, wie oft das Element mindestens und höchstens gefunden werden muss, damit es zu einer Übereinstimmung kommt. Wenn Sie zum Beispiel eine Lücke oder eine Reihe von Wörtern zwischen zwei anderen Elementen aus 0 bis 3 Wörtern definieren möchten, könnten Sie Zwischen 0 und 3 aus der Liste auswählen oder die Zahlen direkt im Dialogfeld eingeben. Die Standardeinstellung ist 'Genau 1'. In manchen Fällen möchten Sie vielleicht ein Element als optional definieren. In diesem Fall erhält das Element die Mindestmenge 0 und eine Höchstmenge größer als 0 (d. h. 0 oder 1, zwischen 0 und 2). Beachten Sie, dass das erste Element in einer Regel nicht optional sein kann, d. h., es kann nicht die Menge 0 besitzen.
- Spalte **Beispiel-Token**. Wenn Sie auf Tokens abrufen klicken, teilt das Programm den Beispieltext in Tokens ein und fügt jene Tokens in diese Spalte ein, die mit den von Ihnen definierten Elementen übereinstimmen. Sie können, falls gewünscht, diese Tokens auch in der Ausgabetabelle ansehen.

Tabelle **Regelausgabe**. Jede Zeile in dieser Tabelle legt fest, wie die TLA-Musterausgabe in den Ergebnissen erscheint. Die Regelausgabe kann Muster aus bis zu sechs Konzept-/Typ-Spaltenpaaren erzeugen, von denen jedes einen *Slot* repräsentiert. Beispielsweise ist das Typmuster `<Location> + <Positive>` ein Muster aus zwei Slots, d. h., es besteht aus zwei Konzept-/Typ-Spaltenpaaren.

Ebenso wie uns Sprache die Freiheit gibt, die gleichen grundlegenden Ideen und Vorstellungen auf unterschiedliche Arten auszudrücken, haben Sie vielleicht einige Regeln definiert, um die gleiche Grundidee zu erfassen. Zum Beispiel vertreten die Sätze *"Paris ist ein Ort, der mir gefällt"* und *"Mir gefallen Paris und Florenz wirklich sehr gut"* die gleiche Grundidee (der Person gefällt Paris), doch diese wird unterschiedlich ausgedrückt und es wären zwei verschiedene Regeln erforderlich, um beide Sätze zu erfassen. Es ist jedoch einfacher, mit den Musterergebnissen zu arbeiten, wenn ähnliche Ideen zusammengefasst werden. Aus diesem Grund könnten Sie zwei unterschiedliche Regeln zur Erfassung dieser beiden Sätze verwenden, aber auch dieselbe Ausgabe für beide Regeln definieren, damit das Typmuster `<Location> + <Positive>` beide Texte erfassen würde. So stellen Sie fest, dass die Ausgabe nicht immer der Struktur oder Reihenfolge der Wörter ähnelt, die im Originaltext gefunden wurde. Des Weiteren könnte ein

solches Typmuster anderen Sätzen entsprechen und Konzeptmuster wie das folgende erzeugen: `paris + gefällt` und `tokio + gefällt`.

Damit Sie die Ausgabe schnell und mit weniger Fehlern definieren können, verwenden Sie das Kontextmenü, um das Element zu wählen, das Sie in der Ausgabe sehen möchten. Alternativ können Sie auch Elemente aus der Regelwert-Tabelle ziehen und in der Ausgabe ablegen. Wenn Sie beispielsweise eine Regel definiert haben, die eine Referenz zu dem Makro `mTopic` in Zeile 2 der Regelwert-Tabelle enthält, und Sie möchten, dass dieser Wert in Ihrer Ausgabe erscheint, können Sie einfach das Element für `mTopic` ziehen und im ersten Spaltenpaar in der Tabelle "Regelausgabe" ablegen. Dadurch werden automatisch die Spalten "Konzept" und "Typ" mit dem von Ihnen ausgewählten Paar ausgefüllt. Wenn Sie möchten, dass die Ausgabe mit dem Typ beginnt, der durch das dritte Element (Zeile 3) der Regelwert-Tabelle definiert wurde, dann ziehen Sie diesen Typ aus der Regelwert-Tabelle in die Zelle Typ 1 in der Ausgabetabelle. Die Tabelle wird aktualisiert und zeigt nun die Zeilenreferenz in Klammern an (3).

Alternativ können Sie diese Referenzen auch manuell in der Tabelle eingeben, indem Sie auf die Zelle in jeder Konzept-Spalte doppelklicken, die in der Ausgabe erscheinen soll und das Zeichen `$` gefolgt von der Zeilennummer eingeben, z. B. `$2`, um auf das in Zeile 2 der Regelwert-Tabelle definierte Element zu verweisen. Wenn Sie die Informationen manuell eingeben, müssen Sie auch die Typ-Spalte definieren und das `#`-Symbol gefolgt von der Zeilennummer eingeben, zum Beispiel `#2`, um auf das in Zeile 2 der Regelwert-Tabelle definierte Element zu verweisen.

Weiterhin können Sie sogar Verfahren kombinieren. Angenommen, in Ihrer Regelwert-Tabelle steht der Typ `<Positive>` in Zeile 4. Sie könnten ihn in die Spalte `Typ 2` ziehen und dann auf die Zelle in der Spalte `Konzept 2` doppelklicken und dann manuell das Wort "*nicht*" davor eingeben. In der Ausgabe-Spalte der Tabelle würde dann `nicht (4)` stehen oder `nicht $4`, wenn Sie sich im Bearbeitungs- oder Quellenmodus befänden. Anschließend könnten Sie mit der rechten Maustaste auf die Spalte "Typ 1" klicken und beispielsweise das Makro `mTopic` auswählen. Diese Ausgabe könnte dann zum Beispiel folgendes Konzeptmuster erzeugen: `Auto + schlecht`.

Die meisten Regeln haben nur eine Ausgabezeile, aber gelegentlich ist mehr als eine Ausgabe möglich oder erwünscht. Definieren Sie in diesem Fall eine Ausgabe pro Zeile in der Regelausgabe-Tabelle.

Wichtig: Denken Sie daran, dass andere linguistische Operationen bei der Extrahierung von TLA-Mustern stattfinden. Wenn die Ausgabe also `t$3\t#3` lautet, bedeutet dies, dass das Muster schließlich das endgültige Konzept für das dritte Element und den endgültigen Typ für das dritte Element anzeigt, sobald sämtliche linguistische Verarbeitung durchgeführt wurde (Synyme und andere Gruppierungen).

- **Ausgabe anzeigen als.** Standardmäßig ist die Option Referenz auf Zeile in Regelwert-Tabelle ausgewählt und die Ausgabe wird durch numerische Referenzen auf die Zeile angezeigt, so wie es auf der Registerkarte "Regelwert" definiert wurde. Wenn Sie zuvor auf "Tokens abrufen" geklickt haben und Tokens in der Spalte "Beispiel-Tokens" in der Regelwert-Tabelle vorhanden sind, können Sie die Ausgabe für diese jeweiligen Tokens ansehen, indem Sie die Option auswählen.

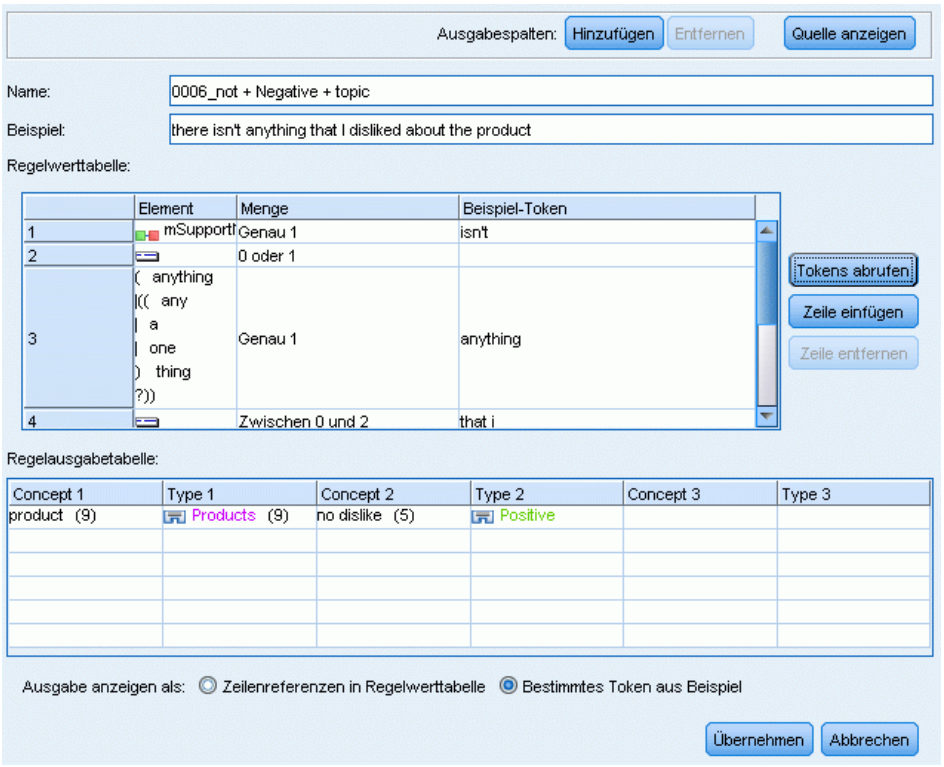
Anmerkung: Wenn nicht genug Konzept-/Typ-Ausgabepaare in der Ausgabetabelle vorhanden sind, können Sie ein weiteres Paar hinzufügen, indem Sie auf die Schaltfläche "Hinzufügen" in der Symbolleiste des Editors klicken. Wenn zum Beispiel aktuell drei Paare angezeigt werden und Sie auf "Hinzufügen" klicken, werden zwei weitere Spalten (Konzept 4 und Typ 4) zur Tabelle hinzugefügt. Das bedeutet, dass Sie nun vier Paare in der Ausgabetabelle für alle Regeln sehen.

Sie können auch nicht verwendete Paare entfernen, solange dieses Paar von keiner anderen Regel im Regelset dieser Bibliothek verwendet wird.

Beispielregel

Angenommen, Ihre Ressourcen enthalten die folgende Textlinkanalyseregel und Sie haben die Extrahierung von TLA-Ergebnissen aktiviert:

Abbildung 19-11
Registerkarte "Textlinkregeln": Regeleditor



Bei der Extrahierung liest die Extrahierungsengine jeden Satz und versucht, die folgende Sequenz abzugleichen:

Element (Zeile)	Beschreibung der Argumente
1	Das Konzept aus einem der durch die Makros `mPos` oder `mNeg` dargestellten Typen oder aus dem Typ `<Ungeklärt>`.
2	Ein Konzept, das einem der durch das Makro `mTopic` dargestellten Typen zugeordnet wurde.
3	Eines der durch das Makro `mBe` dargestellten Wörter.
4	Ein optionales Element, 0 oder 1 Wörter, auch als Wortlücke oder `<Beliebiges Token>` bezeichnet.
5	Ein Konzept, das einem der durch das Makro `mTopic` dargestellten Typen zugeordnet wurde.

Die Ausgabetabelle zeigt, dass von dieser Regel lediglich ein Muster verlangt wird: ein beliebiges Konzept oder ein beliebiger Typ, der dem Makro `mTopic` entspricht, das in Zeile 5 der Regelwert-Tabelle definiert wurde, + ein beliebiges Konzept oder ein beliebiger Typ, der `mPos`,

`mNeg` oder `<Uncertain>` entspricht, wie in Zeile 1 der Regelwert-Tabelle definiert. Beispiel: `Wurst + mag` oder `<Unknown> + <Positive>`.

Erstellen und Bearbeiten von Regeln

Sie können neue Regeln erstellen oder bestehende bearbeiten. Folgen Sie den Anweisungen und Beschreibungen für den Regeleditor. Für weitere Informationen siehe Thema Arbeiten mit Textlinkregeln auf S. 361.

Erstellen neuer Regeln

- Wählen Sie im Menü Extras > Neue Regel aus. Klicken Sie alternativ auf das Symbol "Neue Regel" in der Baumsymbolleiste, um eine neue Regel im Editor zu öffnen.
- Geben Sie einen eindeutigen Namen ein und definieren Sie die Regelwertelemente.
- Klicken Sie, wenn Sie fertig sind, auf Übernehmen, um nach möglichen Fehlern zu suchen.

Bearbeiten von Regeln

- Klicken Sie auf den Regelnamen im Baum. Die Regel öffnet sich rechts im Editorbereich.
- Nehmen Sie die gewünschten Änderungen vor.
- Klicken Sie, wenn Sie fertig sind, auf Übernehmen, um nach möglichen Fehlern zu suchen.

Deaktivieren und Löschen von Regeln

Deaktivieren von Regeln

Wenn Sie möchten, dass eine Regel während der Verarbeitung ignoriert wird, können Sie sie deaktivieren. Lassen Sie beim Löschen und Deaktivieren von Regeln Vorsicht walten.

- Klicken Sie auf den Regelnamen im Baum. Die Regel öffnet sich rechts im Editorbereich.
- Klicken Sie mit der rechten Maustaste auf den Namen.
- Wählen Sie in den Kontextmenüs Deaktivieren. Das Regelsymbol wird grau und die Regel selbst kann nicht mehr bearbeitet werden.

Löschen von Regeln

Wenn Sie eine Regel nicht mehr benötigen, können Sie sie löschen. Lassen Sie beim Löschen und Deaktivieren von Regeln Vorsicht walten.

- Klicken Sie auf den Regelnamen im Baum. Die Regel öffnet sich rechts im Editorbereich.
- Klicken Sie mit der rechten Maustaste auf den Namen.
- Wählen Sie in den Kontextmenüs Löschen. Die Regel wird aus der Liste entfernt.

Fehlersuche, Speichern und Abbrechen

Regeländerungen übernehmen

Wenn Sie auf eine Stelle außerhalb des Regeleditors oder auf Übernehmen klicken, wird die Regel automatisch nach Fehlern durchsucht. Wird ein Fehler gefunden, müssen Sie diesen beheben, bevor Sie in einen anderen Bereich der Anwendung wechseln.

Wenn jedoch weniger schwerwiegende Fehler gefunden werden, erscheint nur eine Warnmeldung. Wenn Ihre Regel zum Beispiel unvollständige Definitionen oder Definitionen ohne Verweis auf Typen oder Makros enthält, wird eine Warnmeldung angezeigt. Wenn Sie auf Übernehmen klicken, erscheint bei allen nicht behobenen Warnungen ein Warnsymbol links neben dem Regelnamen links im Baum.

Durch das Übernehmen einer Regel wird Ihre Regel nicht permanent gespeichert. Durch das Übernehmen wird im Validierungsprozess nach Fehlern und Warnungen gesucht.

Speichern von Ressourcen in einer interaktiven Workbench-Sitzung

- So speichern Sie die Änderungen an Ihren Ressourcen während einer interaktiven Workbench-Sitzung, damit Sie sie aufrufen können, wenn Sie das nächste Mal Ihren Stream ausführen:

  Aktualisieren Sie Ihren Modellierungsknoten, um sicherzustellen, dass Sie das nächste Mal, wenn Sie Ihren Stream ausführen, auf die gleichen Ressouren zurückgreifen können. Für weitere Informationen siehe Thema Aktualisieren von Modellierungsknoten und Speichern in Kapitel 8 auf S. 140. Speichern Sie anschließend Ihren Stream. Speichern Sie Ihren Stream im Hauptbereich von IBM® SPSS® Modeler nach der Aktualisierung des Modellierungsknotens.

- So speichern Sie die Änderungen an Ihren Ressourcen während einer interaktiven Workbench-Sitzung, damit Sie sie in anderen Streams verwenden können:
  - Aktualisieren Sie die verwendete Vorlage oder erstellen Sie eine neue. Für weitere Informationen siehe Thema Erstellen und Aktualisieren von Vorlagen in Kapitel 14 auf S. 276. Dadurch werden die Änderungen für den aktuellen Knoten nicht übernommen (siehe vorheriger Schritt).
  - Alternativ können Sie auch das verwendete TAP aktualisieren. Für weitere Informationen siehe Thema Aktualisierung von Text Analysis Packages in Kapitel 10 auf S. 235.

Speichern von Ressourcen in Template Editor

- Veröffentlichen Sie zunächst die Bibliothek. Für weitere Informationen siehe Thema Bibliotheken veröffentlichen in Kapitel 16 auf S. 307.

- Speichern Sie anschließend die Vorlage über die Menüfolge Datei > Ressourcenvorlage speichern.

Regeländerungen abbrechen

- Wenn Sie die Änderungen verwerfen möchten, klicken Sie im Editorbereich auf Abbrechen.

Verarbeitungsreihenfolge für Regeln

Wenn während der Extrahierung eine Textlinkanalyse ausgeführt wird, wird ein “Satz” (Klausel, Wort, Phrase) nacheinander mit jeder Regel abgeglichen, bis eine Übereinstimmung gefunden wird oder alle Regeln abgearbeitet wurden. Die Position im Baum bestimmt die Reihenfolge, in der Regeln angewendet werden. Es wird empfohlen, dass Sie Ihre Regeln absteigend von der genauesten bis zur allgemeinsten Regel sortieren. Die genauesten Regeln sollten im oberen Bereich des Baums erscheinen. Um die Reihenfolge einer genauen Regel oder eines Regelsets zu ändern, wählen Sie aus dem Kontextmenü des Regel- und Makrobaums Nach oben verschieben oder Nach unten verschieben aus oder verwenden Sie die Schaltflächen “Pfeil-nach-oben” und “Pfeil-nach-unten” in der Symbolleiste.

Wenn Sie sich in der *Quellenansicht* befinden, können Sie die Reihenfolge der Regeln durch Verschieben im Editor nicht ändern. Je weiter oben die Regel in der Quellenansicht erscheint, desto früher wird sie verarbeitet. Es wird dringend empfohlen, Regeln nur im Baum neu zu sortieren, um Probleme beim Kopieren/Einfügen zu vermeiden.

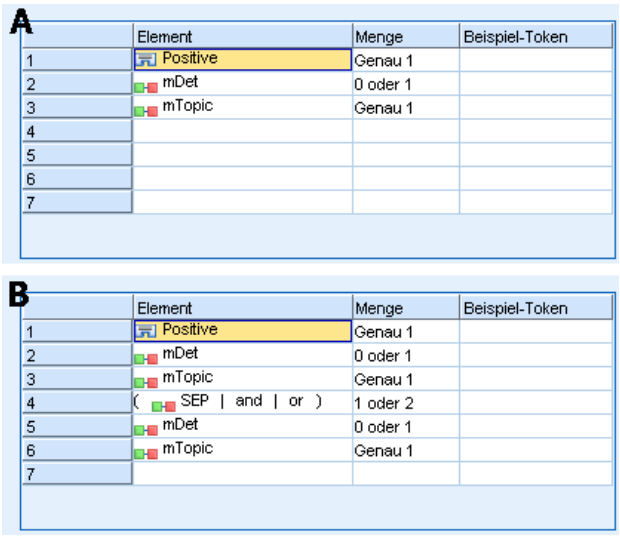
Wichtig: In früheren Versionen von IBM® SPSS® Modeler Text Analytics wurde eine eindeutige, numerische Regel-ID benötigt. Ab Version 15 können Sie die Verarbeitungsreihenfolge nur bestimmen, indem Sie eine Regel im Baum nach oben oder unten verschieben, oder durch ihre Position in der Quellenansicht.

Angenommen, Ihr Text enthält die folgenden beiden Sätze:

Ich mag Sardellen.
Ich mag Sardellen und grüne Paprika.

Nehmen wir zudem an, dass zwei Textlinkanalyseregeln vorhanden sind, die folgende Werte aufweisen:

Abbildung 19-12
2 Beispielregeln



In der Quellenansicht können die Regelwerte wie folgt aussehen:

A: `value = $Positive $mDet?  $mTopic`

B: `value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

Wenn sich Regel **A** weiter oben im Baum befindet als Regel **B**, dann wird Regel **A** zuerst verarbeitet und der Satz *Ich mag Sardellen und grüne Paprika* wird zuerst durch `$Positive $mDet? $mTopic` abgeglichen. Dadurch entsteht eine unvollständige Musterausgabe (`Sardellen + mag`), da der Satz mit einer Regel abgeglichen wurde, die nicht nach zwei `$mTopic`-Übereinstimmungen gesucht hat.

Um die wahre Bedeutung des Texts zu erfassen, muss sich die genauere Regel, in diesem Fall Regel **B**, weiter oben im Baum befinden als die allgemeinere Regel, in diesem Fall Regel **A**.

Arbeiten mit Regelsets (Mehrere Durchgänge)

Ein Regelset ist eine hilfreiche Möglichkeit, zusammengehörende Regeln im Regel- und Makrobaum zusammenzufassen und mehrere Durchgänge zu verarbeiten. Ein Regelset trägt lediglich einen Namen und hat ansonsten keine Definition. Es wird verwendet, um Ihre Regeln in aussagekräftige Gruppen einzuteilen. In einigen Kontexten ist der Text zu inhaltsreich und vielfältig, um in einem Durchgang verarbeitet werden zu können. Wenn Sie beispielsweise mit Sicherheitsdaten arbeiten, kann der Text Verbindungen zwischen einzelnen Personen enthalten, die über Kontaktmethoden (*x rief y an*), Verwandtschaftsgrade (*ys Schwager x*), den Austausch von Geld (*x überwies $100 an y*) usw. erkannt werden. In diesem Fall ist es hilfreich, einen besonderen Satz von Textlinkanalyseregeln zu erstellen, von denen sich jede auf eine bestimmte Beziehung wie beispielsweise die Erkennung von Kontakten, von Verwandtschaftsgraden usw. konzentriert.

Um ein Regelset zu erstellen, wählen Sie “Regelset erstellen” im Kontextmenü des Regel- und Makrobaums oder in der Symbolleiste. Sie können anschließend direkt unter einem Regelset-Knoten im Baum neue Regeln erstellen oder vorhandene Regeln in ein Regelset verschieben.

Wenn Sie eine Extrahierung mithilfe von Ressourcen durchführen, in denen die Regeln in Regelsets zusammengefasst sind, ist die Extrahierungs-Engine gezwungen, mehrere Durchgänge durch den Text durchzunehmen, um in den einzelnen Durchgängen unterschiedliche Muster abzugleichen. So kann ein “Satz” mit einer Regel in jedem Regelset abgeglichen werden, während er ohne Regelset nur mit einer einzigen Regel abgeglichen werden könnte.

Anmerkung: Sie können bis zu 512 Regeln in ein Regelset aufnehmen.

Erstellen neuer Regelsets

- Wählen Sie im Menü Extras > Neues Regelset aus. Klicken Sie alternativ auf das Symbol “Neues Regelset” in der Symbolleiste des Baums. Es erscheint ein Regelset im Regelbaum.
- Fügen Sie dieser Regel neue Regeln hinzu oder verschieben Sie bestehende Regeln in das Set.

Deaktivieren von Regelsets

- Klicken Sie mit der rechten Maustaste auf den Regelsetnamen im Baum.

- Wählen Sie in den Kontextmenüs Deaktivieren. Das Regelsetsymbol wird grau und alle Regeln in diesem Regelset werden ebenfalls deaktiviert und während der Verarbeitung ignoriert.

Löschen von Regelsets

- Klicken Sie mit der rechten Maustaste auf den Regelsetnamen im Baum.
- Wählen Sie in den Kontextmenüs Löschen. Das Regelset und alle darin enthaltenen Regeln werden aus den Ressourcen gelöscht.

Unterstützte Elemente für Regeln und Makros

Folgende Argumente werden für die Werteparameter in Textlinkanalyseregeln und Makros akzeptiert:

Makros

Sie können ein Makro direkt in einer Textlinkanalyseregel oder in einem andere Makro verwenden. Wenn Sie den Makronamen manuell oder in der Quellenansicht eingeben (anstatt den Makronamen aus einem Kontextmenü auszuwählen), stellen Sie sicher, dass Sie dem Namen ein Dollar-Zeichen (`$`) voranstellen, z. B. `$mTopic`. Bei dem Makronamen wird zwischen Groß- und Kleinschreibung unterschieden. Wenn Sie Makros über die Kontextmenüs auswählen, können Sie aus jedem beliebigen Makro wählen, das aktuell auf der Registerkarte “Textlinkregeln” definiert wurde.

Typen

Sie können einen Typ direkt in einer Textlinkanalyseregel oder in einem Makro verwenden. Wenn Sie den Typnamen manuell oder in der Quellenansicht eingeben (anstatt den Typ aus einem Kontextmenü auszuwählen), stellen Sie sicher, dass Sie dem Typnamen ein Dollar-Zeichen (`$`) voranstellen, z. B. `$Person`. Bei dem Typnamen wird zwischen Groß- und Kleinschreibung unterschieden. Wenn Sie die Kontextmenüs verwenden, können Sie aus einem beliebigen Typ aus dem aktuell verwendeten Set von Ressourcen wählen.

Wenn Sie auf einen nicht erkannten Typ verweisen, erhalten Sie eine Warnmeldung und neben der Regel steht ein Warnsymbol im Regel- und Makrobaum, bis Sie den Fehler beheben.

Zeichenfolgen

Um Informationen einzubeziehen, die nie extrahiert wurden, können Sie eine Zeichenfolge definieren, nach der die Extrahierungsengine suchen soll. Allen extrahierten Wörtern oder Ausdrücken wurde ein Typ zugewiesen. Daher können sie nicht in Zeichenfolgen verwendet werden. Wenn Sie ein Wort verwenden, das extrahiert wurde, wird es ignoriert, selbst wenn der Typ `<Unknown>` ist.

Eine Zeichenfolge kann aus einem oder mehreren Wörtern bestehen. Folgende Regeln müssen bei der Definition einer Liste von Zeichenfolgen eingehalten werden:

- Schließen Sie die Liste der Zeichenfolgen in Klammern ein, z. B. (sein). Falls eine Auswahl von Zeichenketten vorhanden ist, muss jede Zeichenkette durch den Operator OR getrennt werden, z. B. `(ein|eine|das)` oder `(sein|ihr|unser)`.
- Geben Sie einzelne Wörter oder zusammengesetzte Wörter an.
- Trennen Sie die Wörter in der Liste mit dem Trennzeichen `|`; dieses Zeichen entspricht dem Boole'schen `ODER`.
- Sollen sowohl die Singularform als auch die Pluralform berücksichtigt werden, geben Sie beide Formen ein. Die Beugung erfolgt nicht automatisch.
- Geben Sie nur Kleinbuchstaben ein.
- Um Zeichenfolgen wiederzuverwenden, definieren Sie sie als Makro und setzen Sie dieses Makro dann in andere Makros und Textlinkanalyseregeln ein.
- Wenn eine Zeichenfolge Punkte oder Bindestriche enthält, müssen Sie sie aufnehmen. Um beispielsweise die Zeichenfolge `z. B.` im Text abzugleichen, müssen Sie die Punkte mit den Punkten `z. B.` als Zeichenfolge eingeben.

Ausschlussoperator

Verwenden Sie `!` als Ausschlussoperator, um zu verhindern, dass ein Ausdruck der Negation einen bestimmten Slot belegt. Sie können einen Ausschlussoperator nur manuell durch Bearbeiten in der Zelle (Doppelklick auf die Zelle in der Regelwert-Tabelle oder der Makrowert-Tabelle) oder in der Quellenansicht hinzufügen. Wenn Sie Ihrer Textlinkanalyseregel beispielsweise `($mTopic @{0,2} !($Positive) $Budget)` hinzufügen, suchen Sie nach einem Text, der (1) einen Ausdruck enthält, der einem der Typen im Makro `mTopic` zugeordnet ist, (2) eine Wortlücke von null bis zu zwei Wörtern lang ist, (3) keine Instanzen eines Ausdrucks enthält, der dem Typ `<Positive>` zugeordnet ist, und (4) einen Ausdruck enthält, der dem Typ`<Budget>` zugeordnet ist. Dadurch würde beispielsweise *Autos sind ein teures Unterfangen* erfasst, aber *Geschäft bietet unglaubliche Angebote* ignoriert.

Um diesen Operator zu verwenden, müssen Sie das Ausrufezeichen und Klammern manuell in der Elementzelle eingeben, indem Sie auf die Zelle doppelklicken.

Wortlücken (<Beliebiges Token>)

Eine Wortlücke, auch als `<Beliebiges Token>` bezeichnet, definiert einen numerischen Bereich von Tokens (Wörter, Ziffern, Buchstaben), die zwischen zwei Elementen vorliegen können. Wortlücken sind sehr hilfreich, wenn sehr ähnliche Ausdrücke abgeglichen werden, die sich aufgrund zusätzlicher Determinatoren, Präpositionalphrasen, Adjektive oder ähnlicher Wörter nur wenig voneinander unterscheiden.

Tabelle 19-1
Beispiel für die Elemente in einer Regelwert-Tabelle ohne eine Wortlücke

#	Element
1	 `Unbekannt`

2	mBeHave
3	Positive

Anmerkung: In der Quellenansicht wird dieser Wert folgendermaßen definiert: `$Unbekannt $mBeHave $Positiv`

Dieser Wert wird mit Sätzen wie "*Das Hotelpersonal war nett.*" abgeglichen, wobei *Hotelpersonal* zu dem Typ `<Unbekannt>` gehört, *war* befindet sich unter dem Makro `mBeHave` und *nett* ist `<Positive>`. Der Satz "*Das Hotelpersonal war sehr nett.*" wird hingegen nicht abgeglichen.

Tabelle 19-2
Beispiel für die Elemente in einer Regelwert-Tabelle mit einer <Beliebiges Token>-Wortlücke

#	Element
1	Unbekannt
2	mBeHave
3	
4	Positive

Anmerkung: In der Quellenansicht wird dieser Wert folgendermaßen definiert: `$Unbekannt $mBeHave @{0,1} $Positive`

Wenn Sie Ihrem Regelwert eine Wortlücke hinzufügen, wird diese sowohl mit "*Das Hotelpersonal war nett.*" als auch "*Das Hotelpersonal war sehr nett.*" abgeglichen.

In der Quellenansicht oder beim Bearbeiten in der Zeile ist die Syntax für eine Wortlücke `@{#,#}`. Dabei bedeutet `@` eine Wortlücke und `{#,#}` definiert die Mindest- und Höchstanzahl von Wörtern, die zwischen dem vorangehenden und dem folgenden Element akzeptiert werden. Die Angabe `@{1,3}` bedeutet beispielsweise, dass eine Übereinstimmung mit den beiden definierten Elementen vorliegt, wenn mindestens ein Wort vorhanden ist, jedoch mehr als drei Wörter zwischen diesen beiden Elementen vorhanden sind. `@{0,3}` bedeutet, dass eine Übereinstimmung zwischen den beiden definierten Elementen vorliegt, wenn 0, 1, 2 oder 3 Wörter vorhanden sind, nicht jedoch mehr als drei Wörter.

Ansicht und Arbeiten im Quellenmodus

Der TLA-Editor generiert für jede Regel und jedes Makro den zugrundeliegenden Quellcode, den der Extraktor zum Abgleichen und Erzeugen einer TLA-Ausgabe verwendet. Falls Sie lieber mit dem Code arbeiten, können Sie diesen Quellcode ansehen und ihn direkt bearbeiten, indem Sie auf die Schaltfläche "Quelle anzeigen" im oberen Bereich des Editors klicken. Die Quellenansicht springt zur/zum aktuell ausgewählten Regel oder Makro. Es wird jedoch empfohlen, die Editorbereiche zu verwenden, um die Gefahr von Fehlern zu minimieren.

Klicken Sie auf Quellenansicht verlassen, wenn Sie mit der Ansicht oder dem Bearbeiten der Quelle fertig sind. Wenn Sie eine ungültige Syntax für eine Regel generieren, müssen Sie diese Syntax korrigieren, bevor Sie die Quellenansicht verlassen.

Wichtig: Wenn Sie in der Quellenansicht bearbeiten, empfehlen wir dringend, immer nur eine Regel oder ein Makro gleichzeitig zu bearbeiten. Validieren Sie die Ergebnisse nach der Bearbeitung eines Makros durch Extrahierung. Wenn Sie mit dem Ergebnis zufrieden sind, empfehlen wir, die Vorlage zu speichern, bevor Sie eine weitere Änderung vornehmen. Wenn Sie mit dem Ergebnis nicht zufrieden sind oder ein Fehler auftritt, kehren Sie zu Ihren gespeicherten Ressourcen zurück.

Makros in der Quellenansicht

```
[macro]
name = macro_name
value = ([type_name|macro_name|literal_string|word_gap])
```

`[macro]`	Jedes Makro muss mit einer Zeile beginnen, die mit [macro] gekennzeichnet ist, um den Start eines Makros zu kennzeichnen.
`name`	Der Name der Makrodefinition. Die Namen müssen eindeutig sein.
`value`	Eine Kombination aus einem oder mehreren Typen, Zeichenfolgen, Wortlücken oder Makros.Für weitere Informationen siehe Thema Unterstützte Elemente für Regeln und Makros auf S. 369. Um Argumente zu kombinieren, müssen Sie Klammern ( ) verwenden, um die Argumente zu einer Gruppe zusammenzufassen, und das Zeichen | verwenden, um den Boole'schen Operator ODER zu kennzeichnen.

Zusätzlich zu den Anweisungen und der Syntax, die im Abschnitt über Makros behandelt wurden, gibt es in der Quellenansicht einige Dinge zu beachten, die bei der Arbeit in der Editoransicht nicht notwendig sind. Makros müssen bei der Arbeit im Quellenmodus außerdem folgende Regeln einhalten:

- Jedes Makro muss mit einer Zeile beginnen, die mit `[macro]` gekennzeichnet ist, um den Start eines Makros zu kennzeichnen.
- Soll ein Element deaktiviert werden, geben Sie eine Raute (#) am Anfang der entsprechenden Zeilen ein.

Beispiel. Dieses Beispiel definiert das Makro `mTopic`. Der Wert für `mTopic` ist das Vorhandensein *eines* der folgenden Typen: `<Product>`, `<Person>`, `<Location>`, `<Organization>`, `<Budget>` oder `<Unknown>`.

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

Regeln in der Quellenansicht

```
[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>)]`	Gibt den Start der Textlinkanalyseregel an und legt eine eindeutige numerische ID fest, die verwendet wird, um die Verarbeitungsreihenfolge zu ermitteln.
`name`	Gibt dieser Textlinkanalyseregel einen eindeutigen Namen.

`value` Liefert die Syntax und die Argumente für den Textabgleich. Für weitere Informationen siehe Thema Unterstützte Elemente für Regeln und Makros auf S. 369.

`output` Das Ausgabeformat für die resultierenden, übereinstimmenden Muster, die im Text festgestellt wurden. Die Ausgabe entspricht nicht immer der exakten, ursprünglichen Poisition der Elemente im Quelltext. Zusätzlich können mehrere Ausgabezeilen für eine bestimmte Textlinkanalyseregel vorhanden sein, wenn jede Ausgabe in eine eigene Zeile gesetzt wird.

Syntax für die Ausgabe:

- Trennen Sie Ausgabeelemente mit dem Tabulator-code `\t` (z. B. `$1\t#1\t$3\t#3`).
- `$` und eine Zahl ruft den Ausdruck auf, der mit dem Argument übereinstimmt, das im Werteparameter an dieser Position definiert ist. `$1` bezieht sich also auf den Begriff, der mit dem ersten für den Wert definierten Argument übereinstimmt.
- `#` und eine Zahl ruft den Typennamen des Elements in dieser Position auf. Wenn ein Objekt eine Liste von Zeichenfolgen ist, wird der Typ `<Unknown>` zugeordnet.
- Der Wert `Null\tNull` erzeugt keine Ausgabe.

Zusätzlich zu den Anweisungen und der Syntax, die im Abschnitt über Regeln behandelt wurden, gibt es in der Quellenansicht einige Dinge zu beachten, die bei der Arbeit in der Editoransicht nicht notwendig sind. Regeln müssen bei der Arbeit im Quellenmodus außerdem folgende Regeln einhalten:

- Sind mindestens zwei Elemente definiert, müssen sie in Klammern eingeschlossen werden, unabhängig davon, ob sie optional sind oder nicht (beispielsweise `($Negative|$Positive)` *oder* `($mCoord|$SEP)?`). `$SEP` steht für ein Komma.
- Das erste Element in einer Textlinkanalyseregel kann kein optionales Element sein. Das Element `value = $mTopic?` oder `value = @{0,1}` als erstes Element wäre beispielsweise nicht zulässig.
- Es ist möglich, einem Token eine Menge (oder Anzahl an Instanzen) zuzuweisen. Dies ist insbesondere dann nützlich, wenn Sie eine Regel für alle Fälle schreiben, anstatt je eine Regel für die einzelnen Fälle zu notieren. Mit der Zeichenfolge `($SEP|and)` können Sie beispielsweise eine Übereinstimmung mit `,` (Komma) oder mit dem Wort `and` ermitteln. Wenn Sie diese Regel mit einer Menge ergänzen (die Zeichenfolge wird zu `($SEP|and){1,2}`), wird in allen nachstehenden Fällen eine Übereinstimmung erkannt: “`,`” “`und`” “ `und`”.
- Leerzeichen zwischen dem Variablen- oder Makronamen und dem Dollar- bzw. Fragezeichen (`$` und `?`) sind in der Textlinkanalyseregel `value` nicht zulässig.
- Leerzeichen sind in der Textlinkanalyseregel `output` nicht zulässig.
- Soll ein Element deaktiviert werden, geben Sie eine Raute (#) am Anfang der entsprechenden Zeilen ein.

Beispiel. Angenommen, Ihre Ressourcen enthalten die folgende TLA-Textlinkanalyseregel und Sie haben die Extrahierung von TLA-Ergebnissen aktiviert:

```
## Jean Doe was the former HR director of IBM in 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
```

Bei der Extrahierung liest die Extrahierungsengine jeden Satz und versucht, die folgende Sequenz abzugleichen:

Position	Beschreibung der Argumente
1	Der Name einer Person (`$Person`),
2	Eines oder zwei der Folgenden: Komma (`$SEP`), Determinator (`$mDet`), Hilfsverb (`$mSupport`), die Zeichenfolgen "`then`" oder "`as`",
3	0 oder 1 Wort (`@{0,1}`)
4	Eine Funktion (`$Function`)
5	Eine der folgenden Zeichenfolgen: "`von`", "`mit`", "`für`", "`in`", "`bis`" oder "`bei`".
6	0 oder 1 Wort (`@{0,1}`)
7	Der Name einer Organisation (`$Organization`)
8	0, 1 oder 2 Wörter (`@{0,2}`)
9	Der Name eines Orts (`$Location`)

Dieses Beispiel einer Textlinkanalyseregel würde folgende Sätze oder Ausdrücke abgleichen:
Max Mustermann, Direktor von IBM in Deutschland
Max Mustermann war der frühere Direktor von IBM in Deutschland
IBM stellte Max Mustermann als Direktor von IBM in Deutschland ein

Dieses Beispiel einer Textlinkanalyseregel würde folgende Ausgabe erzeugen:

```
Max Mustermann  <Person>  Direktor  <Function>  ibm <Organization>  deutschland  <Location>
```

Erläuterung:

- `max mustermann` ist der Ausdruck, der `$1` (dem ersten Element in der Textlinkanalyseregel) entspricht, und `<Person>` ist der Typ für `max mustermann` (`#1`),
- `direktor` ist der Ausdruck, der `$4` (dem vierten Element in der Textlinkanalyseregel) entspricht, und `<Function>` ist der Typ für `direktor` (`#4`),
- `ibm` ist der Ausdruck, der `$7` (dem siebten Element in der Textlinkanalyseregel) entspricht, und `<Organization>` ist der Typ für `ibm`. (`#7`),
- `deutschland` ist der Ausdruck, der `$9` (dem neunten Element in der Textlinkanalyseregel) entspricht, und `<Location>` ist der Typ für `deutschland` (`#9`),

Regelsets in der Quellenansicht

[set(<ID>)]

`[set (<ID>)]`	Gibt den Start des Regelsets an und legt eine eindeutige numerische ID fest, die verwendet wird, um die Verarbeitungsreihenfolge der Sets zu ermitteln.

Beispiel. Der folgende Satz enthält Informationen über Personen, deren Funktion in einem Unternehmen und Fusions-/Übernahmeaktivitäten dieses Unternehmens.

```
IBM has entered into a definitive merger agreement with SPSS, said
Jack Noonan, CEO of SPSS. (IBM hat einen Fusionsvertrag mit SPSS
geschlossen, sagte Jack Noonan, CEO von SPSS.)
```

Sie könnten eine Regel mit mehreren Ausgaben schreiben, um sämtliche möglichen Ausgaben zu berücksichtigen, z. B.:

```
## IBM entered into a definitive merger agreement with SPSS, said
Jack Noonan, CEO of 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
```

Diese Regeln würden folgende beiden Ausgabemuster erzeugen:

- `ibm <Organization> + merges with <ActiveVerb> + spss <Organization>`
- `jack noonan <Person> + ceo <Function> + spss <Organization>`

Wichtig: Denken Sie daran, dass andere linguistische Operationen bei der Extrahierung von TLA-Mustern stattfinden. In diesem Fall wird `merger` während der Synonymgruppierungsphase des Extrahierungsprozesses unter `merges with` gruppiert. Und da `merges with` zum Typ `<ActiveVerb>` gehört, erscheint dieser Typname in der endgültigen TLA-Musterausgabe. Wenn die Ausgabe also `t$3\t#3` lautet, bedeutet dies, dass das Muster schließlich das endgültige Konzept für das dritte Element und den endgültigen Typ für das dritte Element anzeigt, sobald sämtliche linguistische Verarbeitung durchgeführt wurde (Synyme und andere Gruppierungen).

Anstatt komplexe Regeln wie die vorherige zu verfassen, kann es einfacher sein, mit zwei Regeln zu arbeiten. Der erste konzentriert sich auf die Suche nach Fusionen/Übernahmen zwischen den Unternehmen:

```
[set(1)]
## IBM has entered into a definitive merger agreement with 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
```

Diese Regel würde Folgendes liefern: `ibm <Organization> + merges with <ActiveVerb> + spss <Organization>`

Der zweite konzentriert sich auf die Person/die Funktion/das Unternehmen:

```
[set(2)]
## said Jack Noonan, CEO of 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
```

Diese Regel würde Folgendes liefern: `jack noonan <Person> + ceo <Function> + spss <Organization>`

Anhang

A

Ausnahmen bei japanischem Text

Anmerkung: Die in diesem Kapitel behandelten Funktionen stehen nur in IBM® SPSS® Modeler Premium zur Verfügung.

Für Text in japanischer Sprache erfolgen Verarbeitung und Mining zwar ähnlich wie für andere Sprachen in IBM® SPSS® Modeler Text Analytics , jedoch gibt es eine Reihe von Abweichungen. Die geringfügigeren Änderungen werden in dieser Dokumentation direkt mit den Anleitungen für die anderen Sprachen beschrieben. Jedoch werden einige der größeren Unterschiede in diesem Anhang erläutert.

Anmerkung: Es gibt viele neue Verbesserungen seit der japanischen Version von Text-Mining für Clementine 2.2. Weitere Informationen zu diesen Verbesserungen können Sie bei Ihrem Fachhändler erfragen.

Extrahieren und Kategorisieren von japanischem Text

Der Mining-Vorgang für Text in japanischer Sprache gleicht dem Vorgang für andere unterstützte Sprachen. Für weitere Informationen siehe Thema Informationen zum Text-Mining in Kapitel 1 auf S. 2. Jedoch gelten für die japanische Sprache die im Folgenden erläuterten Unterschiede.

Wie Extrahierung funktioniert

Während der Extrahierung von Schlüsselkonzepten und -begriffen aus Ihren Antworten wird bei IBM® SPSS® Modeler Text Analytics die linguistisch basierte Textanalyse angewendet. Diese Methode bietet die Geschwindigkeit und Kosteneffektivität von statistisch basierten Systemen. Sie ermöglicht jedoch ein weitaus höheres Maß an Genauigkeit, während ein deutlich geringeres Maß an Eingriffen seitens des Benutzers erforderlich ist. Linguistisch basierte Textanalyse baut auf einem Forschungsgebiet namens Verarbeitung natürlicher Sprache auf, das auch als Computerlinguistik bekannt ist.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Für Text in japanischer Sprache lässt sich der Unterschied zwischen statistikbasierten und linguistikbasierten Methoden beim Extrahierungsvorgang am Wort 沈む als ein Beispiel illustrieren. Mit diesem Wort können wir Ausdrücke wie 日が沈む finden, übersetzt als *Sonne geht unter*, oder 気分が沈む, übersetzt als *sich traurig fühlen*. Wenn Sie nur statistische Techniken einsetzen, werden 日 (übersetzt als *Sonne*), 気分 (übersetzt als *fühlen*) und 沈む (übersetzt als *traurig*) separat extrahiert. Bei Einsatz der Stimmungsanalyse, die linguistische Verfahren verwendet, werden jedoch nicht nur 日, 気分 und 沈む, sondern auch 気分が沈む (übersetzt als *sich traurig fühlen*) extrahiert und dem Typ <悪い – 悲しみ全般> zugewiesen. Die Verwendung von linguistischen Verfahren durch die Stimmungsanalyse ermöglicht die Extrahierung von bedeutungsvolleren Ausdrücken. Die Analyse und Erfassung von Stimmungen beseitigt die

Mehrdeutigkeit von Texten, weshalb das linguistisch basierte Text-Mining definitionsgemäß den verlässlicheren Ansatz darstellt.

Wenn Sie verstehen, wie der Extrahierungsprozess funktioniert, fällt es Ihnen leichter, bei der Feinabstimmung Ihrer linguistischen Ressourcen (Bibliotheken, Typen, Synonyme und anderer) zentrale Entscheidungen zu treffen. Der Extrahierungsprozess umfasst folgende Schritte:

- Konvertieren von Quelldaten in ein Standardformat
- Ermittlung von Kandidaten
- Ermittlung von Äquivalenzklassen und Integration von Synonymen
- Zuweisung eines Typs
- Indexerstellung und, falls gewünscht, Musterabgleich mit einem Sekundäranalysator

Schritt 1. Konvertieren von Quelldaten in ein Standardformat

Im ersten Schritt werden die importierten Daten in ein einheitliches Format konvertiert, das für weitergehende Analysen genutzt werden kann. Diese Konvertierung erfolgt intern. Ihre Ausgangsdaten werden dabei nicht geändert.

Schritt 2. Ermittlung von Kandidaten

Es ist wichtig zu verstehen, welche Rolle die linguistischen Ressourcen während der linguistischen Extrahierung bei der Ermittlung von Kandidaten spielen. Linguistische Ressourcen kommen jedes Mal zum Einsatz, wenn ein Extrahierungsvorgang ausgeführt wird. Sie liegen in Form von Vorlagen, Bibliotheken und zusammengestellten Ressourcen vor. Bibliotheken bestehen aus Wortlisten, Beziehungen und weiteren Informationen, die eingesetzt werden, um die Extrahierung abzustimmen oder zu spezifizieren. Die zusammengestellten Ressourcen können nicht angezeigt oder bearbeitet werden. Die übrigen Ressourcen können jedoch im Template Editor bzw., wenn eine interaktive Workbench-Sitzung gestartet wurde, im Resource Editor bearbeitet werden.

Zusammengestellte Ressourcen sind interne Kernkomponenten der Extraktor-Engine in SPSS Modeler Text Analytics . Diese Ressourcen umfassen ein allgemeines Wörterbuch, in dem eine Liste von Grundformen mit einem Code für die Wortart (Part of Speech) enthalten ist (Nomen, Verb, Adjektiv usw.). Die Ressourcen beinhalten auch belegte integrierte Typen, die verwendet werden, um den folgenden Typen eine Vielzahl von extrahierten Termen zuzuweisen: <地名>, <組織> oder <人名>. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text auf S. 387.*Anmerkung*: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Zusätzlich zu diesen zusammengestellten Ressourcen sind auch mehrere Bibliotheken im Lieferumfang enthalten. Diese können verwendet werden, um die Typen und Konzeptdefinitionen der zusammengestellten Ressourcen zu ergänzen und Synonyme zu liefern. Diese Bibliotheken—sowie sämtliche benutzerdefinierte Bibliotheken, die Sie erstellen—bestehen aus mehreren Wörterbüchern. Diese umfassen Typ-Wörterbücher, Synonymwörterbücher sowie Ausschlusswörterbücher. Für weitere Informationen siehe Thema Bearbeitungsressourcen für japanischen Text auf S. 383.*Anmerkung*: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Sobald die Daten importiert und konvertiert wurden, beginnt die Extrahierungsengine, Kandidaten für die Extrahierung zu identifizieren. Kandidaten sind Wörter oder Wortgruppen, die verwendet werden, um Konzepte im Text zu ermitteln. Bei der Verarbeitung des Texts werden einzelne Wörter (**Uniterms**) und zusammengesetzte Wörter (**Multiterms**) über Extraktoren auf

der Grundlage von Wortklasse-Mustern (Part of Speech) ermittelt. Der Multiterm 青森りんご, der dem Wortklasse-Muster <地名> + <名詞> entspricht, besteht beispielsweise aus zwei Komponenten. Anschließend werden mithilfe der Stimmungstextlinkanalyse Kandidaten für Stimmungsstichwörter identifiziert.

Angenommen, Sie verfügen über den folgenden japanischen Text: 写真が新鮮で良かった. In diesem Fall würde die Extrahierungsengine den Stimmungstyp 良い – 褒め・賞賛 zuweisen, nachdem（品物） + が + 良い mithilfe einer der Stimmungstextlinkregeln abgeglichen wurde.

Anmerkung: Die Terme aus dem oben genannten zusammengestellten allgemeinen Wörterbuch stellen eine Liste aller Wörter dar, die als Uniterms wahrscheinlich uninteressant sind oder sprachliche Mehrdeutigkeiten aufweisen. Diese Wörter werden von der Extrahierung ausgeschlossen, wenn die Uniterms ermittelt werden. Sie werden jedoch erneut ausgewertet, wenn Wortarten bestimmt oder längere zusammengesetzte Wörter (Multiterms) als Kandidaten geprüft werden.

Schritt 3. Ermittlung von Äquivalenzklassen und Integration von Synonymen

Im Anschluss an die Ermittlung von Uniterms und Multiterms, die als Kandidaten infrage kommen, werden über ein Normalisierungswörterbuch der Software Äquivalenzklassen ermittelt. Bei einer Äquivalenzklasse handelt es sich um eine Grundform eines Ausdrucks oder einer einzelnen Form von zwei Varianten desselben Ausdrucks. Ausdrücke werden Äquivalenzklassen zugeordnet, damit beispielsweise die Begriffe `Nebenwirkung` und 副作用 nicht als unterschiedliche Konzepte betrachtet werden. Um festzustellen, welches Konzept für die betreffende Äquivalenzklasse als Hauptterm verwendet wird, `Nebenwirkung` oder 副作用, werden die folgenden Regeln in der aufgeführten Reihenfolge durch die Extrahierungsengine angewendet:

- Die vom Benutzer festgelegte Form in einer Bibliothek.
- Die häufigste Form, wie von vorkompilierten Ressourcen definiert.

Schritt 4. Zuweisen eines Typs

Anschließend werden den extrahierten Konzepten Typen zugewiesen. Bei einem Typ handelt es sich um Konzepte, die nach semantischen Gesichtspunkten gruppiert werden. Für diesen Schritt werden sowohl zusammengestellte Ressourcen als auch die Bibliotheken verwendet. Zu den Typen gehören beispielsweise übergeordnete Konzepte, positive und negative Wörter, Vornamen, Orte, Organisationen und anderes. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

Japanische Textressourcen verfügen über ein anderes Set an Typen. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text auf S. 387.*Anmerkung*: Extraktion für japanischen Text steht in SPSS Modeler Premium zur Verfügung.

Schritt 5. Indexerstellung und Musterabgleich mit Ereignisextrahierung

Die gesamte Reihe der Datensätze oder Dokumente wird indiziert. Dazu wird ein Zeiger zwischen einer Textstelle und dem bezeichnenden Term für jede Äquivalenzklasse erstellt. Das setzt voraus, dass sämtliche gebeugten Formen, die als Kandidaten für ein Konzept vorkommen, als Grundform für den Kandidaten indiziert werden. Die globale Häufigkeit wird für jede Grundform berechnet.

Mit SPSS Modeler Text Analytics können nicht nur Typen und Konzepte sondern auch Beziehungen ermittelt werden, die zwischen diesen bestehen. Mit diesem Produkt stehen mehrere Algorithmen und Bibliotheken zur Verfügung, über die Beziehungsmuster der Textlinkanalyse

zwischen Typen und Konzepten extrahiert werden können. Sie sind besonders nützlich beim Versuch, bestimmte Meinungen zu entdecken (z. B. Reaktionen auf ein Produkt).

Wie sekundäre Extrahierung funktioniert

Beim Ausführen einer Extrahierung aus japanischem Text erhalten Sie automatisch Konzepte aus den grundlegenden Stichwörtern und acht Basistypen: 人名, 地名, 組織名, 名詞, 形容詞, 動詞, 形容動詞 und その他. Um jedoch die Standardressourcen für japanischen Text optimal zu nutzen, müssen Sie eine der folgenden sekundären Analysen wählen: Stimmung oder Abhängigkeit.

Durch die Wahl eines Sekundär-Analysators können Sie Textlinkanalysemuster extrahieren und Beziehungen zwischen den Begriffen im Text aufdecken. Beim Definieren Ihres Knotens oder der Wahl von Extrahierungsoptionen in einer interaktiven Workbench-Sitzung können Sie nach Wunsch eine sekundäre Analyse für den Extrahierungsprozess hinzufügen.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Sekundäre Analyse. Beim Start einer Extrahierung werden anhand des Standardsatzes an Typen grundlegende Stichwörter extrahiert. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text auf S. 387. Wenn Sie jedoch eine sekundäre Analyse wählen, erhalten Sie mehr und vielfältigere Konzepte, da der Extraktor nun Partikel und Hilfsverben als Teil des Konzepts berücksichtigt. Nehmen Sie beispielsweise an, der Satz 肩の荷が下りた wurde in "*Mir wurde eine große Last von den Schultern genommen*". In diesem Beispiel kann die grundlegende Stichwortextrahierung jedes Konzept wie folgt separat extrahieren: 肩 (*Schultern*), 荷 (*Last*), 下りる (*wurde genommen*), jedoch wird die Beziehung zwischen diesen Wörtern nicht extrahiert. Wenn Sie jedoch die Stimmungsanalyse anwenden, können Sie vielfältigere Konzepte hinsichtlich eines Stimmungstyps extrahieren, z. B. das Konzept =肩の荷が下りた, das als "*eine große Last von den Schultern nehmen*", was dem Typ <良い-安心> zugeordnet ist. Im Fall der Stimmungsanalyse wird auch eine große Anzahl an zusätzlichen Typen eingeschlossen. Des Weiteren ermöglicht Ihnen die Wahl einer sekundären Analyse, auch Textlinkanalysis-Ergebnisse zu generieren.

Anmerkung: Wenn ein Sekundär-Analysator genannt wird, dauert der Extrahierungsprozess länger. Für weitere Informationen siehe Thema Wie sekundäre Extrahierung funktioniert auf S. 380.

- **Abhängigkeitsanalyse**. Die Wahl dieser Option führt zu erweiterten Partikeln für die Extrahierungskonzepte aus der grundlegenden Typ- und Stichwortextrahierung. Sie können auch vielfältigere Musterergebnisse aus einer Abhängigkeits-Textlinkanalyse (TLA) gewinnen.
- **Stimmungsanalyse**. Bei dieser Analyse werden zusätzliche Konzepte und – wann immer möglich – TLA-Musterergebnisse extrahiert. Neben den grundlegenden Typen erhalten Sie auch den Vorteil von über 80 Stimmungstypen, z. B. 嬉しい, 吉報, 幸運, 安心, 幸福 usw. Mithilfe dieser Typen werden Konzepte und Muster im Text durch den Ausdruck von Emotionen, Stimmungen und Meinungen aufgedeckt. Es gibt drei Optionen, die den Fokus für die Stimmungsanalyse festlegen: Alle Stimmungen, Nur repräsentative Stimmung und Nur Schlussfolgerungen.

Optionen für Stimmungsanalyse

Beim Arbeiten mit japanischem Text können Sie nach Wunsch mithilfe der Stimmungsanalyse zusätzliche Konzepte und Typen extrahieren. Diese Analyse umfasst mehr als 80 zusätzliche Typen, mit denen Sie Meinungen, Gefühle und Emotionen aus Ihren Textdaten extrahieren können. Zusätzlich müssen Sie bei der Wahl der Stimmungsanalyse als sekundäre Analyse eine der folgenden Optionen wählen, die der Extrahierungsengine mitteilen, welche Stimmungen extrahiert werden sollen:

- **Alle Stimmungen**
- **Nur repräsentative Stimmung**
- **Nur Schlussfolgerungen**

Bei der Extrahierung beginnt die Stimmungsanalyse durch Zerlegen eines Datensatzes oder Dokuments in Klauseln, von denen jede ein Prädikat enthält. Zum Beispiel wird der Text "4月になったがまだ寒い。", der mit "*Es ist April, aber es ist immer noch kalt.*" übersetzt wurde, vom Analysator als 2 Klauseln interpretiert, obwohl er nur ein Satzendzeichen (。) enthält. Jede Klausel wird dann von der Extrahierungsengine untersucht, um festzustellen, ob sie Ihrer gewählten Option entspricht.

Untersuchen wir anhand des folgenden Beispieltextes die drei Optionen: "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。". Dieser Text wird übersetzt mit: "*Eine Bedienung war nicht freundlich, aber das Zimmer war groß und zufriedenstellend. Auch mit dem Abendessen war ich zufrieden*". Bei der Extrahierung wird der Originaltext in die folgenden Klauseln zerlegt:

- 案内してくれた仲居さんは無愛想だったが、, übersetzt als "*Eine Bedienung war nicht freundlich, aber*"
- 部屋は広くて申し分なかった。, übersetzt als "*Das Zimmer war groß und zufriedenstellend*"
- 夕食も満足。, übersetzt als "*Auch mit dem Abendessen war ich zufrieden.*"

Alle Stimmungen

Diese Option extrahiert alle Stimmungen, Meinungen und Emotionen, die den Ressourcen und Stimmungstextlinkregeln entsprechen. In unserem Beispiel würden die folgenden Konzepte aus dem Beispieltext extrahiert.

Tabelle A-1
Mögliche Ausgabe für das Beispiel mit der Option "Alle Stimmungen"

Konzept	Typ
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>

Anmerkung: In der vorangehenden Tabelle zeigen die zweite und dritte Zeile, wie der Extraktor zwei Konzepte aus derselben Klausel beziehen kann.

Nur repräsentative Stimmung

Diese Option extrahiert nur die repräsentativeren Meinungen oder Emotionen, die in jeder Klausel ausgedrückt werden. Wenn der Text mehrere Meinungen oder Emotionen enthält, wird ein Algorithmus angewendet. Dieser Algorithmus versucht, die Wichtigkeit der ermittelten Stimmungen und die Position der Wörter in einer Klausel zu bestimmen. In einigen Fällen, in denen zwei Stimmungsstichwörter mit derselben Wichtigkeit gefunden werden, wird das Stimmungsstichwort an der letzten Position in der Klausel extrahiert.

部屋は広くて (übersetzt als *Das Zimmer war groß*) wird nicht aus dem Text extrahiert, da der zweite Begriff, 申し分なかった, gegenüber 部屋は広くて in dieser Klausel als wichtiger erachtet wird, wenn der interne Algorithmus und die Wortposition angewendet werden.

Tabelle A-2
Mögliche Ausgabe für den Text mit der Option "Nur repräsentative Stimmung"

Konzept	Typ
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>

Nur Schlussfolgerungen

Diese Option zwingt den Extraktor, ein Stimmungsstichwort zu identifizieren und zu extrahieren, das die Schlussfolgerung des gesamten Datensatzes oder Dokuments repräsentiert. Nicht sämtlicher Text verfügt über eine Schlussfolgernung, in einigen Fällen wird also für einen bestimmten Text mit dieser Option nichts extrahiert. Außerdem ist es für die Analyse bei einem längeren Datensatz oder Dokument zunehmend schwieriger, die Hauptschlussfolgerung zu identifizieren. Obwohl dies selten auftritt, können manchmal auch mehrere Schlussfolgerungen extrahiert werden.

満足 (übersetzt als *zufrieden*) wird als die wichtigste Schlussfolgerung der im Text ausgedrückten Gefühle angesehen.

Tabelle A-3
Mögliche Ausgabe für den Text mit der Option "Nur Schlussfolgerungen"

Konzept	Typ
満足	<満足>

Wie die Kategorisierung funktioniert

Bei der Erstellung von Kategoriemodellen mit IBM® SPSS® Modeler Text Analytics haben Sie die Wahl zwischen verschiedenen Methoden, um Kategorien zu erstellen. Da jeder Datensatz seine besonderen Eigenheiten aufweist, kann die Zahl der Methoden und die Reihenfolge, in der sie angewendet werden, gegebenenfalls variieren. Da sich Ihre Interpretation der Ergebnisse möglicherweise von der einer anderen Person unterscheidet, müssen Sie gegebenenfalls ein wenig mit den unterschiedlichen Methoden experimentieren, um zu erkennen, mit welcher Sie die besten Ergebnisse für Ihre Textdaten erzielen. In SPSS Modeler Text Analytics können Sie

Kategoriemodelle in einer interaktiven Workbench-Sitzung erstellen, in denen Sie Ihre Kategorien weiter untersuchen und abstimmen können.

In diesem Handbuch bezieht sich **Kategorieerstellung** auf die Erstellung von Kategoriedefinitionen und Klassifikation über eine oder mehrere integrierte Methoden und **Kategorisierung** auf das Scoring oder Labeling, einen Prozess, bei dem den Kategoriedefinitionen für jeden Datensatz oder jedes Dokument unverwechselbare IDs (Name/ID/Wert) zugewiesen werden.

Während der Kategorieerstellung werden die extrahierten Konzepte und Typen als Bausteine für Ihre Kategorien verwendet. Bei der Erstellung von Kategorien werden den Kategorien die Datensätze oder Dokumente zugewiesen, die Text enthalten, der einem Element der jeweiligen Kategoriedefinition entspricht.

SPSS Modeler Text Analytics bietet Ihnen mehrere automatisierte Kategorieerstellungsmethoden, mit denen Sie Ihre Dokumente oder Datensätze schnell kategorisieren können. Die einzelnen Verfahren sind für bestimmte Datentypen und Situationen jeweils sehr gut geeignet, doch ist es oftmals nützlich, bei einer Analyse mehrere Verfahren miteinander zu verbinden, um die Dokumente oder Datensätze vollständig zu erfassen. Sie können ein Konzept in mehreren Kategorien erkennen oder redundante Kategorien vorfinden.

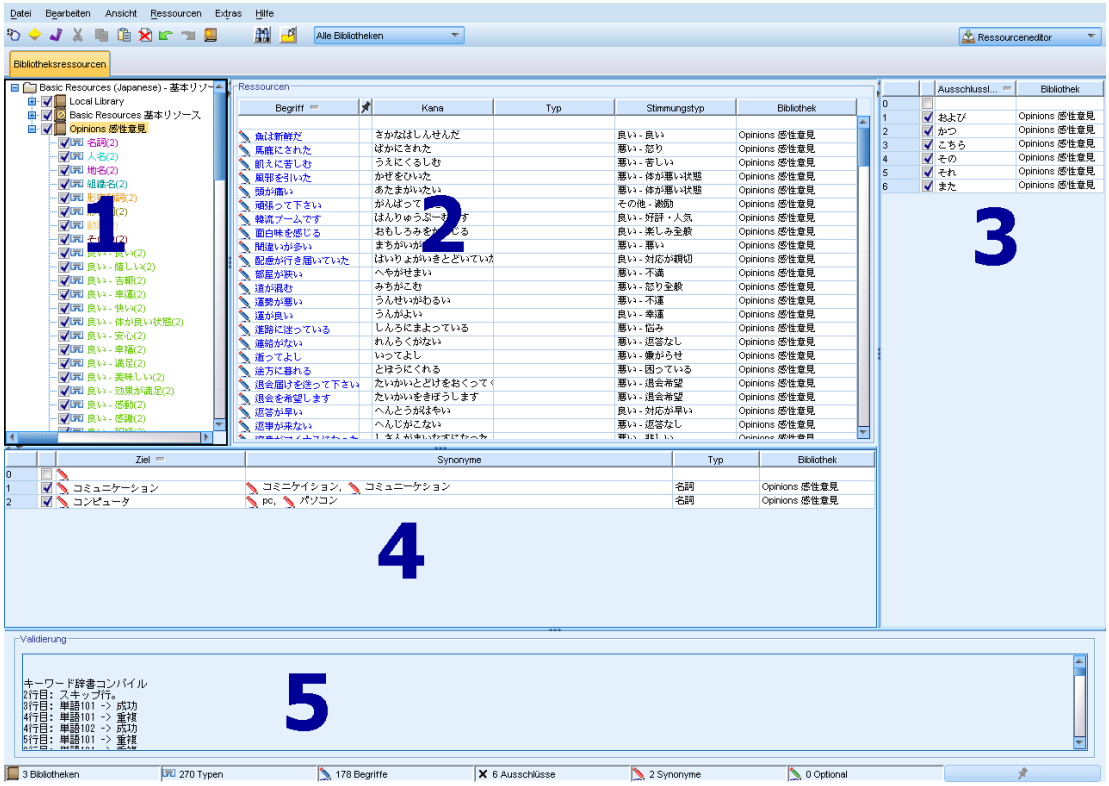
Bearbeitungsressourcen für japanischen Text

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Ab IBM® SPSS® Modeler Text Analytics Version 14stehen eine neue Vorlage und ein neues Textanalysepaket (TAP) für Texte in japanischer Sprache zur Verfügung. Sie können Änderungen an den Ressourcen vornehmen, indem Sie Begriffe hinzufügen und bearbeiten und diese dann an Ihre Daten anpassen. Das Text Analysis Package enthält auch ein Kategorie-Set, das aus Kategorien besteht, die positive Stimmungen, negative Stimmungen und kontextbezogene/allgemeine Stimmungen darstellen.

Sie können Ihre Ressourcen im Resource Editor und im Template Editor bearbeiten. Editoren funktionieren für alle Textsprachen in ähnlicher Weise. Es gibt jedoch einige wichtige Unterschiede für japanischsprachige Texte, die im Folgenden erläutert werden.

Abbildung A-1
Resource Editor-Ansicht für japanischen Text



Die folgenden Punkte umreißen einige der Hauptunterschiede bei der Arbeit mit Ressourcen für japanischen Text. Eine allgemeine Beschreibung der vier Hauptbereiche auf der Registerkarte “Bibliotheksressourcen” finden Sie unter Die Editoroberfläche auf S. 282.

1. Fensterbereich “Bibliothek”. Dieser Bereich befindet sich links oben und funktioniert im Wesentlichen wie für andere Sprachen. Jedoch gibt es wenige Abweichungen, z. B. können keine neuen Typen erstellt oder bestehende Typen umbenannt werden. Für weitere Informationen siehe Thema Mit Bibliotheken arbeiten in Kapitel 16 auf S. 297.

2. Fensterbereich “Fachausdruck” für Typen-Wörterbücher. Dieser Bereich befindet sich rechts neben dem Bibliotheksbaum und unterscheidet sich für japanischen Text deutlich. Neben dem Ausdrucksnamen können Sie auch den Kana-Namen hinzufügen sowie einen oder zwei Typen wählen, mit denen Sie den Fachausdruck assoziieren können. Jedoch können Sie keine flektierten Formen von Fachausdrücken generieren oder Abgleichsoptionen für japanische Fachausdrücke zuordnen, wie dies für andere Sprachen möglich ist. Für weitere Informationen siehe Thema Japanischer Bibliotheksbaum, Typen und Fachausdrucksbereich auf S. 385.

3. Fensterbereich “Austausch-/Synonymwörterbuch”. In japanischen Textressourcen finden Sie eine Registerkarte “Synonym”, in der Sie alle Synonyme für Ihre Ressourcen definieren können. In der Registerkarte “Synonym” befindet sich die zusätzliche Spalte “Typ”, in der Sie für die eingegebenen Synonyme einen Typ festlegen müssen. Für weitere Informationen siehe Thema Verwenden des Synonymwörterbuchs für japanischen Text auf S. 392.*Anmerkung*: Da die Registerkarte “Optionale Elemente” nicht für japanischen Text gilt, ist sie nicht verfügbar.

4. Fensterbereich "Wörterbuch ausschließen". Hier gibt es keine Abweichungen für japanische Textressourcen, außer dass die Verwendung des Platzhalterzeichens * nicht unterstützt wird.

5. Fensterbereich "Validierung". Für japanischen Text gibt es einen zusätzlichen Validierungsbereich, in dem Sie Ihre Ressourcen vor der Extrahierung prüfen können. Beim Extrahieren von japanischem Text kompiliert die Extrahierungsengine die Ressourcen automatisch neu, wenn vor Beginn des Extrahierungsvorgangs Änderungen erkannt werden. Um potenzielle Fehler bei der Extrahierung zu vermeiden, können Sie die Ressourcen vor dem Extrahieren neu kompilieren und validieren, damit Sie etwaige auftretende Fehler korrigieren können. Für weitere Informationen siehe Thema Validieren und Kompilieren von japanischen Ressourcen auf S. 394.

Anmerkung: Es gibt keine bearbeitbaren erweiterten Ressourcen oder Textlinkregeln für japanischen Text, deshalb sind diese Registerkarten nicht verfügbar.

Japanischer Bibliotheksbaum, Typen und Fachausdrucksbereich

Die Arbeit mit Bibliotheken und Typen erfolgt für japanische Ressourcen sehr ähnlich wie für andere Sprachen. Für weitere Informationen siehe Thema Typ-Wörterbücher in Kapitel 17 auf S. 312.

Jedoch gibt es die folgenden wesentlichen Unterschiede:

- Japanische Textressourcen verfügen über ein anderes Set an Typen. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text auf S. 387.
- Typen können nicht erstellt oder umbenannt werden, jedoch lassen sich ihre Eigenschaften bearbeiten. Für weitere Informationen siehe Thema Bearbeiten von japanischen Typeigenschaften auf S. 391.
- Sie können Fachausdrücke hinzufügen und bearbeiten, einschließlich der Angabe eines Kana-Namens für einen Fachausdruck sowie der Zuweisung zu einem Typ und einem sekundären Stimmungstyp. Für weitere Informationen siehe Thema Japanischer Bibliotheksbaum, Typen und Fachausdrucksbereich auf S. 385.

Der Bibliotheksbaum zeigt die Bibliotheken sowie ihre Typ-Wörterbücher an. Wenn Sie eine Bibliothek oder einen Typ im linken Bereich auswählen, zeigt ein Bereich rechts die Fachausdrücke für die ausgewählten Bibliotheken bzw. Typ-Wörterbücher an. Sie können Fachausdrücke zu einem Typ-Wörterbuch direkt im Fachausdrucksbereich oder über das Dialogfeld Fachausdrücke hinzufügen hinzufügen. Bei den hinzugefügten Fachausdrücken kann es sich um einzelne Wörter oder um Wortfolgen handeln. Am Anfang der Liste befindet sich stets eine leere Zeile, in die Sie ein neues Wort eingeben können.

Wenn Sie einen Fachausdruck in einem Typ-Wörterbuch definieren, wird dieser standardmäßig als Nomen betrachtet und erhält automatisch den Typ <名詞>. Sie können jedoch den Typ in einen anderen Basistyp wie <動詞>, <形容詞>, <地名> usw. ändern. Wenn die Extrahierungsengine diesen Fachausdruck als dieselbe Wortklasse wie den Typen findet, dem Sie ihn in der Spalte Typ zugeordnet haben, wird er diesem Typ zugeordnet und extrahiert. Sie können den Fachausdruck zusätzlich einem der Stimmungstypen in der Spalte Stimmungstyp zuweisen. Wenn Sie dann die sekundäre Stimmungsanalyse verwenden, wird Text ein zweites Mal verarbeitet, um Fachausdrücke zu finden und den Stimmungstyp zuzuordnen. Wenn Sie darüber hinaus sowohl einen Stimmungstyp als auch einen Basistyp definieren und die

Extrahierungsengine findet, dass dieser Fachausdruck beiden Typen entspricht, wenn ebenfalls eine sekundäre Stimmungsanalyse durchgeführt wurde, hat der Stimmungstyp Vorrang und wird im Fensterbereich “Extrahierungsergebnisse” und in den Textlinkanalyse-Ergebnissen angezeigt. Beispiel: Wenn ein Verb als Verbtyp <**動詞**> sowie als positive Art von Typ wie “beliebt” extrahiert wurde, würde die Oberfläche diesen Fachausdruck dem positiven Typ zugehörig anzeigen, da die Erfassung von Stimmungen meistens interessanter ist als nur eine Wortklasse.

Abbildung A-2
Fensterbereiche “Bibliothek” und “Fachausdruck” für japanische Ressourcen

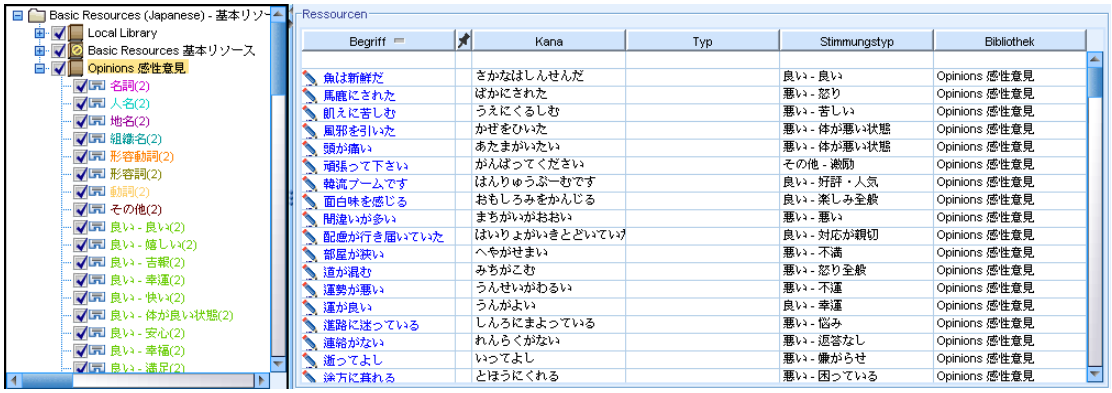


Tabelle A-4
Beschreibungen der Spalten im Fachausdruckbereich

Spaltenname	**Spaltenbeschreibung**
Begriff	Geben Sie ein Wort oder eine Wortfolge in die Zelle ein. In welcher Farbe der Fachausdruck angezeigt wird, hängt von der Farbe des Typs ab, in dem der Ausdruck gespeichert wurde. Die Typfarben können Sie im Dialogfeld “Typeigenschaften” ändern. Für weitere Informationen siehe Thema Bearbeiten von japanischen Typeigenschaften auf S. 391. Im Allgemeinen wird der Fachausdruck in Kanji geschrieben, kann aber auch Kana umfassen. **Wichtig:** Die Eingabe von Verben mit Katakana-Zeichen wird nicht unterstützt.
Erzwingen	Wenn Sie auf ein Reißzweckensymbol klicken und es in dieser Zelle hinzufügen, weisen Sie die Extrahierungs-Engine an, das Vorkommen desselben Ausdrucks in allen anderen Bibliotheken zu ignorieren. Für weitere Informationen siehe Thema Erzwingen von Fachausdrücken in Kapitel 17 auf S. 320. Dies ist für alle Sprachen gleich.
Kana	Geben Sie die Kana-Schreibung des Kanji-Fachausdrucksnamens ein.
Typ	Wählen Sie den Basistypnamen, dem der Fachausdruck zugeordnet werden soll. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text auf S. 387.
Stimmungstyp	Wenn eine sekundäre Analyse ausgeführt wird, wählen Sie den Stimmungsnamen, dem der Ausdruck zugewiesen werden soll. Für weitere Informationen siehe Thema Verfügbare Typen für japanischen Text auf S. 387.
Bibliothek	Wählen Sie die Bibliothek aus, in der Ihr Fachausdruck gespeichert wird. Sie können einen Begriff mit der Maus auf einen anderen Typ im Bibliotheksbaum ziehen, um seine Bibliothek zu ändern.

So fügen Sie einen einzelnen Fachausdruck zu einem Typ-Wörterbuch hinzu:

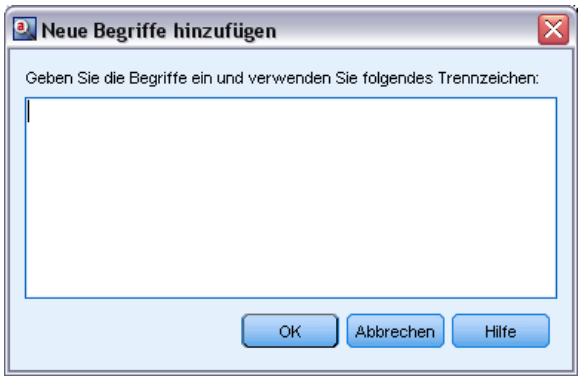
- Wählen Sie im Bibliotheksbaum das Typ-Wörterbuch aus, zu dem Sie den Fachausdruck hinzufügen möchten.

- Geben Sie in der mittleren Liste der Fachausdrücke Ihren Fachausdruck in die erste verfügbare leere Zelle ein und wählen Sie die gewünschten Optionen für diesen Fachausdruck aus.

So fügen Sie mehrere Fachausdrücke zu einem Typ-Wörterbuch hinzu:

- Wählen Sie im Bibliotheksbaumbereich das Typ-Wörterbuch aus, dem Sie die Fachausdrücke hinzufügen möchten.
- Wählen Sie im Menü Extras > Neue Begriffe aus. Das Dialogfeld "Neue Fachausdrücke hinzufügen" wird geöffnet.

Abbildung A-3
Dialogfeld "Neue Fachausdrücke hinzufügen"



- Geben Sie die Begriffe ein, die Sie dem ausgewählten Typ-Wörterbuch hinzufügen möchten. Sie können die Begriffe eingeben oder mehrere Begriffe einfügen. Wenn Sie mehrere Fachausdrücke eingeben, müssen Sie diese mit dem Trennzeichen, das im Dialogfeld "Optionen" definiert ist, trennen oder jeden Begriff in einer neuen Zeile eingeben. Für weitere Informationen siehe Thema Festlegen von Optionen in Kapitel 8 auf S. 135.
- Klicken Sie auf OK, um die Begriffe zum Wörterbuch hinzuzufügen. Das Dialogfeld wird geschlossen und die neuen Begriffe erscheinen im Wörterbuch.

Verfügbare Typen für japanischen Text

Sie können den japanischen Ressourcen keine neuen Typen hinzufügen, jedoch können Sie ihnen Fachausdrücke hinzufügen oder von ihnen entfernen. Die folgenden Tabellen umfassen das Set der aktuell verfügbaren japanischen Typen.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Typen für grundlegende Extrahierung

Bei jedem Start einer Extrahierung werden die nachstehenden Typen verwendet.

Tabelle A-5
Typen für grundlegende Extrahierung

Typen	Beschreibung
名詞	Wörter, die sich auf Dinge beziehen, z. B. “Auto” und “Film”. Personennamen, Ortsnamen und Organisationsnamen werden jedoch separat kategorisiert.
人名	Nomen, die den Namen bestimmter Personen entsprechen, z. B. “Tokugawa” und “Ieyasu”. Kombinationen von Vor- und Nachnamen wie “Tokugawa Ieyasu” sind ebenfalls Personennamen.
地名	Nomen wie “Tokyo” und “London”, die sich auf bestimmte Orte beziehen.
組織名	Nomen, die bestimmte Firmen und Organisationen bezeichnen, z. B. “The Federation of Economic Organizations”.
形容動詞	Wörter wie “still (shizuka)”, die die Charakteristik oder Bedingung eines Gegenstands beschreiben und in Phrasen wie “nicht [Adjektiv] (~ de nai)” und “ein [Adjektiv] Gegenstand (~ na koto)” beschrieben werden können.
形容詞	Wörter wie “lustig (tanoshii)”, die die Charakteristik oder Bedingung eines Gegenstands beschreiben und in Phrasen wie “[Adjektiv] werden (~ku naru)” und “ein [Adjektiv] Gegenstand (~i koto)” verwendet werden können.
動詞	Wörter, die Bewegungen oder Aktionen beschreiben, einschließlich Verben des Typs I (Konsonantenstamm), Verben des Typs II (Vokalstamm) sowie unregelmäßige Verben (sagyou henkaku und kagyou henkaku).
その他	Wörter wie Adverbien, Adjektivkomplemente vor Nomina, Konjunktionen und Interjektionen; Beispiele sind “ganz”, “was auch immer”, “dann” und “danke”.

Typen für Stimmungsanalyse

Immer wenn Sie die sekundäre Analyse zur Stimmungsextrahierung verwenden, erhalten Sie eine große Anzahl an Typen zusätzlich zu den acht Basistypen.

Tabelle A-6
Typen für Stimmungsanalyse

Typen	Beschreibung
良い - 良い	Ausdrücke für generell positive Dinge, die sich als “gut” klassifizieren lassen.
良い - 嬉しい	Beschreibt ein wünschenswertes Ereignis, das eine angenehme Stimulation erzeugt.
良い - 吉報	Beschreibt ein angenehmes Ereignis, das nur durch beträchtlichen Aufwand ermöglicht werden kann.
良い - 幸運	Beschreibt ein glückliches Ereignis, das nur durch Zufall oder ein bemerkenswertes Zusammentreffen ermöglicht werden kann.
良い - 快い	Legt nahe, dass etwas als Stimulus oder Umgebung zur Auslösung einer angenehmen physiologischen Empfindung dient.
良い - 体が良い状態	Beschreibt einen Zustand, in dem der Körper frei von Krankheit, Verletzung oder Müdigkeit ist, oder einen Zustand, in dem sich die physische Verfassung verbessert.
良い - 安心	Legt nahe, dass man ruhig ist und kein Risiko von Unglück oder Schaden besteht.
良い - 幸福	Gibt an, dass man durch eigene Aktionen oder die Umstände der Geburt besonders vorteilhafte Bedingungen oder Zuneigung erlangt hat.
良い - 満足	Beschreibt ein wünschenswertes Ereignis, das den Geist beruhigt.
良い - 美味しい	Gibt an, dass das Essen einen angenehmen Geschmack hat.
良い - 効果が満足	Impliziert, dass ein bestimmter Gegenstand die erwartete Wirkung erzeugt hat.

Typen	Beschreibung
良い - 感動	Legt nahe, dass die Signifikanz, die Bedeutung oder der Wert von etwas erstaunlich positiv ist.
良い - 感謝	Legt nahe, dass man die Aktionen eines anderen auf positive Weise anerkennt.
良い - 祝福	Drückt die Ansicht aus, dass die Situation einer anderen Person vorteilhaft ist (in einem Maß, das für den Sprecher akzeptabel ist).
良い - 喜び全般	Andere positive Ereignisse oder positive Ereignisse mit wenig Verbindung zum Sprecher.
良い - 楽しい	Weist auf Aktivitäten wie Kameradschaft, Unterhaltung und Erholung hin bzw. erwartet diese.
良い - 可笑しい	Bedeutet, dass etwas eine humorvolle Qualität hat, die eine angenehme Stimulation verursacht.
良い - 笑い	Drückt ein Lächeln oder Lachen aus, das durch eine gute und/oder humorvolle Angelegenheit verursacht wurde.
良い - 期待	Sagt voraus, dass in der Zukunft ein positives Ereignis eintreten wird.
良い - 楽しみ全般	Andere erfreuliche Ereignisse und/oder positive Aktivitäten/Verhalten mit wenig Verbindung zum Sprecher.
良い - 金額への賞賛	Deutet darauf hin, dass etwas vom Käuferstandpunkt gesehen einen wünschenswerten Geldwert hat.
良い - 対応が早い	Legt nahe, dass ein Service in der erwarteten Zeit geliefert oder abgeschlossen wurde.
良い - 対応が親切	Legt nahe, dass die Haltung oder das Verhalten eines Serviceanbieters gewissenhaft war.
良い - 説明が良い	Drückt die Idee aus, dass der Typ und/oder die Quantität an Information und/oder die Methode der Informationsbereitstellung angemessen ist.
良い - 対応への賞賛	Ansichten, die von den obigen abweichen, die den Anbieter eines Service loben.
良い - 褒め・ 賞賛	Ansichten, die von den obigen abweichen, die die Merkmale, die Fähigkeiten und/oder den Betrieb eines bestimmten Gegenstands loben.
良い - 好き	Drückt den Wunsch aus, einen bestimmten Gegenstand zu besitzen oder diesem näherzukommen.
良い - 入会希望	Beschreibt den Wunsch, einer bestimmten Gruppe anzugehören bzw. Teil dieser Gruppe zu bleiben.
良い - 買いたい	Impliziert, dass man Geld für den Bezug eines bestimmten Gegenstands aufbringen möchte.
良い - 好評・ 人気	Gibt an, dass die Anzahl der Personen, die einen bestimmten Gegenstand möchten oder schätzen, ein bestimmtes Ziel überschritten hat.
良い - 売れた	Weist darauf hin, dass die Präsenz von Personen, die einen bestimmten Gegenstand kaufen oder die Anzahl oder der Wert der Käufe ein bestimmtes Ziel überschritten hat.
悪い - 悪い	Ausdrücke für generell negative Dinge, die sich als “schlecht” klassifizieren lassen.
悪い - 怒り	Ein deutliches Gefühl des Ärgers, wenn etwas nicht wie geplant läuft.
悪い - 批判	Drückt den Gedanken aus, dass eine andere Person nicht die passende Wahl getroffen hat.
悪い - お叱り	Wörter oder Aktionen, die eine Person so einschüchtern, dass sie den Absichten einer anderen Person folgen.
悪い - 誹謗・ 中傷	Wörter, mit denen eine äußerst schlechte Meinung von einer anderen Person demonstriert wird.
悪い - 軽蔑	Impliziert, dass der Charakter, die Fähigkeiten und/oder sonstigen Qualitäten einer anderen Person zu wünschen übrig lassen.

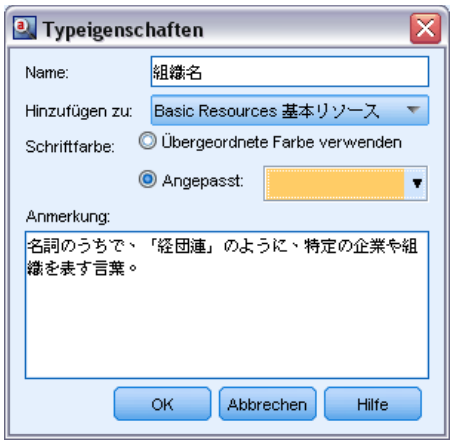
Typen	Beschreibung
悪い - 恨み	Drückt Ärger oder Verstimmung für einen Nachteil aus, der durch eine andere Person verursacht wurde.
悪い - 嫌がらせ	Wörter, die verwendet werden, um Kommunikation zu unterbinden.
悪い - 不満	Ein unangenehmes Gefühl, das sich aus der Unfähigkeit ergibt, den gewünschten Gegenstand oder Zustand zu erreichen.
悪い - 不味い	Gibt an, dass das Essen einen schlechten Geschmack hat.
悪い - 効果が不満	Impliziert, dass etwas nicht die erwartete Wirkung erzeugt hat.
悪い - 金額が不満	Deutet darauf hin, dass etwas vom Käuferstandpunkt gesehen einen nicht erwünschten Geldwert hat.
悪い - 対応への不満	Legt nahe, dass der Anbieter eines Service im Unrecht ist.
悪い - 対応が遅い	Legt nahe, dass ein Service nicht rechtzeitig ausgeführt/abgeschlossen wurde oder dass der Service noch abgeschlossen werden muss.
悪い - 対応が不親切	Bezeichnet ein unangenehmes Gefühl, das durch die Haltung oder das Verhalten eines Serviceanbieters verursacht wurde.
悪い - 説明が悪い	Drückt die Idee aus, dass der Typ und/oder die Quantität an Information und/oder die Methode der Informationsbereitstellung nicht angemessen ist.
悪い - 返答なし	Gibt an, dass der Anbieter eines Service nicht geeignet reagiert, obwohl die Situation eine entsprechende Reaktion verlangt.
悪い - 不快	Legt nahe, dass etwas als Stimulus oder Umgebung zur Auslösung einer negativen physiologischen Empfindung dient.
悪い - 怒り全般	Gefühle des Ärgers, die von den obigen abweichen. Allgemeiner Ärger in der Organisation oder Firma des Sprechers oder Beschreibungen von Ereignissen, die durch den erwähnten Ärger verursacht wurden.
悪い - 悲しい	Ein deutlich unangenehmes Gefühl, wenn jemand etwas verliert oder nicht bekommt.
悪い - 凶報	Drückt den Gedanken aus, dass ein bestimmtes Ziel trotz erheblichem Aufwand nicht erreicht werden konnte.
悪い - 不運	Gibt ein negatives Ergebnis an, das durch ein unglückliches Zusammentreffen oder Pech und nicht durch einen eigenen Fehler verursacht wurde.
悪い - ショック	Legt nahe, dass jemand durch eine unverhergesehene, negative Sache oder ein Ereignis aufgeregt ist und keine geeignete Antwort finden kann.
悪い - 残念	Ein unglückliches Gefühl, das entsteht, wenn ein erwartetes Geschehnis nicht eintritt.
悪い - 落胆	Ein Zustand, in dem einen ein unglückliches, enttäuschtes Gefühl überkommt.
悪い - 諦め	Legt nahe, dass eine negative Sache, die dem Sprecher oder einer anderen Person widerfahren ist, nicht verbessert werden kann.
悪い - 後悔	Drückt den Gedanken aus, dass man in der Vergangenheit versäumt hat, die korrekte Wahl zu treffen, obwohl diese als Option verfügbar war.
悪い - 謝罪	Gibt die Erkenntnis des Sprechers an, dass dieser einer anderen Person Schaden zugefügt hat.
悪い - 淋しい	Drückt den Gedanken aus, dass Kontakt mit anderen selten stattfindet oder dass nur mit wenigen anderen Personen Kontakt möglich ist.
悪い - 哀れみ	Drückt den Gedanken aus, dass die Situation einer anderen Person erheblich schlechter als die des Sprechers ist.
悪い - 悩み	Gibt an, dass eine Auswahl getroffen werden muss, man aber nicht in der Lage ist, aus den verfügbaren Optionen zu wählen.
悪い - 困っている	Drückt den Gedanken aus, dass es keine wirksame Methode gibt, auf eine Situation zu reagieren, die eine Aktion erfordert.

Typen	Beschreibung
悪い - 苦しい	Drückt einen unangenehmen psychologischen Zustand aus, in dem man aufgrund von externen Ursachen oder eigener Fehler nicht normal handeln kann.
悪い - 体が悪い状態	Beschreibt einen Zustand, in dem der Körper krank, verletzt und/oder müde ist, oder einen Zustand, in dem sich die physische Verfassung nicht verbessert.
悪い - 不安	Drückt den Gedanken aus, dass etwas nicht in seinem wünschenswerten Zustand fortfährt oder die Erwartungen nicht erfüllt.
悪い - 恐怖	Legt nahe, dass eine bestimmte Sache wahrscheinlich Schaden oder Verletzungen verursacht.
悪い - 悲しみ全般	Von den obigen abweichende Gefühle der Traurigkeit, z. B. allgemeine Traurigkeit über eine nicht spezifizierte Sache.
悪い - 嫌い	Gibt an, dass man eine Sache auf Abstand halten oder sich von einer Sache entfernen möchte.
悪い - 退会希望	Beschreibt den Wunsch, eine bestimmte Gruppe zu verlassen bzw. dieser Gruppe nicht beizutreten.
悪い - 買いたくない	Legt nahe, dass man eine bestimmte Sache nicht wünscht oder nicht beabsichtigt, für diese Sache zu bezahlen.
悪い - 不評・ 不人気	Gibt an, dass die Anzahl an Personen, die eine bestimmte Sache mögen, nicht einem bestimmten Ziel entspricht bzw. dass viele Personen der erwähnten Sache gegenüber negative Gefühle hegen.
悪い - 売れていない	Weist darauf hin, dass die Abwesenheit von Personen, die einen bestimmten Gegenstand kaufen, oder die Anzahl oder der Wert der Käufe ein bestimmtes Ziel nicht erreicht hat.
その他 - 疑問	Ausdrücke, die Informationen anfordern, zu denen die andere Person weitere Untersuchungen oder Gedanken aufbringen muss.
その他 - 問い合わせ	Ausdrücke, die Informationen anfordern, die bereits im Besitz der anderen Person sind.
その他 - 要望	Ausdrücke, die der anderen Person befehlen, ein Problem zu lösen (wenn die andere Person direkt Schuld ist oder eine niedrigere Position als der Sprecher hat).
その他 - 提案・ 忠告	Ausdrücke, die der anderen Person befehlen, sich besser zu benehmen (wenn die andere Person direkt Schuld ist oder eine niedrigere Position als der Sprecher hat).
その他 - お願い	Ausdrücke, die der anderen Person befehlen, ein Problem zu lösen (wenn die andere Person nicht Schuld ist oder keine niedrigere Position als der Sprecher hat).
その他 - 激励	Ausdrücke, die eine andere Person ermutigen, oder Beschreibungen eines ermutigenden Verhaltens.
その他 - 勧誘	Ausdrücke, die einer anderen Person befehlen, gemeinsam mit dem Sprecher etwas zu unternehmen.
その他 - 驚き	Drückt den Gedanken aus, dass die Plötzlichkeit oder der Grad eines Ereignisses über ein rationales Urteil/Verständnis hinausgeht.
評価なし - 評価なし	Kein Ausdruck der Bewertung.

Bearbeiten von japanischen Typeigenschaften

Sie können zwar in japanischen Ressourcen keine Typen erstellen, jedoch können Sie Typeigenschaften anzeigen und bearbeiten. Beachten Sie, dass die Optionen wie die Abgleichoption und flektierte Formen nicht für japanischen Text gelten.

Abbildung A-4
Dialogfeld "Typeinstellungen" für japanische Textressourcen



Name. Der Name des Typwörterbuchs.

Hinzufügen zu. Dieses Feld gibt die Bibliothek an, in der Sie Ihr neues Typ-Wörterbuch erstellen.

Schriftfarbe. Über dieses Feld können Sie ein Unterscheidungsmerkmal für diesen Typ in Bezug auf die anderen auf der Benutzeroberfläche dargestellten Typen festlegen. Wenn Sie die Option Übergeordnete Farbe verwenden auswählen, wird für dieses Typ-Wörterbuch die Standardtypfarbe verwendet. Die Standardfarbe wird im Dialogfeld "Optionen" festgelegt. Für weitere Informationen siehe Thema Optionen: Registerkarte "Anzeige" in Kapitel 8 auf S. 136. Wenn Sie Benutzerdefiniert auswählen, wählen Sie aus der Dropdown-Liste eine Farbe aus.

Anmerkung. Dieses Feld ist optional und wird für Kommentare und Beschreibungen verwendet.

So können Sie die Typeigenschaften anzeigen bzw. bearbeiten:

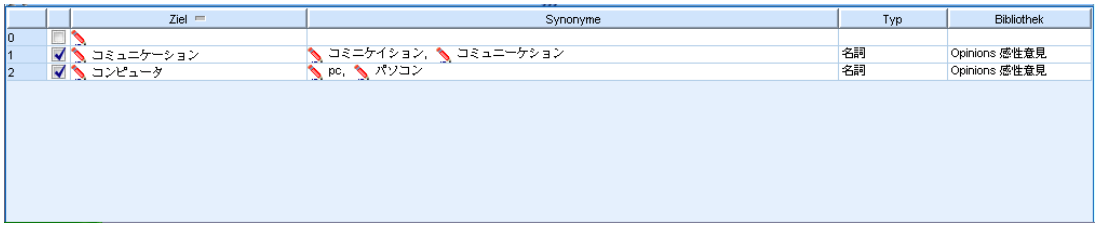
- Wählen Sie die Typeigenschaften aus, die Sie sehen möchten.
- Klicken Sie mit der rechten Maustaste und wählen Sie im Kontextmenü Typeigenschaften aus. Das Dialogfeld "Typ-Eigenschaften" wird geöffnet.
- Nehmen Sie alle erforderlichen Änderungen vor.
- Klicken Sie auf OK, um die Änderungen im Typ-Wörterbuch zu speichern.

Verwenden des Synonymwörterbuchs für japanischen Text

Für japanischen Text enthält das Substitutionswörterbuch nur eine Registerkarte zur Verwaltung Ihrer Synonyme, nämlich die Registerkarte "Synonyme". Synonyme verknüpfen zwei oder mehr Wörter mit derselben Bedeutung. Mithilfe von Synonymen können Sie außerdem Fachausdrücke mit ihren Abkürzungen gruppieren oder auch falsch geschriebene Wörter mit dem Fachausdruck in der richtigen Schreibweise.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Abbildung A-5
Synonymeinträge für japanischen Text



Eine Synonymdefinition setzt sich aus zwei Teilen zusammen. Der Zielausdruck ist der Ausdruck, unter dem die Extrahierungsengine alle Synonyme zusammenfassen soll. Wenn dieser Zielausdruck nicht als Synonym eines anderen Zielausdrucks verwendet wird oder er ausgeschlossen wird, ist es wahrscheinlich, dass dieser Ausdruck das Konzept wird, das im Bereich “Extrahierungsergebnisse” angezeigt wird. Die Liste der Synonyme sind die Ausdrücke, die unter dem Zielbegriff zusammengefasst werden.

Auf der Registerkarte “Synonyme” können Sie in die Leerzeile am Anfang der Tabelle eine Synonymdefinition eingeben. Beginnen Sie, indem Sie den Zielausdruck und seine Synonyme definieren. Sie können auch die Bibliothek auswählen, in der diese Definition gespeichert werden soll. Während der Extrahierung werden alle Fundstellen der Synonyme für die endgültige Extrahierung unter dem Zielausdruck gruppiert. Für weitere Informationen siehe Thema Hinzufügen von Fachausdrücken in Kapitel 17 auf S. 316.

Wenn Sie Ihre Typ-Wörterbücher erstellen, können Sie einen Ausdruck eingeben, für den Ihnen drei oder vier Synonyme einfallen. In diesem Fall könnten Sie alle Ausdrücke und anschließend Ihren Zielausdruck in das Substitutionswörterbuch eingeben und dann die Synonyme durch Ziehen übertragen.

Wichtig: Platzhalter- und Sonderzeichen werden für japanische Textsynonyme nicht unterstützt.

So fügen Sie einen Synonymeintrag hinzu:

- Geben Sie in der leeren Zeile am Anfang der Tabelle in der Registerkarte “Synonym” im Bereich “Substitution” Ihren Zielausdruck in die Zielspalte ein. Der von Ihnen eingegebene Zielausdruck wird farbig angezeigt. Die Farbe stellt den Typ dar, als der der Ausdruck erscheint oder in den er erzwungen wird, sofern dies der Fall ist. Wenn der Ausdruck schwarz dargestellt wird, bedeutet dies, dass er in keinem der Typ-Wörterbücher vorkommt.

- Klicken Sie auf die zweite Zelle rechts neben dem Zielausdruck und geben Sie die Synonyme ein. Trennen Sie die einzelnen Einträge mithilfe des globalen Trennzeichens, das im Dialogfeld “Optionen” definiert ist. Alle eingegebenen Synonyme sollten denselben Typ aufweisen. Für weitere Informationen siehe Thema Festlegen von Optionen in Kapitel 8 auf S. 135. Die von Ihnen eingegebenen Ausdrücke werden farbig angezeigt. Die Farbe stellt den Typ dar, in der der Ausdruck angezeigt wird. Wenn der Ausdruck schwarz dargestellt wird, bedeutet dies, dass er in keinem der Typ-Wörterbücher vorkommt.

- Legen Sie in der dritten Spalte, der Spalte “Typ”, einen Typ für diese Synonyme fest. Das Ziel übernimmt jedoch den Typ, der bei der Extrahierung zugewiesen wurde. Wenn allerdings das Ziel nicht als ein Konzept extrahiert wurde, wird der in dieser Spalte aufgelistete Typ dem Ziel in den Extrahierungsergebnissen zugewiesen.

- Klicken Sie auf die letzte Zelle, um die Bibliothek auszuwählen, in der die Synonymdefinition gespeichert werden soll.

Hinweis: Diese Anweisungen zeigen Ihnen, wie Sie in der Resource Editor-Ansicht oder im Template Editor Änderungen vornehmen können. Beachten Sie, dass Sie solche Feinabstimmungen auch direkt im Bereich “Extrahierungsergebnisse”, im Bereich “Daten”, im Bereich “Kategorien” oder im Dialogfeld “Clusterdefinitionen” in den anderen Ansichten vornehmen können. Für weitere Informationen siehe Thema Extrahierungsergebnisse verfeinern in Kapitel 9 auf S. 158.

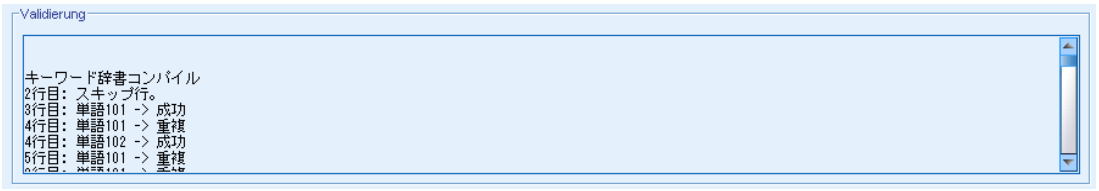
Validieren und Kompilieren von japanischen Ressourcen

Für japanischen Text gibt es einen zusätzlichen Validierungsbereich, in dem Sie Ihre Ressourcen vor der Extrahierung prüfen können. Bevor der Extrahierungsprozess für japanischen Text beginnt, kompiliert die Extrahierungsengine automatisch die Ressourcen, wenn vor dem Start des Extrahierungsprozesses Änderungen erkannt wurden. Wenn bei der Extrahierung ein Fehler gefunden wird, kann der Vorgang möglicherweise nicht korrekt abgeschlossen werden.

Zur Vermeidung von Kompilierungsfehlern wird empfohlen, dass Sie Ihre Ressourcen nach dem Durchführen von Änderungen im Resource Editor oder Template Editor validieren und kompilieren. Wenn Fehlermeldungen auftreten, können Sie Korrekturen ausführen und die Validierung erneut versuchen.

Anmerkung: Extraktion für japanischen Text steht in IBM® SPSS® Modeler Premium zur Verfügung.

Abbildung A-6
Validierungsbereich für japanischen Text



So validieren Sie Ressourcen:

- Wählen Sie in den Menüs Extras > Ressourcen-Validierung. Der Validierungsbereich wird geöffnet, um Kompilierungs- und Fehlermeldungen anzuzeigen.

Weitere Ausnahmen für Japanisch

Interne Ressourcen, die benutzerdefinierte Ressourcen überschreiben

Für japanischen Text umfassen die Standardressourcen einige vorkompilierte, interne Basisressourcen. Diese internen Ressourcen sind nicht editierbar. Dennoch können Sie Resource Editor oder Template Editor verwenden, um einige Änderungen und Feineinstellungen vorzunehmen. In beinahe allen Fällen haben die Fachausdrücke, Synonyme und Ausschlusslisteneinträge, die Sie in Ihren Ressourcen definieren, Vorrang vor den vorkompilierten

internen Ressourcen. Jedoch gibt es mehrere Ausnahmen, wie in einigen der folgenden Beispiele illustriert.

- Es gibt Instanzen, in denen das Hinzufügen von Fachausdrücken zu einem bestimmten Typ keine Wirkung auf die Extrahierungsergebnisse hat. Dies tritt am wahrscheinlichsten auf, wenn die Daten lange Sätze mit mehreren morphologischen Elementen, Interpunktion oder Symbolen enthalten. Da in japanische Textressourcen zudem eine große Anzahl an häufigen Fachausdrücken vorkompiliert ist, gibt es einige häufige Wörter, die stets in eine bestimmte linguistische Definition gezwungen werden.
- Eventuell können Sie Fachausdrücke wie ある, いる oder なる nicht ausschließen, da die Extrahierungsengine die Extrahierung dieser Fachausdrücke immer erzwingt.
- Es ist zwar möglich, den Typ des Fachausdrucks 東京 von <地名> in <名詞> zu ändern, jedoch ignoriert die Extrahierungsengine Ihre Änderung, wenn Sie versuchen, mit dem Stichwort- (Typ-) Wörterbuch den Typ eines Fachausdrucks von <地名> in <動詞> oder in <形容詞> zu ändern.
- Gelegentlich ist es möglich, dass Änderungen, die Sie im Resource Editor oder im Template Editor vornehmen, die Extrahierungsergebnisse von einem Satz beeinflussen, jedoch nicht von einem anderen Satz, da der Extrahierungsvorgang durch Verweisen auf die Kookkurenzwörter in jedem Satz endet.

Probleme mit Katakana-Anzeige in halber Breite

Katakana-Zeichen in halber Breite werden bei der Extrahierung intern in Katakana-Zeichen in voller Breite umgewandelt, erscheinen aber dennoch im Datenbereich in der interaktiven Workbench-Sitzung (nur für Text-Mining-Knoten) in der ursprünglichen halben Breite. Beachten Sie, dass Katakana-Zeichen in halber Breite im Datenbereich nicht markiert werden können. Konvertieren Sie zur Vermeidung dieses Problems alle Ihre Datensätze vor der Verarbeitung in Katakana in voller Breite.

Verwendung von Klein- und Großschreibung

Großbuchstaben werden beim Einlesen in die Anwendung temporär in Kleinbuchstaben umgewandelt. Jedoch zeigt der Datenbereich den Text stets in derselben Schreibung wie im Originaltext an. Klein- und Großbuchstaben werden in diesem Produkt als identisch behandelt.

Anhang

B

Hinweise

Diese Informationen wurden für weltweit angebotene Produkte und Dienstleistungen erarbeitet.

IBM bietet die in diesem Dokument behandelten Produkte, Dienstleistungen oder Merkmale möglicherweise nicht in anderen Ländern an. Informationen zu den in derzeit in Ihrem Land erhältlichen Produkten und Dienstleistungen erhalten Sie bei Ihrem zuständigen IBM-Mitarbeiter vor Ort. Mit etwaigen Verweisen auf Produkte, Programme oder Dienste von IBM soll nicht behauptet oder impliziert werden, dass nur das betreffende Produkt oder Programm bzw. der betreffende Dienst von IBM verwendet werden kann. Stattdessen können alle funktional gleichwertigen Produkte, Programme oder Dienste verwendet werden, die keine geistigen Eigentumsrechte von IBM verletzen. Es obliegt jedoch der Verantwortung des Benutzers, die Funktionsweise von Produkten, Programmen oder Diensten von Drittanbietern zu bewerten und zu überprüfen.

IBM verfügt möglicherweise über Patente oder hat Patentanträge gestellt, die sich auf in diesem Dokument beschriebenen Inhalte beziehen. Durch die Bereitstellung dieses Dokuments werden Ihnen keinerlei Lizenzen an diesen Patenten gewährt. Lizenzanfragen können schriftlich an folgende Adresse gesendet werden:

IBM Director of LicensingIBM CorporationNorth Castle DriveArmonk, NY 10504-1785USA

Bei Lizenzanfragen in Bezug auf DBCS-Daten (Double-Byte Character Set) wenden Sie sich an die für geistiges Eigentum zuständige Abteilung von IBM in Ihrem Land. Schriftliche Anfragen können Sie auch an folgende Adresse senden:

Intellectual Property LicensingLegal and Intellectual Property LawIBM Japan Ltd.1623-14, Shimotsuruma, Yamato-shiKanagawa 242-8502 Japan

Der folgende Absatz gilt nicht für Großbritannien oder andere Länder, in denen derartige Bestimmungen nicht mit dem dort geltenden Recht vereinbar sind. INTERNATIONAL BUSINESS MACHINES ÜBERNIMMT FÜR DIE VORLIEGENDE DOKUMENTATION KEINERLEI GEWÄHRLEISTUNG IRGENDWELCHER ART, WEDER AUSDRÜCKLICH NOCH STILLSCHWEIGEND, EINSCHLIESSLICH, JEDOCH NICHT DARAUF BEGRENZT, DER STILLSCHWEIGENDEN GEWÄHRLEISTUNGEN IN BEZUG AUF DIE NICHTVERLETZUNG VON RECHTEN DRITTER, AUF HANDELSÜBLICHKEIT ODER DIE EIGNUNG FÜR EINEN BESTIMMTEN ZWECK. Einige Staaten lassen bei bestimmten Transaktionen keine Ausschlussklauseln ausdrücklicher oder stillschweigender Gewährleistungen zu, sodass diese Erklärung möglicherweise nicht auf Sie zutrifft.

Diese Informationen können technische Ungenauigkeiten oder typografische Fehler enthalten. An den hierin enthaltenen Informationen werden in regelmäßigen Abständen Änderungen vorgenommen, die in spätere Ausgaben der Publikation eingearbeitet werden. IBM kann jederzeit ohne Vorankündigung Verbesserungen und/oder Veränderungen an den in dieser Publikation beschriebenen Produkten und/oder Programmen vornehmen.

Alle in diesen Ausführungen enthaltenen Verweise auf Websites, die nicht zu IBM gehören, dienen lediglich der Information. Die Nennung bedeutet nicht, dass IBM den Inhalt dieser Websites unterstützt. Das Material auf diesen Websites ist kein Bestandteil des Materials für dieses IBM-Produkt. Sie verwenden diese Websites auf eigenes Risiko.

IBM ist berechtigt, die von Ihnen bereitgestellten Informationen in jeglicher Form zu verwenden bzw. weiterzugeben, die dem Unternehmen geeignet erscheint, ohne dass ihm daraus Verbindlichkeiten Ihnen gegenüber entstehen.

Lizenznehmer dieses Programms, die Informationen dazu benötigen, wie (i) der Austausch von Informationen zwischen unabhängig erstellten Programmen und anderen Programmen und (ii) die gegenseitige Verwendung dieser ausgetauschten Informationen ermöglicht wird, wenden sich an:

IBM Software GroupAttention: Licensing233 S. Wacker Dr.Chicago, IL 60606USA

Derartige Informationen stehen ggf. in Abhängigkeit von den jeweiligen Geschäftsbedingungen sowie in einigen Fällen der Zahlung einer Gebühr zur Verfügung.

Das in diesem Dokument beschriebene lizenzierte Programm und sämtliche dafür verfügbaren lizenzierten Materialien werden von IBM gemäß dem IBM-Kundenvertrag, den internationalen Nutzungsbedingungen für Programmpakete der IBM oder einer anderen zwischen uns getroffenen Vereinbarung bereitgestellt.

Alle in diesem Dokument enthaltenen Leistungsdaten wurden in einer kontrollierten Umgebung ermittelt. Daher können die unter anderen Betriebsumgebungen erzielten Ergebnisse erheblich abweichen. Einige Messungen wurden möglicherweise an Systemen im Entwicklungsstadium vorgenommen und es besteht keine Garantie, dass spätere allgemein verfügbare Systeme dieselben Messwerte aufweisen. Außerdem wurden einige Messwerte möglicherweise mittels Extrapolation geschätzt. Die tatsächlichen Ergebnisse können abweichen. Die Benutzer dieses Dokuments sollten die entsprechenden Daten für ihre jeweilige Umgebung überprüfen.

Informationen zu Drittanbieterprodukten stammen von den Herstellern dieser Produkte, ihren veröffentlichten Verlautbarungen oder aus anderen öffentlich verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher die Richtigkeit der Angaben zu Leistung und Kompatibilität oder anderer Behauptungen in Bezug auf Drittanbieterprodukte nicht bestätigen. Fragen zu den Funktionen von Drittanbieterprodukten sind an die Hersteller dieser Produkte zu richten.

Diese Informationen enthalten Beispiele für Daten und Berichte, die in alltäglichen Betriebsabläufen verwendet werden. Um sie möglichst umfassend darzulegen, enthalten die Beispiele Namen von Einzelpersonen, Unternehmen, Marken und Produkten. Alle diese Namen sind frei erfunden und jegliche Ähnlichkeit mit den von einem tatsächlichen Handelsunternehmen verwendeten Namen und Adressen ist rein zufällig.

Bei der Anzeige dieser digitalen Informationsversion sind die Fotografien und Farbillustrationen möglicherweise nicht sichtbar.

Marken

IBM, das IBM-Logo, ibm.com und SPSS sind Marken von IBM Corporation, die in vielen Ländern weltweit eingetragen sind. Eine aktuelle Liste der IBM-Marken finden Sie im Internet unter *http://www.ibm.com/legal/copytrade.hmtl*.

Index