IBM SPSS Modeler Text Analytics 16 Benutzerhandbuch

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 249 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 16, Release 0, Modifikation 0 von IBM SPSS Modeler Text Analytics und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM SPSS Modeler Text Analytics 16, User's Guide,
herausgegeben von International Business Machines Corporation, USA

Herausgegeben von:
TSC Germany
Kst. 2877
Oktober 2013

Inhaltsverzeichnis

Vorwort

IBM® SPSS Modeler Text Analytics bietet leistungsstarke Textanalysefunktionen, die mithilfe hoch entwickelter linguistischer Technologien und Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) eine schnelle Verarbeitung einer großen Vielfalt an unstrukturierten Textdaten ermöglichen und die Schlüsselkonzepte aus diesem Text extrahieren und ordnen. Darüber hinaus können diese Konzepte mit IBM 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. Wortfolgen, auf intelligente Art und Weise extrahieren. Außerdem ermöglicht die Kenntnis der zugrunde liegenden Sprache eine Klassifizierung von Termen 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. IBM SPSS Modeler Text Analytics wird mit einer Reihe linguistischer Ressourcen wie Wörterbüchern für Terme und Synonyme, Bibliotheken und Vorlagen bereitgestellt. Darüber hinaus können Sie mit diesem Produkt die linguistischen Ressourcen Ihrem Umfeld entsprechend entwickeln und optimieren. Bei der Optimierung 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 Anwendungen für Business Intelligence, Vorhersageanalyse, Finanz- und Strategiemanagement sowie Analysen 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 staatlichen Lehr- und Forschungseinrichtungen weltweit nutzen IBM SPSS-Technologie als Wettbewerbsvorteil für die 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.

Technical Support

Kunden mit Wartungsvertrag können den Technical Support in Anspruch nehmen. Kunden können sich an den Technical Support wenden, wenn sie Hilfe bei der Arbeit mit IBM Produkten oder bei der Installation in einer der unterstützten Hardwareumgebungen benötigen. Zur Kontaktaufnahme mit dem Technical Support besuchen Sie die IBM Website unter http://www.ibm.com/support. Sie müssen bei der Kontaktaufnahme Ihren Namen, Ihre Organisation und Ihre Supportvereinbarung angeben.

Kapitel 1. IBM SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics bietet leistungsstarke Textanalysefunktionen, die mithilfe hoch entwickelter linguistischer Technologien und Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) eine schnelle Verarbeitung einer großen Vielfalt an unstrukturierten Textdaten ermöglichen und die Schlüsselkonzepte aus diesem Text extrahieren und ordnen. Darüber hinaus können diese Konzepte mit IBM 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. Wortfolgen, auf intelligente Art und Weise extrahieren. Außerdem ermöglicht die Kenntnis der zugrunde liegenden Sprache eine Klassifizierung von Termen 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. IBM SPSS Modeler Text Analytics wird mit einer Reihe linguistischer Ressourcen wie Wörterbüchern für Terme und Synonyme, Bibliotheken und Vorlagen bereitgestellt. Darüber hinaus können Sie mit diesem Produkt die linguistischen Ressourcen Ihrem Umfeld entsprechend entwickeln und optimieren. Bei der Optimierung 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 Textmining-Streams für die Echtzeitbewertung unstrukturierter Daten bereitstellen. Die Möglichkeit, diese Streams zu verwenden, gewährleistet erfolgreiche Textmining-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 IBM SPSS Modeler Text Analytics mit IBM SPSS Modeler Solution Publisher auszuführen, fügen Sie das Verzeichnis `<Installationsverzeichnis>/ext/bin/spss.TMWBServer` der Umgebungsvariablen `$LD_LIBRARY_PATH` hinzu.

Automatische Übersetzung unterstützter Sprachen. IBM 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 Textmining-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 Übersetzungsforschung sind.

Upgrade auf IBM SPSS Modeler Text Analytics Version 16

Aktualisierung früherer Versionen von PASW Text Analytics oder Text Mining für Clementine

Vor Installation der IBM SPSS Modeler Text Analytics-Version 16 müssen Sie alle TAP-Dateien, Vorlagen und Bibliotheken in der aktuellen Versionen, die Sie in der neuen Versionen verwenden möchten, speichern und exportieren. Es wird empfohlen, diese Dateien in einem Verzeichnis zu speichern, das bei der Installation der neuesten Version nicht gelöscht oder überschrieben wird.

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

Wichtig! Wenn Sie die aktuelle Version deinstallieren, ohne zuvor die Dateien zu speichern oder zu exportieren, gehen in der vorherigen Version erstellte TAP-Dateien, Vorlagen und öffentlichen Bibliotheken verloren und können deshalb nicht in IBM SPSS Modeler Text Analytics Version 16 verwendet werden.

Informationen zum Textmining

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

Textmining 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. Terme kennen müssen, die die Autoren verwendet haben, um diese Konzepte auszudrücken. Obwohl es sehr große Unterschiede gibt, wird Textmining zuweilen mit Informationsrückgewinnung verwechselt. Das genaue Erfassen und Speichern von Informationen ist zwar eine große Herausforderung, doch Extraktion und Verwaltung von qualitativ hochwertigen Inhalten, Terminologie und Beziehungen, die in den Informationen enthalten sind, stellen entscheidende und heikle Prozesse dar.

Textmining und Data-Mining

Für jeden Textartikel gibt linguistisch basiertes Textmining 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 Textmining 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 in demselben Verzeichnis gespeichert werden. Bei Datenbanken muss das Feld ermittelt werden, das den Text enthält.
2. **Anwenden des Minings auf den Text und Extraktion strukturierter Daten.** Anwenden der Textmining-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 , sodass 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 Extraktion 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, sogenanntes "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 Textmining* bezeichnet.

Beim *linguistisch basierten Textmining* dagegen werden die Prinzipien der Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) - der computerunterstützten Analyse menschlicher Sprachen - auf die Analyse der Wörter, Wortfolgen und der Syntax (Struktur) des Texts angewendet. Ein System, das NLP verwendet, kann Konzepte, u. a. Wortfolgen, 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 Textmining 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 Extraktionsprozess mit Texten in allen Sprachen außer Japanisch dient die Überlegung, wie der jeweilige Ansatz auf eine Abfrage zum Term `Dokumentreproduktion` reagieren würde. Sowohl bei den statistikbasierten als auch bei den linguistisch basierten Lösungen müsste eine Erweiterung 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 Generierung einer Reihe von irrelevanten Ergebnissen führt. Das Verstehen von Sprache beseitigt die Mehrdeutigkeit von Texten, weshalb das linguistisch basierte Textmining definitionsgemäß den verlässlicheren Ansatz darstellt.

Hinweis: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

Die Verwendung linguistisch basierter Verfahren über die Stimmungsanalysefunktion ermöglicht es, bedeutungsreichere Ausdrücke zu extrahieren. Die Analyse und Erfassung von Stimmungen beseitigt die Mehrdeutigkeit von Texten, weshalb das linguistisch basierte Textmining definitionsgemäß den verlässlicheren Ansatz darstellt.

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

- Konvertieren von Quellendaten 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 Quellendaten 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 infrage kommenden Termen

Es ist wichtig zu verstehen, welche Rolle die linguistischen Ressourcen während der linguistischen Extraktion bei der Ermittlung von infrage kommenden Termen spielen. Linguistische Ressourcen kommen jedes Mal zum Einsatz, wenn ein Extraktionsvorgang ausgeführt wird. Sie liegen in Form von Vorlagen, Bibliotheken und kompilierten Ressourcen vor. Bibliotheken bestehen aus Wortlisten, Beziehungen und weiteren Informationen, die eingesetzt werden, um die Extraktion abzustimmen oder zu spezifizieren. Die kompilierten Ressourcen können nicht angezeigt oder bearbeitet werden. Die übrigen Ressourcen können jedoch im Vorlageneditor bzw., wenn eine interaktive Workbenchsitzung gestartet wurde, im Ressourceneditor bearbeitet werden.

Kompilierte Ressourcen sind interne Kernkomponenten der Extraktionsengine in IBM SPSS Modeler Text Analytics. Diese Ressourcen umfassen ein allgemeines Wörterbuch, in dem eine Liste von Grundformen mit einem Code für die Wortart enthalten ist (Nomen, Verb, Adjektiv usw.).

Zusätzlich zu diesen kompilierten Ressourcen sind auch mehrere Bibliotheken im Lieferumfang enthalten. Diese können verwendet werden, um die Typen und Konzeptdefinitionen der kompilierten Ressourcen zu ergänzen und Synonyme zu liefern. Diese Bibliotheken - sowie sämtliche benutzerdefinierte Bibliotheken, die Sie erstellen - bestehen aus mehreren Wörterbüchern. Diese umfassen Typwörterbücher, Synonymwörterbücher sowie Ausschlusswörterbücher.

Sobald die Daten importiert und konvertiert wurden, beginnt die Extraktionsengine, Kandidaten für die Extraktion zu identifizieren. Infrage kommende Terme (Kandidaten) sind Wörter oder Wortgruppen, die verwendet werden, um Konzepte im Text zu ermitteln. Bei der Verarbeitung des Texts werden einzelne Wörter (**Uniterme**) und zusammengesetzte Wörter (**Multiterme**) über Extraktoren auf der Grundlage von Wortartmustern (POS-Muster; POS - Part of Speech) ermittelt. Anschließend werden mithilfe der Stimmungstextlinkanalyse Kandidaten für Stimmungsstichwörter identifiziert.

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

Schritt 3. Ermittlung von Äquivalenzklassen und Integration von Synonymen

Im Anschluss an die Ermittlung von Unitermen und Multitermen, die als Kandidaten infrage kommen, werden über ein Normalisierungswörterbuch der Software Äquivalenzklassen ermittelt. Bei einer Äquivalenzklasse handelt es sich um eine Grundform einer Wortfolge oder einer einzelnen Form von zwei Varianten derselben Wortfolge. Um festzustellen, welches Konzept für die betreffende Äquivalenzklasse als Hauptterm verwendet wird, werden die folgenden Regeln in der aufgeführten Reihenfolge durch die Extraktionsengine 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 kompilierte Ressourcen als auch die Bibliotheken verwendet. Zu den Typen gehören beispielsweise übergeordnete Konzepte, positive und negative Wörter, Vornamen, Orte, Organisationen und anderes. Weitere Informationen finden Sie im Thema „Typwörterbücher" auf Seite 201.

Beachten Sie, dass japanische Textressourcen über ein anderes Set an Typen verfügen.

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 IBM SPSS Modeler Text Analytics.

Funktionsweise der Extraktion

Während der Extraktion 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 Extraktionsprozess eine andere Schrittfolge aus. *Hinweis*: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

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

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

Schritt 1. Konvertieren von Quellendaten 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 infrage kommenden Termen

Es ist wichtig zu verstehen, welche Rolle die linguistischen Ressourcen während der linguistischen Extraktion bei der Ermittlung von infrage kommenden Termen spielen. Linguistische Ressourcen kommen jedes Mal zum Einsatz, wenn ein Extraktionsvorgang ausgeführt wird. Sie liegen in Form von Vorlagen, Bibliotheken und kompilierten Ressourcen vor. Bibliotheken bestehen aus Wortlisten, Beziehungen und weiteren Informationen, die eingesetzt werden, um die Extraktion abzustimmen oder zu spezifizieren. Die kompilierten Ressourcen können nicht angezeigt oder bearbeitet werden. Die übrigen Ressourcen (Vorlagen) können jedoch im Vorlageneditor bzw., wenn eine interaktive Workbenchsitzung gestartet wurde, im Ressourceneditor bearbeitet werden.

Kompilierte Ressourcen sind interne Kernkomponenten der Extraktionsengine 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 reservierte integrierte Typen, die verwendet werden, um den folgenden Typen eine Vielzahl von extrahierten Termen zuzuweisen: `<Location>`, `<Organization>` oder `<Person>`. Weitere Informationen finden Sie im Thema „Integrierte Typen" auf Seite 202.

Zusätzlich zu diesen kompilierten Ressourcen sind auch mehrere Bibliotheken im Lieferumfang enthalten. Diese können verwendet werden, um die Typen und Konzeptdefinitionen der kompilierten 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 Typwörterbücher, Substitutionswörterbücher (Synonyme und optionale Elemente) sowie Ausschlusswörterbücher. Weitere Informationen finden Sie im Thema Kapitel 16, „Arbeiten mit Bibliotheken", auf Seite 191.

Sobald die Daten importiert und konvertiert wurden, beginnt die Extraktionsengine, Kandidaten für die Extraktion zu identifizieren. Infrage kommende Terme (Kandidaten) sind Wörter oder Wortgruppen, die verwendet werden, um Konzepte im Text zu ermitteln. Während der Verarbeitung des Texts werden einzelne Wörter (**Uniterme**), die nicht in den kompilierten Ressourcen enthalten sind, als infrage kommende Terme für die Extraktion betrachtet. Kandidaten, die aus zusammengesetzten Wörtern bestehen (**Multiterme**), werden über Extraktoren auf der Grundlage von Wortartmustern (POS-Muster; POS - Part of

Speech) ermittelt. Der Multiterm Sportwagen, der dem Wortartmuster "Adjektiv-Nomen" entspricht, besteht beispielsweise aus zwei Komponenten. Der Multiterm schneller Sportwagen, der dem Wortartmuster "Adjektiv-Adjektiv-Nomen" entspricht, besteht aus drei Komponenten.

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

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

Schritt 3. Ermittlung von Äquivalenzklassen und Integration von Synonymen

Im Anschluss an die Ermittlung von Unitermen und Multitermen, 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 einer Wortfolge oder einer einzelnen Form von zwei Varianten desselben Wortfolge. Wortfolgen 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 Extraktionsengine 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 kompilierte 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. Weitere Informationen finden Sie im Thema „Typwörterbücher” auf Seite 201.

Schritt 5. Indexerstellung

Das gesamte Set der Datensätze oder Dokumente wird indiziert. Dazu wird ein Zeiger zwischen einer Textposition 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 Ereignisextraktion.

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.

Funktionsweise der Kategorisierung

Bei der Erstellung von Kategoriemodellen mit IBM SPSS Modeler Text Analytics haben Sie die Wahl zwischen verschiedenen Verfahren, 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 Verfahren experimentieren, um zu erkennen, mit welchem Sie die besten Ergebnisse für Ihre Textdaten erzielen. In IBM SPSS Modeler Text Analytics können Sie Kategoriemodelle in einer Workbenchsitzung erstellen, in der Sie eine Untersuchung oder weitere Optimierung Ihrer Kategorien vornehmen können.

In diesem Handbuch bezieht sich **Kategorieerstellung** auf die Generierung von Kategoriedefinitionen und Klassifikation über mindestens eine integrierte Methode und **Kategorisierung** bezieht sich auf den Scoring- oder Beschriftungsprozess, bei dem den Kategoriedefinitionen für jeden Datensatz oder jedes Dokument eindeutige 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.

IBM SPSS Modeler Text Analytics bietet Ihnen mehrere Methoden zur automatisierten Kategorieerstellung, mit denen Sie Ihre Dokumente oder Datensätze schnell kategorisieren können.

Gruppierungsverfahren

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 kombinieren, um die Dokumente oder Datensätze vollständig zu erfassen. Möglicherweise erkennen Sie ein Konzept in mehreren Kategorien oder finden redundante Kategorien vor.

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. Weitere Informationen finden Sie im Thema „Konzeptwurzelableitung” auf Seite 124. 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. Dieses Verfahren wird empfohlen, wenn die Konzepte dem semantischen Netz bekannt und nicht zu mehrdeutig sind. Es ist weniger hilfreich, wenn der Text eine 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 gleichgeordnete Elemente 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 diesem Release nur für englischen Text verfügbar. Weitere Informationen finden Sie im Thema „Semantische Netze” auf Seite 126.

Konzepteinbeziehung. Dieses Verfahren erstellt Kategorien durch die Gruppierung von Multiterm-Konzepten (zusammengesetzte Wörter) basierend darauf, ob sie Wörter enthalten, die Subsets oder Supersets

eines Worts in dem anderen sind. So wird beispielsweise `Sitz` mit `Kindersitz`, `Sitzheizung` und `Kindersitzgurt` zu einer Gruppe zusammengefasst. Weitere Informationen finden Sie im Thema „Konzepteinbeziehung" auf Seite 125.

Kookkurrenz. Dieses Verfahren erstellt Kategorien aus Kookkurrenzen im Text. Dahinter steht folgende Überlegung: Wenn Konzepte oder Konzeptmuster häufig gemeinsam 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. Weitere Informationen finden Sie im Thema „Kookkurrenzregeln" auf Seite 127.

Minimale Anzahl an Dokumenten. 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 Textminingknoten arbeiten, um die Möglichkeiten der Textanalyse in Ihre Streams aufzunehmen. IBM SPSS Modeler Text Analytics enthält mehrere Textminingknoten, mit denen genau dies möglich ist. Diese Knoten sind auf der IBM SPSS Modeler Text Analytics-Registerkarte der Knotenpalette gespeichert.

Folgende Knoten sind enthalten:

- Der **Quellenknoten für die Dateiliste** generiert eine Liste der Dokumentnamen als Eingabe für den Textminingprozess. 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 Textminingknoten verwendet werden. Weitere Informationen finden Sie im Thema „Dateilistenknoten" auf Seite 11.
- Der **Web-Feed-Quellenknoten** ermöglicht es, Text aus Web-Feeds einzulesen, beispielsweise aus Blogs oder Newsfeeds in RSS- oder HTML-Formaten, und diese Daten im Textminingprozess 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 Textminingknoten verwendet werden. Weitere Informationen finden Sie im Thema „Web-Feed-Knoten" auf Seite 13.
- Über den **Textminingknoten** werden mit linguistischen Methoden Schlüsselkonzepte aus dem Text extrahiert. Der Textminingknoten ermöglicht es, mit diesen Konzepten und anderen Daten Kategorien zu erstellen. Daneben können Beziehungen und Zuordnungen 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. Weitere Informationen finden Sie im Thema „Textmining-Modellierungsknoten" auf Seite 20.
- Der **Textlinkanalyseknoten** extrahiert Konzepte und ermittelt auch Beziehungen zwischen Konzepten, die auf bekannten Mustern innerhalb des Texts beruhen. Mithilfe der Musterextraktion 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. Über eine interaktive Workbenchsitzung im Textmining-Modellierungsknoten wird auch TLA ermöglicht. Weitere Informationen finden Sie im Thema „Textlinkanalyseknoten" auf Seite 51.

- 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 Textmining 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. Weitere Informationen finden Sie im Thema „Übersetzungsknoten" auf Seite 61.
- Bei der Durchführung von Textmining aus externen Dokumenten kann mithilfe des **Textmining-Ausgabeknotens** eine HTML-Seite generiert werden, die Links zu den Dokumenten enthält, aus denen Konzepte extrahiert wurden. Weitere Informationen finden Sie im Thema „Datei-Viewer-Knoten" auf Seite 65.

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 Exploration 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.
- **Marktforschung.** Verwendung in der Marktanalyse zur Ermittlung von zentralen Themen in Umfragen mit offenen Antworten.
- **Analyse von Blogs und Web-Feeds.** Exploration und Erstellung von Modellen unter Verwendung der zentralen Konzepte aus Newsfeeds, Blogs usw.
- **CRM.** Erstellen von Modellen mithilfe aller Kundenberührungspunkte wie E-Mails, Transaktionen und Umfragen.

Kapitel 2. Einlesen von Quellentext

Daten für Textmining 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 Textminingprozess zu erstellen. Weitere Informationen finden Sie im Thema „Dateilistenknoten".
- 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 Textminingprozess zu formatieren. Weitere Informationen finden Sie im Thema „Web-Feed-Knoten" auf Seite 13.
- Um Text in einem der von IBM SPSS Modeler verwendeten Standarddatenformate einzulesen, wie beispielsweise einer Datenbank mit mindestens einem Textfeld für Kundenkommentare, kann jeder der Standardquellenknoten von IBM SPSS Modeler verwendet werden. Weitere Informationen hierzu finden Sie in der IBM 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 Textminingprozess 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 Textminingpalette 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 Stamms, 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 Textminingknoten verwendet werden.

Sie finden diesen Knoten auf der IBM SPSS Modeler Text Analytics-Registerkarte der Knotenpalette am unteren Rand des IBM SPSS Modeler-Fensters. Weitere Informationen finden Sie im Thema „IBM SPSS Modeler Text Analytics-Knoten" auf Seite 9.

Wichtig! Jegliche Verzeichnis- und Dateinamen, die Zeichen enthalten, die nicht in der lokalen Rechnercodierung 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 bei einer französischen Ländereinstellung.

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 IBM SPSS Modeler Text Analytics und IBM SPSS Modeler Text Analytics Server befinden.

- **Hinweis**: Führen Sie kein Upgrade auf Adobe Reader Version 10 oder höher durch, 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 ein Upgrade 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.
- Adobe 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 IBM SPSS Modeler Text Analytics keinen Einfluss hat.
- Wenn für die Sicherheitseinstellung einer Adobe PDF im Dialogfeld "Dokumenteigenschaften" auf der Registerkarte "Sicherheit" der Adobe PDF angegeben ist, dass das Kopieren oder Entnehmen von Inhalt nicht zulässig ist, kann das Dokument außerdem weder gefiltert noch 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 IBM SPSS Modeler Text Analytics Server (lokal oder fern) 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 einer fernen Instanz von IBM SPSS Modeler Text Analytics Server verbunden sind und ein Stream mit einem Dateilistenknoten vorhanden ist, müssen sich die Daten auf demselben Computer wie die IBM SPSS Modeler Text Analytics Server-Instanz befinden oder der Server-Computer muss Zugriff auf den Ordner haben, in dem die Quellendaten im Dateilistenknoten gespeichert sind.

Dateilistenknoten: Registerkarte "Einstellungen"

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

Hinweis: Die Textminingextraktion kann auf nicht unter Microsoft Windows laufenden Plattformen keine Microsoft Office- und Adobe PDF-Dateien verarbeiten. XML-, HTML- oder Textdateien können jedoch verarbeitet werden.

Jegliche Verzeichnis- und Dateinamen, die Zeichen enthalten, die nicht in der lokalen Rechnercodierung 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 bei einer französischen Ländereinstellung.

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 einzuschließende(r) 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:

Tabelle 1. Dateitypfilter nach Dateierweiterung

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

Hinweis: Weitere Informationen finden Sie im Thema „Dateilistenknoten” auf Seite 11.

Wenn Sie über Dateien ohne Erweiterung oder mit einem abschließenden Punkt als Erweiterung verfügen (z. B. `File01` oder `File01.`), verwenden Sie die Option **Keine Erweiterung**, um diese auszuwählen.

Wichtig! Ab Version 14 ist die Option "Liste der Verzeichnisse" 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 Textmining

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

Wenn Sie den Dateilistenknoten verwenden, dann stellen Sie sicher, dass im Textmining- 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.

Nehmen wir als Beispiel an, dass ein Dateilistenknoten mit einem Textminingknoten verbunden wurde, um Text zu liefern, der sich in externen Dokumenten befindet:

1. **Dateilistenknoten (Registerkarte "Einstellungen").** Zuerst fügten wir diesen Knoten zum Stream hinzu, um anzugeben, wo die Textdokumente gespeichert sind. Wir wählten das Verzeichnis aus mit allen Dokumenten, in denen wir eine Textdatensuche durchführen wollten.
2. **Textminingknoten (Registerkarte "Felder").** Anschließend fügten wir dem Dateilistenknoten einen Textminingknoten 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** enthält, sowie andere Einstellungen. Weitere Informationen finden Sie im Thema „Verwenden des Textminingknotens in einem Stream” auf Seite 31.

Weitere Informationen zur Verwendung des Textminingknotens finden Sie in „Textmining-Modellierungsknoten” auf Seite 20.

Web-Feed-Knoten

Mit dem Web-Feed-Knoten können Textdaten aus Web-Feeds für den Textminingprozess 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 ein Set verlinkter Artikel enthält, z. B. Nachrichtenquellen und Blogs. Da es sich bei RSS um ein Standardformat handelt, werden verlinkte Artikel automatisch identifiziert und im resultierenden Datenstrom als einzelne Datensätze behandelt. Es ist keine weitere Eingabe erforderlich, um wichtige Textdaten und Datensätze aus dem Feed identifizieren zu können, es sei denn, Sie möchten ein Filterverfahren auf den Text anwenden.
- **HTML-Format.** Auf der Registerkarte "Eingabe" können Sie mindestens eine URL zu HTML-Seiten definieren. Anschließend können Sie auf der Registerkarte "Datensätze" den Datensatzanfangstag definieren und diejenigen Tags festlegen, die den Zielinhalt begrenzen, und diese Tags den Ausgabefeldern Ihrer Wahl zuweisen (Beschreibung, Titel, Änderungsdatum usw.). Weitere Informationen finden Sie im Thema „Web-Feed-Knoten: Registerkarte "Datensätze"” auf Seite 15.

Wichtig! Wenn Sie versuchen, Informationen über das Web durch einen Proxy-Server abzurufen, müssen Sie den Proxy-Server in der Datei `net.properties` für die Client- und Serverversion 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 Web zugreifen oder eine Lizenz für SDL Software as a Service (SaaS) abrufen, da diese Verbindungen über Java aufgebaut werden. In der Clientversion befindet sich diese Datei standardmäßig im Verzeichnis *C:\Programme\IBM\SPSS\Modeler\16\jre\lib\net.properties*. In der Serverversion befindet sich diese Datei standardmäßig im Verzeichnis *C:\Programme\IBM\SPSS\Modeler\16\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 Textminingknoten ausgewählt werden.

Sie finden diesen Knoten auf der IBM SPSS Modeler Text Analytics-Registerkarte der Knotenpalette am unteren Rand des IBM SPSS Modeler-Fensters. Weitere Informationen finden Sie im Thema „IBM SPSS Modeler Text Analytics-Knoten” auf Seite 9.

Web-Feed-Knoten: Registerkarte "Eingabe"

Die Registerkarte "Eingabe" wird zur Angabe mindestens einer Internetadresse bzw. URL verwendet, um die Textdaten zu erfassen. Im Zusammenhang mit Textmining 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.

Folgende Parameter können festgelegt werden:

URLs eingeben oder einfügen. 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 ein. 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 ein Set verlinkter Artikel enthält, z. B. Nachrichtenquellen und Blogs. Da es sich bei RSS um ein Standardformat handelt, werden verlinkte Artikel automatisch identifiziert und im resultierenden Datenstrom als einzelne Datensätze behandelt. Es ist keine weitere Eingabe erforderlich, um wichtige Textdaten und Datensätze aus dem Feed identifizieren zu können, es sei denn, Sie möchten ein Filterverfahren auf den Text anwenden.
- *HTML-Format.* Auf der Registerkarte "Eingabe" können Sie mindestens eine URL zu HTML-Seiten definieren. Anschließend können Sie auf der Registerkarte "Datensätze" den Datensatzanfangstag 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. Weitere Informationen finden Sie im Thema „Web-Feed-Knoten: Registerkarte "Datensätze"".

Anzahl der letzten Eingaben, die pro URL gelesen werden. 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 der nachgelagerten Extraktion in einem Textmining-oder Textlinkanalyseknoten aus.

Vorherige Web-Feeds, wenn möglich, speichern und wiederverwenden. Mit dieser Option werden Web-Feeds durchsucht und die verarbeiteten Ergebnisse im Cache zwischengespeichert. Wenn sich die Inhalte eines Feeds dann bei nachfolgenden Streamausführungen 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 Wiederverwendung möglich. Sie können diese Web-Feed-Caches in der Sitzungstabelle von IBM SPSS Text Analytics Administration Console verwalten. Im Benutzerhandbuch zu IBM SPSS Text Analytics Administration Console 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 relevante Informationen zu den einzelnen Datensätzen. Wenn Sie wissen, dass ein Nicht-RSS-Feed (HTML) Text enthält, der auf mehrere Datensätze verteilt ist, müssen Sie hier den Datensatzanfangstag angeben, damit der Text nicht als ein Datensatz behandelt wird. Obwohl RSS-Feeds standardisiert sind und daher auf dieser Registerkarte keine Tagangaben 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.

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 den Anfangstag definieren, der den Beginn eines neuen Datensatzes kennzeichnet, einen Tag für das Änderungsdatum oder einen 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 dem Feld "Suchen" können Sie auf dieser Seite bestimmte Tags oder Informationen suchen, die sie kopieren und unten in die Tabelle einfügen können. Bei dem Feld "Suchen" muss die Groß-/Kleinschreibung nicht beachtet werden. 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 "Vorschau" HTML-Tags definieren.

Datensatzanfangstag (nicht RSS). 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 Blogeintrags). Wenn Sie für einen Nicht-RSS-Feed keinen Anfangstag definieren, wird die gesamte Seite als ein einziger Datensatz behandelt und der gesamte Inhalt wird im Feld **Beschreibung** ausgegeben. Als **Änderungsdatum** 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 Anfangstag angeben. Geben Sie lediglich den Anfangstag 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. Mögliche Ausgabefelder für Nicht-RSS-Feeds (HTML-Formate)

Name des Ausgabefelds	Erwarteter Taginhalt
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 Datensatzbegrenzer festgelegt wurde).
Autor	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 einen Tag in die Tabelle eingeben, wird dieser Tag beim Durchsuchen des Feeds als ein minimaler 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 3. 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 ein Filterverfahren auf den Inhalt von RSS-Feeds anzuwenden. Diese Registerkarte gilt nicht für HTML-Feeds. Eventuell möchten Sie eine Filterung durch-

führen, wenn der Feed viel Text in Form von Kopfzeilen, 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.

Inhaltsfilterung. Wenn Sie kein Bereinigungsverfahren verwenden möchten, wählen Sie **Keine** aus. Wählen Sie andernfalls **RSS Content Cleaner** aus.

RSS Content Cleaner-Optionen. Wenn Sie **RSS Content Cleaner** auswählen, 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 Einzelzeichenwörtern enthalten.
- **Zeilen verwerfen, die bestimmte Tags enthalten**. 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 Textmining

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

Wenn Sie den Web-Feed-Knoten verwenden, müssen Sie im Textmining- oder Textlinkanalyseknoten sicherstellen, dass angegeben ist, dass das Textfeld den **tatsächlichen Text** enthält, um anzugeben, dass diese Feeds direkt auf die Artikel oder Blogeinträge verweisen.

Wichtig! Wenn Sie versuchen, Informationen über das Web durch einen Proxy-Server abzurufen, müssen Sie den Proxy-Server in der Datei `net.properties` für die Client- und Serverversion 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 Web zugreifen oder eine Lizenz für SDL Software as a Service (SaaS) abrufen, da diese Verbindungen über Java aufgebaut werden. In der Clientversion befindet sich diese Datei standardmäßig im Verzeichnis *C:\Programme\IBM\SPSS\Modeler\16\jre\lib\net.properties*. In der Serverversion befindet sich diese Datei standardmäßig im Verzeichnis *C:\Programme\IBM\SPSS\Modeler\16\ext\bin\spss.TMWBServer\jre\lib\net.properties*.

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

Als Beispiel stellen wir eine Verbindung zwischen einem Web-Feed-Knoten und einem Textminingknoten her, um Textdaten aus einem RSS-Feed direkt in den Textminingprozess einzugeben.

1. **Web-Feed-Knoten (Registerkarte "Eingabe").** Zuerst haben wir diesen Knoten zum 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.
2. **Textminingknoten (Registerkarte "Felder").** Im nächsten Schritt haben wir einen Textminingknoten 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.

3. **Textminingknoten (Registerkarte "Modell").** Als Nächstes haben wir auf der Registerkarte "Modell" den Erstellungsmodus und die Ressourcen ausgewählt. In diesem Beispiel haben wir festgelegt, dass mithilfe der Standardressourcenvorlage direkt aus diesem Knoten ein Konzeptmodell erstellt werden soll.

Weitere Informationen zur Verwendung des Textminingknotens finden Sie in „Textmining-Modellierungsknoten" auf Seite 20.

Kapitel 3. Mining nach Konzepten und Kategorien

Der Textmining-Modellierungsknoten wird verwendet, um eines der beiden Textmining-Modellnuggets zu generieren:

- *Konzeptmodellnuggets* finden und extrahieren hervorstechende Konzepte in strukturierten oder unstrukturierten Textdaten.
- *Kategoriemodellnuggets* 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 Modellnuggets 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 "Anmeldeprobleme" in Ihre Modelle aufnehmen.

Außerdem ist der Textmining-Modellierungsknoten vollständig in IBM SPSS Modeler integriert, sodass die Bereitstellung von Textmining-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 Textmining-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 Textminingmodellergebnissen in Streams steigert erwiesenermaßen die Genauigkeit von Vorhersagedatenmodellen.

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

In IBM 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 Exploration und Modellerstellung fundiertere Entscheidungen zu treffen.

Konzepte und Konzeptmodellnuggets

Während des Extraktionsprozesses werden die Textdaten gescannt und analysiert, um interessante oder relevante Einzelwörter, z. B. `Wahl` oder `Frieden`, sowie Wortfolgen, z. B. `Präsidentschaftswahl`, `Wahl des Präsidenten` oder `Friedensverträge`, zu ermitteln. Diese Wörter Wortfolgen werden auch als *Terme* bezeichnet. Unter Verwendung der linguistischen Ressourcen werden die relevanten Terme extrahiert. Ähnliche Terme werden dabei unter einem übergeordneten Term zusammengefasst, der als **Konzept** bezeichnet wird.

So kann ein Konzept gegebenenfalls aus mehreren zugrunde liegenden Termen bestehen. Dies hängt von dem betreffenden Text sowie von den verwendeten linguistischen Ressourcen ab. Nehmen wir zum Beispiel an, wir hätten eine Umfrage zur Mitarbeiterzufriedenheit 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 z. B. die Terme `Lohn`, `Einkommen` und `Verdienst`. Diese Terme werden unter `Gehalt` gruppiert, da die Extraktionsengine 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 Terme enthalten sind, so behandelt, als würden sie das Wort `Gehalt` beinhalten.

Um herauszufinden, welche Terme unter einem Konzept zusammengefasst sind, können Sie das Konzept im Rahmen einer interaktiven Workbenchsitzung untersuchen oder prüfen, welche Synonyme im Konzeptmodell angezeigt werden. Weitere Informationen finden Sie im Thema „Zugrunde liegende Terme in Konzeptmodellen” auf Seite 35.

Ein **Konzeptmodellnugget** 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 Termen 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.

Weitere Informationen finden Sie im Thema „Textminingnugget: Konzeptmodell” auf Seite 32.

Kategorien und Kategoriemodellnuggets

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 Set von vordefinierten Kategorien aus einem Text Analysis Package laden. Die manuelle Erstellung und Optimierung von Kategorien kann ausschließlich über die interaktive Workbench erfolgen. Weitere Informationen finden Sie im Thema „Textminingknoten: Registerkarte "Modell"” auf Seite 24.

Ein **Kategoriemodellnugget** enthält ein Set von Kategorien mit den zugehörigen Deskriptoren. Das Modell kann genutzt werden, um ein Set von Dokumenten oder Datensätzen auf der Grundlage des darin enthaltenen Texts 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. Kategoriemodellnuggets können beispielsweise verwendet werden, um die wesentlichen Anschauungen in Umfragen mit offenen Antworten oder in einem Set von Blogeinträgen zu ermitteln.

Weitere Informationen finden Sie im Thema „Textminingnugget: Kategoriemodell” auf Seite 42.

Textmining-Modellierungsknoten

Über den Textminingknoten 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 Kategoriemodellnugget zu erstellen. Bei der Ausführung dieses Modellierungsknotens führt eine interne linguistische Extraktionsengine die Extraktion und Anordnung der Konzepte, Muster und/oder Kategorien mithilfe von Methoden für die Verarbeitung natürlicher Sprache durch.

Sie können den Textminingknoten ausführen und über die Option **Direkt generieren** automatisch ein Konzept- oder Kategoriemodellnugget 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 optimieren können, sondern auch eine Textlinkanalyse durchführen und Cluster untersuchen können. Weitere Informationen finden Sie im Thema „Textminingknoten: Registerkarte "Modell"" auf Seite 24.

Sie finden diesen Knoten auf der IBM SPSS Modeler Text Analytics-Registerkarte der Knotenpalette am unteren Rand des IBM SPSS Modeler-Fensters. Weitere Informationen finden Sie im Thema „IBM SPSS Modeler Text Analytics-Knoten" auf Seite 9.

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

Hinweis: Dieser Knoten ersetzt den Textextraktionsknoten für alle Benutzer sowie den alten Textminingknoten für japanische Benutzer, der in früheren Versionen von Textmining für Clementine angeboten wurde. Wenn Sie über ältere Streams verfügen, in denen diese Knoten oder Modellnuggets genutzt werden, ist es erforderlich, die betreffenden Streams mithilfe des neuen Textminingknotens erneut zu erstellen.
Hinweis: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

Textminingknoten: 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 Datasets die Verwendung eines Stichprobenknotens weiter oben im Stream in Erwägung; auf diese Weise lässt sich die Verarbeitungsdauer verkürzen. Weitere Informationen finden Sie im Thema „Stichprobenziehung weiter oben im Stream zur Zeitersparnis" auf Seite 31.

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:

- **Tatsächlicher 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 Extraktion 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 Extraktionsprozess oder Teile des Extraktionsprozesses zu überspringen. Er ermöglicht das Definieren von Trennzeichen für Terme, 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. Weitere Informationen finden Sie im Thema „Dokumenteinstellungen der Registerkarte "Felder"" auf Seite 22.

- **XML-Text.** 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-Text** des Dialogfelds "Dokumenteinstellungen" explizit die XML-Elemente angeben, die den Text enthalten, der während des Extraktionsprozesses gelesen werden soll. Weitere Informationen finden Sie im Thema „Dokumenteinstellungen der Registerkarte "Felder"".

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 Extraktionsmodus 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 Extraktionsprozess 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 Extraktionen 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 Extraktionsgröße vermeiden.

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

Eingabecodierung. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld **Pfadnamen zu Dokumenten** enthält. Es bestimmt die Standardtextcodierung. Für alle Sprachen außer Japanisch erfolgt eine Konvertierung von der angegebenen bzw. erkannten Codierung in `ISO-8859-1`. Selbst wenn Sie also eine andere Codierung angeben, wird diese vor der Verarbeitung von der Extraktionsengine in `ISO-8859-1` konvertiert. Alle Zeichen, die nicht in die `ISO-8859-1`-Codierungsdefinition passen, werden in Leerzeichen umgewandelt. Für japanischen Text können Sie eine von mehreren Codierungsoptionen auswählen: `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"

Formatierung als gegliederter Text

Wenn Sie den gesamten Extraktionsprozess oder Teile des Extraktionsprozesses 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 Terme 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 Extraktionsengine kann durch explizite Deklarationen ersetzt werden. In einer Bibliografiedatei, in der die Schlüsselwortfelder durch Trennzeichen getrennt sind, z. B. durch ein Semikolon (;) oder ein Komma (,), genügt es, die Zeichenfolge zwischen zwei Trennzeichen zu extrahieren. Daher können Sie den gesamten Extraktionsprozess überspringen und stattdessen spezielle Regeln zum Umgang definieren, um Trennzeichen für Terme zu deklarieren, dem extrahierten Text Typen zuzuweisen oder einen minimalen Häufigkeitswert für die Extraktion festzulegen.

Verwenden Sie beim Deklarieren von gegliederten Textelementen die folgenden Regeln:

- Pro Zeile kann nur ein Feld, Tag oder Element deklariert werden. Sie müssen nicht in den Daten vorhanden sein.
- Bei Deklarationen muss die Groß-/Kleinschreibung beachtet werden.
- 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 den Tag ohne das Attribut oder die spitze schließende Klammer (>) wie folgt hinzu: `<title`
- Fügen Sie nach dem Feld bzw. Tagnamen 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 der Tag mehrere Terme enthält und zum Definieren der einzelnen Terme 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 Typnamen 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 einen 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 Terme 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 Backslash-Zeichen 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 bibliografische Felder vorliegen:

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

Falls sich der Extraktionsprozess in diesem Beispiel auf den Autor sowie die Kurzdarstellung 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 stattdessen 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 während der Extraktion gescannt; hierbei werden allerdings die standardmäßige linguistische Verarbeitung und die Typzuweisung herangezogen.

Formatierung als XML-Text

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

Wichtig! Wenn Sie den Extraktionsvorgang überspringen und Regeln für Trennzeichen für Terme festlegen, dem extrahierten Text Typen zuweisen oder einen Häufigkeitswert für extrahierte Terme festlegen möchten, verwenden Sie die im Folgenden beschriebene Option **Gegliederter Text**.

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

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

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 verschachtelten Tags (`Verkehrssignale` und `Straßenschilder sind hilfreich`) während des Extraktionsvorgangs gescannt. Der Text `Das Erlernen der Vorschriften ist hilfreich` wird jedoch ignoriert, da der Tag `<p>` weder ausdrücklich deklariert wurde noch in einem deklarierten Tag verschachtelt ist.

Textminingknoten: Registerkarte "Modell"

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

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.

Erstellungsmodus. Legt fest, wie die Modellnuggets erstellt werden, wenn ein Stream mit diesem Textminingknoten 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 optimieren können, sondern auch eine Textlinkanalyse 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, Kategorien erstellen und optimieren, die linguistischen Ressourcen (Vorlagen, Synonyme, Typen, Bibliotheken usw.) optimieren und Kategoriemodellnuggets erstellen können. Weitere Informationen finden Sie im Thema „Interaktiv erstellen" auf Seite 25.
- **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 Bearbei-

tung erforderlich. Wenn Sie diese Option auswählen, werden modellspezifische Optionen angezeigt, über die Sie festlegen können, welche Art von Modell Sie erstellen möchten. Weitere Informationen finden Sie im Thema „Direkt generieren" auf Seite 27.

Ressourcen kopieren von. Beim Textmining basiert die Extraktion 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 des Texts während der Extraktion, um die Konzepte, Typen und manchmal Muster zu erhalten. Sie können Ressourcen aus einer Ressourcenvorlage oder einem Text Analysis Package in diesen Knoten kopieren. Wählen Sie eine aus 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 ein aktualisiertes TAP verwenden möchten, müssen Sie sie bzw. es hier oder in einer interaktiven Workbenchsitzung neu laden. Um Ihnen die Arbeit zu erleichtern, werden die Uhrzeit und das Datum im Knoten angezeigt, zu der bzw. an dem die Ressourcen kopiert und geladen wurden. Weitere Informationen finden Sie im Thema „Kopieren von Ressourcen aus TAPs und Vorlagen" auf Seite 27.

Textsprache. Bestimmt die Sprache des für das Mining verwendeten Texts. Die in den Knoten kopierten Ressourcen steuern die angezeigten Sprachoptionen. Sie können die Sprache auswählen, auf die die Ressourcen abgestimmt wurden, oder die Option **ALLE** auswählen. Es wird empfohlen, für die Textdaten die exakte Sprache anzugeben. Wenn Sie sich jedoch nicht sicher sind, können Sie die Option **ALLE** auswählen. Die Option **ALLE** ist für japanischen Text nicht verfügbar. Diese Option **ALLE** verlängert die Ausführungsdauer, da mithilfe der automatischen Spracherkennung zunächst alle Dokumente und Datensätze gescannt werden, um die Textsprache zu ermitteln. Wenn Sie diese Option auswählen, werden alle Datensätze oder Dokumente, die in einer unterstützten und lizenzierten Sprache vorliegen, durch die Extraktionsengine mit den internen Wörterbüchern in der jeweiligen Sprache gelesen. Weitere Informationen finden Sie im Thema „Language Identifier" auf Seite 225. 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 Textmining-Modellierungsknotens können Sie einen Erstellungsmodus für Ihre Modellnuggets auswählen. Wenn Sie **Interaktiv erstellen** auswählen, wird eine interaktive Schnittstelle geöffnet, wenn Sie den Stream ausführen. In dieser interaktiven Workbench können Sie Folgendes tun:

- Die Extraktionsergebnisse einschließlich Konzepten und Typen extrahieren und untersuchen, um hervorstechende Ideen in Ihren Textdaten zu erkennen.
- Verwenden 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.
- Optimieren Sie Ihre linguistischen Ressourcen (Ressourcenvorlagen, Bibliotheken, Wörterbücher, Synonyme usw.), damit Sie Ihre Ergebnisse mithilfe eines iterativen Prozesses, in dem Konzepte extrahiert, untersucht und optimiert werden, verbessern können.
- Führen Sie eine Textlinkanalyse (TLA) durch und verwenden Sie die erkannten TLA-Muster, um bessere Kategoriemodellnuggets zu erstellen. Der Textlinkanalyseknoten bietet nicht dieselben Untersuchungs- und Modellierungsfunktionen.
- Generieren Sie Cluster zur Ermittlung neuer Beziehungen und untersuchen Sie Beziehungen zwischen Konzepten, Typen, Mustern und Kategorien im Visualisierungsbereich.
- Generieren Sie optimierte Kategoriemodellnuggets für die Modellpalette in IBM SPSS Modeler und verwenden Sie sie in anderen Streams.

Arbeit der Sitzung (Kategorien, TLA, Ressourcen etc.) aus letzter Knotenaktualisierung verwenden. Bei einer interaktiven Workbenchsitzung können Sie den Knoten mit Sitzungsdaten (Extraktionsparameter, 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 inaktiviert, da keine Sitzungsdaten gespeichert werden konnten. Informationen dazu, wie der Knoten mit Sitzungsdaten aktualisiert und so die Verwendung dieser Option ermöglicht wird, finden Sie in „Aktualisieren von Modellierungsknoten und Speichern" auf Seite 90.

Wenn Sie eine Sitzung *mit* dieser Option starten, stehen die Extraktionseinstellungen, Kategorien, Ressourcen und sämtliche anderen Arbeiten von der letzten Knotenaktualisierung im Rahmen einer interaktiven Workbenchsitzung 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 inaktiviert 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.

Hinweis: Wenn Sie den Quellenknoten für Ihren Stream ändern, nachdem Extraktionsergebnisse mit der Option **Arbeit der Sitzung verwenden...** im Cache gespeichert wurden, müssen Sie eine neue Extraktion ausführen, sobald die interaktive Workbenchsitzung gestartet wird, damit Sie aktualisierte Extraktionsergebnisse erhalten.

Extraktion überspringen und zwischengespeicherte Daten und Ergebnisse wiederverwenden. In der interaktiven Workbenchsitzung können sämtliche Extraktionsergebnisse und Daten wiederverwendet werden. Diese Option ist besonders nützlich, wenn Sie Zeit sparen und Extraktionsergebnisse wiederverwenden möchten, anstatt nach dem Start der Sitzung zu warten, bis eine ganz neue Extraktion durchgeführt wurde. Um diese Option verwenden zu können, muss dieser Knoten im Rahmen einer interaktiven Workbenchsitzung aktualisiert und die Option **Arbeit der Sitzung behalten und Textdaten mit Extraktionsergebnissen für Wiederverwendung zwischenspeichern** ausgewählt worden sein. Informationen dazu, wie der Knoten mit Sitzungsdaten aktualisiert und so die Verwendung dieser Option ermöglicht wird, finden Sie in „Aktualisieren von Modellierungsknoten und Speichern" auf Seite 90.

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

- **Extraktionsergebnisse werden für Erstellung von Kategorien verwendet.** Mit dieser Option wird die interaktive Workbench in der Ansicht "Kategorien und Konzepte" gestartet und gegebenenfalls eine Extraktion durchgeführt. In dieser Ansicht können Sie Kategorien erstellen und ein Kategoriemodell generieren. Außerdem können Sie eine andere Ansicht wählen. Weitere Informationen finden Sie im Thema Kapitel 8, „Modus "Interaktive Workbench"", auf Seite 79.
- **Ergebnisse der Textlinkanalyse (TLA) werden untersucht.** Mit dieser Option wird die Ansicht "Textlinkanalyse" aufgerufen und zunächst die Extraktion durchgeführt. Dann werden Beziehungen zwischen Konzepten im Text identifiziert, beispielsweise Meinungen und andere Links. Es ist erforderlich, dass eine Vorlage oder ein Text Analysis Package ausgewählt wird, die bzw. das TLA-Musterregeln enthält, damit diese Option angezeigt wird und Ergebnisse ausgegeben werden. Wenn Sie mit größeren Datasets arbeiten, kann die TLA-Extraktion einige Zeit dauern. In diesem Fall könnten Sie erwägen, einen vorgeordneten Stichprobenknoten zu verwenden. Weitere Informationen finden Sie im Thema Kapitel 12, „Untersuchen von Textlinkanalyse", auf Seite 163.
- **Kowort-Cluster werden analysiert.** Mit dieser Option wird die Clusteransicht aufgerufen und veraltete Extraktionsergebnisse werden aktualisiert. In dieser Ansicht können Sie eine Kowort-Cluster-Analyse durchführen, die zu einer Reihe von Clustern führt. Unter Kowort-Clustering 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. Weitere Informationen finden Sie im Thema Kapitel 8, „Modus "Interaktive Workbench"", auf Seite 79.

Direkt generieren

Auf der Registerkarte "Modell" des Textmining-Modellierungsknotens können Sie einen Erstellungsmodus für Ihre Modellnuggets auswählen. Wenn Sie die Option **Direkt generieren** auswählen, können Sie die Optionen im Knoten festlegen und anschließend einfach den Stream ausführen. Bei der Ausgabe handelt es sich um ein Konzeptmodellnugget, 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.

Maximale Anzahl an in Modell einzuschließende 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 auf der Basis der höchsten Häufigkeit auswählen. Höchste Anzahl an Konzepten.** Gibt - ausgehend von dem Konzept mit der höchsten Häufigkeit - die Anzahl der zu markierenden Konzepte an. Häufigkeit bezieht sich in diesem Fall darauf, wie häufig die betreffenden Konzepte (und alle zugrunde liegenden Terme) in sämtlichen Dokumenten/Datensätzen insgesamt auftauchen. Dieser Wert kann höher sein, als der Datensatzwert, da ein Konzept mehrmals in einem Datensatz vorkommen kann.
- **Konzepte abwählen, 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.

Für Scoring-Geschwindigkeit optimieren. Diese Option ist standardmäßig aktiviert und stellt sicher, dass das erstellte Modell kompakt ist und das Scoring mit hoher Geschwindigkeit durchführt. Wenn diese Option inaktiviert wird, wird ein sehr viel größeres Modell erstellt, für das das Scoring langsamer durchführt wird. Durch das größere Modell wird jedoch sichergestellt, dass Scores, die ursprünglich im generierten Konzeptmodell angezeigt werden, mit den Scores identisch sind, die man erhält, wenn derselbe Text mit dem Modellnugget gescort wird.

Kopieren von Ressourcen aus TAPs und Vorlagen

Beim Textmining basiert die Extraktion 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 des Texts während der Extraktion, 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 Textminingknoten befinden, können Sie auch ein *Text Analysis Package* (TAP) auswählen.

Standardmäßig werden Ressourcen von der Basisvorlage für die lizenzierte Sprache für Ihr Produkt in den Knoten kopiert, wenn Sie den Knoten dem Erstellungsbereich 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 Textminingknoten aktualisieren und die Option **Arbeit der Sitzung verwenden** auswählen. Weitere Informationen zur Option **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.

Hinweis: Sie können keine TAPs in den Textlinkanalyseknoten laden.

Ressourcenvorlagen

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 Textmining-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 Text Analysis Package laden, indem Sie entweder **Ressourcenvorlage** oder **Text Analysis Package** auswählen und anschließend auf **Laden** klicken. Für Vorlagen können Sie die Vorlage dann im Dialogfeld "Ressourcenvorlage laden" auswählen.

Hinweis: Wenn die gewünschte Vorlage in der Liste nicht angezeigt wird, Sie jedoch über eine exportierte Kopie auf Ihrem Computer verfügen, können Sie sie jetzt importieren. Über dieses Dialogfeld können Sie auch Exporte für die gemeinsame Nutzung mit anderen Benutzern durchführen. Weitere Informationen finden Sie im Thema „Import und Export von Vorlagen” auf Seite 187.

Text Analysis Packages (TAPs)

Bei einem Text Analysis Package (TAP) handelt es sich um ein vordefiniertes Set von Bibliotheken und erweiterten linguistischen und nicht linguistischen Ressourcen sowie um mindestens ein Set vordefinierter Kategorien. IBM SPSS Modeler Text Analytics bietet mehrere vordefinierte TAPs für Text in englischer Sprache und auch für Text in japanischer Sprache. Jedes TAP ist dabei auf eine bestimmte Domäne feinabgestimmt. 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. Weitere Informationen finden Sie im Thema „Laden von Text Analysis Packages” auf Seite 150. *Hinweis*: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

Hinweis: 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 Textmining-Modellierungsknoten mit diesen aktuellen Änderungen aktualisieren. In diesem Fall würden Sie die Option **Arbeit der Sitzung verwenden** auf der Registerkarte "Modell" des Textmining-Modellierungsknotens auswählen.

Wenn Sie **Arbeit der Sitzung verwenden** auswählen, ist die Schaltfläche **Laden** im Knoten inaktiviert, um darauf hinzuweisen, dass diese Ressourcen aus der interaktiven Workbenchsitzung anstelle 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 Workbenchsitzung in der Ansicht "Ressourceneditor" bearbeiten oder zwischen ihnen wechseln. Weitere Informationen finden Sie im Thema „Aktualisieren von Knotenressourcen nach dem Laden” auf Seite 185.

Textminingknoten: 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 Extraktionsprozesses fest. Sie stellen jedoch nur einen Teil der Ihnen zur

Verfügung stehenden Optionen dar. Zudem werden die Extraktionsergebnisse auch von einer Reihe linguistischer Ressourcen und Optionen beeinflusst. Diese werden über die Ressourcenvorlage gesteuert, die Sie auf der Registerkarte "Modell" auswählen. Weitere Informationen finden Sie im Thema „Textminingknoten: Registerkarte "Modell"" auf Seite 24.

Hinweis: Wenn Sie auf der Registerkarte "Modell" den Modus **Interaktiv erstellen** mit gespeicherten interaktiven Workbenchinformationen ausgewählt haben, ist die gesamte Registerkarte inaktiviert. In diesem Fall werden die Extraktionseinstellungen von der letzten gespeicherten Workbenchsitzung übernommen.

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

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

Hinweis: 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.

Extraktion beschränken auf Konzepte mit globaler Häufigkeit von mindestens: [n]. Gibt an, wie oft ein Wort oder eine Wortfolge mindestens im Text vorkommen muss, damit es bzw. sie extrahiert wird. So begrenzt der Wert 5 die Extraktion auf diejenigen Wörter oder Wortfolgen, 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 Extraktionsergebnissen (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 Extraktionsergebnissen finden. Würden Sie jedoch die Extraktion auf eine globale Häufigkeit von 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 Texts 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 Terme enthält (was normalerweise der Fall ist) und ein Term für die Vorlage im Text gefunden wird, wird der Term unabhängig von seiner Häufigkeit indiziert.

Nehmen wir beispielsweise an, dass Sie die Vorlage "Grundlegende Ressourcen" verwenden, die unter dem Typ `<Location>` in der Kernbibliothek "los angeles" enthält, wenn Ihr Dokument Los Angeles nur einmal enthält, gehört Los Angeles zur Liste der Konzepte. Um dies zu verhindern, müssen Sie einen Filter festlegen, um nur Konzepte anzuzeigen, die mindestens so oft vorkommen wie im Feld **Extraktion beschränken auf Konzepte mit globaler Häufigkeit von mindestens: [n]** angegeben.

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Interpunktionsfehlern (zum Beispiel ungeeignete Verwendung) während der Extraktion, 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 Fuzzy-Gruppierungsverfahren an, das hilft, häufig falsch geschriebene Wörter oder ähnlich geschriebene Wörter unter einem Konzept zu gruppieren. Der Algorithmus für Fuzzy-Gruppierung entfernt alle Vokale (außer dem ersten) und doppelte/dreifache Konsonanten temporär aus extrahierten Wörtern und vergleicht sie, um festzustellen, ob sie gleich sind, sodass `Modellierung` und `Modelierung` zusammen

gruppiert werden würden. Wenn jedoch jeder Term einem anderen Typ (ausschließlich des Typs <Unknown>) zugewiesen ist, wird das Fuzzy-Gruppierungsverfahren nicht angewendet.

Sie können auch die minimal erforderliche Zahl von *Stamm*zeichen definieren, bevor Fuzzy-Gruppierung eingesetzt wird. Die Anzahl der Stammzeichen in einem Term berechnet sich aus der Summe aller Zeichen abzüglich aller Zeichen, die Beugungsendungen und - bei zusammengesetzten Termen - Determinatoren und Präpositionen bilden. So würde beispielsweise der Term 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 sich herausstellt, dass bestimmte Wörter später falsch eingruppiert werden, können Sie einzelne Wortpaare aus dem Verfahren ausschließen, indem Sie sie auf der Registerkarte "Erweiterte Ressourcen" im Bereich **Fuzzy-Gruppierung: Ausnahmen** explizit deklarieren. Weitere Informationen finden Sie im Thema „Fuzzy-Gruppierung” auf Seite 218.

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

Nicht linguistische Entitäten extrahieren. Diese Option extrahiert nicht linguistische Entitäten wie beispielsweise Telefonnummern, Personalausweisnummern, Uhrzeiten, Datumsangaben, Währungen, Ziffern, Prozentsätze, E-Mail-Adressen und HTTP-Adressen. Sie können bestimmte Typen von nicht linguistischen Entitäten im Abschnitt **Nicht linguistische Entitäten: Konfiguration** der Registerkarte "Erweiterte Ressourcen" ein- bzw. ausschließen. Durch Inaktivierung unnötiger Entitäten vergeudet die Extraktionsengine keine Verarbeitungszeit. Weitere Informationen finden Sie im Thema „Konfiguration” auf Seite 222.

Großbuchstaben-Algorithmus. Diese Option extrahiert einfache und zusammengesetzte Terme, die sich nicht in den integrierten Wörterbüchern befinden, solange der erste Buchstabe des Terms 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 Texts 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 Terme abzugleichen, die dem Typ <Person> zugeordnet sind. Wird beispielsweise *doe* gefunden und anfänglich dem Typ <Unknown> zugeordnet, überprüft die Extraktionsengine, ob ein zusammengesetzter Term 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 Uniterme extrahiert werden.

Maximale Füllwörter in zusammengesetzten Konzepten. Diese Option gibt die maximale Anzahl von Füllwörtern an, die für die Anwendung des Permutationsverfahrens vorhanden sein müssen. Dieses Permutationsverfahren gruppiert ähnliche Wortfolgen, 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 als auch Vertreter des Unternehmens extrahiert wurden. In diesem Fall würden beide extrahierte Terme in der endgültigen Konzeptliste zusammen gruppiert, da beide Terme als gleich betrachtet werden, wenn des ignoriert wird.

Hinweis: Um die Extraktion von Ergebnissen der Textlinkanalyse zu ermöglichen, müssen Sie die Sitzung mit der Option **Ergebnisse der Textlinkanalyse (TLA) werden untersucht** beginnen und außerdem Ressourcen mit TLA-Definitionen auswählen. Sie können TLA-Ergebnisse auch später während einer interaktiven Workbenchsitzung über das Dialogfeld "Extraktionseinstellungen" extrahieren. Weitere Informationen finden Sie im Thema „Extrahieren von Daten” auf Seite 94.

Für japanischen Text

Dieses Dialogfeld enthält abweichende Optionen für japanischen Text, da der Extraktionsprozess einige Unterschiede aufweist. Für die Arbeit mit japanischem Text müssen Sie auch eine Vorlage oder ein Text Analysis Package für die japanische Sprache auf der Registerkarte "Modell" dieses Knotens auswählen. Weitere Informationen finden Sie im Thema „Kopieren von Ressourcen aus TAPs und Vorlagen” auf Seite 27.

Hinweis: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

Sekundäre Analyse. Bei einer Extraktion werden grundlegende Schlüsselwörter unter Verwendung des Standardsets von Typen extrahiert. Wenn Sie jedoch einen Sekundäranalysator auswählen, können Sie auch mehrere oder reichhaltigere Konzepte erhalten, da der Extraktor nun Partikel und Hilfsverben als Teil des Konzepts beinhaltet. Im Fall der Stimmungsanalyse wird auch eine große Anzahl an zusätzlichen Typen eingeschlossen. Des Weiteren ermöglicht Ihnen die Wahl eines Sekundäranalysators, auch Ergebnisse für die Textlinkanalyse zu generieren.

Hinweis: Wenn ein Sekundäranalysator aufgerufen wird, dauert der Extraktionsprozess länger.

- **Abhängigkeitsanalyse**. Die Wahl dieser Option führt zu erweiterten Partikeln für die Extraktionskonzepte aus der grundlegenden Typ- und Stichwortextraktion. Sie können auch vielfältigere Musterergebnisse aus einer Abhängigkeitstextlinkanalyse (TLA) gewinnen.
- **Stimmungsanalyse**. Bei dieser Analyse werden zusätzliche Konzepte und - wann immer möglich - TLA-Musterergebnisse extrahiert. Neben den grundlegenden Typen können Sie zusätzlich mehr als 80 Stimmungstypen nutzen. 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**.
- **Kein Sekundäranalysator**. Diese Option schaltet sämtliche Sekundäranalysatoren aus. Diese Option ist ausgeblendet, wenn die Option **Ergebnisse der Textlinkanalyse (TLA) werden untersucht** auf der Registerkarte "Modell" ausgewählt wurde, da ein Sekundäranalysator erforderlich ist, um TLA-Ergebnisse zu erhalten. Wenn Sie diese Option auswählen, später aber die Option **Ergebnisse der Textlinkanalyse (TLA) werden untersucht** aktivieren, tritt während der Streamausführung 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 Workbenchsitzung. Je größer der Umfang der Daten, desto mehr Zeit nehmen Extraktion und Kategorisierung in Anspruch. Für effizienteres Arbeiten können Sie einen der Stichprobenknoten von IBM SPSS Modeler weiter oben als Ihren Textminingknoten im Stream hinzufügen. Ziehen Sie mithilfe dieses Stichprobenknotens eine Zufallsstichprobe mit einem kleineren Subset 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 Dataset ausgeführt haben und die Ergebnisse Ihren Vorstellungen entsprechen, können Sie dasselbe Verfahren anwenden, um Kategorien für das gesamte Dataset zu erstellen. Im Anschluss können Sie nach Dokumenten oder Datensätzen suchen, die nicht in die von Ihnen definierten Kategorien fallen, und nach Bedarf Anpassungen vornehmen.

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

Verwenden des Textminingknotens in einem Stream

Der Textmining-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 und den Knoten für feste Dateien. Für Text in externen Dokumenten kann ein Dateilistenknoten verwendet werden.

Beispiel 1: Dateilistenknoten und Textminingknoten zum direkten Erstellen eines Konzeptmodellnuggets

Das folgende Beispiel zeigt die Verwendung des Dateilistenknotens mit dem Textmining-Modellierungsknoten zum Generieren des Konzeptmodellnuggets. Weitere Informationen zur Verwendung des Dateilistenknotens finden Sie in „Dateilistenknoten" auf Seite 11.

1. **Dateilistenknoten (Registerkarte "Einstellungen").** Zuerst fügten wir diesen Knoten zum Stream hinzu, um anzugeben, wo die Textdokumente gespeichert sind. Wir wählten das Verzeichnis aus mit allen Dokumenten, in denen wir eine Textdatensuche durchführen wollten.
2. **Textminingknoten (Registerkarte "Felder").** Anschließend fügten wir dem Dateilistenknoten einen Textminingknoten 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** enthält, sowie andere Einstellungen. Weitere Informationen finden Sie im Thema „Verwenden des Textminingknotens in einem Stream" auf Seite 31.
3. **Textminingknoten (Registerkarte "Modell")**. Als Nächstes haben wir auf der Registerkarte "Modell" den Erstellungsmodus zum Generieren eines Konzeptmodells direkt über diesen Knoten gewählt. Sie können eine andere Ressourcenvorlage auswählen oder die grundlegenden Ressourcen beibehalten.

Beispiel 2: Excel-Datei- und Textminingknoten zur interaktiven Erstellung eines Kategoriemodells

Dieses Beispiel zeigt, wie eine interaktive Workbenchsitzung auch über den Textminingknoten gestartet werden kann. Weitere Informationen zur interaktiven Workbench finden Sie in Kapitel 8, „Modus "Interaktive Workbench"", auf Seite 79.

1. **Excel-Quellenknoten (Registerkarte "Daten").** Zuerst fügten wir diesen Knoten zum Stream hinzu, um anzugeben, wo der Text gespeichert ist.
2. **Textminingknoten (Registerkarte "Felder").** Als Nächstes haben wir einen Textminingknoten hinzugefügt und angeschlossen. Auf dieser ersten Registerkarte haben wir unser Eingabeformat festgelegt. Wir haben einen Feldnamen des Quellenknotens und die Option **Textfeld enthält: Tatsächlicher Text** ausgewählt.
3. **Textminingknoten (Registerkarte "Modell").** Als Nächstes haben wir auf der Registerkarte "Modell" die interaktive Erstellung eines Kategoriemodellnuggets und die Verwendung von Extraktionsergebnissen zur automatischen Erstellung von Kategorien ausgewählt. In diesem Beispiel haben wir eine Kopie von Ressourcen und ein Set von Kategorien aus einem Text Analysis Package geladen.
4. **Interaktive Workbenchsitzung**. Als Nächstes haben wir den Stream ausgeführt. Die Benutzerschnittstelle der interaktiven Workbench wurde angezeigt. Nach einer Extraktion begannen wir, unsere Daten zu untersuchen und die Kategorien zu verbessern.

Textminingnugget: Konzeptmodell

Ein Textmining-Konzeptmodellnugget wird jedes Mal erstellt, wenn Sie einen Textmining-Modellknoten erfolgreich ausführen und dafür auf der Registerkarte "Modell" die Option **Modell direkt generieren** ausgewählt haben. Ein Textmining-Konzeptmodellnugget wird für die Echtzeitermittlung von Schlüsselkonzepten in anderen Textdaten, wie beispielsweise Notizen aus einem Call-Center, verwendet.

Das Konzeptmodellnugget 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 Textmining-Modellnugget enthält, werden den Daten gemäß dem Erstellungsmodus, der auf der Registerkarte "Modell" des Textmining-Modellierungsknotens ausgewählt wurde, vor der Erstellung des Modells neue Felder hinzugefügt. Weitere Informationen finden Sie im Thema „Konzeptmodell: Registerkarte "Modell"" auf Seite 33.

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

Die Textmining-Modellnuggets werden nach der Generierung in der Palette der Modellnuggets gespeichert (diese befindet sich rechts oben im IBM SPSS Modeler-Fenster auf der Registerkarte "Modelle").

Anzeigen von Ergebnissen

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

Hinzufügen von Modellen zu Streams

Um das Modellnugget Ihrem Stream hinzuzufügen, klicken Sie auf das Symbol in der Palette der Modellnuggets und dann auf den Streamerstellungsbereich, 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 Vorhersagen zu erstellen.

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.

Hinweis: Wenn Sie stattdessen ein Kategoriemodellnugget generiert haben, enthält diese Registerkarte andere Informationen. Weitere Informationen finden Sie im Thema „Kategoriemodellnugget: Registerkarte "Modell"" auf Seite 43.

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. die extrahierte übergeordnete Wortfolge. In einigen Fällen werden als Konzept der Konzeptname sowie einige weitere zugrunde liegende Terme, die mit diesem Konzept verbunden sind, angegeben. Um zu sehen, welche zugrunde liegenden Terme zu einem Konzept gehören, zeigen Sie den Bereich "Zugrunde liegende Terme" auf dieser Registerkarte an und wählen Sie das betreffende Konzept aus, um die entsprechenden Terme unten im Dialogfeld anzuzeigen. Weitere Informationen finden Sie im Thema „Zugrunde liegende Terme in Konzeptmodellen" auf Seite 35.

Globalwert. Globalwert (Häufigkeit) bezieht sich in diesem Fall darauf, wie häufig die betreffenden Konzepte (und alle zugrunde liegenden Terme) 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 Vorkommen 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 zugrunde liegenden Terme).

- **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. Weitere Informationen finden Sie im Thema „Typwörterbücher” auf Seite 201.

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 markieren.** Aktiviert alle Kontrollkästchen für die ausgewählten Zeilen in der Tabelle und berücksichtigt dabei die Konzepte für das Scoring.
- **Markierung für ausgewählte Elemente aufheben.** Inaktiviert 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.** Inaktiviert 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". Weitere Informationen finden Sie im Thema „Optionen zum Einschließen von Konzepten für das Scoring”.

Optionen zum Einschließen von Konzepten für das Scoring

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

Abbildung 1. Symbolleistenschaltfläche "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.

Folgende Optionen stehen zur Auswahl:

Konzepte auf der Basis der höchsten Häufigkeit auswählen. Höchste 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 zugrunde

liegenden Terme) in sämtlichen Dokumenten/Datensätzen insgesamt auftauchen. Dieser Wert kann höher sein, als der Datensatzwert, da ein Konzept mehrmals in einem Datensatz vorkommen kann.

Konzepte auf der Basis der Dokumentenanzahl auswählen. 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 (einschließlich aller zugrunde liegenden Terme).

Dem Typ zugewiesene Konzepte auswählen. 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 Extraktionsprozesses 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. Weitere Informationen finden Sie im Thema „Typwörterbücher" auf Seite 201.

Konzepte abwählen, 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.

Konzepte abwählen, die folgendem Typ zugewiesen sind. Hebt die Markierung von Konzepten auf, die mit dem in der Dropdown-Liste ausgewählten Typ übereinstimmen.

Zugrunde liegende Terme in Konzeptmodellen

Sie können die zugrunde liegenden Terme anzeigen, die für die in der Tabelle ausgewählten Konzepte definiert sind. Sie können die Tabelle für zugrunde liegende Terme in einem geteilten Fensterbereich unten im Bildschirm anzeigen, indem Sie auf die Umschaltfläche für zugrunde liegende Terme in der Symbolleiste klicken.

Diese zugrunde liegenden Terme umfassen die Synonyme, die in den linguistischen Ressourcen definiert sind (unabhängig davon, ob sie im Text gefunden wurden) sowie extrahierte Plural-/Singularformen, die im Text zur Generierung des Modellnuggets gefunden wurden, permutierte Terme, Terme aus Fuzzy-Gruppierungen usw.

Abbildung 2. Symbolleistenschaltfläche "Zugrunde liegende Terme anzeigen"

Hinweis: Sie können die Liste der zugrunde liegenden Terme nicht bearbeiten. Diese Liste wird durch Substitutionen, Synonymdefinitionen (im Substitutionswörterbuch), Fuzzy-Gruppierung u. a. generiert - Verfahren, die alle in den linguistischen Ressourcen definiert sind. Nehmen Sie Änderungen an der Gruppierung von Termen unter einem Konzept oder deren Behandlung direkt in den Ressourcen vor (Bearbeitung ist im Ressourceneditor in der interaktiven Workbench bzw. im Vorlageneditor möglich, anschließend erneutes Laden in den Knoten). Führen Sie dann den Stream erneut aus, um ein neues Modellnugget mit den aktualisierten Ergebnissen zu erhalten.

Wenn Sie mit der rechten Maustaste auf die Zelle eines zugrunde liegenden Terms 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).

Hinweis: Diese Registerkarte wird nur angezeigt, wenn das Modellnugget im Erstellungsbereich platziert wird. Sie ist nicht vorhanden, wenn Sie direkt in der Modellpalette auf dieses Dialogfeld zugreifen.

Scoring-Modus: Konzepte als Datensätze

Mit diesem Scoring-Modus wird für jedes `Konzept/Dokument`-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 4. Ausgabefelder für "Konzepte als Datensätze"

Feld	Beschreibung
`Konzept`	Enthält den im Textdatenfeld gefundenen Namen des extrahierten Konzepts.
`Typ`	Speichert den Konzepttyp als vollständigen Typnamen, beispielsweise *Location* oder *Person*. Bei einem **Typ** handelt es sich um Konzepte, die nach semantischen Gesichtspunkten gruppiert werden. Weitere Informationen finden Sie im Thema „Typwörterbücher" auf Seite 201.
`Anzahl`	Zeigt an, wie häufig dieses Konzept (einschließlich der zugrunde liegenden Terme) im Texthauptteil (Datensatz/Dokument) vorkommt.

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

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 neues Feld (Spalte) für jedes auf der Registerkarte "Modell" ausgewählte Konzept (Auswahl durch Setzen eines Häkchens). Der Wert für jedes Konzeptfeld hängt davon ab, ob Sie auf dieser Registerkarte **Flags** oder **Anzahl** als Feldwert auswählen.

Feldwerte. Wählen Sie aus, ob das neue Feld für jedes Konzept eine Anzahl oder einen Flagwert enthalten soll.

- **Flags.** Diese Option wird verwendet, um Flags mit zwei verschiedenen Werten in der Ausgabe zu erhalten, z. B. *Ja/Nein*, *Wahr/Falsch*, *W/F* oder *1* und *2*. Die Speichertypen werden automatisch so festgelegt, dass sie die ausgewählten Werte widerspiegeln. Wenn Sie beispielsweise numerische Werte für die Flags eingeben, werden sie automatisch als Ganzzahl behandelt. Als Speichertypen für Flags sind "Zeichenfolge", "Ganzzahl", "Reelle Zahl" und "Datum/Uhrzeit" möglich. Geben Sie einen Flagwert für **Wahr** und für **Falsch** ein.
- **Häufigkeiten.** 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 Zeichenfolge einzufügen. Wählen Sie **Suffix**, um die Erweiterung am Ende der Zeichenfolge einzufügen.

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Interpunktionsfehlern (zum Beispiel ungeeignete Verwendung) während der Extraktion, 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.

Hinweis: 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).

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

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:

- **Tatsächlicher 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 Extraktion 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 Extraktionsprozess oder Teile des Extraktionsprozesses zu überspringen. Er ermöglicht das Definieren von Trennzeichen für Terme, 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. Weitere Informationen finden Sie im Thema „Dokumenteinstellungen der Registerkarte "Felder"" auf Seite 22.
- **XML-Text.** 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-Text** des Dialogfelds "Dokumenteinstellungen" explizit die XML-Elemente angeben, die den Text enthalten, der während des Extraktionsprozesses gelesen werden soll. Weitere Informationen finden Sie im Thema „Dokumenteinstellungen der Registerkarte "Felder"" auf Seite 22.

Eingabecodierung. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld **Pfadnamen zu Dokumenten** enthält. Es bestimmt die Standardtextcodierung. Für alle Sprachen außer Japanisch erfolgt eine Konvertierung von der angegebenen bzw. erkannten Codierung in `ISO-8859-1`. Selbst wenn Sie also eine andere Codierung angeben, wird diese vor der Verarbeitung von der Extraktionsengine in `ISO-8859-1` konvertiert. Alle Zeichen, die nicht in die `ISO-8859-1`-Codierungsdefinition passen, werden in Leerzeichen umgewandelt. Für japanischen Text können Sie eine von mehreren Codierungsoptionen auswählen: `SHIFT_JIS`, `EUC_JP`, `UTF-8` oder `ISO-2022-JP`.

Textsprache. Gibt die Sprache des Texts an, für den 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.

Verwenden von Konzeptmodellnuggets in einem Stream

Beim Verwenden eines Textmining-Modellierungsknotens können Sie ein Konzeptmodellnugget oder ein Kategoriemodellnugget (durch eine interaktive Workbenchsitzung) generieren. Das folgende Beispiel zeigt die Verwendung eines Konzeptmodells in einem einfachen Stream.

Beispiel: Statistikdateiknoten mit Konzeptmodellnugget

Das folgende Beispiel zeigt die Verwendung des Textmining-Konzeptmodellnuggets.

Abbildung 3. Beispielstream: Statistikdateiknoten mit einem Textmining-Konzeptmodellnugget

1. **Statistikdateiknoten (Registerkarte "Daten").** Zuerst fügten wir diesen Knoten zum Stream hinzu, um anzugeben, wo die Textdokumente gespeichert sind.

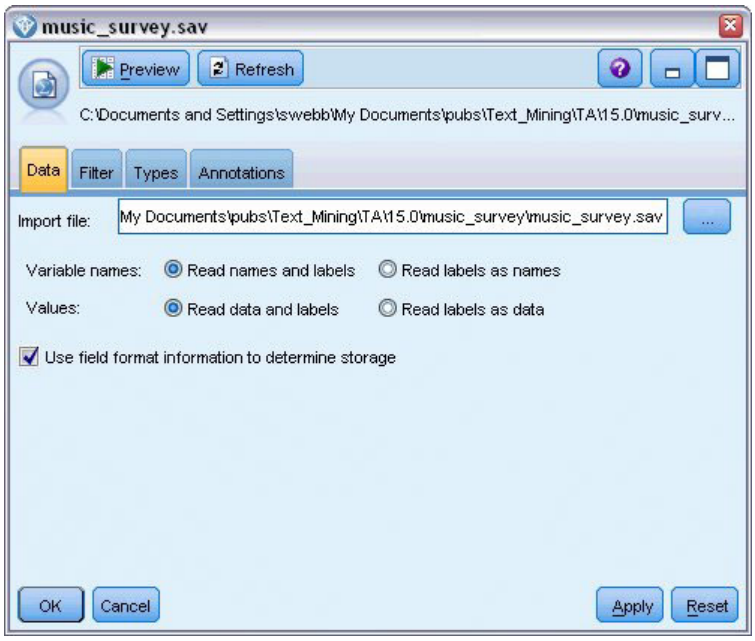


Abbildung 4. Dialogfeld des Statistikdateiknotens: Registerkarte "Daten"

2. **Textmining-Konzeptmodellnugget (Registerkarte "Modell").** Als Nächstes haben wir ein Konzeptmodellnugget hinzugefügt und mit dem Statistikdateiknoten verbunden. Wir haben die Konzepte ausgewählt, die wir für das Scoring unserer Daten verwenden wollten.

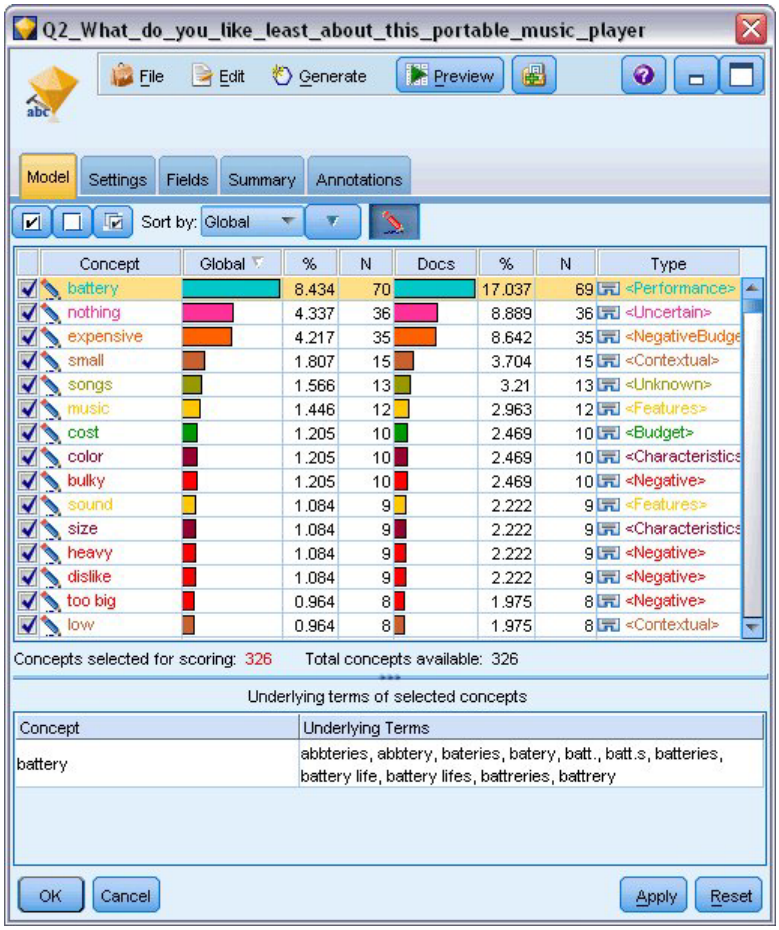


Abbildung 5. Dialogfeld des Textmining-Modellnuggets: Registerkarte "Modell"

3. **Textmining-Konzeptmodellnugget (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 auf der Registerkarte "Modell" ausgewählt wurde. Jeder Feldname wird aus dem Konzeptnamen und dem Präfix "*Concept_*" gebildet.

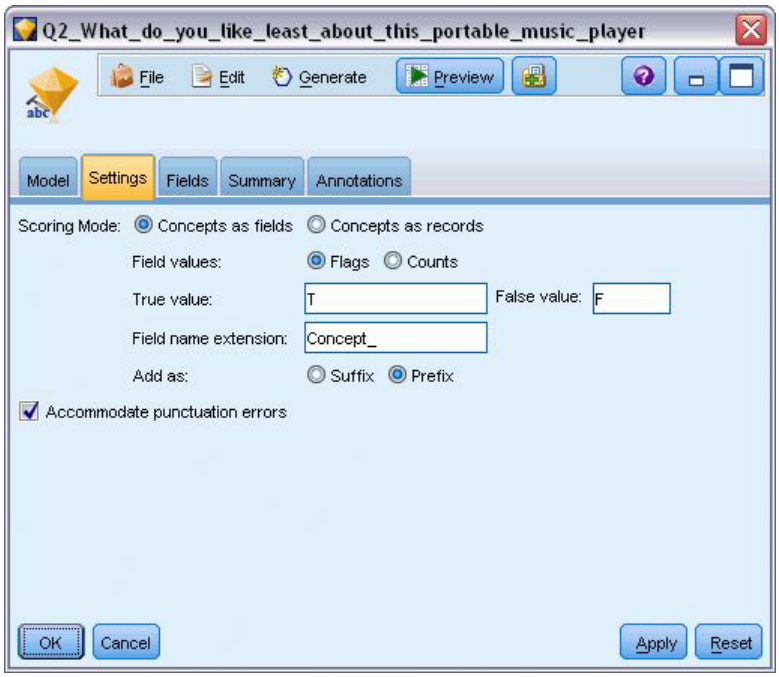


Abbildung 6. Dialogfeld des Textmining-Konzeptmodellnuggets: Registerkarte "Einstellungen"

4. **Textmining-Konzeptmodellnugget (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 außerdem die Option **Textfeld enthält: Tatsächlicher Text** ausgewählt.

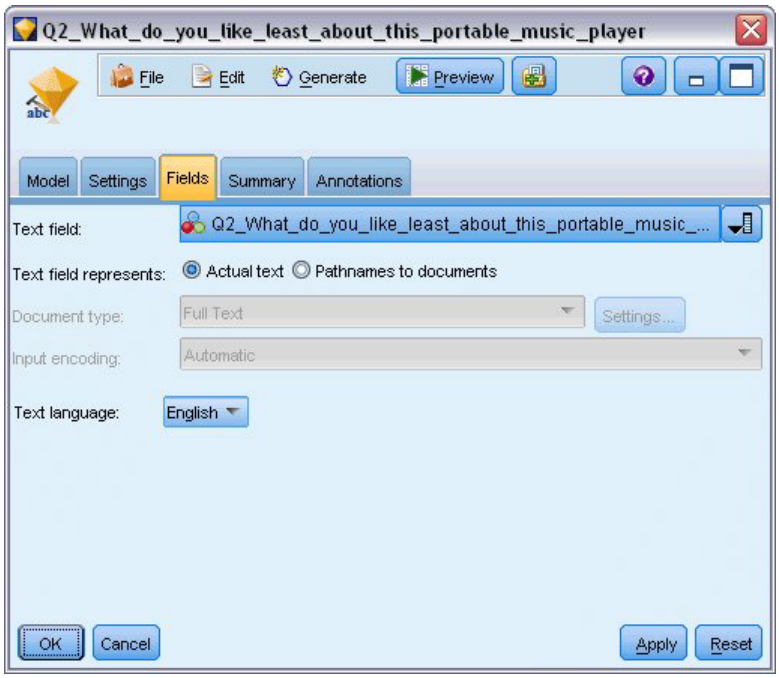
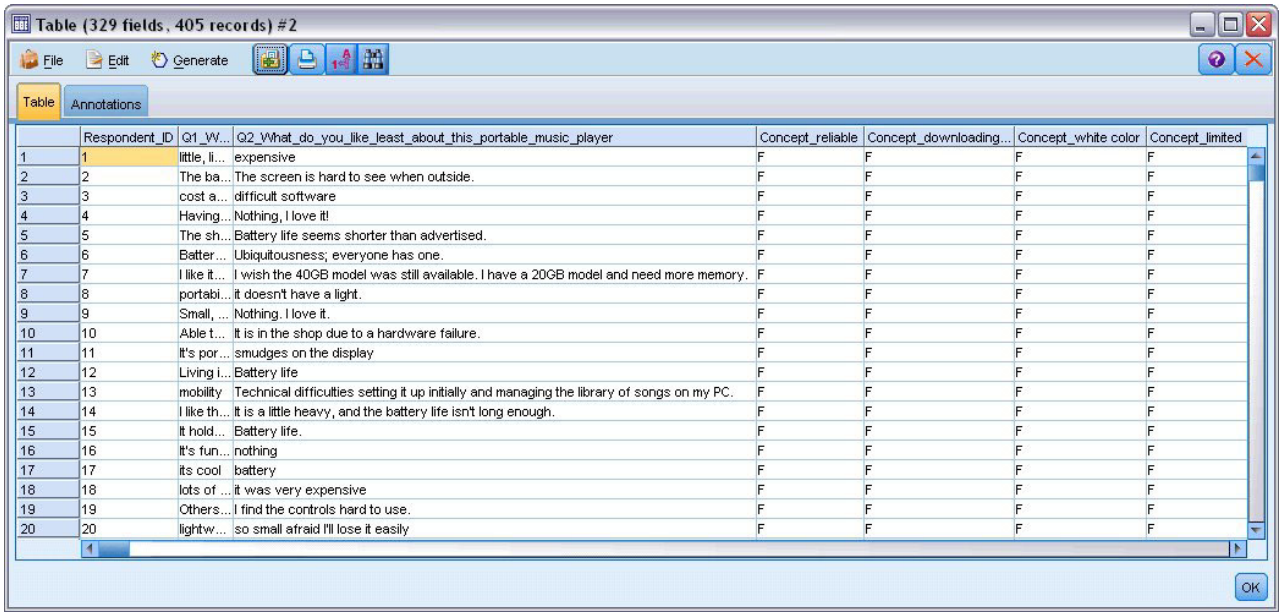


Abbildung 7. Dialogfeld des Textmining-Konzeptmodellnuggets: Registerkarte "Felder"

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



	Respondent_ID	Q1_W...	Q2_What_do_you_like_least_about_this_portable_music_player	Concept_reliable	Concept_downloading...	Concept_white color	Concept_limited
1	1	little, li...	expensive	F	F	F	F
2	2	The ba...	The screen is hard to see when outside.	F	F	F	F
3	3	cost a...	difficult software	F	F	F	F
4	4	Having...	Nothing, I love it!	F	F	F	F
5	5	The sh...	Battery life seems shorter than advertised.	F	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	portabi...	it doesn't have a light.	F	F	F	F
9	9	Small, ...	Nothing. I love it.	F	F	F	F
10	10	Able t...	It is in the shop due to a hardware failure.	F	F	F	F
11	11	It's por...	smudges on the display	F	F	F	F
12	12	Living i...	Battery life	F	F	F	F
13	13	mobility	Technical difficulties setting it up initially and managing the library of songs on my PC.	F	F	F	F
14	14	I like th...	It is a little heavy, and the battery life isn't long enough.	F	F	F	F
15	15	It hold...	Battery life.	F	F	F	F
16	16	It's fun...	nothing	F	F	F	F
17	17	its cool	battery	F	F	F	F
18	18	lots of ...	it was very expensive	F	F	F	F
19	19	Others...	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	F

Abbildung 8. Tabellenausgabe, in der geblättert wird, um die Konzeptflags anzuzeigen

Textminingnugget: Kategoriemodell

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

Wenn Sie im Modellierungsknoten eine interaktive Workbenchsitzung starten, können Sie die Extraktionsergebnisse untersuchen, die Ressourcen optimieren und eine Optimierung Ihrer Kategorien vornehmen, bevor Sie Kategoriemodelle generieren. Wenn Sie einen Stream ausführen, der ein Textmining-Modellnugget enthält, werden den Daten gemäß dem Erstellungsmodus, der auf der Registerkarte "Modell" des Textmining-Modellierungsknotens ausgewählt wurde, vor der Erstellung des Modells neue Felder hinzugefügt. Weitere Informationen finden Sie im Thema „Kategoriemodellnugget: Registerkarte "Modell"" auf Seite 43.

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

Die Textmining-Modellnuggets werden nach der Generierung in der Palette der Modellnuggets gespeichert (diese befindet sich rechts oben im IBM SPSS Modeler-Fenster auf der Registerkarte "Modelle").

Anzeigen von Ergebnissen

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

Hinzufügen von Modellen zu Streams

Um das Modellnugget Ihrem Stream hinzuzufügen, klicken Sie auf das Symbol in der Palette der Modellnuggets und dann auf den Streamerstellungsbereich, 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 Vorhersagen zu erstellen.

Kategoriemodellnugget: 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 mindestens einen der Deskriptoren oder zugrunde liegende Terme enthält, wird das Dokument bzw. der Datensatz der Kategorie zugewiesen, welcher der Deskriptor angehört. Diese zugrunde liegenden Terme 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 Modellnuggets gefunden wurden, permutierte Terme, Terme aus Fuzzy-Gruppierungen usw.

Hinweis: Wenn Sie stattdessen ein Konzeptmodell generiert haben, enthält diese Registerkarte andere Ergebnisse. Weitere Informationen finden Sie im Thema „Konzeptmodell: Registerkarte "Modell"" auf Seite 33.

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:

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

Tabelle 5. Deskriptorsymbole

	Konzepte		TLA-Muster
	Typen		Kategorieregeln

- **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, wird das Kontrollkästchen mit einem blauen Hintergrund angezeigt, um darauf hinzuweisen, dass in den untergeordneten Kategorien 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 markieren.** Aktiviert alle Kontrollkästchen für die ausgewählten Zeilen in der Tabelle.
- **Markierung für ausgewählte Elemente aufheben.** Inaktiviert 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 in der Symbolleiste verwenden.
- **Alle Markierungen aufheben.** Inaktiviert alle Kontrollkästchen in der Tabelle. Wenn Sie eine Kategorie inaktivieren, wird diese in der Endausgabe nicht verwendet. Sie können auch das entsprechende leere Kontrollkästchensymbol 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.

Kategoriemodellnugget: 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).

Hinweis: Diese Registerkarte wird nur im Knotendialogfeld angezeigt, wenn das Modellnugget im Erstellungsbereich oder in einem Stream platziert wird. Sie ist nicht vorhanden, wenn Sie direkt in der Modellpalette auf dieses Nugget zugreifen.

Scoring-Modus: 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 Flagwert für **Wahr** und für **Falsch** ein, z. B. *Ja/Nein*, *Wahr/Falsch*, *W/F* oder *1* und *2*. Die Speichertypen werden automatisch so festgelegt, dass sie die ausgewählten Werte widerspiegeln. Wenn Sie beispielsweise numerische Werte für die Flags eingeben, werden sie automatisch als Ganzzahl behandelt. Als Speichertypen für Flags sind "Zeichenfolge", "Ganzzahl", "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 Zeichenfolge einzufügen. Wählen Sie **Suffix**, um die Erweiterung am Ende der Zeichenfolge einzufügen.

Wenn eine Unterkategorie nicht ausgewählt wurde. 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 **Zugehörige Deskriptoren vollständig aus Scoring ausschließen** bewirkt, dass Deskriptoren von Unterkategorien ohne Häkchen (nicht ausgewählt) beim Scoring ignoriert und nicht verwendet werden.
- Die Option **Deskriptoren mit jenen in der übergeordneten Kategorie aggregieren** bewirkt, 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 Interpunktionsfehlern (zum Beispiel ungeeignete Verwendung) während der Extraktion, 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.

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

Scoring-Modus: Kategorien als Datensätze

Mit dieser Option wird für jedes `Kategorie/Dokument`-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 6. Ausgabefelder für "Kategorien als Datensätze"

Feld "Neue Ausgabe"	Beschreibung
`Kategorie`	Enthält den Namen der Kategorie, der das Textdokument zugewiesen wurde. Wenn die Kategorie die Unterkategorie einer anderen Kategorie ist, 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 wurde. 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 **Zugehörige Deskriptoren vollständig aus Scoring ausschließen** bewirkt, dass Deskriptoren von Unterkategorien ohne Häkchen (nicht ausgewählt) beim Scoring ignoriert und nicht verwendet werden.

- Die Option **Deskriptoren mit jenen in der übergeordneten Kategorie aggregieren** bewirkt, 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 Interpunktionsfehlern (zum Beispiel ungeeignete Verwendung) während der Extraktion, 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.

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

Kategoriemodellnugget: Andere Registerkarten

Die Registerkarten "Felder" und "Einstellungen" für das Kategoriemodellnugget sind dieselben wie für das Konzeptmodellnugget.

- Registerkarte "Felder". Weitere Informationen finden Sie im Thema „Konzeptmodell: Registerkarte "Felder"" auf Seite 37.
- Registerkarte "Übersicht". Weitere Informationen finden Sie im Thema „Konzeptmodell: Registerkarte "Übersicht"" auf Seite 38.

Verwenden von Kategoriemodellnuggets in einem Stream

Das Textmining-Kategoriemodellnugget wird aus einer interaktiven Workbenchsitzung generiert. Sie können dieses Modellnugget in einem Stream verwenden.

Beispiel: Statistikdateiknoten mit Kategoriemodellnugget

Das folgende Beispiel zeigt die Verwendung des Textmining-Modellnuggets.

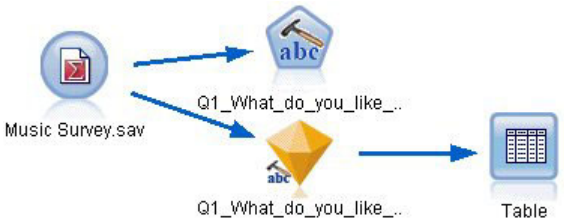


Abbildung 9. Beispielstream: Statistikdateiknoten mit einem Textmining-Kategoriemodellnugget

1. **Statistikdateiknoten (Registerkarte "Daten").** Zuerst fügten wir diesen Knoten zum Stream hinzu, um anzugeben, wo die Textdokumente gespeichert sind.

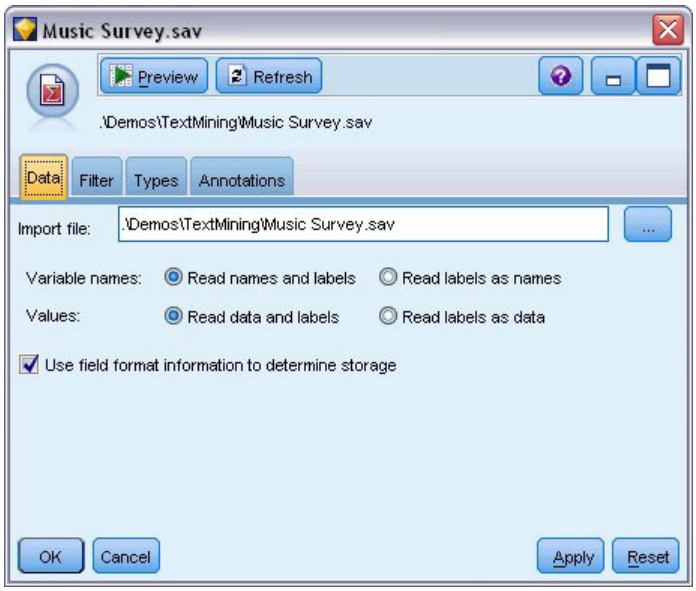


Abbildung 10. Dialogfeld des Statistikdateiknotens: Registerkarte "Daten"

2. **Textmining-Kategoriemodellnugget (Registerkarte "Modell").** Als Nächstes haben wir ein Kategoriemodellnugget hinzugefügt und mit dem Statistikdateiknoten verbunden. Wir haben die Kategorien ausgewählt, die wir für das Scoring unserer Daten verwenden wollten.

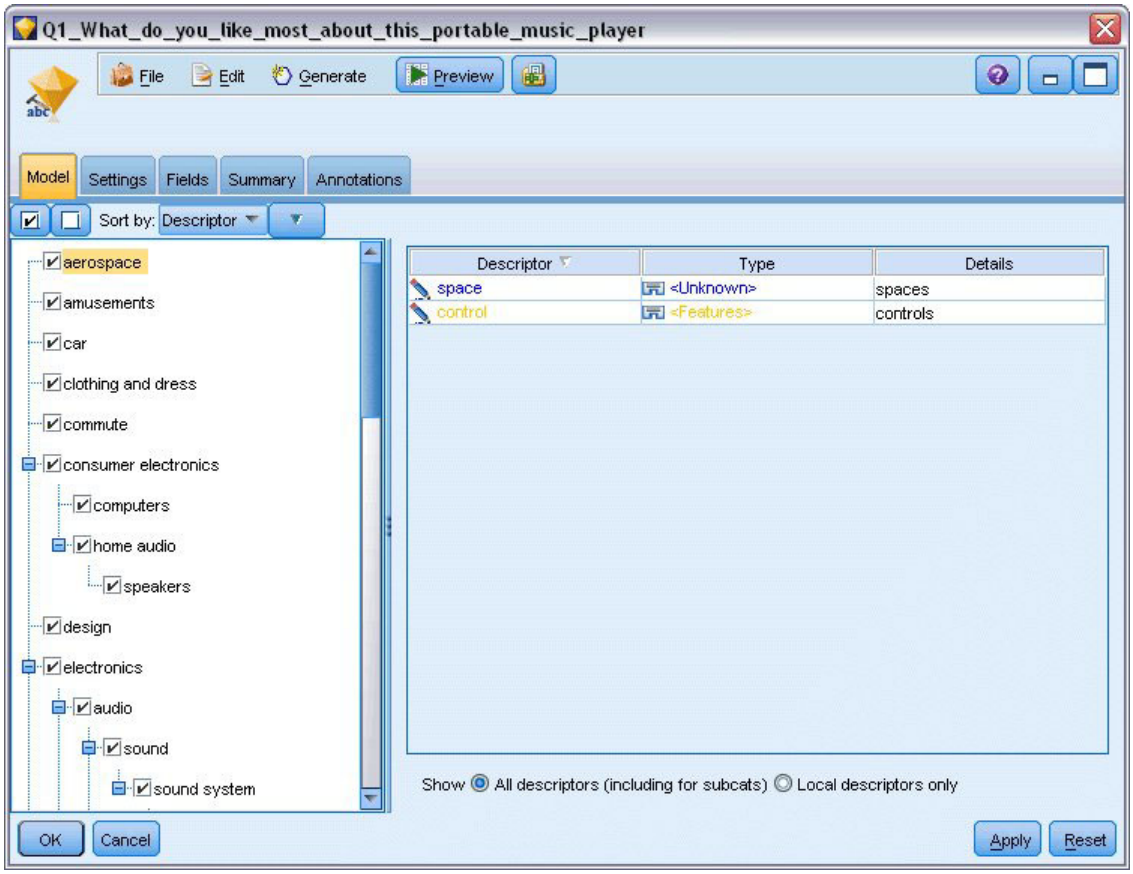


Abbildung 11. Dialogfeld des Textmining-Modellnuggets: Registerkarte "Modell"

3. **Textmining-Modellnugget (Registerkarte "Einstellungen").** Als Nächstes haben wir das Ausgabeformat **Kategorien als Felder** definiert.

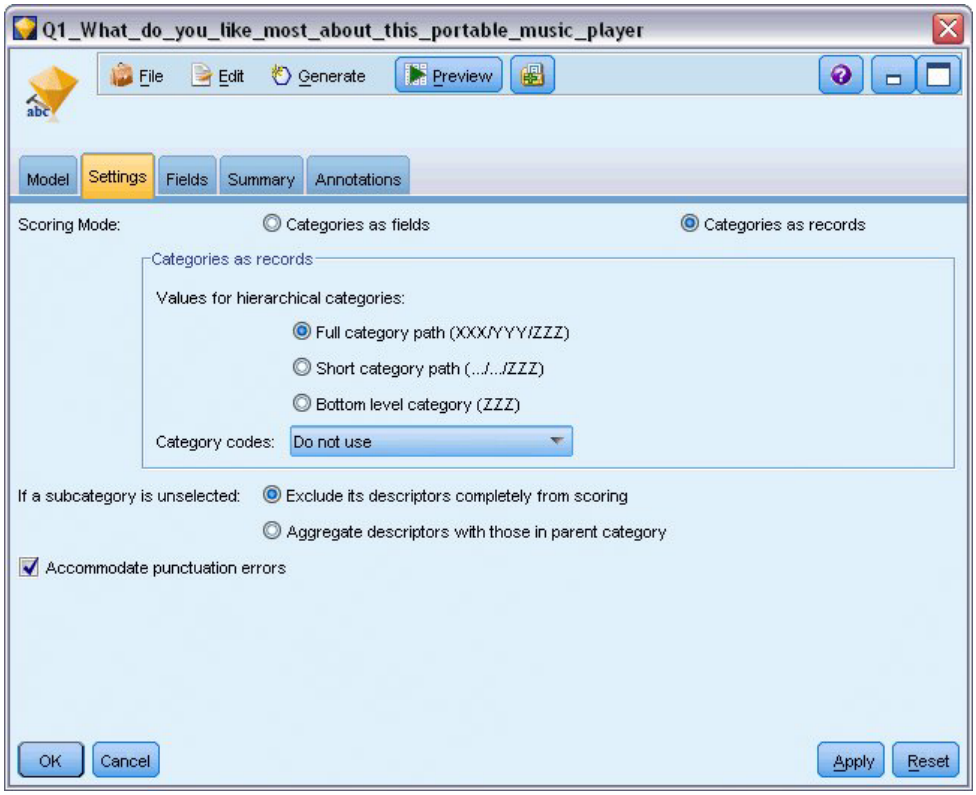


Abbildung 12. Dialogfeld des Kategoriemodellnuggets: Registerkarte "Einstellungen"

4. **Textmining-Kategoriemodellnugget (Registerkarte "Felder").** Anschließend haben wir die Textfeldvariable ausgewählt. Dies ist der Feldname aus dem Statistikdateiknoten. Dann wählten wir die Option **Textfeld enthält: Tatsächlicher Text** sowie andere Einstellungen aus.

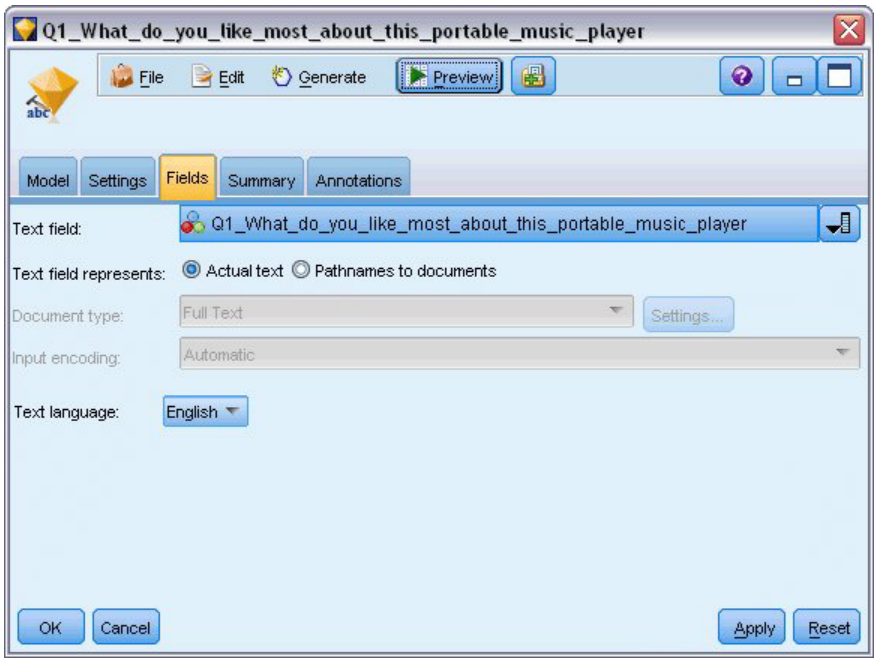
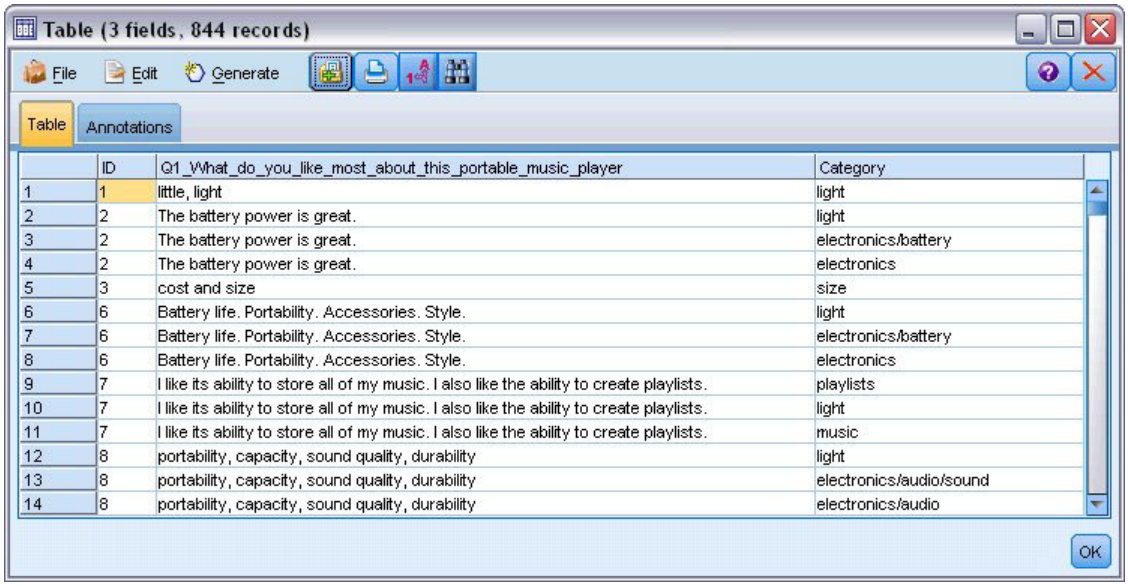


Abbildung 13. Dialogfeld des Textmining-Modellnuggets: Registerkarte "Felder"

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



	ID	Q1_What_do_you_like_most_about_this_portable_music_player	Category
1	1	little, light	light
2	2	The battery power is great.	light
3	2	The battery power is great.	electronics/battery
4	2	The battery power is great.	electronics
5	3	cost and size	size
6	6	Battery life. Portability. Accessories. Style.	light
7	6	Battery life. Portability. Accessories. Style.	electronics/battery
8	6	Battery life. Portability. Accessories. Style.	electronics
9	7	I like its ability to store all of my music. I also like the ability to create playlists.	playlists
10	7	I like its ability to store all of my music. I also like the ability to create playlists.	light
11	7	I like its ability to store all of my music. I also like the ability to create playlists.	music
12	8	portability, capacity, sound quality, durability	light
13	8	portability, capacity, sound quality, durability	electronics/audio/sound
14	8	portability, capacity, sound quality, durability	electronics/audio

Abbildung 14. Tabellenausgabe

Kapitel 4. Mining für Textlinks

Textlinkanalyseknoten

Über den Textlinkanalyseknoten (TLA) wird die Konzeptextraktion beim Textmining 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 boolesche 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 Ausgabedaten neu strukturiert werden. Weitere Informationen finden Sie im Thema Kapitel 19, „Textlinkregeln", auf Seite 227.

Der Textlinkanalyseknoten eröffnet einen direkteren Weg, TLA-Musterergebnisse in Ihrem Text zu ermitteln und die Ergebnisse anschließend dem Dataset im Stream hinzuzufügen. Doch der Textlinkanalyseknoten stellt nicht die einzige Methode zur Textlinkanalyse dar. Sie können dazu auch eine interaktive Workbenchsitzung im Textmining-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. Weitere Informationen finden Sie im Thema Kapitel 12, „Untersuchen von Textlinkanalyse", auf Seite 163. Tatsächlich ist die Nutzung des Textminingknotens zur Extraktion von TLA-Ergebnissen eine hervorragende Methode, um Vorlagen im Hinblick auf Ihre Daten zu untersuchen und eine entsprechende Optimierung vorzunehmen. Anschließend können die Vorlagen direkt im TLA-Knoten verwendet werden.

Die Ausgabe kann in bis zu sechs Slots, oder Teilen, dargestellt werden. Japanische Muster werden nur als ein oder zwei Slots ausgegeben. Weitere Informationen finden Sie im Thema „Ausgabe des TLA-Knotens" auf Seite 55.

Sie finden diesen Knoten auf der IBM SPSS Modeler Text Analytics-Registerkarte der Knotenpalette am unteren Rand des IBM SPSS Modeler-Fensters. Weitere Informationen finden Sie im Thema „IBM SPSS Modeler Text Analytics-Knoten" auf Seite 9.

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

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

Textlinkanalyseknoten: 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:

- **Tatsächlicher 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 Extraktion 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 Extraktionsprozess oder Teile des Extraktionsprozesses zu überspringen. Er ermöglicht das Definieren von Trennzeichen für Terme, 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. Weitere Informationen finden Sie im Thema „Dokumenteinstellungen der Registerkarte "Felder"" auf Seite 22.
- **XML-Text.** 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-Text** des Dialogfelds "Dokumenteinstellungen" explizit die XML-Elemente angeben, die den Text enthalten, der während des Extraktionsprozesses gelesen werden soll. Weitere Informationen finden Sie im Thema „Dokumenteinstellungen der Registerkarte "Felder"" auf Seite 22.

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 Extraktionsmodus 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 Extraktionsprozess 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 Extraktionen 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 Extraktionsgröße vermeiden.

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

Eingabecodierung. Diese Option ist nur verfügbar, wenn Sie angegeben haben, dass das Textfeld **Pfadnamen zu Dokumenten** enthält. Es bestimmt die Standardtextcodierung. Für alle Sprachen außer Japanisch erfolgt eine Konvertierung von der angegebenen bzw. erkannten Codierung in `ISO-8859-1`. Selbst wenn Sie also eine andere Codierung angeben, wird diese vor der Verarbeitung von der Extraktionsengine in `ISO-8859-1` konvertiert. Alle Zeichen, die nicht in die `ISO-8859-1`-Codierungsdefinition passen, werden in Leerzeichen umgewandelt. Für japanischen Text können Sie eine von mehreren Codierungsoptionen auswählen: `SHIFT_JIS`, `EUC_JP`, `UTF-8` oder `ISO-2022-JP`.

Ressourcen kopieren von. Beim Textmining basiert die Extraktion 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 des Texts während der Extraktion, 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 Extraktion. Klicken Sie auf **Laden** und wählen Sie die Vorlage aus, 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. Weitere Informationen finden Sie im Thema „Kopieren von Ressourcen aus TAPs und Vorlagen” auf Seite 27.

Textsprache. Bestimmt die Sprache des für das Mining verwendeten Texts. Die in den Knoten kopierten Ressourcen steuern die angezeigten Sprachoptionen. Sie können die Sprache auswählen, auf die die Ressourcen abgestimmt wurden, oder die Option **ALLE** auswählen. Es wird empfohlen, für die Textdaten die exakte Sprache anzugeben. Wenn Sie sich jedoch nicht sicher sind, können Sie die Option **ALLE** auswählen. Die Option **ALLE** ist für japanischen Text nicht verfügbar. Diese Option **ALLE** verlängert die Ausführungsdauer, da mithilfe der automatischen Spracherkennung zunächst alle Dokumente und Datensätze gescannt werden, um die Textsprache zu ermitteln. Wenn Sie diese Option auswählen, werden alle Datensätze oder Dokumente, die in einer unterstützten und lizenzierten Sprache vorliegen, durch die Extraktionsengine mit den internen Wörterbüchern in der jeweiligen Sprache gelesen. Weitere Informationen finden Sie im Thema „Language Identifier” auf Seite 225. 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 "Modell"

Die Registerkarte "Modell" enthält eine einzige Option, die sich auf die Geschwindigkeit und die Genauigkeit des Extraktionsprozesses auswirkt.

Für Scoring-Geschwindigkeit optimieren. Diese Option ist standardmäßig aktiviert und stellt sicher, dass das erstellte Modell kompakt ist und das Scoring mit hoher Geschwindigkeit durchführt. Durch das Abwählen dieser Option wird ein Modell erstellt, das das Scoring langsamer durchführt aber die gesamte Konzepttypkonsistenz sicherstellt, d. h., das Modell stellt sicher, dass ein bestimmtes Konzept immer nur einem Typ zugewiesen wird.

Textlinkanalyseknoten: Registerkarte "Experten"

In diesem Knoten ist die Extraktion von TLA-Musterergebnissen (TLA - Text Link Analysis) 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 Extraktionsprozesses fest. Zudem werden die Extraktionsergebnisse von einer Reihe von linguistischen Ressourcen und Optionen beeinflusst. Diese werden über die ausgewählte Ressourcenvorlage gesteuert.

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

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Interpunktionsfehlern (zum Beispiel ungeeignete Verwendung) während der Extraktion, 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 Fuzzy-Gruppierungsverfahren an, das hilft, häufig falsch geschriebene Wörter oder ähnlich geschriebene Wörter unter einem Konzept zu gruppieren. Der Algorithmus für Fuzzy-Gruppierung entfernt alle Vokale (außer dem ersten) und doppelte/dreifache Konsonanten temporär aus extrahierten Wörtern und vergleicht sie, um festzustellen, ob sie gleich sind, sodass `Modellierung` und `Modelierung` zusammen gruppiert werden würden. Wenn jedoch jeder Term einem anderen Typ (ausschließlich des Typs `<Unknown>`) zugewiesen ist, wird das Fuzzy-Gruppierungsverfahren nicht angewendet.

Sie können auch die minimal erforderliche Zahl von *Stamm*zeichen definieren, bevor Fuzzy-Gruppierung eingesetzt wird. Die Anzahl der Stammzeichen in einem Term berechnet sich aus der Summe aller Zeichen abzüglich aller Zeichen, die Beugungsendungen und - bei zusammengesetzten Termen - Determinatoren und Präpositionen bilden. So würde beispielsweise der Term `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 sich herausstellt, dass bestimmte Wörter später falsch eingruppiert werden, können Sie einzelne Wortpaare aus dem Verfahren ausschließen, indem Sie sie auf der Registerkarte "Erweiterte Ressourcen" im Bereich **Fuzzy-Gruppierung: Ausnahmen** explizit deklarieren. Weitere Informationen finden Sie im Thema „Fuzzy-Gruppierung” auf Seite 218.

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

Nicht linguistische Entitäten extrahieren. Diese Option extrahiert nicht linguistische Entitäten wie beispielsweise Telefonnummern, Personalausweisnummern, Uhrzeiten, Datumsangaben, Währungen, Ziffern, Prozentsätze, E-Mail-Adressen und HTTP-Adressen. Sie können bestimmte Typen von nicht linguistischen Entitäten im Abschnitt **Nicht linguistische Entitäten: Konfiguration** der Registerkarte "Erweiterte Ressourcen" ein- bzw. ausschließen. Durch Inaktivierung unnötiger Entitäten vergeudet die Extraktionsengine keine Verarbeitungszeit. Weitere Informationen finden Sie im Thema „Konfiguration” auf Seite 222.

Großbuchstaben-Algorithmus. Diese Option extrahiert einfache und zusammengesetzte Terme, die sich nicht in den integrierten Wörterbüchern befinden, solange der erste Buchstabe des Terms 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 Texts 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 Terme abzugleichen, die dem Typ <Person> zugeordnet sind. Wird beispielsweise *doe* gefunden und anfänglich dem Typ <Unknown> zugeordnet, überprüft die Extraktionsengine, ob ein zusammengesetzter Term 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 Uniterme extrahiert werden.

Maximale Füllwörter in zusammengesetzten Konzepten. Diese Option gibt die maximale Anzahl von Füllwörtern an, die für die Anwendung des Permutationsverfahrens vorhanden sein müssen. Dieses Permutationsverfahren gruppiert ähnliche Wortfolgen, 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` als auch `Vertreter des Unternehmens` extrahiert wurden. In diesem Fall würden beide extrahierte Terme in der endgültigen Konzeptliste zusammen gruppiert, da beide Terme als gleich betrachtet werden, wenn `des` ignoriert wird.

Für japanischen Text

Bei japanischem Text können Sie auswählen, welcher Sekundäranalysator angewendet werden soll.

Hinweis: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

Sekundäre Analyse. Bei einer Extraktion werden grundlegende Schlüsselwörter unter Verwendung des Standardsets von Typen extrahiert. Wenn Sie jedoch einen Sekundäranalysator auswählen, können Sie auch mehrere oder reichhaltigere Konzepte erhalten, da der Extraktor nun Partikel und Hilfsverben als Teil des Konzepts beinhaltet. Im Fall der Stimmungsanalyse wird auch eine große Anzahl an zusätzlichen Typen eingeschlossen. Des Weiteren ermöglicht Ihnen die Wahl eines Sekundäranalysators, auch Ergebnisse für die Textlinkanalyse zu generieren.

Hinweis: Wenn ein Sekundäranalysator aufgerufen wird, dauert der Extraktionsprozess länger.

- **Abhängigkeitsanalyse**. Die Wahl dieser Option führt zu erweiterten Partikeln für die Extraktionskonzepte aus der grundlegenden Typ- und Stichwortextraktion. Sie können auch vielfältigere Musterergebnisse aus einer Abhängigkeitstextlinkanalyse (TLA) gewinnen.
- **Stimmungsanalyse**. Bei dieser Analyse werden zusätzliche Konzepte und - wann immer möglich - TLA-Musterergebnisse extrahiert. Neben den grundlegenden Typen können Sie zusätzlich mehr als 80 Stimmungstypen nutzen. 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 Textmining 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.

Hinweis: Textlinkanalyse-Musterregeln für japanischen Text erzeugen nur Musterergebnisse in einem oder zwei Slots.

Hinweis: Bestehende Streams, die einen Textlinkanalyseknoten aus einer früheren Version als 5.0 enthalten, sind möglicherweise erst nach der Aktualisierung der Knoten wieder 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 Leistung 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. Weitere Informationen finden Sie im Thema „Übersetzungseinstellungen” auf Seite 63.

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 Optionsfolge in den Menüs aus: **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 wählen Sie die folgende Optionsfolge in den Menüs aus: **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 15. Beispiel: Statistikdateiknoten mit Textlinkanalyseknoten

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

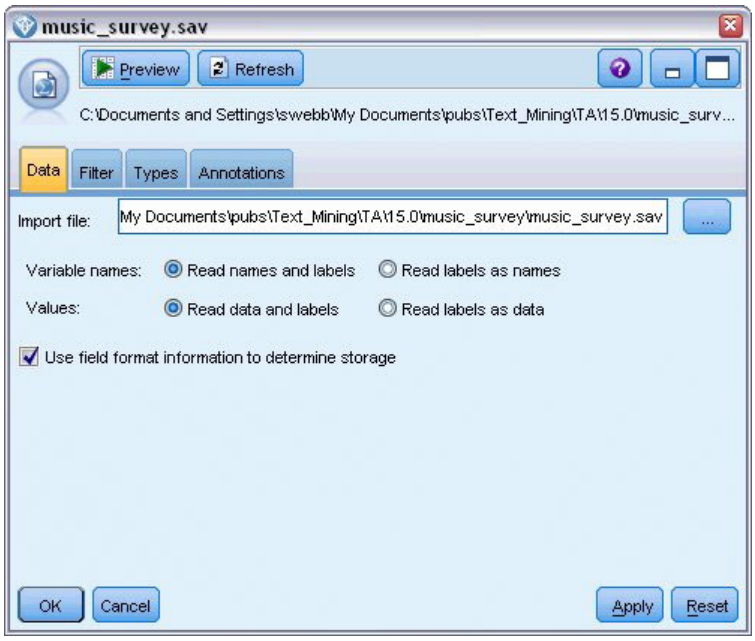


Abbildung 16. Dialogfeld des Statistikdateiknotens: Registerkarte "Daten"

2. **Textlinkanalyseknoten (Registerkarte "Felder").** Anschließend fügten wir diesen Knoten zum 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.

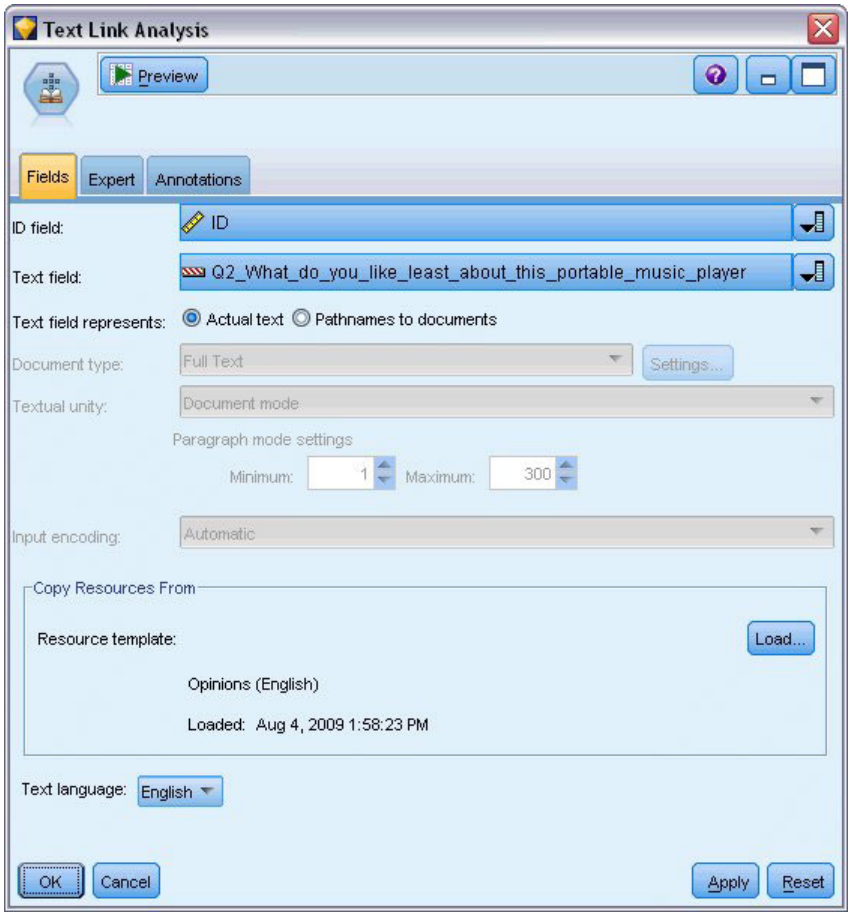
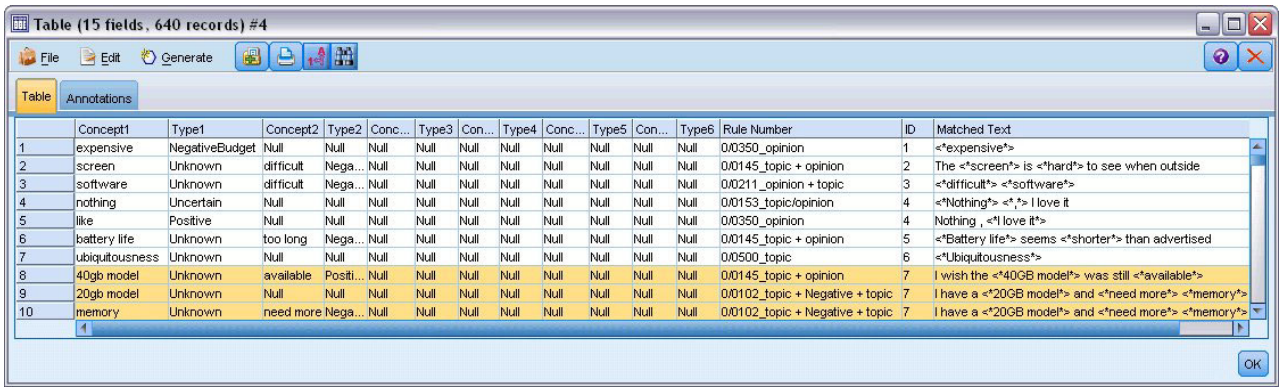


Abbildung 17. Dialogfeld des Textlinkanalyseknotens: Registerkarte "Felder"

3. **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.



	Concept1	Type1	Concept2	Type2	Conc...	Type3	Con...	Type4	Conc...	Type5	Con...	Type6	Rule Number	ID	Matched Text
1	expensive	NegativeBudget	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	1	<*expensive*>
2	screen	Unknown	difficult	Nega...	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	Nega...	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	Nega...	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	available	Positi...	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 more	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>

Abbildung 18. Tabellenausgabeknoten

Kapitel 5. Übersetzen von Text für die Extraktion

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 Textmining 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 Textmining in einer dieser Sprachen fügen Sie einfach einen Übersetzungsknoten vor dem Textmining-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 IBM SPSS Modeler Text Analytics-Registerkarte der Knotenpalette am unteren Rand des IBM SPSS Modeler-Fensters. Weitere Informationen finden Sie im Thema „IBM SPSS Modeler Text Analytics-Knoten" auf Seite 9.

Zwischenspeichern der Übersetzung. Wenn Sie die Übersetzung zwischenspeichern, 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 die folgende Optionsfolge in den Menüs aus: **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 wählen Sie die folgende Optionsfolge in den Menüs aus: **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 Web durch einen Proxy-Server abzurufen, müssen Sie den Proxy-Server in der Datei `net.properties` für die Client- und Serverversion 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 Web zugreifen oder eine Lizenz für SDL Software as a Service (SaaS) abrufen, da diese Verbindungen über Java aufgebaut werden. In der Clientversion befindet sich diese Datei standardmäßig im Verzeichnis *C:\Programme\IBM\SPSS\Modeler\16\jre\lib\net.properties*. In der Serverversion befindet sich diese Datei standardmäßig im Verzeichnis *C:\Programme\IBM\SPSS\Modeler\16\ext\bin\spss.TMWBServer\jre\lib\net.properties*.

Übersetzungsknoten: Registerkarte "Übersetzung"

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 Zeichenfolgenfeld 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:

- **Tatsächlicher 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. Weitere Informationen finden Sie im Thema „Dateilistenknoten" auf Seite 11.

Eingabecodierung. Wählen Sie die Codierung des Quellentexts aus. Sie können mit der Auswahl der Option **Automatisch** beginnen. Wenn Sie jedoch bemerken, dass einige Dateien nicht korrekt verarbeitet werden, sollten Sie die tatsächliche Codierung hier aus der Liste auswählen. Die Option "Automatisch" erkennt die Codierung bei kurzen Texten wie kurzen Datenbanksätzen eventuell nicht richtig. Die Textausgabe von diesem Knoten wird in `UTF-8` codiert.

Einstellungen. Gibt die Übersetzungseinstellungen für den Stream an.

- **Sprachpaarverbindung**. Wählen Sie das zu verwendende Sprachpaar aus. Verfügbare Sprachpaare werden automatisch in dieser Liste angezeigt, nachdem Sie im Dialogfeld **Übersetzungseinstellungen** den Link zum SDL-Service eingerichtet haben. Weitere Informationen finden Sie im Thema „Übersetzungseinstellungen" auf Seite 63.
- **Übersetzungsgenauigkeit**. Geben Sie die gewünschte Genauigkeit an, indem Sie einen Wert von 1 bis 3 auswählen, 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 verwenden** und geben Sie den Namen unter **Name des Wörterbuchs** ein. Zur Verwendung mehrerer Wörterbücher müssen Sie die Namen mit Kommas trennen.
- **Zuvor übersetzten Text, wenn möglich, 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 Beschriftungsnamen für die nächste Ausführung gespeichert. Diese Option ist nur verfügbar, wenn Sie eine SDL-Übersetzungssprache ausgewählt haben.

Hinweis: Wenn der Text im Stream gespeichert wird, können Sie auch das Caching 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.

- **Beschriftung.** Wenn Sie die Option **Nach Möglichkeit vorherigen Übersetzungstext speichern und wiederverwenden** auswählen, geben Sie einen Beschriftungsnamen für die Ergebnisse an. Diese Beschriftung wird verwendet, um den vorherigen Übersetzungstext zu identifizieren. Wenn keine Beschriftung angegeben ist, wird bei Ausführung des Streams eine Warnung zu den Streameigenschaften 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 wiederverwenden können. Sobald Sie hier eine Verbindung definiert haben, können Sie beim Übersetzen rasch eine Sprachpaarverbindung auswählen, ohne erneut alle Verbindungseinstellungen eingeben zu müssen.

Eine Sprachpaarverbindung gibt die Quellen- und Zielsprache sowie die URL-Verbindungsdetails für den Server an. Beispiel: *Chinesisch - Englisch* bedeutet, dass der Quellentext in Chinesisch ist und die resultierende Übersetzung in Englisch sein wird. Jede Verbindung, auf die Sie über die SDL-Onlineservices zugreifen möchten, muss manuell definiert werden.

Wichtig! Wenn Sie versuchen, Informationen über das Web durch einen Proxy-Server abzurufen, müssen Sie den Proxy-Server in der Datei `net.properties` für die Client- und Serverversion 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 Web zugreifen oder eine Lizenz für SDL Software as a Service (SaaS) abrufen, da diese Verbindungen über Java aufgebaut werden. In der Clientversion befindet sich diese Datei standardmäßig im Verzeichnis *C:\Programme\IBM\SPSS\Modeler\16\jre\lib\net.properties*. In der Serverversion befindet sich diese Datei standardmäßig im Verzeichnis *C:\Programme\IBM\SPSS\Modeler\16\ext\bin\spss.TMWBServer\jre\lib\net.properties*.

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

Benutzername. 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 für diese Verbindung anzuzeigen.

Verwenden des Übersetzungsknotens

Fügen Sie vor einem Textminingknoten im Stream einen Übersetzungsknoten ein, um Konzepte aus unterstützten Übersetzungssprachen wie Arabisch, Chinesisch oder Persisch zu extrahieren.

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 Textminingknoten hinzugefügt und die Ausgabe befände sich in dem Verzeichnis, in dem auch der Übersetzungstext gespeichert ist.

Kapitel 6. Durchsuchen von Text aus externen Quellen

Datei-Viewer-Knoten

Wenn Sie ein Textmining für mehrere Dokumente durchführen, können Sie die vollständigen Pfadnamen der Dateien direkt in Ihre Textmining-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 einer einzelnen Datei zusammenführen zu müssen.

Der Datei-Viewer-Knoten kann ein besseres Verständnis der Ergebnisse aus der Textextraktion 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. Weitere Informationen finden Sie im Thema „Verwenden des Datei-Viewer-Knotens" auf Seite 66.

Sie finden diesen Knoten auf der IBM SPSS Modeler Text Analytics-Registerkarte der Knotenpalette am unteren Rand des IBM SPSS Modeler-Fensters. Weitere Informationen finden Sie im Thema „IBM SPSS Modeler Text Analytics-Knoten" auf Seite 9.

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

Einstellungen für Datei-Viewer-Knoten

Sie können die folgenden Einstellungen für den Datei-Viewer-Knoten angeben.

Dokumentfeld. 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 19. Stream, der die Verwendung des Datei-Viewer-Knotens erläutert

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

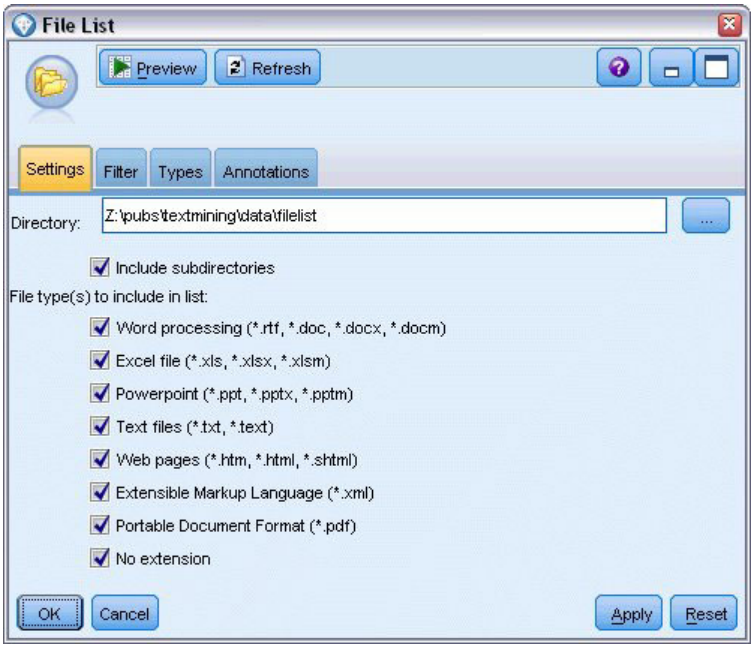


Abbildung 20. Dialogfeld des Dateilistenknotens: Registerkarte "Einstellungen"

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

Abbildung 21. Dialogfeld des Datei-Viewer-Knotens: Registerkarte "Einstellungen"

3. **Dialogfeld "Dateiviewerausgabe".** Als Nächstes führten wir den Stream aus, der die Liste der Dokumente in einem neuen Fenster ausgibt.

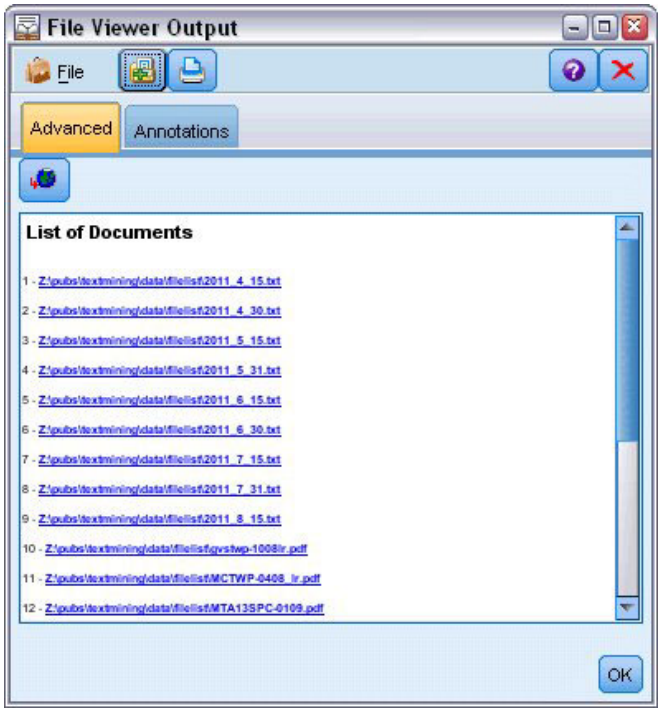


Abbildung 22. Datei-Viewer-Ausgabe

4. 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 Scripts

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

Dateilistenknoten: filelistnode

Sie können die Eigenschaften in der folgenden Tabelle für Scripts verwenden. Der Knoten selbst heißt `filelistnode`.

Tabelle 7. Dateilistenknoten - Scripteigenschaften

Scripteigenschaften	Datentyp
`path`	*Zeichenfolge*
`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*

Hinweis: Der Parameter "Liste erstellen" ist nicht mehr verfügbar und alle Scripts mit dieser Option werden automatisch in die Ausgabe "Dateien" umgewandelt.

Web-Feed-Knoten: webfeednode

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

Tabelle 8. Scripteigenschaften des Web-Feed-Knotens

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
`urls`	*string1 string2 ...stringn*	Alle URLs werden in der Listenstruktur angegeben. Durch "\n" getrennte URL-Liste
`recent_entries`	*Flag*	
`limit_entries`	*Ganzzahl*	Anzahl der aktuellen, pro URL zu lesenden Einträge.
`use_previous`	*Flag*	Zum Speichern und Wiederverwenden von Web-Feed-Cache.
`use_previous_label`	*Zeichenfolge*	Name des gespeicherten Web-Cache.
`start_record`	*Zeichenfolge*	Nicht-RSS-Anfangstag.

Tabelle 8. Scripteigenschaften des Web-Feed-Knotens (Forts.)

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
`url` *n* `.title`	*Zeichenfolge*	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 Anfangstag mit dem Titel des Inhalts.
`url` *n* `.short_description`	*Zeichenfolge*	Entspricht `url` *n* `.title`.
`url` *n* `.description`	*Zeichenfolge*	Entspricht `url` *n* `.title`.
`url` *n* `.authors`	*Zeichenfolge*	Entspricht `url` *n* `.title`.
`url` *n* `.contributors`	*Zeichenfolge*	Entspricht `url` *n* `.title`.
`url` *n* `.published_date`	*Zeichenfolge*	Entspricht `url` *n* `.title`.
`url` *n* `.modified_date`	*Zeichenfolge*	Entspricht `url` *n* `.title`.
`html_alg`	`None` `HTMLCleaner`	Methode zur Inhaltsfilterung.
`discard_lines`	*Flag*	Kurze Zeilen verwerfen. Mit `min_words` verwendet
`min_words`	*Ganzzahl*	Minimale Anzahl an Wörtern.
`discard_words`	*Flag*	Kurze Zeilen verwerfen. Mit `min_avg_len` verwendet
`min_avg_len`	*Ganzzahl*	
`discard_scw`	*Flag*	Zeilen mit vielen Wörtern aus einzelnen Zeichen verwerfen. Mit `max_scw` verwendet
`max_scw`	*Ganzzahl*	Maximaler Anteil (0-100 Prozent) an Einzelzeichenwörtern in einer Zeile
`discard_tags`	*Flag*	Zeilen mit bestimmten Tags verwerfen.
`Tags`	*Zeichenfolge*	Sonderzeichen müssen durch ein Backslash-Zeichen (\) als Escapezeichen entwertet werden.
`discard_spec_words`	*Flag*	Zeilen mit bestimmten Zeichenfolgen verwerfen.
`words`	*Zeichenfolge*	Sonderzeichen müssen durch ein Backslash-Zeichen (\) als Escapezeichen entwertet werden.

Textminingknoten: TextMiningWorkbench

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

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

Tabelle 9. Textmining-Modellierungsknoten - Scripteigenschaften

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
`text`	*Feld*	
`method`	`ReadText` `ReadPath`	
`docType`	*Ganzzahl*	Zulässige Werte (0,1,2), wobei 0 = `Volltext`, 1 = `Gegliederter Text` und 2 = `XML`

Tabelle 9. Textmining-Modellierungsknoten - Scripteigenschaften (Forts.)

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
encoding	Automatic "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	*Ganzzahl*	Zulässige Werte (0,1), wobei 0 = Absatz und 1 = Dokument
para_min	*Ganzzahl*	
para_max	*Ganzzahl*	
mtag	*Zeichenfolge*	Enthält sämtliche mtag-Einstellungen (aus dem Dialogfeld "Einstellungen" für XML-Dateien)
mclef	*Zeichenfolge*	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	*Zeichenfolge*	
use_partitioned_data	*Flag*	Wenn ein Partitionsfeld definiert ist, werden nur die Trainingsdaten für die Modellerstellung verwendet.
model_output_type	Interactive Model	Interactive ergibt ein Kategoriemodell. Model ergibt ein Konzeptmodell.
use_interactive_info	*Flag*	Nur zum interaktiven Erstellen in einer Workbenchsitzung.
reuse_extraction_results	*Flag*	Nur zum interaktiven Erstellen in einer Workbenchsitzung.
interactive_view	Categories TLA Clusters	Nur zum interaktiven Erstellen in einer Workbenchsitzung.
extract_top	*Ganzzahl*	Dieser Parameter wird verwendet bei model_type = Concept
use_check_top	*Flag*	
check_top	*Ganzzahl*	
use_uncheck_top	*Flag*	
uncheck_top	*Ganzzahl*	

Tabelle 9. Textmining-Modellierungsknoten - Scripteigenschaften (Forts.)

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung

`language`	`de`
	`en`
	`es`
	`fr`
	`it`
	`ja`
	`nl`
	`pt`	

`frequency_limit`	*Ganzzahl*	Ab Version 14.0 nicht mehr verwendet.
`concept_count_limit`	*Ganzzahl*	Extraktion beschränken 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`	*Ganzzahl*	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`	*Ganzzahl*	Maximale Füllwörter in zusammengesetzten Konzepten (Standard ist 3). Für japanischen Text nicht verfügbar.

`jp_algorithmset Nur Schlussfolgerungen Nur repräsentative Stimmung Alle Stimmungen`	`0`
	`1`
	`2`
		Nur für die Extraktion von japanischem Text.
		
		Hinweis: Verfügbar in IBM SPSS Modeler Premium.
		`0` = Sekundäre Stimmungsextraktion
		`1` = Abhängigkeitsextraktion
		`2` = Kein Sekundäranalysator festgelegt.

`jp_algorithm_sense_mode`	`0`
	`1`
	`2`
		Nur für die Extraktion von japanischem Text.
		
		Hinweis: Verfügbar in IBM SPSS Modeler Premium.
		`0` = Nur Schlussfolgerungen
		`2` = Nur repräsentative Stimmung
		`3` = Alle Stimmungen.

Textmining-Modellnugget: TMWBModelApplier

Sie können die Eigenschaften in der folgenden Tabelle für Scripts verwenden. Der Knoten als solcher heißt `TMWBModelApplier`.

Tabelle 10. Textmining-Modellnugget - Eigenschaften

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
`scoring_mode`	`Fields` `Records`	
`field_values`	`Flags` `Counts`	Diese Option steht im Kategoriemodellnugget nicht zur Verfügung. Setzen Sie `Flags` auf `TRUE` oder `FALSE`.

Tabelle 10. Textmining-Modellnugget - Eigenschaften (Forts.)

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
true_value	*Zeichenfolge*	Definieren Sie den Wert der Flags als "True".
false_value	*Zeichenfolge*	Definieren Sie den Wert der Flags als "False".
extension_concept	*Zeichenfolge*	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	*Zeichenfolge*	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 "Zugehörige Deskriptoren vollständig aus Scoring ausschließen" bewirkt, 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 aggregieren" bewirkt, 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 nicht mehr verwendet
text	*Feld*	
method	ReadText ReadPath	
docType	*Ganzzahl*	Zulässige Werte (0,1,2), wobei 0 = Volltext, 1 = Gegliederter Text und 2 = XML

Tabelle 10. Textmining-Modellnugget - Eigenschaften (Forts.)

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
encoding	Automatic "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 Scripts zu definieren oder zu aktualisieren. Der Knoten als solcher heißt textlinkanalysis.

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

Tabelle 11. Scripteigenschaften von Textlinkanalyseknoten (TLA-Knoten)

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
id_field	*Feld*	
text	*Feld*	
method	ReadText ReadPath	
docType	*Ganzzahl*	Zulässige Werte (0,1,2), wobei 0 = Volltext, 1 = Gegliederter Text und 2 = XML
encoding	Automatic "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	*Ganzzahl*	Zulässige Werte (0,1), wobei 0 = Absatz und 1 = Dokument
para_min	*Ganzzahl*	
para_max	*Ganzzahl*	
mtag	*Zeichenfolge*	Enthält sämtliche mtag-Einstellungen (aus dem Dialogfeld "Einstellungen" für XML-Dateien)

Tabelle 11. Scripteigenschaften von Textlinkanalyseknoten (TLA-Knoten) (Forts.)

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
mclef	*Zeichenfolge*	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	*Ganzzahl*	Extraktion beschränken 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	*Ganzzahl*	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	*Ganzzahl*	Maximale Füllwörter in zusammengesetzten Konzepten (Standard ist 3). Für japanischen Text nicht verfügbar.
jp_algorithmset Nur Schlussfolgerungen Nur repräsentative Stimmung Alle Stimmungen	0 1 2	Nur für die Extraktion von japanischem Text. *Hinweis*: Verfügbar in IBM SPSS Modeler Premium. 0 = Sekundäre Stimmungsextraktion 1 = Abhängigkeitsextraktion 2 = Kein Sekundäranalysator festgelegt.
jp_algorithm_sense_mode	0 1 2	Nur für die Extraktion von japanischem Text. *Hinweis*: Verfügbar in IBM SPSS Modeler Premium. 0 = Nur Schlussfolgerungen 2 = Nur repräsentative Stimmung 3 = Alle Stimmungen.

Übersetzungsknoten: translatenode

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

Tabelle 12. Übersetzungsknoten-Eigenschaften

Scripteigenschaften	Datentyp	Eigenschaftsbeschreibung
text	*Feld*	
method	ReadText ReadPath	

Tabelle 12. Übersetzungsknoten-Eigenschaften (Forts.)

Scripteigenschaften	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	*Zeichenfolge*	
lw_port	*Ganzzahl*	
url	*Zeichenfolge*	URL des Übersetzungsservers
apiKey	*Zeichenfolge*	
user_id	*Zeichenfolge*	
lpid	*Ganzzahl*	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	

Tabelle 12. Übersetzungsknoten-Eigenschaften (Forts.)

Scripteigenschaften	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	*Ganzzahl*	Dient zur Angabe des gewünschten Genauigkeitsgrads für den Übersetzungsprozess. Wählen Sie einen Wert von 1 bis 3 aus.
use_previous_translation	*Flag*	Gibt an, dass bereits Übersetzungsergebnisse aus einer früheren Ausführung vorliegen, die wiederverwendet werden können.
translation_label	*Zeichenfolge*	Geben Sie eine Beschriftung ein, um die Übersetzungsergebnisse zu kennzeichnen, die wiederverwendet werden können.

Kapitel 8. Modus "Interaktive Workbench"

Sie können über einen Textmining-Modellierungsknoten eine interaktive Workbenchsitzung während der Streamausführung starten. In dieser Workbench können Sie die wichtigsten Konzepte aus Ihren Textdaten extrahieren, Kategorien aufbauen und Textlinkanalysemuster und -cluster untersuchen und Kategoriemodelle generieren. In diesem Kapitel finden Sie einen allgemeinen Überblick über die Workbenchschnittstelle sowie die wichtigsten Elemente, mit denen Sie in einer Workbenchsitzung arbeiten:

- **Extraktionsergebnisse.** Nach der Durchführung einer Extraktion sind dies die Schlüsselwörter und -wortfolgen, 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 Extraktionsergebnissen, 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 Ihre linguistischen Ressourcen Musterregeln für die Textlinkanalyse (TLA) enthalten oder Sie 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 einen Sekundäranalysator wählen und TLA-Extraktion aktivieren. *Hinweis*: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.
- **Linguistische Ressourcen.** Der Extraktionsvorgang 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.

Ansicht "Kategorien und Konzepte"

Die Anwendungsschnittstelle besteht aus mehreren Ansichten. Die Ansicht "Kategorien und Konzepte" ist das Fenster, in dem Sie Kategorien erstellen und untersuchen sowie die Extraktionsergebnisse 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 Extraktionsergebnisse, die als Bausteine (sogenannte Deskriptoren) für Ihre Kategorien verwendet werden können.

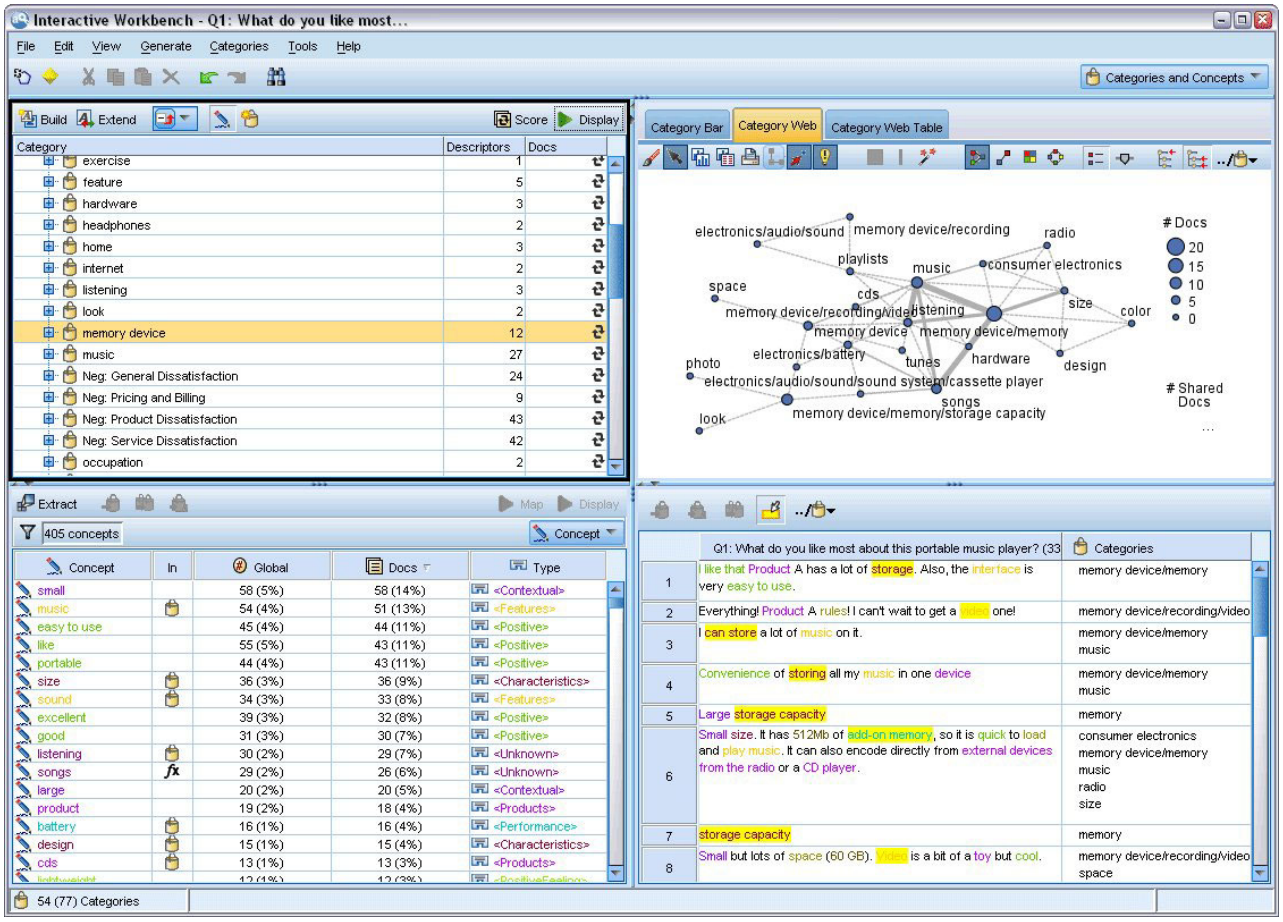


Abbildung 23. Kategorie- und Konzeptansicht

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. Weitere Informationen finden Sie im Thema Kapitel 10, „Kategorisieren von Textdaten”, auf Seite 107.

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. Weitere Informationen finden Sie im Thema Kapitel 10, „Kategorisieren von Textdaten”, auf Seite 107. Nicht alle automatischen Verfahren stehen für alle Sprachen zur Verfügung.

Wenn Sie eine Zeile im Fensterbereich auswählen, können Sie Informationen zu den entsprechenden Dokumenten/Datensätzen bzw. Deskriptoren im Datenbereich und im Visualisierungsbereich anzeigen.

Bereich "Extraktionsergebnisse"

Dieser Bereich befindet sich links unten und zeigt die Ergebnisse der Extraktion an. Wenn Sie eine Extraktion ausführen, liest die Extraktionsengine die Textdaten, identifiziert die relevanten Konzepte und weist jedem davon einen Typ zu. **Konzepte** sind Wörter bzw. Wortfolgen, die aus Ihren Textdaten extrahiert wurden.

Typen sind semantische Gruppierungen von Konzepten, die in Form von Typwörterbüchern gespeichert sind. Wenn die Extraktion abgeschlossen ist, werden die Konzepte und Typen im Bereich "Extraktionsergebnisse" mit Farbcodierung angezeigt. Weitere Informationen finden Sie im Thema „Extraktionsergebnisse: Konzepte und Typen” auf Seite 93.

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

Textmining ist ein iterativer Prozess, in dem Extraktionsergebnisse dem Kontext der Textdaten gemäß überprüft, für die Gewinnung neuer Ergebnisse optimiert und dann neu bewertet werden. Die Extraktionsergebnisse können durch Bearbeiten der linguistischen Ressourcen optimiert werden. Diese Optimierung kann teilweise direkt im Bereich "Extraktionsergebnisse" oder im Datenbereich vorgenommen werden, aber auch direkt in der Ansicht "Ressourceneditor". Weitere Informationen finden Sie im Thema „Ressourceneditoransicht” auf Seite 86.

Visualisierungsbereich

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 Optimierung 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. Weitere Informationen finden Sie im Thema „Kategoriendiagramme und Grafiken” auf Seite 169.

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 Datenbereich 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 farbcodierte Elemente bewegen, um eine QuickInfo mit dem Namen des Konzepts, unter dem das betreffende Element extrahiert wurde, und dem Typ, dem es zugewiesen wurde, anzuzeigen. Weitere Informationen finden Sie im Thema „Datenbereich” auf Seite 116.

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.

So verwenden Sie die Suchfunktion

1. Wählen Sie in den Menüs in der Ansicht "Kategorien und Konzepte" die Optionsfolge **Bearbeiten > Suchen** aus. Die Symbolleiste "Suchen" wird über dem Kategoriebereich und den Visualisierungsbereichen angezeigt.
2. Geben Sie die Wortfolge, nach der Sie suchen möchten, in das Textfeld ein. Mit den Schaltflächen in 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.
3. 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.
4. Um nach der nächsten Übereinstimmung zu suchen, klicken Sie erneut auf den Namen des Fensterbereichs.

Clusteransicht

In der Clusteransicht können Sie die in Ihren Textdaten gefundenen Clusterergebnisse erstellen und untersuchen. **Cluster** sind Gruppierungen von Konzepten, die durch Clusteralgorithmen generiert werden. Als Grundlage für die Generierung 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 Terme) in demselben Dokument bzw. Datensatz vorkommen. Weitere Informationen finden Sie im Thema Kapitel 11, „Analyse von Clustern”, auf Seite 155.

Sie können Cluster erstellen und mithilfe einer Reihe von Diagrammen und Grafiken untersuchen, um Beziehungen zwischen Konzepten aufzudecken, deren Ermittlung ansonsten zu zeitaufwendig 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. Weitere Informationen finden Sie im Thema „Clusterdefinitionen” auf Seite 160.

Sie können Änderungen an den Einstellungen für das Clustering vornehmen, um die Ergebnisse zu beeinflussen. Weitere Informationen finden Sie im Thema „Erstellen von Clustern” auf Seite 156.

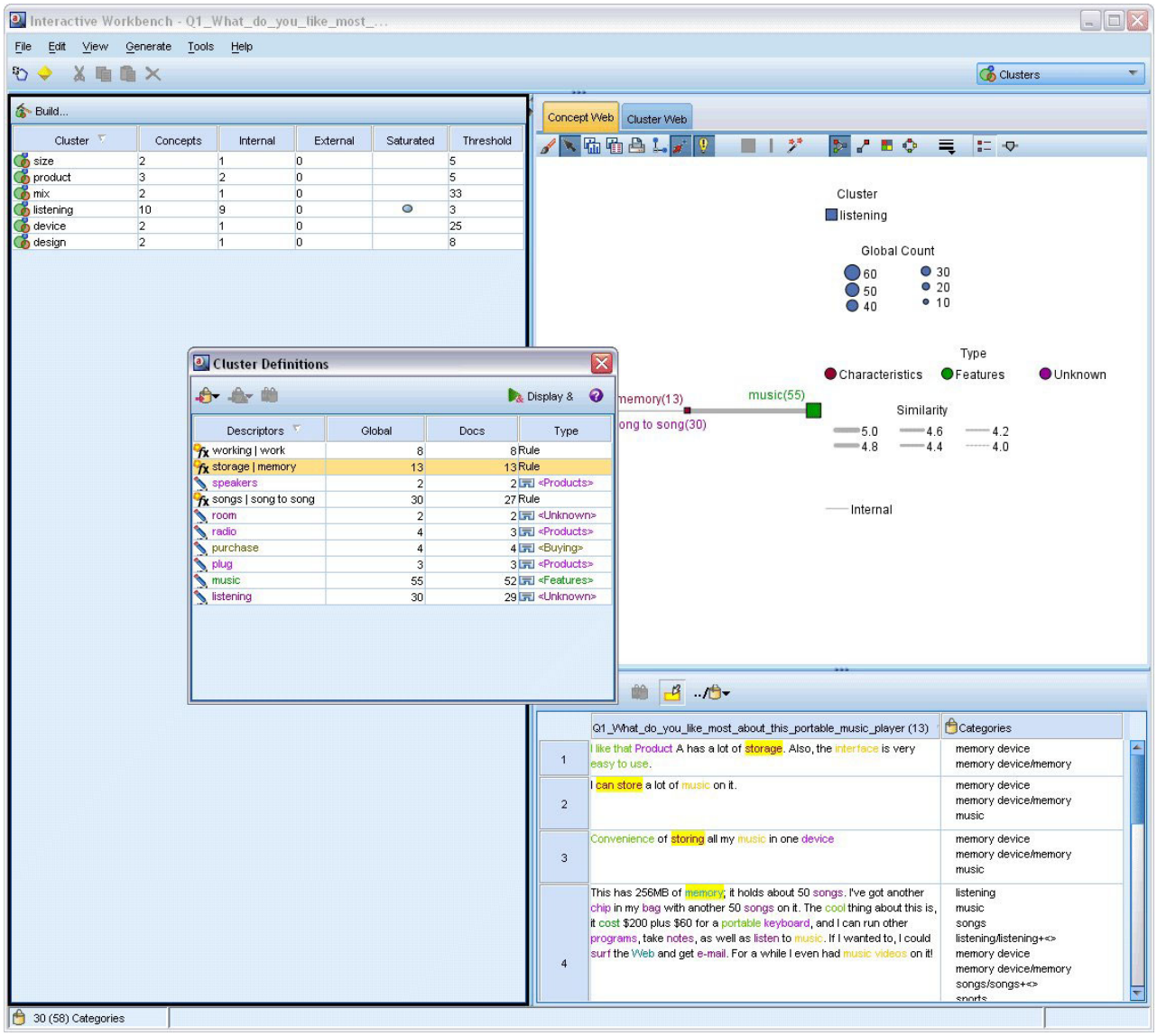


Abbildung 24. Clusteransicht

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 **Erstellen** können Sie Clustering-Ergebnisse erstellen. Cluster werden durch einen Clusteralgorithmus gebildet, der versucht, Konzepte zu identifizieren, die häufig gemeinsam auftreten.

Bei jeder Ausführung der Extraktion 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. Weitere Informationen finden Sie im Thema „Untersuchen von Clustern” auf Seite 159.

Visualisierungsbereich

Dieser Fensterbereich befindet sich rechts oben und bietet zwei Clustering-Perspektiven: 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 sowie die verknüpften Konzepte außerhalb des Clusters anzeigt.
- **Clusternetzdiagramm.** Netzdiagramm, das die Verknüpfungen von den ausgewählten Clustern zu anderen Clustern sowie Zusammenhänge zwischen diesen anderen Clustern anzeigt.

Hinweis: 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). Weitere Informationen finden Sie im Thema „Clusterdiagramme” auf Seite 171.

Datenbereich

Der Datenbereich befindet sich rechts unten und ist standardmäßig ausgeblendet. Über den Fensterbereich "Cluster" können keine Ergebnisse aus dem Datenbereich 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 Datenbereich nur der zugehörige Text angezeigt. Sobald Sie eine Auswahl getroffen haben, können Sie durch Klicken auf die Schaltfläche **Anzeige &** den Datenbereich durch die Dokumente bzw. Datensätze füllen, 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 farbcodierte Elemente bewegen, um das Konzept anzuzeigen, unter dem das betreffende Element extrahiert wurde, und den Typ, dem es zugewiesen wurde. Der Datenbereich kann mehrere Spalten enthalten, doch die Spalte "Textfeld" wird immer angezeigt. Sie trägt den Namen des während der Extraktion verwendeten Textfelds bzw. den Namen eines Dokuments, wenn sich die Textdaten in vielen verschiedenen Dateien befinden. Weitere Spalten sind verfügbar. Weitere Informationen finden Sie im Thema „Datenbereich” auf Seite 116.

Textlinkanalyseansicht

In der Ansicht "Textlinkanalyse" 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 Daten- bzw. Visualisierungsbereich untersuchen und sogar 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. Weitere Informationen finden Sie im Thema Kapitel 19, „Textlinkregeln”, auf Seite 227.

Wenn Sie TLA-Musterergebnisse extrahieren lassen, werden die Ergebnisse in dieser Ansicht gezeigt. Wenn Sie dies nicht durch entsprechende Auswahl festgelegt haben, müssen Sie die Schaltfläche **Extrahieren** verwenden und die Option zur Aktivierung der Extraktion von Mustern auswählen.

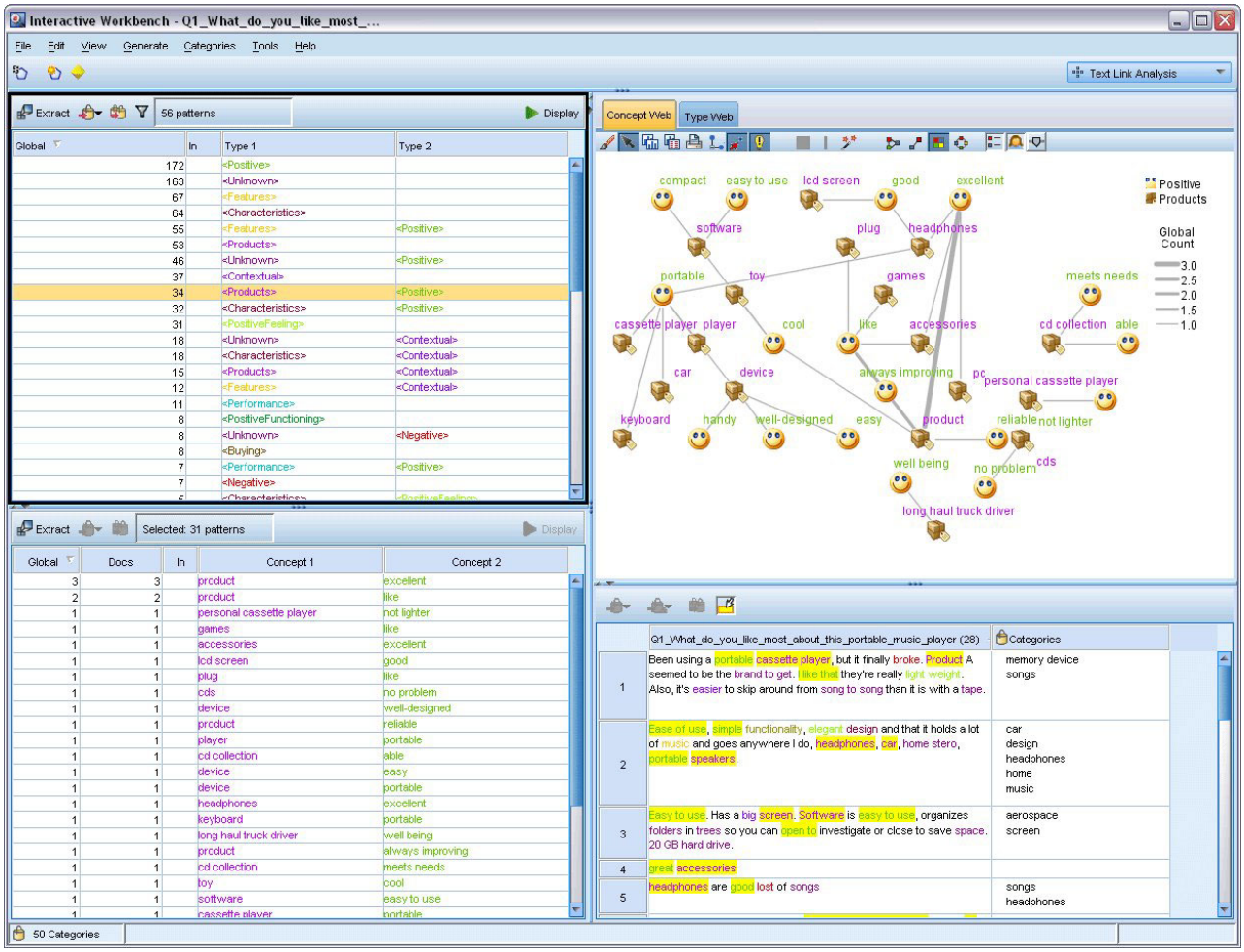


Abbildung 25. Ansicht "Textlinkanalyse"

Die Ansicht "Textlinkanalyse" gliedert sich in vier Fensterbereiche, die jeweils aus- bzw. eingeblendet werden können, indem Sie ihren Namen im Menü "Ansicht" auswählen. Weitere Informationen finden Sie im Thema Kapitel 12, „Untersuchen von Textlinkanalyse”, auf Seite 163.

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 lediglich bis zu einem oder zwei Typen oder Konzepten sind. Die in den linguistischen Ressourcen definierte TLA-Musterregel legt die Komplexität der Musterergebnisse fest. Weitere Informationen finden Sie im Thema Kapitel 19, „Textlinkregeln”, auf Seite 227. *Hinweis*: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

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

- **Typmuster.** Im Fensterbereich "Typmuster" finden Sie extrahierte Muster, die aus mindestens zwei verwandten Typen bestehen, die einer TLA-Musterregel entsprechen. Typmuster werden als `<Organization> + <Location> + <Positive>` angezeigt, 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 derzeit im oberhalb gelegenen Fensterbereich "Typmuster" ausgewählten Typmuster. Konzeptmuster weisen folgende Struktur auf: `Hotel + Paris + herrlich`.

Wie bei den Extraktionsergebnissen in der Ansicht "Kategorien und Konzepte" können Sie hier die Ergebnisse überprüfen. Wenn Sie die Typen und Konzepte, aus denen diese Muster bestehen, weiter optimieren möchten, können Sie diese Optimierung im Bereich "Extraktionsergebnisse" in der Ansicht "Kategorien und Konzepte" oder direkt im Ressourceneditor vornehmen und Ihre Muster erneut extrahieren.

Visualisierungsbereich

Dieser Fensterbereich befindet sich rechts oben in der Ansicht "Textlinkanalyse" und enthält ein Netzdiagramm der ausgewählten 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.** 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 Vorkommen in der ausgewählten Tabelle an.
- **Typdiagramm.** 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 Vorkommen in der ausgewählten Tabelle an. Knoten werden entweder durch eine Typfarbe oder durch ein Symbol dargestellt.

Weitere Informationen finden Sie im Thema „Textlinkanalyse-Diagramme” auf Seite 172.

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 Datenbereich 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 farbcodierte Elemente bewegen, um eine QuickInfo mit dem Namen des Konzepts, unter dem das betreffende Element extrahiert wurde, und dem Typ, dem es zugewiesen wurde, anzuzeigen. Weitere Informationen finden Sie im Thema „Datenbereich” auf Seite 116.

Ressourceneditoransicht

IBM SPSS Modeler Text Analytics erfasst mithilfe einer robusten Extraktionsengine 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 "Ressourceneditor" können Sie die zum Extrahieren von Konzepten verwendeten linguistischen Ressourcen anzeigen und optimieren, sie zu Typen zusammenfassen, Muster in den Textdaten er-

kennen und vieles mehr. IBM SPSS Modeler Text Analytics bietet einige vorkonfigurierte Ressourcenvorlagen. In einigen Sprachen können Sie auch die Ressourcen in einem Text Analysis Package verwenden. Weitere Informationen finden Sie im Thema „Verwendung von Text Analysis Packages” auf Seite 148.

Da diese Ressourcen möglicherweise nicht immer perfekt an den Kontext Ihrer Daten angepasst sind, können Sie im Ressourceneditor Ihre eigenen Ressourcen für einen bestimmten Kontext oder eine bestimmte Domäne erstellen, bearbeiten und verwalten. Weitere Informationen finden Sie im Thema Kapitel 16, „Arbeiten mit Bibliotheken”, auf Seite 191.

Um die Optimierung Ihrer linguistischen Ressourcen zu vereinfachen, können Sie über Kontextmenüs im Bereich "Extraktionsergebnisse" und im Datenbereich häufig verwendete Wörterbuchaufgaben direkt aus der Ansicht "Kategorien und Konzepte" heraus durchführen. Weitere Informationen finden Sie im Thema „Optimieren von Extraktionsergebnissen” auf Seite 101.

Hinweis: Die Schnittstelle für Ressourcen, die auf japanischen Text eingestellt sind, unterscheidet sich geringfügig. Extraktion für japanischen Text steht in IBM SPSS Modeler Premium zur Verfügung.

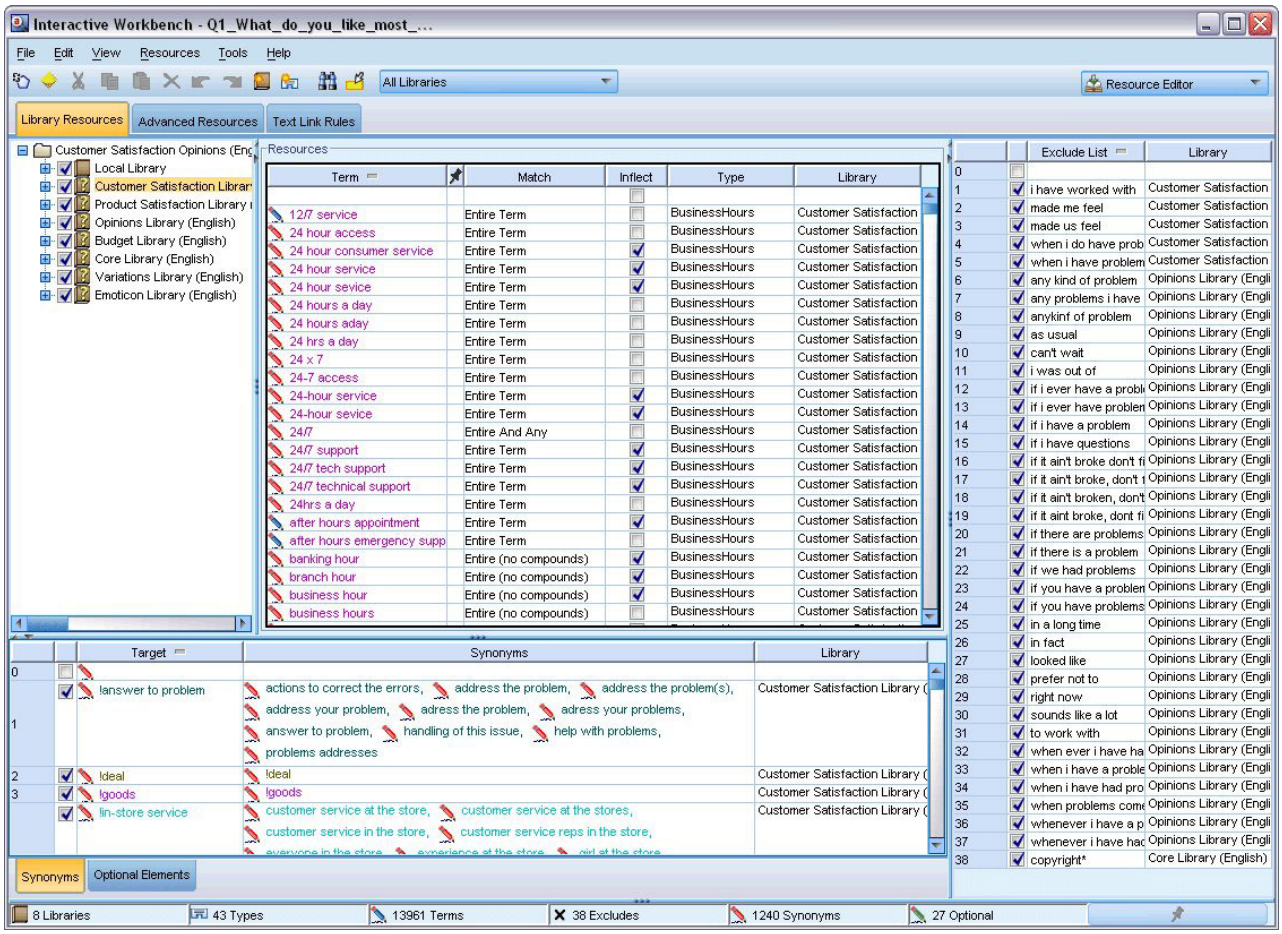


Abbildung 26. Ansicht "Ressourceneditor"

Bei den in der Ansicht "Ressourceneditor" durchgeführten Operationen geht es um die Verwaltung und Optimierung der linguistischen Ressourcen. Diese Ressourcen werden in Form von Vorlagen und Bibliotheken gespeichert. Die Ansicht "Ressourceneditor" gliedert sich in vier Bereiche: den Fensterbereich "Bibliotheksbaum", den Fensterbereich "Typwörterbuch", den Fensterbereich "Substitutionswörterbuch" und den Fensterbereich "Ausschlusswörterbuch".

Hinweis: Weitere Informationen finden Sie im Thema „Editorschnittstelle” auf Seite 181.

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 Benutzerschnittstelle verwendeten Farben.
- **Klänge.** Diese Registerkarte enthält Optionen für Tonsignale.

So bearbeiten Sie Optionen

1. Wählen Sie in den Menüs die Optionsfolge **Tools > Optionen** aus. Das Dialogfeld "Optionen" wird geöffnet.
2. Wählen Sie die Registerkarte mit den zu ändernden Informationen aus.
3. Ändern Sie die Optionen nach Bedarf.
4. Klicken Sie auf **OK**, um die Änderungen zu speichern.

Optionen: Registerkarte "Sitzung"

Auf dieser Registerkarte können Sie einige Grundeinstellungen festlegen

Datenbereich und Kategoriediagrammanzeige. Diese Optionen beeinflussen, wie Daten im Datenbereich und im Visualisierungsbereich in der Ansicht "Kategorien und Konzepte" dargestellt werden.

- **Anzeigegrenzwert für Datenbereich und Kategorienetz.** 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 zur Anzeigezeit anzeigen.** Wenn diese Option ausgewählt ist, werden Dokumente oder Datensätze gescort, wenn Sie auf die Schaltfläche "Anzeigen" klicken, sodass Kategorien, zu denen sie gehören, im Datenbereich in der Spalte "Kategorien" und in den Kategoriediagrammen angezeigt werden können. In einigen Fällen, insbesondere bei größeren Datasets, kann es sinnvoll sein, diese Option zu inaktivieren, 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 Kategorien- und Konzeptansicht, kopieren.** Beim Hinzufügen eines Dokuments oder Datensatzes aus dem Datenbereich in dieser Ansicht werden entweder **Nur Konzepte** oder **Konzepte und Muster** kopiert.
- **In Textlinkanalyseansicht, kopieren.** Beim Hinzufügen eines Dokuments oder Datensatzes aus dem Datenbereich in dieser Ansicht werden entweder **Nur Muster** oder **Konzepte und Muster** kopiert.

Ressourceneditor-Begrenzer. 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 "Anzeigen"

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

Hinweis: Um zu einem klassischen Erscheinungsbild oder einem aus einem früheren Release des Produkts zu wechseln, öffnen Sie das Dialogfeld "Benutzeroptionen" im Menü "Tools" im Hauptfenster von IBM SPSS Modeler.

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 Datenbereich sichtbar sind.
- **Hervorhebungshintergrund.** Hintergrundfarbe für die Textauswahl (bei der Auswahl von Elementen in den Fensterbereichen oder von Text im Datenbereich).
- **Extraktionsbedingter Hintergrund.** Hintergrundfarbe der Bereiche "Extraktionsergebnisse", "Muster" und "Cluster", die anzeigt, dass Änderungen an den Bibliotheken vorgenommen wurden und eine Extraktion erforderlich ist.
- **Hintergrund für Kategoriefeedback.** Hintergrundfarbe für Kategorien, die nach einer Operation angezeigt werden.
- **Standardtyp.** Standardfarbe für Typen und Konzepte, die im Datenbereich und im Bereich "Extraktionsergebnisse" angezeigt werden. Diese Farbe gilt für alle benutzerdefinierten Typen, die Sie im Ressourceneditor erstellen. Sie können diese Standardfarbe für Ihre benutzerdefinierten Typwörterbücher außer Kraft setzen, indem Sie die Eigenschaften für diese Typwörterbücher im Ressourceneditor bearbeiten. Weitere Informationen finden Sie im Thema „Erstellen von Typen” auf Seite 203.
- **Gestreifte Tabelle 1.** Die erste der beiden Farben, die sich in der Tabelle im Dialogfeld "Erzwungene Terme 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 Terme bearbeiten" abwechseln, damit die einzelnen Zeilen besser auseinandergehalten werden können.

Hinweis: Wenn Sie auf die Schaltfläche **Auf Standardwerte 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.

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

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.

1. Wählen Sie in den Menüs des Internet Explorer folgende Optionen aus:

   Extras > **Internetoptionen...**

2. Klicken Sie auf die Registerkarte **Erweitert**.
3. Führen Sie einen Bildlauf nach unten zum Abschnitt **Sicherheit** durch.
4. Aktivieren Sie **Ausführung aktiver Inhalte in Dateien auf dem lokalen Computer zulassen**.

Generieren von Modellnuggets und Modellierungsknoten

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

- **Textmining-Modellierungsknoten.** Bei einem aus einer interaktiven Workbenchsitzung generierten Modellierungsknoten handelt es sich um einen Textminingknoten, 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 Textminingknoten verfügen oder eine neue Version erstellen möchten. Weitere Informationen finden Sie im Thema Kapitel 3, „Mining nach Konzepten und Kategorien", auf Seite 19.
- **Kategoriemodellnugget.** Bei einem aus einer interaktiven Workbenchsitzung generierten Modellnugget handelt es sich um ein Kategoriemodellnugget. In der Ansicht "Kategorien und Konzepte" muss mindestens eine Kategorie vorliegen, um ein Kategoriemodellnugget generieren zu können. Weitere Informationen finden Sie im Thema „Textminingnugget: Kategoriemodell" auf Seite 42.

So generieren Sie einen Textmining-Modellierungsknoten

1. Wählen Sie in den Menüs die Optionsfolge **Generieren > Modellierungsknoten generieren** aus. Ein Textmining-Modellierungsknoten wird unter Verwendung aller derzeit in der Workbenchsitzung gültigen Einstellungen zum Arbeitserstellungsbereich hinzugefügt. Der Knoten wird nach dem Textfeld benannt.

So generieren Sie ein Kategoriemodellnugget:

1. Wählen Sie in den Menüs die Optionsfolge **Generieren > Modell generieren** aus. Ein Modellnugget wird mit dem Standardnamen direkt in 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 Workbenchsitzung abgeschlossen haben und Ihre Arbeit speichern möchten. Wenn Sie den Modellierungsknoten aktualisieren, wird der Inhalt der Workbenchsitzung wieder in dem Textminingknoten gespeichert, von dem die interaktive Workbenchsitzung 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

1. Wählen Sie in den Menüs die Optionsfolge **Datei > Modellierungsknoten aktualisieren** aus. Der Modellierungsknoten wird mit den Erstellungs- und Extraktionseinstellungen 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 Ausgangsmodellierungsknoten speichern sowie Bibliotheken zur Wiederverwendung in anderen Sitzungen veröffentlichen. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von

Bibliotheken" auf Seite 196. Nach dem Speichern wird das Sitzungsfenster geschlossen und die Sitzung wird aus Output 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 Output Manager im IBM 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 in Output Manager im IBM SPSS Modeler-Fenster auswählen.

So schließen Sie eine Workbenchsitzung

1. Wählen Sie in den Menüs die Optionsfolge **Datei > Schließen** aus.

Tastatureingabehilfen

Die interaktive Workbenchschnittstelle enthält Tastenkombinationen, 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 Tastenkombinationen für die alternative Navigation behandelt. Es gibt andere Tastenkombinationen für die IBM SPSS Modeler-Schnittstelle.

Tabelle 13. Allgemeine Tastenkombinationen

Tastenkombination	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 Extraktion in den Ansichten "Kategorien und Konzepte" sowie "Textlinkanalyse".
Strg+F	Zeigt die Suchsymbolleiste im Ressourceneditor/Vorlageneditor 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 den ausgewählten Cluster.
Strg+R	Öffnet das Dialogfeld "Terme hinzufügen" im Ressourceneditor/Vorlageneditor.
Strg+T	Öffnet das Dialogfeld "Typeigenschaften" zum Erstellen eines neuen Typs im Ressourceneditor/Vorlageneditor.
Strg+V	Fügt den Inhalt der Zwischenablage ein.
Strg+X	Schneidet die ausgewählten Elemente aus dem Ressourceneditor/Vorlageneditor 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.
F8	Verschiebt den Fokus auf die Fensterteiler (zum Ändern der Größe).
F10	Erweitert das Haupt-Dateimenü.
Aufwärtspfeil, Abwärtspfeil	Dient zur vertikalen Größenänderung, wenn der Teilungsbalken ausgewählt ist.

Tabelle 13. Allgemeine Tastenkombinationen (Forts.)

Tastenkombination	Funktion
Linkspfeil, Rechtspfeil	Dient zur horizontalen Größenänderung, wenn der Teilungsbalken ausgewählt ist.
Pos1, Ende	Maximiert bzw. minimiert die Fenstergröße, wenn der Teilungsbalken ausgewählt ist.
Tabulator	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.

Tastenkombinationen für Dialogfelder

Einige der Tastenkombinationen und das Sprachausgabeprogramm 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 Sprachausgabeprogramm zu starten. In der folgenden Tabelle finden Sie eine vollständige Liste mit speziellen Tastenkombinationen und das Sprachausgabeprogramm.

Tabelle 14. Tastenkombinationen für Dialogfelder

Tastenkombination	Funktion
Tabulator	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. Extrahieren von Konzepten und Typen

Wenn Sie einen Stream ausführen, der die interaktive Workbench startet, erfolgt automatisch eine Extraktion der Textdaten in dem Stream. Das Endergebnis dieser Extraktion ist eine Reihe von Konzepten, Typen und, falls TLA-Muster in den linguistischen Ressourcen vorhanden sind, von Mustern. Im Bereich "Extraktionsergebnisse" können Sie Konzepte und Typen anzeigen und mit ihnen arbeiten. Weitere Informationen finden Sie im Thema „Funktionsweise der Extraktion” auf Seite 5.

Um die Extraktionsergebnisse zu optimieren, können Sie die linguistischen Ressourcen verändern und eine erneute Extraktion durchführen. Weitere Informationen finden Sie im Thema „Optimieren von Extraktionsergebnissen” auf Seite 101. Die Ressourcen und Parameter im Dialogfeld "Extrahieren" bestimmen, wie die Ergebnisse extrahiert und geordnet werden. Mit den Extraktionsergebnissen können Sie den Großteil, wenn nicht sogar alle, Ihrer Kategoriedefinitionen festlegen.

Extraktionsergebnisse: Konzepte und Typen

Während des Extraktionsprozesses werden sämtliche Textdaten untersucht und die relevanten Konzepte ermittelt, extrahiert und entsprechenden Typen zugewiesen. Nach Abschluss der Extraktion werden die Ergebnisse in dem Bereich "Extraktionsergebnisse" 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 **Extraktionsergebnisse** 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 Erstellen der Kategorien.

Textmining ist ein iterativer Prozess, bei dem Extraktionsergebnisse dem Kontext der Textdaten gemäß überprüft, für die Erzeugung neuer Ergebnisse optimiert und dann neu ausgewertet 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 "Extraktionsergebnisse"vom Datenbereich, vom Dialogfeld "Kategoriedefinitionen" oder vom Dialogfeld "Clusterdefinitionen" aus optimieren. Weitere Informationen finden Sie im Thema „Optimieren von Extraktionsergebnissen” auf Seite 101. Außerdem können Sie Optimierungen direkt in der Ansicht "Ressourceneditor" vornehmen. Weitere Informationen finden Sie im Thema „Ressourceneditoransicht” auf Seite 86.

Nach der Optimierung können Sie erneut extrahieren, um die neuen Ergebnisse anzuzeigen. Wenn Sie Ihre Extraktionsergebnisse von Anfang an optimieren, gehen Sie sicher, dass Sie bei jeder neuen Extraktion 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 Wortfolgen (z. B. `Wahl zur Präsidentschaft`, `Wahl des Präsidenten`, oder `Verhandlungen um Frieden`) im Text zu ermitteln. Diese Wörter Wortfolgen werden auch als *Terme* bezeichnet. Die relevanten Terme werden mithilfe der linguistischen Ressourcen extrahiert und anschließend werden ähnliche Terme unter einem übergeordneten Term, einem **Konzept**, gruppiert.

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

Standardmäßig werden die Konzepte in Kleinbuchstaben angezeigt und in absteigender Reihenfolge entsprechend der Dokumentanzahl (Dokumentspalte) 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 Typeigenschaften im Ressourceneditor definiert. Weitere Informationen finden Sie im Thema „Typwörterbücher" auf Seite 201.

Bei Verwendung von Konzepten, Typen oder Mustern in einer Kategoriedefinition wird in der sortierbaren Spalte **In** ein Symbol angezeigt.

Typen

Typen sind semantische Gruppierungen von Konzepten. Beim Extrahieren wird den Konzepten ein Typ zugewiesen, um das Gruppieren ähnlicher Konzepte zu erleichtern. Im Lieferumfang von IBM SPSS Modeler Text Analytics sind mehrere integrierte Typen enthalten, z. B. `<Location>`, `<Organization>`, `<Person>`, `<Positive>`, `<Negative>` usw. Der Typ `<Ort>` gruppiert z. B. geografische Stichwörter und Orte. Dieser Typ würde Konzepten wie z. B. `Chicago`, `Paris` und `Tokio` zugewiesen. Für die meisten Sprachen gilt: Konzepte , die in keinem Typwörterbuch gefunden, aber aus dem Text extrahiert werden, erhalten automatisch den Typ `<Unknown>`Weitere Informationen finden Sie im Thema „Integrierte Typen" auf Seite 202.

Wenn Sie die Ansicht "Typ" auswählen, werden die extrahierten Typen standardmäßig in absteigender Reihenfolge nach globaler Häufigkeit angezeigt. Sie sehen außerdem, dass die Typen farblich gekennzeichnet sind, um ihre Unterscheidung zu erleichtern. Farben sind Teil der Typeigenschaften. Weitere Informationen finden Sie im Thema „Erstellen von Typen" auf Seite 203. Sie können auch eigene Typen erstellen.

Muster

Außerdem können Muster aus Ihren Textdaten extrahiert werden. Es muss jedoch eine Bibliothek vorhanden sein, die einige Musterregeln für die Textlinkanalyse (TLA) im Ressourceneditor enthält. Sie müssen darüber hinaus die Extraktion dieser Muster in der Knoteneinstellung für IBM SPSS Modeler Text Analytics oder im Dialogfeld "Extrahieren" über die Option **Musterextraktion für Textlinkanalyse aktivieren** auswählen. Weitere Informationen finden Sie im Thema Kapitel 12, „Untersuchen von Textlinkanalyse", auf Seite 163.

Extrahieren von Daten

Wenn eine Extraktion nötig ist, wird der Bereich "Extraktionsergebnisse" 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 Extraktionsergebnisse haben, linguistische Ressourcen geändert haben und die Extraktionsergebnisse aktualisieren müssen oder eine Sitzung neu geöffnet haben, in der Sie die Extraktionsergebnisse nicht gespeichert haben (**Tools > Optionen**).

Hinweis: Wenn Sie den Quellenknoten für Ihren Stream ändern, nachdem Extraktionsergebnisse mit der Option **Arbeit der Sitzung verwenden...** im Cache gespeichert wurden, müssen Sie eine neue Extraktion ausführen, sobald die interaktive Workbenchsitzung gestartet wird, damit Sie aktualisierte Extraktionsergebnisse erhalten.

Wenn Sie eine Extraktion durchführen, erscheint ein Fortschrittsanzeiger, der den Status der Extraktion anzeigt. Währenddessen liest die Extraktionsengine alle Textdaten, identifiziert die relevanten Terme und Muster, extrahiert sie und weist sie einem Typ zu. Dann versucht die Engine, synonyme Terme unter einem Leitausdruck, einem Konzept, zu gruppieren. Wenn der Vorgang abgeschlossen ist, werden die resultierenden Konzepte, Typen und Muster im Bereich "Extraktionsergebnisse" angezeigt.

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

Hinweis: Die für den Extraktionsprozess benötigte Zeit steht in direkter Beziehung zur Größe Ihres Datasets. Sie haben jederzeit die Möglichkeit, einen vorgeordneten Stichprobenknoten einzufügen oder die Konfiguration Ihres Computers zu optimieren.

Daten extrahieren

1. Wählen Sie in den Menüs die Optionsfolge **Tools > Extrahieren** aus. Alternativ können Sie auf die Symbolleistenschaltfläche **Extrahieren** klicken.
2. Wenn Sie das Dialogfeld "Extraktionseinstellungen" immer anzeigen lassen, wird es angezeigt, damit Sie Änderungen vornehmen können. Weitere Informationen zu Deskriptoren für jede Einstellung finden Sie in diesem Thema.
3. Klicken Sie auf **Extrahieren**, um die Extraktion zu starten. Sobald die Extraktion beginnt, öffnet sich die Statusanzeige. Nach der Extraktion werden die Ergebnisse im Bereich "Extraktionsergebnisse" dargestellt. Standardmäßig werden die Konzepte in Kleinbuchstaben angezeigt und in absteigender Reihenfolge entsprechend der Dokumentanzahl (Dokumentspalte) 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 Extraktionsergebnisse auch optimieren, indem Sie mit den linguistischen Ressourcen arbeiten. Weitere Informationen finden Sie im Thema „Optimieren von Extraktionsergebnissen” auf Seite 101.

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

Das Dialogfeld "Extraktionseinstellungen" enthält einige grundlegende Extraktionsoptionen.

Musterextraktion 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 Extraktionsdauer erheblich verlängern. Weitere Informationen finden Sie im Thema Kapitel 12, „Untersuchen von Textlinkanalyse”, auf Seite 163.

Interpunktionsfehler korrigieren. Diese Option normalisiert Text mit Interpunktionsfehlern (zum Beispiel ungeeignete Verwendung) während der Extraktion, 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 Fuzzy-Gruppierungsverfahren an, das hilft, häufig falsch geschriebene Wörter oder ähnlich geschriebene Wörter unter einem Konzept zu gruppieren. Der Algorithmus für Fuzzy-Gruppierung entfernt alle

Vokale (außer dem ersten) und doppelte/dreifache Konsonanten temporär aus extrahierten Wörtern und vergleicht sie, um festzustellen, ob sie gleich sind, sodass Modellierung und Modelierung zusammen gruppiert werden würden. Wenn jedoch jeder Term einem anderen Typ (ausschließlich des Typs <Unknown>) zugewiesen ist, wird das Fuzzy-Gruppierungsverfahren nicht angewendet.

Sie können auch die minimal erforderliche Zahl von *Stamm*zeichen definieren, bevor Fuzzy-Gruppierung eingesetzt wird. Die Anzahl der Stammzeichen in einem Term berechnet sich aus der Summe aller Zeichen abzüglich aller Zeichen, die Beugungsendungen und - bei zusammengesetzten Termen - Determinatoren und Präpositionen bilden. So würde beispielsweise der Term 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 sich herausstellt, dass bestimmte Wörter später falsch eingruppiert werden, können Sie einzelne Wortpaare aus dem Verfahren ausschließen, indem Sie sie auf der Registerkarte "Erweiterte Ressourcen" im Bereich **Fuzzy-Gruppierung: Ausnahmen** explizit deklarieren. Weitere Informationen finden Sie im Thema „Fuzzy-Gruppierung” auf Seite 218.

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

Nicht linguistische Entitäten extrahieren. Diese Option extrahiert nicht linguistische Entitäten wie beispielsweise Telefonnummern, Personalausweisnummern, Uhrzeiten, Datumsangaben, Währungen, Ziffern, Prozentsätze, E-Mail-Adressen und HTTP-Adressen. Sie können bestimmte Typen von nicht linguistischen Entitäten im Abschnitt **Nicht linguistische Entitäten: Konfiguration** der Registerkarte "Erweiterte Ressourcen" ein- bzw. ausschließen. Durch Inaktivierung unnötiger Entitäten vergeudet die Extraktionsengine keine Verarbeitungszeit. Weitere Informationen finden Sie im Thema „Konfiguration” auf Seite 222.

Großbuchstaben-Algorithmus. Diese Option extrahiert einfache und zusammengesetzte Terme, die sich nicht in den integrierten Wörterbüchern befinden, solange der erste Buchstabe des Terms 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 Texts 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 Terme abzugleichen, die dem Typ <Person> zugeordnet sind. Wird beispielsweise *doe* gefunden und anfänglich dem Typ <Unknown> zugeordnet, überprüft die Extraktionsengine, ob ein zusammengesetzter Term 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 Uniterme extrahiert werden.

Maximale Füllwörter in zusammengesetzten Konzepten. Diese Option gibt die maximale Anzahl von Füllwörtern an, die für die Anwendung des Permutationsverfahrens vorhanden sein müssen. Dieses Permutationsverfahren gruppiert ähnliche Wortfolgen, 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 als auch Vertreter des Unternehmens extrahiert wurden. In diesem Fall würden beide extrahierte Terme in der endgültigen Konzeptliste zusammen gruppiert, da beide Terme als gleich betrachtet werden, wenn des ignoriert wird.

Indexoption für Konzeptkarte Gibt an, dass Sie den Kartenindex zur Zeit der Extraktion erstellen möchten, damit die Konzeptkarten später schneller dargestellt werden können. Um die Indexeinstellungen zu bearbeiten, klicken Sie auf **Einstellungen**. Weitere Informationen finden Sie im Thema „Erstellen von Konzeptkartenindizes” auf Seite 101.

Dieses Dialogfeld vor dem Start einer Extraktion immer anzeigen. Legen Sie fest, ob Sie das Dialogfeld "Extraktionseinstellungen" bei jeder Extraktion anzeigen möchten, ob Sie es nie anzeigen möchten (außer beim Aufruf über das Menü "Tools") oder ob Sie bei jeder Extraktion gefragt werden möchten, ob Sie Änderungen an den Extraktionseinstellungen vornehmen wollen.

Für japanischen Text

Das Dialogfeld "Extraktionseinstellungen" enthält einige grundlegende Extraktionsoptionen für die japanische Textsprache. Standardmäßig sind die im Dialogfeld ausgewählten Einstellungen mit denen auf der Registerkarte "Experten" des Textmining-Modellierungsknotens identisch. Für die Arbeit mit japanischem Text müssen Sie den Text als Eingabe verwenden sowie eine japanische Sprachvorlage oder ein japanisches Text Analysis Package auf der Registerkarte "Modell" des Textminingknotens auswählen. Weitere Informationen finden Sie im Thema „Kopieren von Ressourcen aus TAPs und Vorlagen” auf Seite 27.

Hinweis: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

Sekundäre Analyse. Bei einer Extraktion werden grundlegende Schlüsselwörter unter Verwendung des Standardsets von Typen extrahiert. Wenn Sie jedoch einen Sekundäranalysator auswählen, können Sie auch mehrere oder reichhaltigere Konzepte erhalten, da der Extraktor nun Partikel und Hilfsverben als Teil des Konzepts beinhaltet. Im Fall der Stimmungsanalyse wird auch eine große Anzahl an zusätzlichen Typen eingeschlossen. Des Weiteren ermöglicht Ihnen die Wahl eines Sekundäranalysators, auch Ergebnisse für die Textlinkanalyse zu generieren.

Hinweis: Wenn ein Sekundäranalysator aufgerufen wird, dauert der Extraktionsprozess länger.

- **Abhängigkeitsanalyse**. Die Wahl dieser Option führt zu erweiterten Partikeln für die Extraktionskonzepte aus der grundlegenden Typ- und Stichwortextraktion. Sie können auch vielfältigere Musterergebnisse aus einer Abhängigkeitstextlinkanalyse (TLA) gewinnen.
- **Stimmungsanalyse**. Bei dieser Analyse werden zusätzliche Konzepte und - wann immer möglich - TLA-Musterergebnisse extrahiert. Neben den grundlegenden Typen können Sie zusätzlich mehr als 80 Stimmungstypen nutzen. 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**.
- **Kein Sekundäranalysator**. Diese Option schaltet sämtliche Sekundäranalysatoren aus. Diese Option ist nicht auswählbar, wenn die Option **Musterextraktion für Textlinkanalyse aktivieren** ausgewählt wurde, da ein Sekundäranalysator erforderlich ist, um TLA-Ergebnisse zu erhalten.

Musterextraktion 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 Extraktionsdauer erheblich verlängern. Zusätzlich muss ein Sekundäranalysator ausgewählt werden, um TLA-Musterergebnisse extrahieren zu können. Weitere Informationen finden Sie im Thema Kapitel 12, „Untersuchen von Textlinkanalyse”, auf Seite 163.

Filtern von Extraktionsergebnissen

Wenn Sie mit sehr großen Datasets arbeiten, kann der Extraktionsprozess Millionen von Ergebnissen liefern. Durch diese Menge ist eine effektive Überprüfung der Ergebnisse für viele Benutzer mühsam. Um sich daher auf die interessantesten zu konzentrieren, können Sie diese Ergebnisse über das Dialogfeld "Filter" im Bereich "Extraktionsergebnisse" filtern.

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

Nach Häufigkeit filtern. Mit diesem Filter werden nur Ergebnisse mit einem bestimmten globalen Häufigkeitswert 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 15. 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.
Exakte Übereinstimmung	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 "Extraktionsergebnisse" angezeigte Ergebnisse

Hier einige Beispiele, wie die Ergebnisse auf der Grundlage von Filtern in der Symbolleiste des Bereichs "Extraktionsergebnisse" in Englisch angezeigt werden könnten.

Tabelle 16. Beispiele für Filterfeedback

Filterfeedback	Beschreibung
	In der Symbolleiste wird die Anzahl der Ergebnisse angezeigt. Da kein Textübereinstimmungsfilter gesetzt war und die Höchstzahl nicht erreicht wurde, werden keine weiteren Symbole angezeigt.
	In der Symbolleiste wird angezeigt, 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.
	In der Symbolleiste wird angezeigt, dass die Ergebnisse über einen Übereinstimmungstextfilter beschränkt wurden. Dies wird durch ein Lupensymbol angezeigt.

Ergebnisse filtern

1. Wählen Sie in den Menüs die Optionsfolge **Tools > Filter** aus. Das Filterdialogfeld wird geöffnet.
2. Wählen und optimieren Sie die Filter, die Sie verwenden möchten.

3. Klicken Sie auf **OK**, um die Filter anzuwenden und die neuen Ergebnisse im Bereich "Extraktionsergebnisse" anzuzeigen.

Untersuchen von Konzeptkarten

Sie können eine Konzeptkarte erstellen, um festzustellen, wie Konzepte miteinander in Zusammenhang stehen. Wenn Sie ein einzelnes Konzept auswählen und auf **Zuordnen** 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 generiert werden. Dies kann einige Minuten dauern. Wenn Sie den Index generiert haben, müssen Sie erst wieder einen Index erstellen, wenn Sie eine neue Extraktion vornehmen. Wenn der Index bei jeder Extraktion automatisch erstellt werden soll, wählen Sie in den Extraktionseinstellungen diese Option aus. Weitere Informationen finden Sie im Thema „Extrahieren von Daten” auf Seite 94.

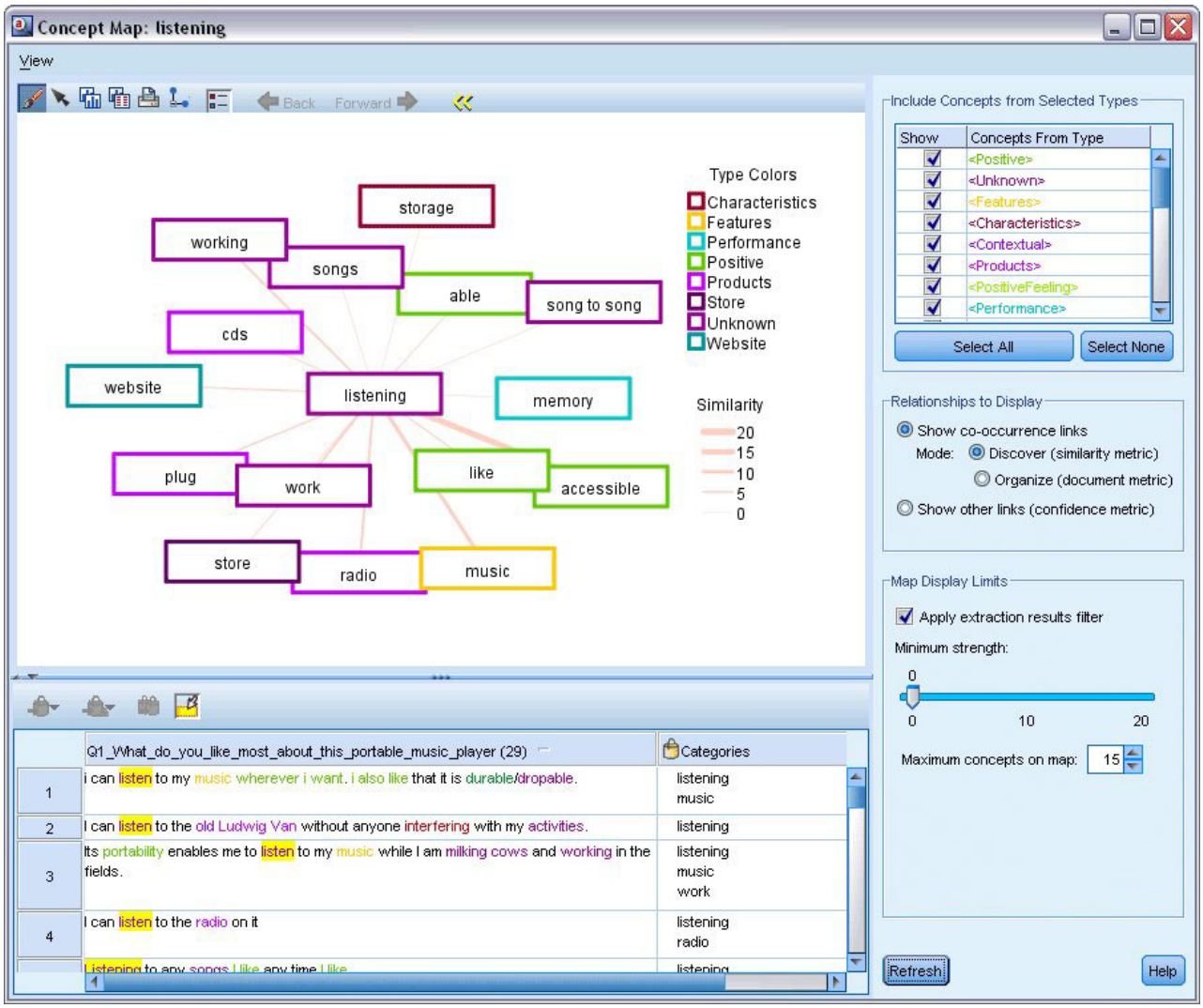


Abbildung 27. Eine Konzeptkarte für das ausgewählte Konzept

So zeigen Sie eine Konzeptkarte an

1. Wählen Sie im Bereich "Extraktionsergebnisse" ein einzelnes Konzept aus.
2. Klicken Sie in der Symbolleiste dieses Bereichs auf die Schaltfläche **Zuordnen**. Wenn der Kartenindex bereits generiert war, wird die Konzeptkarte in einem separaten 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.
3. 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.
4. In der oberen Symbolleiste werden einige allgemeine Zuordnungstools bereitgestellt, wie z. B. 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. Weitere Informationen finden Sie im Thema „Verwenden von Diagrammsymbolleisten und Paletten” auf Seite 173.
5. 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, inaktivieren Sie den Typ in der Tabelle.

Karteneinstellungen: Anzuzeigende Beziehungen

Kookkurrenzzusammenhänge zeigen. Wenn Sie Kookkurrenzzusammenhänge anzeigen möchten, wählen Sie diesen Modus aus. Der Modus wirkt sich darauf aus, wie die Stärke des Zusammenhangs berechnet wurde.

- *Erkennen (Ähnlichkeitsmetrik)*. Mit dieser Metrik wird die Stärke des Zusammenhangs 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. Mit der folgenden Formel werden Gleitkommawerte in Ganzzahlen konvertiert.

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

Abbildung 28. Formel des Ähnlichkeitskoeffizienten

In dieser Formel 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 Stärke der Zusammenhänge bei dieser Metrik wird durch die reine Anzahl der Kookurrenzen 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, eine 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 Optimie-

rung der Ressourcen (insbesondere Synonymie) bzw. die Disambiguierung (Begriffsklärung) erleichtert. Kurzbeschreibungen zu den einzelnen Gruppierungsverfahren finden Sie hier: „Erweiterte linguistische Einstellungen" auf Seite 120

Hinweis: Beachten Sie, dass keine angezeigt werden, wenn diese nicht ausgewählt wurden, als der Index erstellt wurde, oder wenn keine Beziehungen gefunden werden. Weitere Informationen finden Sie im Thema „Erstellen von Konzeptkartenindizes".

Karteneinstellungen: Kartenanzeigegrenzen

Extraktionsergebnisfilter anwenden. Wenn Sie nicht alle Konzepte verwenden wollen, können Sie den Filter in den Extraktionsergebnissen verwenden, um das Gezeigte einzugrenzen. Wählen Sie dann diese Option aus. IBM SPSS Modeler Text Analytics sucht nach zugehörigen Konzepten unter Verwendung dieser gefilterten Menge. Weitere Informationen finden Sie im Thema „Filtern von Extraktionsergebnissen" auf Seite 97.

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

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

Erstellen von Konzeptkartenindizes

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

Gruppierungsverfahren. Wählen Sie mindestens ein Verfahren aus. Eine kurze Beschreibung dieser Verfahren finden Sie in „Linguistische Verfahren" auf Seite 123. Nicht alle Verfahren stehen für alle Textsprachen zur Verfügung.

Paarbildung spezifischer Konzepte verhindern. Wählen Sie dieses Kontrollkästchen aus, um den Vorgang der Gruppierung oder Paarbildung von zwei Konzepten in der Ausgabe zu unterbinden. Klicken Sie zum Erstellen oder Verwalten von Konzeptpaaren auf **Paare verwalten**. Weitere Informationen finden Sie im Thema „Verwalten von Linkausnahmepaaren" auf Seite 123.

Die Erstellung des Index kann mehrere Minuten in Anspruch nehmen. Wenn Sie den Index generiert haben, müssen Sie erst wieder einen Index generieren, wenn Sie eine neue Extraktion 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 Extraktionseinstellungen auswählen. Weitere Informationen finden Sie im Thema „Extrahieren von Daten" auf Seite 94.

Optimieren von Extraktionsergebnissen

Die Extraktion 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 Textmining und Kategorisierung unverzichtbar sind, gewährleistet das Optimieren Ihrer Extraktionsergebnisse von Anfang an, dass Sie bei jeder erneuten Extraktion genau dieselben Ergebnisse in Ihren Kategoriedefinitionen erhalten. So wird die Zuweisung von Datensätzen und Dokumenten zu Ihren Kategorien genauer und wiederholbar.

Die Extraktionsergebnisse dienen als Bausteine für Ihre Kategorien. Beim Erstellen von Kategorien mithilfe der Extraktionsergebnisse werden Datensätze und Dokumente automatisch Kategorien zugewiesen, wenn sie Text enthalten, der mit mindestens einem Kategoriedeskriptor übereinstimmt. Sie können zwar

vor dem Optimieren der linguistischen Ressourcen mit der Kategorisierung beginnen, aber es ist nützlich, die Extraktionsergebnisse vorher mindestens einmal zu überprüfen.

Beim Überprüfen Ihrer Ergebnisse stoßen Sie möglicherweise auf Elemente, die die Extraktionsengine 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 Extraktionsergebnissen 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. Weitere Informationen finden Sie im Thema „Hinzufügen von Synonymen" auf Seite 103.
- **Den falschen Typen zugeordnete Konzepte.** Angenommen, die Konzepte in Ihren Extraktionsergebnissen werden bei einem Typ aufgeführt und Sie möchten sie einem anderen Typ zuweisen. Ein weiteres Beispiel: Sie finden in den Extraktionsergebnissen 15 Konzepte zu Pflanzen und Sie möchten alle einem neuen Typ mit der Bezeichnung `<Pflanze>` hinzufügen. Für die meisten Sprachen gilt: Konzepte , die in keinem Typwörterbuch gefunden, aber aus dem Text extrahiert werden, erhalten automatisch den Typ `<Unknown>` Sie können Konzepte Typen hinzufügen. Weitere Informationen finden Sie im Thema „Hinzufügen von Konzepten zu Typen" auf Seite 104.
- **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 Extraktion ausschließen. Weitere Informationen finden Sie im Thema „Ausschließen von Konzepten von der Extraktion" auf Seite 105.
- **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. Weitere Informationen finden Sie im Thema „Fuzzy-Gruppierung" auf Seite 218. Fuzzy-Gruppierung ist für japanischen Text nicht verfügbar.
- **Nicht extrahierte Konzepte.** Angenommen, Sie erwarten die Extraktion bestimmter Konzepte, stellen jedoch bei der Überprüfung des Datensatz- oder Dokumenttexts fest, dass einige Wörter oder Wortfolgen 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 Wortfolgen als Teil einer Kategoriedefinition verwenden. Um das Konzept zu extrahieren, können Sie die Aufnahme eines Terms in ein Typwörterbuch erzwingen. Weitere Informationen finden Sie im Thema „Erzwingen der Extraktion von Wörtern" auf Seite 106.

Viele dieser Änderungen können Sie direkt im Bereich "Extraktionsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" vornehmen, indem Sie mindestens ein Element auswählen und durch Klicken mit der rechten Maustaste auf die Kontextmenüs zugreifen.

Wenn Sie die Änderungen vorgenommen haben, wechselt die Hintergrundfarbe des Bereichs und zeigt so an, dass zum Anzeigen der Änderungen eine erneute Extraktion erforderlich ist. Weitere Informationen finden Sie im Thema „Extrahieren von Daten" auf Seite 94. Beim Arbeiten mit größeren Datensätzen kann es effizienter sein, erst nach mehreren Änderungen statt nach jeder einzelnen erneut zu extrahieren.

Hinweis: In der Ansicht "Ressourceneditor" (Ansicht > Ressourceneditor) können Sie die Gesamtmenge der für die Erzeugung der Extraktionsergebnisse verwendeten bearbeitbaren 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. Weitere Informationen finden Sie im Thema Kapitel 16, „Arbeiten mit Bibliotheken", auf Seite 191.

Hinzufügen von Synonymen

Synonyme verknüpfen zwei oder mehr Wörter mit derselben Bedeutung. Synonyme werden häufig auch verwendet, um Terme 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 Extraktion erkannt werden.

Im ersten Schritt legen Sie das Zielkonzept oder Hauptkonzept fest. Das **Zielkonzept** ist das Wort oder die Wortfolge, unter dem bzw. der Sie alle synonymen Terme in den endgültigen Ergebnissen gruppieren möchten. Während der Extraktion werden die Synonyme unter diesem Zielkonzept gruppiert. Im zweiten Schritt werden alle Synonyme für dieses Konzept ermittelt. Bei der endgültigen Extraktion werden alle Synonyme durch das Zielkonzept substituiert. Ein Term 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 Ressourceneditor in Synonymwörterbücher aufgenommen. Um den gesamten Inhalt der Synonymwörterbücher anzuzeigen oder eine beträchtliche Anzahl von Änderungen vorzunehmen, sollten Sie direkt im Ressourceneditor arbeiten. Weitere Informationen finden Sie im Thema „Substitutions-/Synonymwörterbücher” auf Seite 208.

Neue Synonyme werden automatisch in der ersten Bibliothek im Bibliotheksbaum in der Ansicht "Ressourceneditor" gespeichert — standardmäßig ist dies die *lokale Bibliothek*.

Hinweis: Wenn Sie vergeblich nach einer Synonymdefinition in den Kontextmenüs oder direkt im Ressourceneditor suchen, erhalten Sie möglicherweise eine Übereinstimmung mithilfe eines internen Fuzzy-Gruppierungsverfahrens. Weitere Informationen finden Sie im Thema „Fuzzy-Gruppierung” auf Seite 218.

So erstellen Sie ein neues Synonym

1. Wählen Sie die Konzepte, für die Sie ein neues Synonym erstellen möchten, im Bereich "Extraktionsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Zu Synonym hinzufügen> Neu** aus. Das Dialogfeld "Synonym erstellen" wird geöffnet.
3. Geben Sie im Textfeld "Ziel" ein Zielkonzept ein. Unter diesem Konzept werden alle Synonyme gruppiert.
4. Um weitere Synonyme aufzunehmen, geben Sie sie in das Listenfeld "Synonyme" ein. Trennen Sie die synonymen Terme mit dem globalen Trennzeichen. Weitere Informationen finden Sie im Thema „Optionen: Registerkarte "Sitzung"” auf Seite 88.
5. Beim Arbeiten mit japanischem Text legen Sie einen Typen für diese Synonyme fest, indem Sie den Typnamen im Feld **Synonyme vom Typ** auswählen. Das Ziel übernimmt jedoch den Typ, der bei der Extraktion zugewiesen wurde. Wenn allerdings das Ziel nicht als Konzept extrahiert wurde, wird der in dieser Spalte aufgelistete Typ dem Ziel in den Extraktionsergebnissen zugewiesen.

   Hinweis: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.
6. Klicken Sie auf **OK**, um die Änderungen anzuwenden. Das Dialogfeld wird geschlossen und die Hintergrundfarbe im Bereich "Extraktionsergebnisse" wechselt und zeigt so an, dass zum Anzeigen der Änderungen eine erneute Extraktion erforderlich ist. Nehmen Sie alle gewünschten Änderungen vor der erneuten Extraktion vor.

So fügen Sie einem Synonym Konzepte hinzu

1. Wählen Sie die Konzepte, die Sie einer vorhandenen Synonymdefinition hinzufügen möchten, im Bereich "Extraktionsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Zu Synonym hinzufügen >** aus. 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.
3. 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.

Hinzufügen von Konzepten zu Typen

Bei der Extraktion werden die extrahierten Konzepte Typen zugewiesen, damit Terme mit Gemeinsamkeiten gruppiert werden. IBM SPSS Modeler Text Analytics enthalten. Weitere Informationen finden Sie im Thema „Integrierte Typen” auf Seite 202. Für die meisten Sprachen gilt: Konzepte , die in keinem Typwörterbuch gefunden, aber aus dem Text extrahiert werden, erhalten automatisch den Typ `<Unknown>`

Wenn Sie Ihre Ergebnisse prüfen, finden Sie eventuell einige Konzepte, die in einem Typ erscheinen und 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 neuen 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 `<Organisation>` erhalten. In diesem Fall fügen Sie dieses Konzept dem Typ `<Organisation>` hinzu.

Wenn Sie einen Typ erstellen oder Konzepte als Terme einer Termliste eines Typs hinzufügen, werden die Änderungen in Typwörterbüchern in den Bibliotheken der linguistischen Ressourcen im Ressourceneditor aufgezeichnet. Um den Inhalt dieser Bibliotheken anzuzeigen oder eine beträchtliche Anzahl von Änderungen vorzunehmen, sollten Sie direkt im Ressourceneditor arbeiten. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen” auf Seite 204.

So fügen Sie einem Typ ein Konzept hinzu

1. Wählen Sie die Konzepte, die Sie einem vorhandenen Typ hinzufügen möchten, im Bereich "Extraktionsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.
2. Klicken Sie mit der rechten Maustaste, um das Kontextmenü aufzurufen.
3. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Zu Typ hinzufügen>** aus. 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.

4. 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 Terme zu den Typen hinzugefügt.

Hinweis: In japanischem Text gibt es einige Instanzen, bei denen die Typänderung eines Terms nicht den Typ ändert, dem er zum Schluss in der endgültigen Extraktionsliste zugewiesen wird. Dies liegt an internen Wörterbüchern, die bei der Extraktion für einige grundlegende Terme Vorrang haben.

Hinweis: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

So erstellen Sie einen neuen Typ

1. Wählen Sie die Konzepte, für die Sie einen neuen Typ erstellen möchten, im Bereich "Extraktionsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Zu Typ hinzufügen> Neu** aus. Das Dialogfeld "Typ-Eigenschaften" wird geöffnet.
3. Geben Sie einen neuen Namen für diesen Typ im Textfeld "Name" ein und nehmen Sie die gewünschten Änderungen in den anderen Feldern vor. Weitere Informationen finden Sie im Thema „Erstellen von Typen” auf Seite 203.
4. Klicken Sie auf **OK**, um die Änderungen anzuwenden. Das Dialogfeld wird geschlossen und die Hintergrundfarbe im Bereich "Extraktionsergebnisse" wechselt und zeigt so an, dass zum Anzeigen der Änderungen eine erneute Extraktion erforderlich ist. Nehmen Sie alle gewünschten Änderungen vor der erneuten Extraktion vor.

Ausschließen von Konzepten von der Extraktion

Bei der Überprüfung der Ergebnisse entdecken Sie gegebenenfalls gelegentlich unerwünschte Konzepte oder Konzepte, die von automatisierten Kategorieerstellungsverfahren 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 es aus der endgültigen Extraktion auszuschließen. In der Regel sind Konzepte in dieser Liste Füllwörter oder -wortfolgen zur Verbesserung des Textflusses, die keine Bedeutung tragen und die Extraktionsergebnisse unnötig anfüllen. Wenn Sie diese Konzepte in das Ausschlusswörterbuch aufnehmen, verhindern Sie ihre Extraktion.

Durch das Ausschließen von Konzepten werden bei der nächsten Extraktion alle Variationen des ausgeschlossenen Konzepts aus den Extraktionsergebnissen entfernt. Wenn dieses Konzept bereits als Deskriptor in einer Kategorie angezeigt wird, bleibt es nach der erneuten Extraktion mit der Anzahl null in der Kategorie.

Im Falle eines Ausschlusses werden diese Änderungen im Ressourceneditor in einem Ausschlusswörterbuch aufgezeichnet. Um alle Ausschlussdefinitionen anzuzeigen und sie direkt zu bearbeiten, sollten Sie direkt im Ressourceneditor arbeiten. Weitere Informationen finden Sie im Thema „Ausschlusswörterbücher” auf Seite 212.

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

Hinweis: Extraktion für japanischen Text ist in IBM SPSS Modeler Premium verfügbar.

So schließen Sie Konzepte aus

1. Wählen Sie die Konzepte, die Sie von der Extraktion ausschließen möchten, im Bereich "Extraktionsergebnisse", im Datenbereich, im Dialogfeld "Kategoriedefinitionen" oder im Dialogfeld "Clusterdefinitionen" aus.

2. Klicken Sie mit der rechten Maustaste, um das Kontextmenü aufzurufen.
3. Wählen Sie **Aus Extraktion ausschließen** aus. Das Konzept wird im Ressourceneditor dem Ausschlusswörterbuch hinzugefügt und die Hintergrundfarbe im Bereich "Extraktionsergebnisse" wechselt und zeigt so an, dass zum Anzeigen der Änderungen eine erneute Extraktion erforderlich ist. Nehmen Sie alle gewünschten Änderungen vor der erneuten Extraktion vor.

Anmerkung. Ausgeschlossene Wörter werden automatisch in der ersten Bibliothek im Bibliotheksbaum im Ressourceneditor gespeichert — standardmäßig ist dies die *lokale Bibliothek*.

Erzwingen der Extraktion von Wörtern

Wenn Sie nach der Extraktion die Textdaten im Datenbereich überprüfen, stellen Sie möglicherweise fest, dass einige Wörter oder Wortfolgen 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 Wortfolgen als Teil einer Kategoriedefinition verwenden.

Damit diese Wörter und Wortfolgen extrahiert werden, können Sie die Aufnahme eines Terms in ein Typwörterbuch erzwingen. Weitere Informationen finden Sie im Thema „Erzwingen von Termen” auf Seite 207.

Wichtig! Einen Term in einem Wörterbuch als erzwungen zu markieren, ist kein absolut sicheres Verfahren. Das heißt, obwohl ein Term explizit einem Wörterbuch hinzugefügt worden ist, ist er möglicherweise nach einer erneuten Extraktion nicht im Bereich "Extraktionsergebnisse" vorhanden oder er wird nicht genau so angezeigt, wie Sie ihn deklariert haben. Dies kommt selten vor, ist aber möglich, wenn ein Wort oder eine Wortfolge bereits als Teil einer längeren Wortfolge extrahiert wurde. Um das zu verhindern, wenden Sie die Abgleichsoption **Gesamt (keine Zusammensetzungen)** auf diesen Term im Typwörterbuch an. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen” auf Seite 204.

Kapitel 10. Kategorisieren von Textdaten

In der Ansicht "Kategorien und Konzepte" können Sie **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.

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 Kategoriestrukturen, früher Coderahmen genannt, mit hierarchischen Kategorien importieren und diese hierarchischen Kategorien auch im Produkt erstellen.

Hierarchische Kategorien ermöglichen Ihnen die Erstellung einer Baumstruktur mit mindestens einer Unterkategorie, die eine genauere Gruppierung von Elementen, beispielsweise verschiedenen Konzept- oder Themenbereichen, gestattet. 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 obersten 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.

- **Bereich "Kategorien".** In diesem Fensterbereich können Sie Kategorien erstellen und verwalten. Weitere Informationen finden Sie im Thema „Fensterbereich "Kategorien"" auf Seite 109.
- **Bereich "Extraktionsergebnisse".** In diesem Fensterbereich können Sie mit den extrahierten Konzepten und Typen arbeiten. Weitere Informationen finden Sie im Thema „Extraktionsergebnisse: Konzepte und Typen" auf Seite 93.
- **Visualisierungsbereich.** In diesem Fensterbereich können Sie die Kategorien und ihre Interaktionen visuell untersuchen. Weitere Informationen finden Sie im Thema „Kategoriendiagramme und Grafiken" auf Seite 169.
- **Datenbereich.** In diesem Fensterbereich können Sie den Text untersuchen und überprüfen, der in Dokumenten und Datensätzen enthalten ist, die Ihrer Auswahl entsprechen. Weitere Informationen finden Sie im Thema „Datenbereich" auf Seite 116.

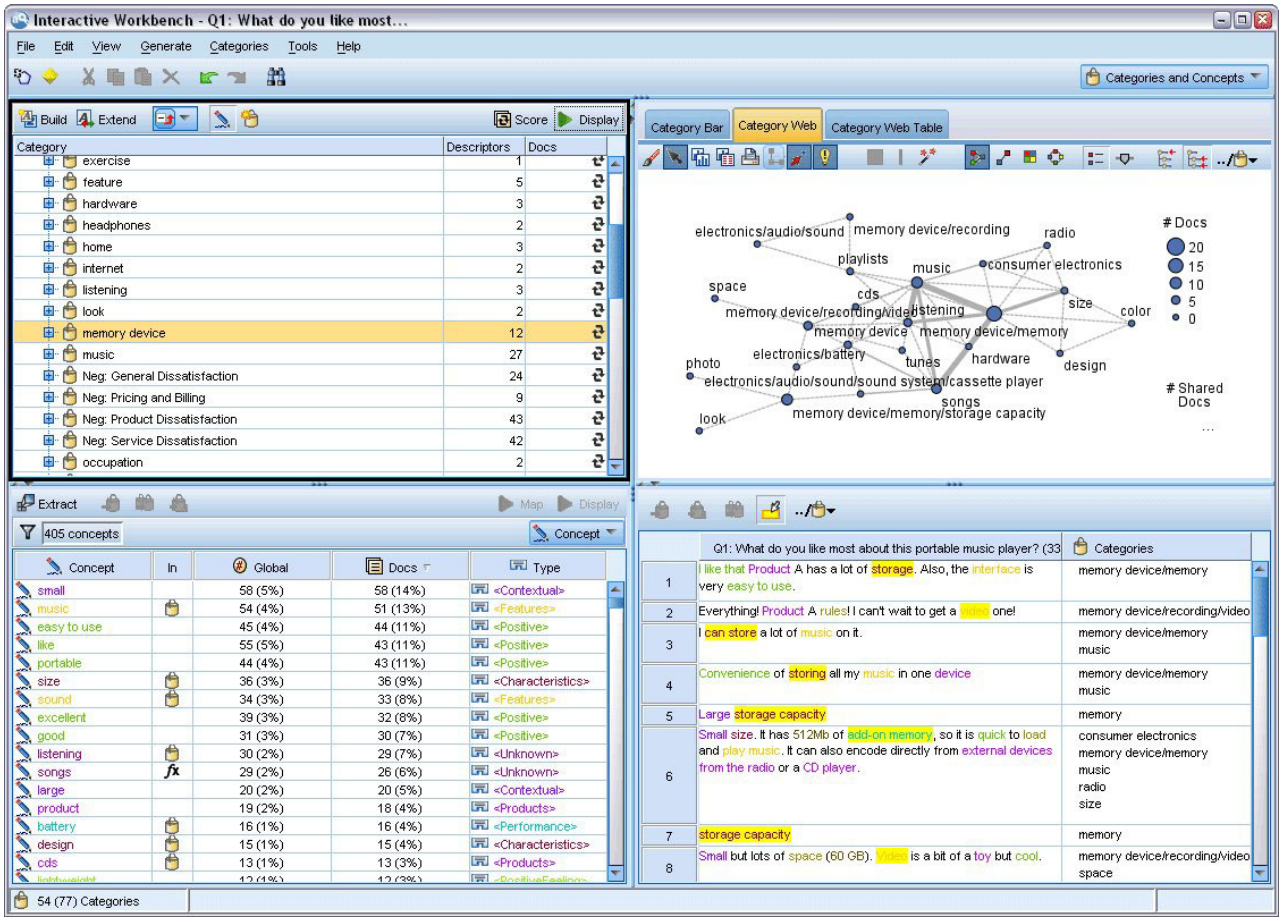


Abbildung 29. Ansicht "Kategorien und Konzepte"

Sie können zwar mit einem Set von Kategorien 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 Extraktionsergebnissen (Konzepte, Typen und Muster) generiert 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 Optimierung von Kategorien ist allerdings nur über die interaktive Workbench möglich. Weitere Informationen finden Sie im Thema „Textminingknoten: Registerkarte "Modell"" auf Seite 24. Sie können Kategoriedefinitionen manuell erstellen, indem Sie die Extraktionsergebnisse durch Ziehen und Ablegen in die Kategorien übertragen. Sie können diese Kategorien oder jede leere Kategorie anreichern, indem Sie einer Kategorie Kategorieregeln hinzufügen, Ihre eigenen vordefinierten Kategorien verwenden. Sie können diese Aktionen auch miteinander kombinieren.

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.

Fensterbereich "Kategorien"

Im Fensterbereich "Kategorien" können Sie Ihre Kategorien erstellen und verwalten. Dieser Fensterbereich befindet sich links oben in 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. Weitere Informationen finden Sie im Thema „Erstellen von Kategorien” auf Seite 118.

Immer wenn eine Kategorie erstellt oder aktualisiert wird, können die Dokumente oder Datensätze durch Klicken 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

In der Baumtabelle in diesem Bereich wird das Set von Kategorien, Unterkategorien und Deskriptoren angezeigt. Der Baum besitzt auch mehrere Spalten mit Informationen für jedes Baumelement. Folgende Spalten stehen gegebenenfalls zur Anzeige zur Verfügung:

- **Code.** Listet den Codewert für jede Kategorie auf. Diese Spalte ist standardmäßig ausgeblendet. Sie können diese Spalte über folgende Optionsfolge in den Menüs anzeigen: **Ansicht > Kategoriebereich**.
- **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 denen 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 Optionsfolge in den Menüs anzeigen oder ausblenden: **Ansicht > Kategoriebereich > 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 oder in den Menüs die Optionsfolge **Ansicht > Kategoriedefinitionen** auswählen, wird das Dialogfeld "Kategoriedefinitionen" geöffnet und alle Elemente (die sogenannten **Deskriptoren**), die zu seiner Definition gehören (z. B. Konzepte, Typen, Muster und Kategorieregeln), werden angezeigt. Weitere Informationen finden Sie im Thema „Erläuterung von Kategorien” auf Seite 115. 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

In der Spalte **Dokumente** in der Kategoriebaumtabelle wird die Anzahl von Dokumenten oder Datensätzen angezeigt, die dieser bestimmten Kategorie zugeordnet werden. Es wird ein Symbol in der Spalte angezeigt, 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 Datasets einige Zeit in Anspruch nehmen kann.

Auswahl von Kategorien im Baum

Wenn Sie eine Auswahl im Baum treffen, können Sie nur gleichgeordnete Kategorien 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 im Datenbereich und im Visualisierungsbereich

Wenn Sie eine Zeile in der Tabelle auswählen, können Sie auf die Schaltfläche **Anzeigen** klicken, um den Visualisierungsbereich und den Datenbereich entsprechend Ihrer Auswahl mit Informationen zu aktualisieren. Wenn ein Fensterbereich nicht sichtbar ist, kann er durch Klicken auf **Anzeigen** aufgerufen werden.

Optimieren 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 Extraktionsergebnisse 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. Weitere Informationen finden Sie im Thema „Bearbeiten und Optimieren von Kategorien" auf Seite 151.

Methoden und Strategien zur Erstellung von Kategorien

Wenn Sie noch keine Extraktion durchgeführt haben oder die Extraktionsergebnisse nicht auf dem neuesten Stand sind, werden Sie automatisch bei Verwendung einer dieser Verfahren zur Erstellung oder zur Erweiterung einer Kategorie zu einer Extraktion 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.

Beachten Sie als Unterstützung zur Erstellung der besten Kategorien die folgenden Abschnitte:
- **Methoden für die Kategorieerstellung**
- **Strategien für die Kategorieerstellung**
- **Tipps zur Erstellung von Kategorien**

Methoden für die Kategorieerstellung

Da jedes Dataset anders ist, können sich die Anzahl der Kategorieerstellungsmethoden und die Reihenfolge, in der sie angewendet werden, im Laufe der Zeit ändern. Da zusätzlich Ihre Textminingziele von Dataset zu Dataset unterschiedlich sein können, müssen Sie möglicherweise mit den verschiedenen Methoden experimentieren, um zu sehen, welche die besten Ergebnisse für die jeweiligen Textdaten hervorbringt. Keines der automatischen Verfahren ergibt eine perfekte Kategorisierung Ihrer Daten, weshalb empfehlen wird, mindestens ein automatisches Verfahren anzuwenden, das gut mit Ihren Daten funktioniert.

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. Weitere Informationen finden Sie im Thema „Erstellen von Kategorien" auf Seite 118.
- **Automatische Erweiterungsverfahren.** Verschiedene linguistische Verfahren stehen zur Verfügung, um vorhandene Kategorien durch Hinzufügen und Verbessern von Deskriptoren zu erweitern, sodass sie mehr Datensätze erfassen. Weitere Informationen finden Sie im Thema „Erweitern von Kategorien" auf Seite 129.
- **Manuelle Verfahren.** Es gibt verschiedene manuelle Methoden wie Drag-and-Drop. Weitere Informationen finden Sie im Thema „Manuelle Erstellung von Kategorien" auf Seite 132.

Strategien für die Kategorieerstellung

Die folgende Liste der Strategien ist keinesfalls vollständig, kann Ihnen jedoch Möglichkeiten aufzeigen, wie Sie bei der Erstellung Ihrer Kategorien vorgehen können.

- Wenn Sie den Textminingknoten definieren, wählen Sie ein Kategorienset aus einem Text Analysis Package (TAP) aus, 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 Kategorieerstellungseinstellungen (**Kategorien > Erstellungseinstellungen**) bearbeiten. Öffnen Sie das Dialogfeld **Erweiterte Einstellungen: Linguistik**, wählen Sie die Kategorieeingabeoption **Nicht verwendete Extraktionsergebnisse** aus und erstellen Sie die zusätzlichen Kategorien.
- Wenn Sie den Knoten definieren, wählen Sie in der Ansicht "Kategorien und Konzepte" der interaktiven Workbench ein Kategorienset aus einem TAP aus. 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.
- Erstellen Sie Kategorien automatisch über die erweiterten linguistischen Einstellungen (**Kategorien > Kategorien erstellen**). Optimieren 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 Generieren von Deskriptoren aus Kategorienamen ausgewählt wurde, später das Dialogfeld "Kategorien erweitern" verwenden und die Option **Leere Kategorien mit aus dem Kategorienamen generierten Deskriptoren erweitern** auswählen. Erweitern Sie dann diese Kategorien ein zweites Mal, aber verwenden Sie dieses Mal die Gruppierungsverfahren.
- 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 optimieren, sodass sie andere zugehörige Deskriptoren einschließen und so mehr Datensätzen entsprechen.

Es wird empfohlen, nach der Anwendung dieser Verfahren die resultierenden Kategorien zu prüfen und manuelle Verfahren zu 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. Weitere Informationen finden Sie im Thema „Bearbeiten und Optimieren von Kategorien" auf Seite 151.

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 das Textdokument bzw. der 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 wie *liberal* und *konservativ* oder *links* und *rechts* sowie speziellere Kategorien wie *sozialliberal*, *fiskalkonservativ* usw. erstellen. 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 Textminings 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 in den folgenden Absätzen beschrieben.

Da Konzepte nicht nur sich selbst, sondern auch eine Reihe zugrunde liegender Terme 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 Terme für ein bestimmtes Konzept zu erfahren, klicken Sie auf den Konzeptnamen in den Extraktionsergebnissen der Ansicht "Kategorien und Konzepte". Wenn Sie mit der Maus über den Konzeptnamen fahren, erscheint eine QuickInfo, in der alle bei der letzten Extraktion in Ihrem Text gefundenen zugrunde liegenden Terme angezeigt werden. Nicht alle Konzepte haben zugrunde liegende Terme. Falls beispielsweise `Auto` und `Fahrzeug` als Synonyme gelten, aber `Auto` als Konzept und `Fahrzeug` als zugrunde liegender Term 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 Terme) 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 Extraktion 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 "*Apfel und Ananas sind die Besten*", zur Extraktion 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 Termen) 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`, `Apfelmus` oder `französischer Apfelkuchen` enthalten.

Sie können auch alle Dokumente oder Datensätze mit Konzepten mit der gleichen Typisierung erfassen, indem Sie einen Typ direkt als Deskriptor verwenden, z. B. `<Obst>`. Hinweis: Ein Stern (*) kann nicht mit Typen verwendet werden.

Weitere Informationen finden Sie im Thema „Extraktionsergebnisse: Konzepte und Typen” auf Seite 93.

Textlinkanalysemuster (TLA-Muster) als Deskriptoren

Verwenden Sie ein TLA-Musterergebnis als Deskriptor, wenn Sie feinere, nuancierte Ideen erfassen möchten. Wenn während einer TLA-Extraktion Text analysiert wird, wird der Text satz- oder teilsatzweise 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. Weitere Informationen finden Sie im Thema „Typ- und Konzeptmuster” auf Seite 165.

Beispielsweise könnten in dem Satz "*Das Zimmer war nicht sonderlich sauber*", die folgenden Konzepte extrahiert werden: `Zimmer` und `sauber`. Falls jedoch die TLA-Extraktion in den Extraktionseinstellungen inaktiviert 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 booleschen Operatoren automatisch einer Kategorie zugewiesen werden. Sie könnten beispielsweise einen Ausdruck schreiben, der Folgendes 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 booleschen Operatoren `&`, `|` und `!()` unterschiedliche Ideen auszudrücken. Nähere Informationen zur Syntax dieser Regeln und wie sie geschrieben und bearbeitet werden, finden Sie in „Verwenden von Kategorieregeln" auf Seite 134.

- Verwenden Sie eine Kategorieregel mit dem booleschen Operator `&` (AND), um Dokumente oder Datensätze zu finden, in denen mindestens zwei Konzepte vorkommen. Die zwei oder mehr durch &-Operatoren verbundenen Konzepte müssen nicht in demselben Satz oder in derselben Wortfolge auftreten, sondern können an beliebiger Stelle in demselben 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 mit dem Text "*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 booleschen Operator `!()` ((NOT) 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 anscheinend zusammengehören, jedoch nicht 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 booleschen Operator `|` (OR) als Deskriptor, um Dokumente oder Datensätze mit einem von 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 Terme 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 Terme 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>]`.

Hinweis: Sie können sowohl `+` als auch `&` in Kategorieregeln verwenden, wenn Sie TLA-Muster in diese Regeln einschließen. Weitere Informationen finden Sie im Thema „Verwenden von TLA-Mustern in Kategorieregeln" auf Seite 136.

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 17. Übereinstimmungen zwischen Beispieldatensätzen und Deskriptoren.

Deskriptor	A	B	C	D	E	Erläuterung
`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-Extraktion 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 allen Datensätzen, in denen eine TLA-Ausgabe mit `sauber` gefunden wurde, unabhängig davon, ob `sauber` mit einem anderen Konzept wie `Zimmer` und in einer beliebigen Slotposition 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 Beschriftung 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 der Beschriftung *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 Befrage, 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 beispielsweise Kategorien automatisch mithilfe von Kategorieerstellungsmethoden (z. B. 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. Weitere Informationen finden Sie im Thema Kapitel 12, „Untersuchen von Textlinkanalyse", auf Seite 163. 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. Weitere Informationen finden Sie im Thema „Verwenden von Kategorieregeln" auf Seite 134.

Kategorieeigenschaften

Neben Deskriptoren verfügen Kategorien zudem über Eigenschaften, die Sie bearbeiten können, um Kategorien umzubenennen oder eine Beschriftung oder Anmerkung hinzuzufügen.

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.
- **Beschriftung.** Die Verwendung von Beschriftungen ist nützlich beim Erstellen aussagekräftigerer Kategoriebeschreibungen zur Verwendung in anderen Produkten oder anderen Tabellen bzw. Grafiken. Wenn Sie die Option zur Anzeige der Beschriftung auswählen, wird die Beschriftung auf der Benutzerschnittstelle 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 erstellen" generiert 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 in den Menüs die Optionsfolge **Kategorien > Zu Anmerkung hinzufügen** auswählen.

Datenbereich

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 Optionsfolge **Ansicht > Fenster > Daten** in den Menüs aus.

Im Datenbereich wird eine Zeile pro Dokument bzw. Datensatz entsprechend der Auswahl im Bereich "Kategorien", "Extraktionsergebnisse" bzw. im Dialogfeld "Kategoriedefinitionen" angezeigt. Die Anzeige erfolgt bis zu einer bestimmten Anzeigegrenze. Standardmäßig ist die Anzahl der im Datenbereich angezeigten Dokumente bzw. Datensätze begrenzt, damit Sie die Daten schneller sehen können. Sie können diese Einstellung jedoch im Optionsdialogfeld ändern. Weitere Informationen finden Sie im Thema „Optionen: Registerkarte "Sitzung"" auf Seite 88.

Datenbereich anzeigen und aktualisieren

Die Anzeige des Datenbereichs wird nicht automatisch aktualisiert, da die automatische Datenaktualisierung bei größeren Datasets 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 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 Datasets 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.

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 Farbcodierung entspricht den Typen, die den Konzepten zugewiesen sind. Alternativ können Sie die Maus über farbcodierte 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 gegebenenfalls 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 Datenbereich 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**. Weitere Informationen finden Sie im Thema „Optionen: Registerkarte "Sitzung"" auf Seite 88.
- **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. Weitere Informationen finden Sie im Thema „Kategorierelevanz".
- **Kategorieanzahl.** Führt die Anzahl der Kategorien auf, denen ein Datensatz angehört.

Kategorierelevanz

Um bessere Kategorien zu erstellen, können Sie die Relevanz der Dokumente oder 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. Weitere Informationen finden Sie im Thema „Datenbereich" auf Seite 116.

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. Weitere Informationen finden Sie im Thema „Datenbereich” auf Seite 116.

Je niedriger die Rangnummer eines Datensatzes, desto besser passt er zur ausgewählten Kategorie oder desto relevanter ist er für die ausgewählte 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 in den Menüs die Optionsfolge **Kategorien > Zu Anmerkung hinzufügen** auswählen.

Erstellen von Kategorien

Sie können zwar über Kategorien aus einem Text Analysis Package verfügen, aber Sie können auch mit einigen linguistischen und häufigkeitsbasierten Verfahren automatisch Kategorien erstellen. Über das Dialogfeld "Kategorien erstellen: Einstellungen" können Sie die automatisierten linguistischen und häufigkeitsbasierten Verfahren anwenden, um Kategorien aus Konzepten oder Konzeptmustern zu erstellen.

In der Regel können Kategorien aus unterschiedlichen Deskriptoren (Typen, Konzepte, TLA-Muster, Kategorieregeln) erstellt werden. Wenn Sie Kategorien mit automatisierten Verfahren zur Erstellung von Kategorien erstellen, wird jede erstellte Kategorie nach einem Konzept oder Konzeptmuster (je nach Auswahl) benannt und enthält ein Set von Deskriptoren. Diese Deskriptoren können aus Kategorieregeln oder Konzepten bestehen und enthalten alle zugehörigen Konzepte, die von den Verfahren erkannt werden.

Nach der Erstellung 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. Weitere Informationen finden Sie im Thema „Bearbeiten und Optimieren von Kategorien” auf Seite 151.

Wichtig! In früheren Releases wurden Kookkurrenz- und Synonymregeln von eckigen Klammern umgeben. In diesem Release zeigen nun eckige Klammern ein Musterergebnis für die Textlinkanalyse an. Stattdessen stehen Kookkurrenz- und Synonymregeln in runden Klammern, z. B. `(Lautsprechersysteme|Lautsprecher)`.

So erstellen Sie Kategorien

1. Wählen Sie in den Menüs die Optionsfolge **Kategorien > Kategorien erstellen** aus. Ein Nachrichtenfeld wird angezeigt, es sei denn, Sie haben durch Auswahl der entsprechenden Option festgelegt, dasss die Aufforderung nicht angezeigt werden soll.
2. Wählen Sie aus, ob Sie die Kategorie jetzt erstellen oder zuerst die Einstellungen bearbeiten möchten.

- Klicken Sie auf **Jetzt erstellen**, um mit den aktuellen Einstellungen die Erstellung von Kategorien zu beginnen. Die standardmäßig ausgewählten Einstellungen sind für den Beginn der Kategorisierung oft ausreichend. Der Kategorieerstellungsprozess wird gestartet und ein Dialogfeld über den Fortschritt wird angezeigt.
- Klicken Sie auf **Bearbeiten**, um die Erstellungseinstellungen zu überprüfen und zu ändern.

Hinweis: Die maximale Anzahl an Kategorien, die angezeigt werden können, ist 10.000. Es wird eine Warnung angezeigt, wenn diese Anzahl erreicht oder überschritten wird. Wenn dies geschieht, sollten Sie die Option "Kategorien erstellen" bzw. "Kategorien erweitern" ändern, um die Anzahl der erstellten Kategorien zu verringern.

Eingaben

Die Kategorien werden aus Deskriptoren erstellt, die aus Typmustern oder Typen abgeleitet werden. In der Tabelle können Sie die einzelnen Typen oder Muster auswählen, die in den Kategorieerstellungsprozess aufgenommen werden.

Typmuster. Wenn Sie Typmuster auswählen, werden die Kategorien aus Mustern anstelle aus einzelnen Typen und Konzepten erstellt. So werden Datensätze oder Dokumente kategorisiert, die ein Konzeptmuster enthalten, das zum ausgewählten Typmuster gehört. Wenn Sie in der Tabelle demnach das Typmuster `<Budget>` und `<Positive>` auswählen, können Kategorien wie `Kosten & <Positive>` oder `Sätze & hervorragend` erzeugt werden.

Wenn Typmuster als Eingabe zur automatisierten Kategorieerstellung verwendet werden, können die Verfahren 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 eine 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 `<Unknown>` vorausgewählt. Das führt dazu, dass alle Muster, die den Typ `<Unknown>` (für nicht japanischen Text) enthalten, ausgewählt werden. . In der Tabelle werden die Typen in absteigender Reihenfolge angezeigt, beginnend mit der größten Anzahl an Datensätzen oder Dokumenten (Dokumentanzahl).

Typen. Wenn Sie Typen auswählen, werden die Kategorien aus Konzepten erstellt, die zu den ausgewählten 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 die Erstellung uninteressanter Kategorien vermeiden. In der Tabelle werden die Typen in absteigender Reihenfolge angezeigt, beginnend mit der größten Anzahl an Datensätzen oder Dokumenten (Dokumentanzahl). Typen aus der `Opinions` Library werden standardmäßig in der Typtabelle inaktiviert.

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 `<Unknown> + <Positive>` auswählen, erhalten Sie möglicherweise die Kategorie `Obst + <Positive>` mit ein oder zwei Arten von Obst, wie beispielsweise `Obst + lecker` und `Apfel + gut`. Bei diesem zweiten Ergebnis werden nur zwei Konzeptmuster angezeigt, da die anderen Vorkommen von Obst nicht unbedingt positiv qualifiziert sind. Und während dies möglicherweise für Ihre aktuellen Textdaten brauchbar ist, kann es für Longitudinalstudien, in denen verschiedene Sets von Doku-

menten verwendet werden, sinnvoll sein, manuell weitere Deskriptoren hinzuzufügen, wie `Zitrusfrucht + positiv`, oder aber Typen zu verwenden. Wenn Sie ausschließlich Typen als Eingabe verwenden, können Sie alle möglichen Arten von Obst finden.

Verfahren

Da jedes Dataset anders ist, kann sich die Anzahl der Methoden und die Reihenfolge, in der sie angewendet werden, im Laufe der Zeit ändern. Da Ihre Textminingziele von Dataset zu Dataset unterschiedlich sein können, müssen Sie möglicherweise mit den verschiedenen Verfahren experimentieren, um zu sehen, welches die besten Ergebnisse für die jeweiligen Textdaten hervorbringt.

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 der Erstellung 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 aus und klicken Sie auf die Schaltfläche "Erweiterte Einstellungen", um die Einstellungen für die ausgewählten Verfahren anzuzeigen. Keines der automatischen Verfahren ergibt eine perfekte Kategorisierung Ihrer Daten, weshalb empfehlen wird, mindestens ein automatisches Verfahren anzuwenden, das gut mit Ihren Daten funktioniert.Ein gleichzeitiger Einsatz von linguistischen und häufigkeitsbasierten Verfahren ist bei der Erstellung nicht möglich.

- **Erweiterte linguistische Verfahren**. Weitere Informationen finden Sie in „Erweiterte linguistische Einstellungen”.
- **Erweiterte häufigkeitsbasierte Verfahren**. Weitere Informationen finden Sie in „Erweiterte Einstellungen für Häufigkeit” auf Seite 128.

Erweiterte linguistische Einstellungen

Beim Erstellen von Kategorien haben Sie die Wahl zwischen einigen erweiterten linguistischen Verfahren für die Kategorieerstellung, 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 Dataset 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 Textminingziele von Dataset zu Dataset unterschiedlich sein können, müssen Sie möglicherweise mit den verschiedenen Verfahren experimentieren, um zu sehen, welches die besten Ergebnisse für die jeweiligen Textdaten hervorbringt. Keines der automatischen Verfahren ergibt eine perfekte Kategorisierung Ihrer Daten, weshalb empfehlen wird, mindestens ein automatisches Verfahren anzuwenden, das gut mit Ihren Daten funktioniert.

Die folgenden Bereiche und Felder sind im Dialogfeld "Erweiterte Einstellungen: Linguistik" verfügbar:

Ein- und Ausgabe

Kategorieeingabe. Auswahl, woraus die Kategorien erstellt werden:

- **Nicht verwendete Extraktionsergebnisse**. Diese Option aktiviert Kategorien, die aus Extraktionsergebnissen 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 Extraktionsergebnisse**. Diese Option aktiviert unter Verwendung der Extraktionsergebnisse zu erstellende Kategorien. Dies ist am sinnvollsten, wenn noch keine oder nur sehr wenige Kategorien vorhanden sind.

Kategorieausgabe. Auswahl der allgemeinen Struktur für die zu erstellenden 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.
- **Kategorien glätten (nur Einzelebene)**. Mit dieser Option wird nur eine Kategorieebene erstellt, d. h., es werden keine Unterkategorien generiert.

Gruppierungsverfahren

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 kombinieren, um die Dokumente oder Datensätze vollständig zu erfassen. Möglicherweise erkennen Sie ein Konzept in mehreren Kategorien oder finden redundante Kategorien vor.

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. Weitere Informationen finden Sie im Thema „Konzeptwurzelableitung" auf Seite 124. 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. Dieses Verfahren wird empfohlen, wenn die Konzepte dem semantischen Netz bekannt und nicht zu mehrdeutig sind. Es ist weniger hilfreich, wenn der Text eine 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 gleichgeordnete Elemente 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 diesem Release nur für englischen Text verfügbar. Weitere Informationen finden Sie im Thema „Semantische Netze" auf Seite 126.

Konzepteinbeziehung. Dieses Verfahren erstellt Kategorien durch die Gruppierung von Multiterm-Konzepten (zusammengesetzte Wörter) basierend darauf, ob sie Wörter enthalten, die Subsets oder Supersets eines Worts in dem anderen sind. So wird beispielsweise `Sitz` mit `Kindersitz`, `Sitzheizung` und `Kindersitzgurt` zu einer Gruppe zusammengefasst. Weitere Informationen finden Sie im Thema „Konzepteinbeziehung" auf Seite 125.

Kookkurrenz. Dieses Verfahren erstellt Kategorien aus Kookkurrenzen im Text. Dahinter steht folgende Überlegung: Wenn Konzepte oder Konzeptmuster häufig gemeinsam 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. Weitere Informationen finden Sie im Thema „Kookkurrenzregeln" auf Seite 127.

Minimale Anzahl an Dokumenten. 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.

Paarbildung spezifischer Konzepte verhindern. Wählen Sie dieses Kontrollkästchen aus, um den Vorgang der Gruppierung oder Paarbildung von zwei Konzepten in der Ausgabe zu verhindern. Klicken Sie zum Erstellen oder Verwalten von Konzeptpaaren auf **Paare verwalten...**. Weitere Informationen finden Sie im Thema „Verwalten von Linkausnahmepaaren" auf Seite 123.

Wenn möglich mit Platzhaltern verallgemeinern. Wählen Sie diese Option aus, um dem Produkt zu ermöglichen, für Kategorien, in denen ein Stern als Platzhalter verwendet wird, allgemeine Regeln aufzustellen. Beispielsweise könnte anstelle der Erzeugung mehrerer Deskriptoren wie `[Apfel vom Bioladen + .]` und `[Apfelmus + .]` 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 die Erstellung von Kategorien

Neben der Auswahl der anzuwendenden Gruppierungsverfahren 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 erstellen" 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 in mehr als einer Kategorie ermöglichen. Wenn diese Option ausgewählt ist, können Deskriptoren in mehr als einer der Kategorien verwendet werden, die als nächste erstellt werden. Diese Option ist im Allgemeinen ausgewählt, da Elemente häufig oder "natürlich" in zwei oder mehr Kategorien fallen, sodass 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 dies könnte je nach vorhandenem Datentyp 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 einer anderen basierend auf `Hersteller` angezeigt werden. 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.

Doppelte Kategorienamen auflösen durch. Wählen Sie aus, 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 Linkausnahmepaaren

Bei der Kategorieerstellung, beim Clustering und beim Zuordnen von Konzepten gruppieren die internen Algorithmen Wörter anhand bekannter Zuordnungen. Damit zwei Konzepte nicht zu einem Paar verbunden oder miteinander verknüpft werden können, aktivieren Sie diese Funktion im Dialogfeld **Kategorien erstellen: Erweiterte Einstellungen**, im Dialogfeld **Cluster erstellen** und im Dialogfeld **Konzeptkartenindex-Einstellungen** und klicken Sie auf die Schaltfläche **Paare verwalten**.

Im anschließend angezeigten Dialogfeld für die Verwaltung von Linkausnahmen 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 Clustering 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 Paar als separate Zeile in der Tabelle hinzufügen.

Linguistische Verfahren

Beim Erstellen bzw. der Erweiterung von Kategorien haben Sie die Wahl zwischen einigen erweiterten linguistischen Verfahren für die Kategorieerstellung, 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 der Erstellung 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 Dataset eindeutig ist und sich die Anzahl der Methoden und die Reihenfolge, in der sie angewendet werden, daher im Laufe der Zeit ändern kann. Da Ihre Textminingziele von Dataset zu Dataset unterschiedlich sein können, müssen Sie möglicherweise mit den verschiedenen Verfahren experimentieren, um zu sehen, welches die besten Ergebnisse für die jeweiligen Textdaten hervorbringt. Keines der automatischen Verfahren ergibt eine perfekte Kategorisierung Ihrer Daten, weshalb empfehlen wird, mindestens ein automatisches Verfahren anzuwenden, das gut mit Ihren Daten funktioniert.

Im Folgenden sind die wichtigsten automatischen linguistischen Verfahren für die Kategorieerstellung aufgeführt:

- **Konzeptwurzelableitung.** Bei diesem Verfahren werden Kategorien erstellt, indem mit einem Konzept verwandte Konzepte durch die Analyse einzelner Konzeptkomponenten hinsichtlich ihrer morphologischen Verwandtschaft gefunden werden. Weitere Informationen finden Sie im Thema „Konzeptwurzelableitung" auf Seite 124. 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. Weitere Informationen finden Sie im Thema „Konzepteinbeziehung" auf Seite 125.
- **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. Weitere Informationen finden Sie im Thema „Semantische Netze" auf Seite 126. Diese Option ist nur für englischen Text verfügbar.

- **Kookkurrenz.** Dieses Verfahren erstellt Kookkurrenzregeln, die verwendet werden können, um eine neue Kategorie zu erstellen oder eine Kategorie zu erweitern, oder die als Eingabe für ein anderes Kategorieverfahren verwendet werden können. Weitere Informationen finden Sie im Thema „Kookkurrenzregeln" auf Seite 127.

Konzeptwurzelableitung

Hinweis: Dieses Verfahren 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 in einem Konzept 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.

Hinweis: Sie können verhindern, dass Konzepte miteinander gruppiert werden, indem Sie sie explizit angeben. Weitere Informationen finden Sie im Thema „Verwalten von Linkausnahmepaaren" auf Seite 123.

Termkomponentenbildung und Bildung der Grundform

Bei Anwendung der Verfahren zur Konzeptwurzelableitung bzw. Konzepteinbeziehung werden die Terme 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 Terme geladen und auf der Grundlage von Trennzeichen wie Leerzeichen, Bindestriche und Apostrophe in Komponenten aufgespalten. Der Term `Systemadministrator` beispielsweise wird wie folgt in Komponenten aufgespalten: `{Administrator, System}`.

Einige Teile des ursprünglichen Terms werden jedoch möglicherweise nicht verwendet. Diese werden als "Stoppwörter" bezeichnet. Im Englischen gehören beispielsweise folgende Wörter zu diesen ignorierbaren Komponenten: `a`, `and`, `as`, `by`, `for`, `from`, `in`, `of`, `on`, `or`, `the`, `to` und `with`.

So weist der Term `examination of the data` das Komponentenset `{data, examination}` auf; `of` und `the` werden als ignorierbar betrachtet. Die Reihenfolge der Komponenten in einem Komponentenset ist nicht von Bedeutung. Daher können folgende drei Terme äquivalent sein: `Hustenmedikament für Kinder`, `Kindermedikament gegen Husten` und `Medikament gegen Husten bei Kindern`, da alle dasselbe Komponentenset `{Kind, Husten, Medikament}` aufweisen. Die einzelnen Termpaare werden als äquivalent identifiziert. Die zugehörigen Konzepte werden zu einem neuen Konzept zusammengeführt, das alle drei Terme referenziert.

Da die Komponenten eines Terms außerdem gebeugt sein können, werden intern sprachspezifische Regeln angewendet, um äquivalente Terme unabhängig von ihren Flexionsabweichungen zu ermitteln. Somit können die Terme `Waldschutz` und `Schutz der Wälder` als äquivalent identifiziert werden, da die Grundform im Singular `Wald` lautet.

Funktionsprinzipien der Konzeptwurzelableitung

Nachdem die Terme 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 Terme 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.

Das Set von 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 Terme. 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 erstellt, indem mit Algorithmen für lexikalische Reihen Konzepte identifiziert werden, die in anderen Konzepten enthalten sind. Dahinter steht folgende Idee: Wenn Wörter in einem Konzept ein Subset 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 Terme 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. Weitere Informationen finden Sie im Thema „Konzeptwurzelableitung" auf Seite 124. Anschließend analysiert der Einbeziehungsalgorithmus die Komponentensets. Bei jedem Komponentenset sucht der Algorithmus nach einem weiteren Komponentenset, bei dem es sich um ein Subset des ersten Komponentensets handelt.

Wenn Ihnen beispielsweise das Konzept `kontinentales Frühstück`, das das Komponentenset {`Frühstück`, `kontinental`} aufweist, und das Konzept `Frühstück`, das das Komponentenset {`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 Bereich "Extraktionsergebnisse" 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 Terme 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.

Hinweis: Sie können verhindern, dass Konzepte miteinander gruppiert werden, indem Sie sie explizit angeben. Weitere Informationen finden Sie im Thema „Verwalten von Linkausnahmepaaren" auf Seite 123.

Semantische Netze

In diesem Release 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 Terme 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: `eating apple`, `dessert apple` und `granny smith`, da das semantische Netz folgende Informationen enthält: 1) `dessert apple` ist ein Synonym von `eating apple` und 2) `granny smith` ist eine Sorte von `eating apple` (d. h. ein Hyponym von `eating apple`).

Isoliert betrachtet sind viele Konzepte, insbesondere Uniterme, mehrdeutig. Das Konzept `buffet` beispielsweise kann eine Art Mahlzeit oder ein Möbelstück bezeichnen. Wenn das Set der Konzepte `meal`, `furniture` und `buffet` beinhaltet, ist der Algorithmus gezwungen, zu entscheiden, ob `buffet` in dieselbe Kategorie eingeordnet werden soll wie `meal` oder `furniture`. 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 `apple pie` eine Art von Kuchen (`pie`) ist, aber nur das semantische Netz erkennt, dass `tart` auch eine Art von Kuchen (`pie`) 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 werden geliefert. 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 `Danish pastry` bis zu `coffee roll` (übergeordnet), dann `bun` (über-übergeordnet) und aufwärts bis `bread`.

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 erkennen und zusammenfassen, die im Set von Dokumenten bzw. Datensätzen eng miteinander verwandt sind. Dahinter steht folgende Überlegung: Wenn Konzepte häufig gemeinsam in Dokumenten und Datensätzen gefunden werden, ist diese Kookkurrenz Ausdruck einer zugrunde liegenden Beziehung, die wahrscheinlich in Ihren Kategoriedefinitionen von Nutzen ist. Dieses Verfahren erstellt Kookkurrenzregeln, die verwendet werden können, um eine neue Kategorie zu erstellen bzw. eine Kategorie zu erweitern, oder die als Eingabe eines anderen Kategorieverfahrens verwendet werden können. Zwei Konzepte weisen eine starke Kookkurrenz auf, wenn sie häufig zusammen in einem Set von Datensätzen vorkommen und selten einzeln in den anderen Datensätzen vorkommen. Dieses Verfahren kann bei größeren Datasets 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`, `Sandwich` häufiger zusammen als getrennt vorkommen, werden sie gemeinsam in eine Konzeptkookkurrenzregel (`Erdnussbutter & Gelee & Sandwich`) aufgenommen.

Wichtig! In früheren Releases wurden Kookkurrenz- und Synonymregeln von eckigen Klammern umgeben. In diesem Release zeigen nun eckige Klammern ein Musterergebnis für die Textlinkanalyse 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 mindestens zwei Konzepte zu finden, die häufig gemeinsam vorkommen. Zwei oder mehr Konzepte weisen eine starke Kookkurrenz auf, wenn sie häufig zusammen in einem Set von Dokumenten bzw. Datensätzen vorkommen und selten einzeln in den anderen Dokumenten oder Datensätzen vorkommen.

Wenn kookkurrierende Konzepte gefunden werden, wird eine Kategorieregel erstellt. Diese Regeln bestehen aus mindestens zwei Konzepten, die über den booleschen Operator & verbunden sind. Bei diesen Regeln handelt es sich um logische Anweisungen, die ein Dokument oder einen Datensatz automatisch in eine Kategorie einordnen, wenn die in der Regel enthaltenen Konzepte alle in dem betreffenden Dokument oder Datensatz kookurrieren.

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 Dokumenten.** Die minimale Anzahl 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 zwei Datensätzen gefunden werden (und keines der beiden Konzepte in irgendeinem anderen Datensatz vorkommt). Wenn **Minimale Anzahl an Dokumenten** auf 2 gesetzt ist (der Standardwert), erstellt das Kookkurrenzverfahren eine Kategorieregel (Apfel und Birne). Wenn der Wert auf 3 erhöht wird, wird die Regel nicht mehr erstellt.

Hinweis: Bei kleinen Datasets (< 1000 Antworten) finden Sie möglicherweise keine Kookurrenzen mit den Standardeinstellungen. Versuchen Sie in diesem Fall, den Wert für den Suchabstand zu erhöhen.

Hinweis: Sie können verhindern, dass Konzepte miteinander gruppiert werden, indem Sie sie explizit angeben. Weitere Informationen finden Sie im Thema „Verwalten von Linkausnahmepaaren" auf Seite 123.

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) erstellen, das als eine gegebene Datensatz- oder Dokumentanzahl übersteigend 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 Vorkommen im gesamten Text.

Das Gruppieren häufig auftretender Objekte kann interessante Ergebnisse bringen, da dies eine übliche oder signifikante Antwort angeben kann. Das Verfahren ist für die nicht verwendeten Extraktionsergebnisse, nachdem andere Verfahren angewendet wurden, sehr nützlich. Eine andere Anwendung besteht in der direkten Ausführung dieses Verfahrens nach der Extraktion, 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. Weitere Informationen finden Sie im Thema „Erweitern von Kategorien" auf Seite 129.

Anstelle dieses Verfahrens könnten Sie die Konzepte oder Konzeptmuster nach absteigenden Nummern von Datensätzen oder Dokumenten in den Extraktionsergebnissen sortieren und dann die obersten in den Kategorienbereich ziehen, um die entsprechenden Kategorien zu erstellen.

Die folgenden Felder sind im Dialogfeld "Erweiterte Einstellungen: Häufigkeiten" verfügbar:

Kategoriedeskriptoren generieren bei. Wählen Sie die Eingabeart für Deskriptoren aus. Weitere Informationen finden Sie im Thema „Erstellen von Kategorien" auf Seite 118.

- **Konzeptebene**. Wenn Sie diese Option auswählen, werden Konzept- oder Konzeptmusterhäufigkeiten verwendet. Konzepte werden verwendet, wenn Typen als Eingabe für die Kategorieerstellung ausgewählt 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 Typ- oder Typmusterhäufigkeiten verwendet. Typen werden verwendet, wenn Typen als Eingabe für die Kategorieerstellung ausgewählt wurden. Typmuster werden verwendet, wenn Typmuster ausgewählt wurden. Wenn Sie dieses Verfahren auf die Typebene anwenden, erhalten Sie einen raschen Einblick in die Art der vorhandenen Informationen.

Mindestanzahl Dok. für Elemente mit eigener Kategorie. Mithilfe dieser Option können Sie Kategorien aus häufig auftretenden Elementen erstellen. Diese Option beschränkt die Ausgabe auf ausschließlich 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 Elemente 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 Extraktionsergebnisse**. Diese Option aktiviert Kategorien, die aus Extraktionsergebnissen 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 Extraktionsergebnisse**. Diese Option aktiviert unter Verwendung der Extraktionsergebnisse zu erstellende Kategorien. Dies ist am sinnvollsten, wenn noch keine oder nur sehr wenige Kategorien vorhanden sind.

Doppelte Kategorienamen auflösen durch. Wählen Sie aus, 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.

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 generierten Deskriptoren erweitern** generiert Deskriptoren mit den Wörtern in den Kategorienamen. Je beschreibender ein Kategoriename, um so besser daher die Ergebnisse.

Hinweis: 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 und Ablegen 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 der die Kategorien sehr beschreibende Namen hatten

- Nach der Optimierung 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 generierten Deskriptoren erweitern** eine Erweiterung vornehmen, um ein erstes Set von 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 Erstellungs- und Erweiterungsgruppierungsverfahren ähnliche Algorithmen zugrunde liegen, führt das Erweitern direkt nach der Erstellung 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 mindestens zwei Kategorieregeln in einer Kategorie erzeugen, die mit genau dem gleichen Set von Dokumenten übereinstimmen, da Regeln während des Prozesses unabhängig erstellt werden. Wenn gewünscht, können Sie die Kategorien prüfen und Redundanzen durch manuelle Bearbeitung der Kategoriebeschreibung entfernen. Weitere Informationen finden Sie im Thema „Bearbeiten von Kategoriedeskriptoren” auf Seite 152.

So erweitern Sie Kategorien

1. Wählen Sie im Fensterbereich "Kategorien" die Kategorien aus, die Sie erweitern möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Kategorien > Kategorien erweitern** aus. Wenn Sie nicht die Option markiert haben, dass diese Aufforderung nie angezeigt werden soll, wird ein Nachrichtenfeld angezeigt.
3. Wählen Sie aus, ob Sie die Kategorie jetzt erstellen 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 es wird 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, sodass Sie sie schnell erkennen. Der Text "Erweitert" bleibt, bis Sie erneut erweitern bzw. die Kategorie anderweitig bearbeiten oder über das Kontextmenü löschen.

Hinweis: Die maximale Anzahl an Kategorien, die angezeigt werden können, ist 10.000. Es wird eine Warnung angezeigt, wenn diese Anzahl erreicht oder überschritten wird. Wenn dies geschieht, sollten Sie die Option "Kategorien erstellen" bzw. "Kategorien erweitern" ändern, um die Anzahl der erstellten Kategorien zu verringern.

Diese Verfahren, die für die Erstellung 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. In der interaktiven Workbench stehen die Konzepte und Typen die unter einer Kategorie gruppiert wurden, bei der nächsten Erstellung von Kategorien weiterhin zur Verfügung. Das bedeutet, dass Sie ein Konzept in mehreren Kategorien erkennen oder redundante Kategorien vorfinden können.

Die folgenden Bereiche und Felder sind im Dialogfeld "Kategorien erweitern: Einstellungen" verfügbar:

Erweitern mit. Auswahl, welche Eingabe zur Erweiterung der Kategorien verwendet wird:

- **Nicht verwendete Extraktionsergebnisse**. Diese Option aktiviert Kategorien, die aus Extraktionsergebnissen 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 Extraktionsergebnisse**. Diese Option aktiviert unter Verwendung der Extraktionsergebnisse zu erstellende Kategorien. Dies ist am sinnvollsten, wenn noch keine oder nur sehr wenige Kategorien vorhanden sind.

Gruppierungsverfahren

Eine kurze Beschreibung dieser Verfahren finden Sie in „Erweiterte linguistische Einstellungen" auf Seite 120. Zu diesen Verfahren 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 **Minimale Anzahl 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 nicht linguistische 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.

Paarbildung spezifischer Konzepte verhindern. Wählen Sie dieses Kontrollkästchen aus, um den Vorgang der Gruppierung oder Paarbildung von zwei Konzepten in der Ausgabe zu verhindern. Klicken Sie zum Erstellen oder Verwalten von Konzeptpaaren auf **Paare verwalten...**. Weitere Informationen finden Sie im Thema „Verwalten von Linkausnahmepaaren" auf Seite 123.

Wenn möglich: Wählen Sie aus, ob Sie einfach erweitern und/oder die Deskriptoren mithilfe von Platzhaltern verallgemeinern 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 Sterns als Platzhalter. Beispielsweise könnte anstelle der Erzeugung mehrerer Deskriptoren wie `[Apfel vom Bioladen + .]` und `[Apfelmus + .]` 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** auszuwählen 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.

  Hinweis: Durch die Auswahl dieser Option wird die Option **Semantisches Netz** inaktiviert. Dies ist darauf zurückzuführen, 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 Elementen, um die ein Deskriptor erweitert wird. Definieren Sie bei der Erweiterung eines Deskriptors um Elemente (Konzepte, Typen und andere Ausdrücke) die maximale Anzahl an Elementen, die einem einzelnen Deskriptor hinzugefügt werden können. Wenn Sie als Grenzwert 10 festlegen, können einem vorhandenen Deskriptor höchstens 10 zusätzliche Elemente hinzugefügt werden. Wenn mehr als 10 Elemente hinzugefügt werden sollen, beendet das Verfahren das Hinzufügen neuer Elemente nach dem zehnten Element. Dies kann eine Deskriptorliste verkürzen, garantiert aber nicht, dass die interessantesten Elemente zuerst verwendet wurden. Eventuell möchten Sie die Größe der Erweiterung ohne Abstriche an der Qualität verringern, indem Sie die Option **Wenn möglich mit Platzhaltern verallgemeinern** verwenden. Diese Option gilt nur für Deskriptoren, die die booleschen Operatoren & (AND) bzw. ! (NOT) enthalten.

Unterkategorien ebenfalls erweitern. Mit dieser Option werden auch alle Unterkategorien unter den ausgewählten Kategorien erweitert.

Leere Kategorien mit aus dem Kategorienamen generierten 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 Kategorie, basierend auf den Wörtern, aus denen der Name der Kategorie besteht, 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 generieren, die es wiederum der Kategorie ermöglichen, Datensätze zu erfassen, die diese Deskriptoren enthalten. Diese Option ist vor allem dann hilfreich, wenn Sie Kategorien von anderer Stelle importieren oder Kategorien mit langen beschreibenden Namen manuell erstellen.

Deskriptoren generieren als. Diese Option ist nur verfügbar, wenn die vorherige Option ausgewählt wurde.

- **Konzepte.** Wählen Sie diese Option aus, um die resultierenden Deskriptoren in der Form von Konzepten zu erzeugen, ungeachtet dessen, ob sie aus dem Quellentext extrahiert wurden.
- **Muster.** Wählen Sie diese Option aus, 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 Methoden zur automatisierten Kategorieerstellung und des Regeleditors können Sie Kategorien auch manuell definieren. Es gibt folgende manuelle Methoden:

- Erstellen einer leeren Kategorie, der Sie nacheinander Elemente hinzufügen. Weitere Informationen finden Sie im Thema „Erstellen neuer Kategorien bzw. Umbenennen von Kategorien".
- Ziehen von Termen, Typen und Mustern in den Kategorienbereich. Weitere Informationen finden Sie im Thema „Erstellen von Kategorien durch Ziehen und Ablegen" auf Seite 133.

Erstellen neuer Kategorien bzw. Umbenennen von Kategorien

Sie können leere Kategorien erstellen, um Konzepte und Typen hinzuzufügen. Außerdem können Sie die Kategorien umbenennen.

So erstellen Sie eine neue, leere Kategorie

1. Rufen Sie den Fensterbereich "Kategorien" auf.
2. Wählen Sie in den Menüs die Optionsfolge **Kategorien > Leere Kategorie erstellen** aus. Das Dialogfeld "Kategorieeigenschaften" wird geöffnet.
3. Geben Sie im Namensfeld einen Namen für diese Kategorie ein.

4. 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. Weitere Informationen finden Sie im Thema „Hinzufügen von Deskriptoren zu Kategorien” auf Seite 152.

So benennen Sie eine Kategorie um

1. Wählen Sie eine Kategorie aus und wählen Sie die Optionsfolge **Kategorien > Kategorie umbenennen** aus. Das Dialogfeld "Kategorieeigenschaften" wird geöffnet.
2. Geben Sie im Namensfeld einen neuen Namen für diese Kategorie ein.
3. 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

Das Drag-and-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 "Extraktionsergebnisse" in den Bereich "Kategorien".
- Extrahierte Konzepte aus dem Datenbereich in den Bereich "Kategorien".
- Ganze Zeilen aus dem Datenbereich in den Bereich "Kategorien". Damit wird eine Kategorie erstellt, die aus allen extrahierten Konzepten und Mustern in dieser Zeile besteht.

Hinweis: Der Bereich "Extraktionsergebnisse" unterstützt die Mehrfachauswahl, um das Ziehen und Ablegen mehrerer Elemente zu vereinfachen.

Wichtig! Sie können keine Konzepte aus dem Datenbereich ziehen, wenn diese nicht aus dem Text extrahiert wurden. Wenn Sie die Extraktion eines Konzepts erzwingen möchten, das Sie in Ihren Daten gefunden haben, müssen Sie das Konzept einem Typ hinzufügen. Führen Sie dann die Extraktion erneut aus. Die neuen Extraktionsergebnisse enthalten das soeben hinzugefügte Konzept. Sie können es dann in Ihrer Kategorie verwenden. Weitere Informationen finden Sie im Thema „Hinzufügen von Konzepten zu Typen” auf Seite 104.

So erstellen Sie Kategorien durch Ziehen und Ablegen:

1. Wählen Sie im Bereich "Extraktionsergebnisse" oder im Datenbereich mindestens ein Konzept, ein Muster, einen Typ, einen Datensatz oder einen Teil eines Datensatzes aus.
2. Ziehen Sie das Element bei gedrückter Maustaste in eine bestehende Kategorie oder in den Fensterbereich, um eine neue Kategorie anzulegen.
3. 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**. Weitere Informationen finden Sie im Thema „Festlegen von Optionen” auf Seite 88.

Hinweis: Die daraus resultierende Kategorie wurde automatisch benannt. Wenn Sie einen Namen ändern möchten, können Sie die Kategorie umbenennen. Weitere Informationen finden Sie im Thema „Erstellen neuer Kategorien bzw. Umbenennen von Kategorien” auf Seite 132.

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 booleschen Operatoren automatisch einer Kategorie zugewiesen werden. Sie könnten beispielsweise einen Ausdruck schreiben, der Folgendes bedeutet: *Schließe alle Datensätze, die das extrahierte Konzept* `Botschaft` *enthalten, nicht jedoch* `Argentinien` *, in diese Kategorie ein*.

Während manche Kategorieregeln bei der Erstellung von Kategorien mithilfe von Gruppierungsverfahren wie *Kookkurrenz* und *Konzeptwurzelableitung* (**Kategorien > Erstellungseinstellungen > 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, sodass jedes Dokument, das mit dieser Regel übereinstimmt, oder jeder Datensatz, der mit dieser Regel übereinstimmt, über das Scoring dieser Kategorie zugewiesen wird.

Kategorieregeln helfen dabei, die Qualität und Produktivität Ihrer Textmining-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 boolescher Logik kombinieren.

Die Möglichkeit zur Erstellung dieser Regeln verbessert die Codierungsgenauigkeit, Effizienz und Produktivität, indem sie es Ihnen ermöglicht, Ihr Fachwissen in die Extraktionstechnologie des Produkts einzubringen.

Hinweis: Beispiele für das Abgleichen von Regeln mit Text finden Sie in „Beispiele für Kategorieregeln” auf Seite 140.

Kategorieregelsyntax

Während manche Kategorieregeln bei der Erstellung von Kategorien mithilfe von Gruppierungsverfahren wie *Kookkurrenz* und *Konzeptwurzelableitung* (**Kategorien > Erstellungseinstellungen > 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, das mit dieser Regel übereinstimmt, oder jeder Datensatz, der mit dieser Regel übereinstimmt, automatisch über das Scoring dieser Kategorie zugewiesen.

Hinweis: Beispiele für das Abgleichen von Regeln mit Text finden Sie in „Beispiele für Kategorieregeln” auf Seite 140.

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 Bereich "Extraktionsergebnisse", Bereichen zur Textlinkanalyse oder dem Datenbereich in den Regeleditor übertragen oder sie über die Kontextmenüs hinzufügen.

Wenn Konzepte, Typen und Muster erkannt werden, wird neben dem Text ein Symbol angezeigt.

Tabelle 18. Extraktionssymbole

Symbol	Beschreibung
	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 19. Unterstützte Syntax

Zeichen	Beschreibung

`&`	Der boolesche Operator "and". Beispielsweise enthält `a & b` a *und* `b` wie etwa bei:
	`- Invasion & Vereinigte Staaten`
	`- 2016 & Olympiade`
	`- gut & Apfel`

`|`	Der boolesche Operator "or" ist einschließend. 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 boolesche Operator "not". Beispielsweise enthält `!(a)` nicht `a` wie etwa bei: `!(gut & Hotel)`, `Anschlag & !(Österreich)` oder `!(Gold) & !(Kupfer)`
`*`	Ein Platzhalter, der je nach Verwendung für alles von einem einzelnen Zeichen bis zu einem ganzen Wort stehen kann. Weitere Informationen finden Sie im Thema „Platzhalter in Kategorieregeln" auf Seite 138.
`()`	Ein Trennzeichen für Ausdrücke. Jeder Ausdruck innerhalb der Klammern wird zuerst ausgewertet.
`+`	Der Musterconnector, der zur Bildung eines reihenfolgespezifischen Musters verwendet wird. Wenn vorhanden, müssen eckige Klammern verwendet werden. Weitere Informationen finden Sie im Thema „Verwenden von TLA-Mustern in Kategorieregeln" auf Seite 136.
`[]`	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. Weitere Informationen finden Sie im Thema „Verwenden von TLA-Mustern in Kategorieregeln" auf Seite 136. Verwenden Sie keine eckige Klammern, wenn Sie nach Übereinstimmungen von Konzepten und Typen anstelle von Mustern suchen. *Hinweis*: 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 das Kookkurrenzverfahren 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`.

Entwerten von Zeichen durch einen umgekehrten Schrägstrich als Escapezeichen

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 entwertet, die andernfalls 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 "and" enthält (`&`), ist der umgekehrte Schrägstrich bei einer Eingabe in den Regeleditor erforderlich, also: `r\&d`.

Verwenden von TLA-Mustern in Kategorieregeln

Textlinkanalysemuster 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 Konzeptextraktionsergebnisse und gleichen Dokumente und Datensätze nur auf der Basis der extrahierten Ergebnisse von Textlinkanalysemustern ab.

Wichtig! Um Dokumente mithilfe von TLA-Mustern in Ihren Kategorieregeln abzugleichen, müssen Sie eine Extraktion mit aktivierter Textlinkanalyse ausgeführt haben. Die Kategorieregel sucht nach den bei diesem Prozess gefundenen Übereinstimmungen. Wenn Sie die Untersuchung von TLA-Ergebnissen auf der Registerkarte "Modell" Ihres Textminingknotens nicht aktiviert haben, können Sie die TLA-Extraktion in den Extraktionseinstellungen während der interaktiven Sitzung aktivieren und anschließend eine neue Extraktion vornehmen. Weitere Informationen finden Sie im Thema „Extrahieren von Daten” auf Seite 94.

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. Wenn 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 Textlinkanalyseansicht direkt einer Kategorie hinzufügen, erscheint es nicht in eckigen Klammern. Wird ein Muster jedoch innerhalb einer Kategorieregel 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 Standardmeinungsressourcen 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 Ansicht "Textlinkanalyse" 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 Textlinkanalyseregeln 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 den 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 Opinions Library zugeordnet sind, standardmäßig in der Ausgabe als letzte präsentiert, wie z. B. `Apfel + schlecht`.

Tabelle 20. Mustersyntax und Verwendung von booleschen 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 also Folgendes verwenden: `!([ ])`.
`[a]`	Enthält ein Muster, in dem mindestens ein Element `a` ist, unabhängig von seiner Position im Muster. Übereinstimmungen für `[Geschäft]` können beispielsweise `[Geschäft + gut]` oder nur `[Geschäft + .]` sein.
`[a + b]`	Enthält ein Konzeptmuster. Beispiel: `[Geschäft + gut]`. *Hinweis:* Wenn Sie dieses Muster nur erfassen wollen, ohne andere Elemente hinzuzufügen, wird empfohlen, das Muster direkt zur Kategorie hinzuzufügen und es nicht zum Erstellen einer Regel zu verwenden.
`[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> + <Negativ>]`. *Hinweis:* Wenn Sie dieses Muster nur erfassen wollen, ohne andere Elemente hinzuzufügen, wird empfohlen, das Muster direkt zur Kategorie hinzuzufügen und es nicht zum Erstellen einer Regel zu verwenden.
`[<A> & <B>]`	Enthält ein beliebiges Typmuster mit Typ `<A>` und Typ `<B>`. Zum Beispiel `[<Budget> & <Negativ>]`. Dieses TLA-Muster wird nie extrahiert; bei dieser Schreibweise gleicht es jedoch tatsächlich `[<Budget> + <Negativ>]|[<Negativ> + <Budget>]`. Die Reihenfolge der übereinstimmenden Elemente ist unwichtig. Es können sich zusätzliche Elemente im Muster befinden, mindestens jedoch `<Budget>` und `<Negative>`.

Tabelle 20. Mustersyntax und Verwendung von booleschen Operatoren (Forts.)

Ausdruck	Resultiert in der Übereinstimmung mit Dokumenten oder Datensätzen, auf die Folgendes zutrifft:
`[a + .]`	Enthält ein Muster, in dem `a` das einzige Konzept ist und keine anderen Slots für dieses Muster einen Inhalt haben. Beispiel: `[Geschäft + .]` stimmt mit dem Konzeptmuster überein, 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, einschließlich positiver Aussagen über ein Geschäft. Bei Verwendung von `[Geschäft + .]` werden jedoch nur die Datensatzmusterergebnisse abgeglichen, die `Geschäft` darstellen, und keine anderen Beziehungen oder Meinungen, d. h., sie würden nicht mit `Geschäft + großartig` übereinstimmen. *Hinweis:* Wenn Sie dieses Muster nur erfassen wollen, ohne andere Elemente hinzuzufügen, wird empfohlen, das Muster direkt zur Kategorie hinzuzufügen und es nicht zum Erstellen einer Regel zu verwenden.
`[<A> + <>]`	Enthält ein Muster, in dem `<A>` der einzige Typ ist. `[<Budget> + <>]` stimmt z. B. mit dem Muster überein, dessen einzige Ausgabe ein Konzept vom Typ `<Budget>` ist. *Hinweis:* Sie können `<>` nur dann verwenden, um einen leeren Typ zu kennzeichnen, wenn Sie es hinter das +-Symbol im Typmuster stellen, z. B. `[<Budget> + <>]`, nicht jedoch `[Preis + <>]`. *Hinweis:* Wenn Sie dieses Muster nur erfassen wollen, ohne andere Elemente hinzuzufügen, wird empfohlen, das Muster direkt zur Kategorie hinzuzufügen und es nicht zum Erstellen einer Regel zu verwenden.
`[a + !(b)]`	Enthält mindestens ein Muster, das das Konzept `a` einschließt, nicht jedoch das Konzept `b`. Muss mindestens ein Muster einschließen. Beispiel: `[Preis + !(hoch)]` oder für Typen `[!(<Obst>|<Gemüse>) + <Positive>]`
`!([<A> & <B>])`	Enthält kein bestimmtes Muster. Beispiel: `!([<Budget> & <Negative>])`.

Hinweis: Beispiele für das Abgleichen von Text durch Regeln finden Sie in „Beispiele für Kategorieregeln" auf Seite 140.

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 oder am Beginn eines Worts (Konto* eröffnen) noch als eigene Komponente (Konto * eröffnen). Sie können Platzhalter außerdem nicht in Typnamen verwenden. Beispiel: `Wort* Wort` wie `Apfel* Rezept` entspricht nicht Apfelmus-Rezept und auch keinem anderen Begriff. Jedoch würde `Apfel* *` den Einträgen *Apfelmus-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 Term überein, da der Stern zwischen zwei anderen Wörtern erscheint. Jedoch würde `Apfel *` den Einträgen *Apfel auf Toast*, *Apfel*, *Apfel im Schlafrock* usw. entsprechen.

Tabelle 21. Verwendung von Platzhaltern

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` - `Apfelmus` - `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: - `Apfelmus` - `Apfel & Orange`
`*produkt*`	Enthält ein Konzept mit der Buchstabenfolge `produkt`, das aber beliebig viele Buchstaben als Präfix, Suffix oder beides aufweisen kann. 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 Zustellung + langsam` - `wichtige Zustellung + spät`

Tabelle 21. Verwendung von Platzhaltern (Forts.)

Ausdruck	Resultiert in der Übereinstimmung mit Dokumenten oder Datensätzen, auf die Folgendes zutrifft:
`Veranstaltung *`	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: - `warmes 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 Reservierungen + gut`

Hinweis: Beispiele für das Abgleichen von Regeln mit Text finden Sie in „Beispiele für Kategorieregeln".

Beispiele für Kategorieregeln

Werfen Sie einen Blick auf folgendes Beispiel, das veranschaulichen soll, wie Regeln auf der Basis der Syntax, mit der sie ausgedrückt werden, auf unterschiedliche Weise mit Datensätzen abgeglichen werden.

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 22. Aus Beispiel extrahierte Konzepte und Typen

Extrahiertes Konzept	Art des Konzepts
`Geldbeutel`	`<Unknown>`
`fehlend`	`<Negative>`
`5 Euro`	`<Currency>`
`Decke`	`<Unknown>`
`Picknickbereich`	`<Unknown>`

Aus dem Beispiel extrahierte TLA-Muster

Tabelle 23. Aus Beispiel extrahierte TLA-Musterausgabe

Extrahierte Konzeptmuster	Extrahierte Typmuster	Aus Datensatz
`Picknickbereich + .`	`<Unknown> + <>`	Datensatz B
`Geldbeutel + .`	`<Unknown> + <>`	Datensatz A
`Decke + fehlend`	`<Unknown> + <Negative>`	Datensatz B
`5 Euro + .`	`<Currency> + <>`	Datensatz B
`5 Euro + fehlend`	`<Currency> + <Negative>`	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 24. 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 & <Currency>`	Übereinstimmung mit Datensatz A und B, da beide das extrahierte Konzept `fehlend` und ein Konzept enthalten, das mit dem Typ `<Currency>` übereinstimmt. Dies entspricht: `(fehlend & <Currency>)`
`<Currency> & fehlend`	Übereinstimmung mit Datensatz A und B, da beide das extrahierte Konzept `fehlend` und ein Konzept enthalten, das mit dem Typ `<Currency>` übereinstimmt. Dies entspricht: `(<Currency> & 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 Zeichen `&` zeigt an, dass die Reihenfolge beim Abgleich 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.
`[fehlend + <Currency>]`	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 Termen (`5 Euro + fehlend`) oder auf Typen (`<Currency> + <Negative>`) basiert, Konzepte und Typen jedoch nicht vermischt.

Tabelle 24. Beispielregeln (Forts.)

Regelsyntax	Ergebnis
`[<Currency> + <Negative>]`	Übereinstimmung mit A, aber nicht mit B, da kein TLA-Muster aus Datensatz B extrahiert wurde. Dies entspricht folgender TLA-Ausgabe: `<Currency> + <Negative>`
`[<Negative> + <Currency>]`	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* (`<Currency>`) die erste Slotposition und die *Meinung* (`<Negative>`) die zweite Slotposition.

Erstellen von Kategorieregeln

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 Terme, 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** auswählen.

Sie können Kontextmenüs und Drag-and-Drop verwenden oder Konzepte, Typen und Muster manuell in den Editor eingeben. Danach kombinieren Sie diese mit booleschen 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 Bereich "Extraktionsergebnisse" oder dem Datenbereich in den Regeleditor übertragen. Achten Sie genau auf die Syntax dieser Regeln, um Fehler zu vermeiden. Weitere Informationen finden Sie im Thema „Kategorieregelsyntax” auf Seite 134.

Hinweis: Beispiele für das Abgleichen von Regeln mit Text finden Sie in „Beispiele für Kategorieregeln” auf Seite 140.

So erstellen Sie eine Regel

1. Wenn Sie noch keine Daten extrahiert haben oder Ihre Extraktion veraltet ist, extrahieren Sie die Daten jetzt. Weitere Informationen finden Sie im Thema „Extrahieren von Daten” auf Seite 94.

   Hinweis: Wenn Sie eine Extraktion so filtern, dass keine Konzepte mehr sichtbar sind, wird eine Fehlernachricht angezeigt, wenn Sie versuchen, eine Kategorieregel zu erstellen oder zu bearbeiten. Um dies zu verhindern, ändern Sie Ihren Extraktionsfilter, sodass Konzepte verfügbar sind.
2. Wählen Sie im Fensterbereich "Kategorien" die Kategorie aus, der Ihre Regel hinzugefügt werden soll.
3. Wählen Sie in den Menüs die Optionsfolge **Kategorien > Regel erstellen** aus. Im Fenster wird der Kategorieregel-Editorbereich geöffnet.
4. 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.
5. 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 Bereich "Extraktionsergebnisse" 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. Weitere Informationen finden Sie im Thema „Kategorieregelsyntax” auf Seite 134.

- Wählen Sie die Konzepte, Typen oder Muster aus, die Sie Regeln hinzufügen möchten, und verwenden Sie die Menüs.
- Boolesche Operatoren hinzufügen, um Elemente in Ihrer Regel miteinander zu verknüpfen. Über die Schaltflächen in der Symbolleiste können Sie die booleschen Operatoren "and" (**&**), "or" (**|**) und "not" (**!()**) sowie runde Klammern (**()**) und eckige Klammern für Muster (**[ ]**) Ihrer Regel hinzufügen.

6. Klicken Sie auf die Schaltfläche **Regel testen**, um zu überprüfen, ob Ihre Regel korrekt formuliert ist. Weitere Informationen finden Sie im Thema „Kategorieregelsyntax" auf Seite 134. Die Anzahl der gefundenen Dokumente oder 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 Fehlernachrichten. Ein rotes Fragezeichen neben dem Typ, Muster oder Konzept zeigt an, dass das Element mit keinen bekannten Extraktionen übereinstimmt. Falls dies der Fall ist, wird die Regel keine Datensätze finden.
7. Um einen Teil der Regel zu testen, wählen Sie den betreffenden Teil aus und klicken Sie auf **Auswahl testen**.
8. Nehmen Sie gegebenenfalls die erforderlichen Änderungen vor und testen Sie die Regel erneut, wenn Sie Probleme festgestellt haben.
9. 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. Weitere Informationen finden Sie im Thema „Kategorieregelsyntax" auf Seite 134.

Wenn Sie eine Regel nicht mehr benötigen, können Sie sie löschen.

So bearbeiten Sie Regeln

1. Wählen Sie die Regel im Dialogfeld "Kategoriedefinition" auf der Registerkarte "Deskriptoren" aus.
2. Wählen Sie in den Menüs die Optionsfolge **Kategorien > Regel bearbeiten** aus oder doppelklicken Sie auf den Regelnamen. Der Editor wird geöffnet und die ausgewählte Regel wird angezeigt.
3. Nehmen Sie Änderungen an der Regel über Extraktionsergebnisse und die Schaltflächen in der Symbolleiste vor.
4. Testen Sie die Regel erneut, um sicherzustellen, dass sie die erwarteten Ergebnisse liefert.
5. Klicken Sie auf **Speichern & Schließen**, um Ihre Regel erneut zu speichern und den Editor zu schließen.

So löschen Sie eine Regel

1. Wählen Sie die Regel im Dialogfeld "Kategoriedefinition" auf der Registerkarte "Deskriptoren" aus.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Löschen** aus. Die Regel wird aus der Kategorie gelöscht.

Import und Export vordefinierter 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 Workbenchsitzung auch in eine Microsoft Excel-Datei (**.xls*, **.xlsx*) exportieren. Wenn Sie Ihre Kategorien exportieren, können Sie einige zusätzliche Informationen einschließen oder ausschließen, z. B. Deskriptoren und Scores. Weitere Informationen finden Sie im Thema „Export von Kategorien" auf Seite 148.

Wenn Ihre vordefinierten Kategorien keine Codes haben bzw. Sie neue Codes wünschen, können Sie automatisch ein neues Set von Codes für das Kategorienset im Bereich "Kategorien" generieren, indem Sie

in den Menüs die Optionsfolge **Kategorien > Kategorien verwalten > Codes automatisch generieren** auswählen. Damit werden alle bestehenden Codes entfernt und automatisch neu nummeriert.

Import vordefinierter Kategorien

Sie können Ihre vordefinierten 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 auswählen, dass das Produkt das Format automatisch erkennen soll. Die folgenden Formate werden unterstützt:

- **Flaches Listenformat**: Weitere Informationen finden Sie im Thema „Flaches Listenformat" auf Seite 145.
- **Kompaktes Format:**: Weitere Informationen finden Sie im Thema „Kompaktes Format" auf Seite 146.
- **Eingerücktes Format:**: Weitere Informationen finden Sie im Thema „Eingerücktes Format" auf Seite 147.

So importieren Sie vordefinierte Kategorien

1. Wählen Sie in den Menüs der interaktiven Workbench die Optionsfolge **Kategorien > Kategorien verwalten > Vordefinierte Kategorien importieren** aus. Ein Assistent für den Import vordefinierter Kategorien wird angezeigt.
2. Wählen Sie in der Dropdown-Liste "Suchen in" das Laufwerk und den Ordner aus, in dem sich die Datei befindet.
3. Wählen Sie die Datei aus der Liste aus. Der Name der Datei wird im Feld "Dateiname" angezeigt.
4. Wählen Sie aus der Liste das Arbeitsblatt aus, das die vordefinierten Kategorien enthält. Der Name des Arbeitsblatts wird im Feld "Arbeitsblatt" angezeigt.
5. Klicken Sie auf **Weiter**, um zu beginnen, das Datenformat auszuwählen.
6. Wählen Sie das Format für Ihre Datei aus oder wählen Sie die Option aus, mit der angegeben wird, dass das Produkt versuchen soll, das Format automatisch zu erkennen. Die automatische Erkennung funktioniert am besten bei den häufigsten Formaten.
   - **Flaches Listenformat**: Weitere Informationen finden Sie im Thema „Flaches Listenformat" auf Seite 145.
   - **Kompaktes Format:**: Weitere Informationen finden Sie im Thema „Kompaktes Format" auf Seite 146.
   - **Eingerücktes Format:**: Weitere Informationen finden Sie im Thema „Eingerücktes Format" auf Seite 147.
7. Klicken Sie auf **Weiter**, um die zusätzlichen Importoptionen zu definieren. Wenn Sie die automatische Formaterkennung wählen, werden Sie zum letzten Schritt geführt.
8. Wenn mindestens eine Zeile Spaltenüberschriften oder andere irrelevante Informationen enthält, 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.
9. Wenn Ihre Datei Kategoriecodes enthält, wählen Sie die Option **Enthält Kategoriecodes** aus. Damit unterstützen Sie den Assistenten bei der korrekten Erkennung Ihrer Daten.
10. Überprüfen Sie die farbcodierten 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.
11. Um das Set von Kategorien und Unterkategorien zu überprüfen, das importiert wird, und um zu definieren, wie Deskriptoren für diese Kategorien erstellt werden, klicken Sie auf **Weiter**.
12. 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 angezeigt werden.

13. Wählen Sie aus, wie bereits vorhandene Kategorien in Ihrer Sitzung gehandhabt werden sollen.
    - **Alle vorhandenen Kategorien ersetzen**. Diese Option löscht alle vorhandenen Kategorien. Anschließend werden die neu 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 Möglichkeit (Option: **Zusammenführen**) besteht darin, alle importierten Kategorien mit vorhandenen Kategorien zusammenzuführen, wenn sie einen gemeinsamen Kategorienamen haben. Eine andere Möglichkeit (Option: **Von Import ausschließen**) besteht darin, den Import von Kategorien zu untersagen, wenn eine Kategorie mit demselben Namen vorhanden ist.
14. **Schlüsselwörter als Deskriptoren importieren** ist eine Option für den Import der Schlüsselwörter, die in Ihren Daten als Deskriptoren für die zugeordnete Kategorie identifiziert werden.
15. **Kategorien durch Ableiten von Deskriptoren erweitern** ist eine Option, die Deskriptoren aus den Wörtern generiert, 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 der 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 generieren, mit deren Hilfe die Kategorie Datensätze erfassen kann, die diese Deskriptoren enthalten.
    - Im Feld **Von** können Sie auswählen, von welchem Text die Deskriptoren abgeleitet werden, den Namen der Kategorien und Unterkategorien und/oder den Wörtern in den Anmerkungen.
    - Im Feld **Als** können Sie auswählen, ob diese Deskriptoren in der Form von Konzepten oder TLA-Mustern erstellt werden. Wenn keine TLA-Extraktion erfolgt ist, sind die Optionen für **Muster** in diesem Assistenten inaktiviert.
16. Um die vordefinierten Kategorien in den Bereich "Kategorien" zu importieren, klicken Sie auf **Fertigstellen**.

Flaches Listenformat

Im flachen Listenformat gibt es nur eine Ausgangsebene (höchste Ebene) von Kategorien ohne Hierarchie, d. h. ohne Unterkategorien oder Teilnetze. Kategoriennamen befinden sich in einer einzelnen Spalte.

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 jederzeit Codes mithilfe dieser Option (**Kategorien > Kategorien verwalten > Codes automatisch erzeugen**) generieren.
- 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 der einzelnen Kategorien 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 25. Flaches Listenformat mit Codes, Schlüsselwörtern und Anmerkungen

Spalte A	Spalte B	Spalte C
Kategoriecode (*optional*)	Kategoriename	Anmerkung
	_Deskriptor-/Schlüsselwortliste (*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.

Die folgenden Informationen können sich in einer Datei dieses Formats befinden:

- Die *erforderliche* Spalte **Codeebene** enthält Zahlen, die die hierarchische Position für die nachfolgenden Informationen in dieser Zeile angeben. 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 bis zur gewünschten Kategorientiefe beliebig weitergeführt werden.
- Die optionale Spalte **Codes** enthält Werte, die eine Kategorie eindeutig angeben. 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 jederzeit Codes mithilfe dieser Option (**Kategorien > Kategorien verwalten > Codes automatisch erzeugen**) generieren.
- 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 der einzelnen Kategorien 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 26. Beispiel für kompaktes Format mit Codes

Spalte A	Spalte B	Spalte C
Hierarchische Codeebene	Kategoriecode (*optional*)	Kategoriename
Hierarchische Codeebene	Unterkategoriecode (*optional*)	Unterkategoriename

Tabelle 27. Beispiel für kompaktes Format ohne Codes

Spalte A	Spalte B
Hierarchische Codeebene	Kategoriename
Hierarchische Codeebene	Unterkategoriename

Eingerücktes Format

Im eingerückten Dateiformat ist der Inhalt hierarchisch strukturiert, d. h., die Datei enthält Kategorien und mindestens eine Ebene 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.

- **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 28. Eingerückte Struktur mit Codes

Spalte A	**Spalte B**	**Spalte C**	**Spalte D**
Kategoriecode (*optional*)	Kategoriename		
	Unterkategoriecode (*optional*)	Unterkategoriename	
		Unter-Unterkategoriecode (*optional*)	Unter-Unterkategoriename

Tabelle 29. 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 jederzeit Codes mithilfe dieser Option (**Kategorien > Kategorien verwalten > Codes automatisch erzeugen**) generieren.
- 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 der einzelnen Kategorien 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.

Export von Kategorien

Sie können die vorhandenen Kategorien in einer geöffneten interaktiven Workbenchsitzung auch 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 Scorewert **Docs.** exportieren möchten.

Tabelle 30. Optionen für den Kategorieexport

Immer exportieren...	Optional exportieren...
• Kategoriecodes, falls vorhanden • Kategorienamen (und Unterkategorienamen) • Codeebenen, falls vorhanden (*Flaches/Kompaktes* Format) • Spaltenüberschriften (*Flaches/Kompaktes* Format)	• **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 TAP-Datei (Text Analysis Package) zu erstellen, da das TAP-Format alle derzeit 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. Weitere Informationen finden Sie im Thema „Verwendung von Text Analysis Packages".

So exportieren Sie vordefinierte Kategorien

1. Wählen Sie in den Menüs der interaktiven Workbench die Optionsfolge **Kategorien > Kategorien verwalten > Kategorien exportieren** aus. Ein Assistent für den Export von Kategorien wird angezeigt.
2. Wählen Sie die Position aus und geben Sie den Namen der Datei ein, die exportiert wird.
3. Geben Sie einen Namen für die Ausgabedatei in das Textfeld "Dateiname" ein.
4. Um das Format auszuwählen, in das Sie Ihre Kategoriendaten exportieren möchten, klicken Sie auf **Weiter**.
5. Wählen Sie das Format aus den folgenden Optionen aus:
   - Flaches/Kompaktes Listenformat: Weitere Informationen finden Sie im Thema „Flaches Listenformat" auf Seite 145. Eine flache Liste enthält keine Unterkategorien. Weitere Informationen finden Sie im Thema „Kompaktes Format" auf Seite 146. Kompaktes Listenformat enthält hierarchische Kategorien.
   - **Eingerücktes Format:**: Weitere Informationen finden Sie im Thema „Eingerücktes Format" auf Seite 147.
6. Um mit der Auswahl des zu exportierenden Inhalts zu beginnen und die vorgeschlagenen Daten zu prüfen, klicken Sie auf **Weiter**.
7. Prüfen Sie den Inhalt für die exportierte Datei.
8. Aktivieren oder inaktivieren Sie zusätzliche zu exportierende Inhaltseinstellungen wie **Anmerkungen** oder **Deskriptornamen**.
9. Um die Kategorien zu exportieren, klicken Sie auf **Fertigstellen**.

Verwendung von Text Analysis Packages

Ein Text Analysis Package, auch TAP genannt, dient als Vorlage für die Kategorisierung von Textantworten. 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 Codierung einer großen Anzahl von Datensätzen erforderlich sind. Mithilfe der linguistischen Ressourcen werden Textdaten analysiert und dem Mining-Verfah-

ren unterzogen, um die wichtigsten Konzepte zu extrahieren. Auf der 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 Kategorienset besteht im Wesentlichen aus vordefinierten Kategorien, Kategoriecodes, Deskriptoren für jede Kategorie und schließlich aus einem Namen für das gesamte Kategorienset. Deskriptoren sind linguistische Elemente (Konzepte, Typen, Muster und Regeln), so z. B. der Term *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 ein Set von Bibliotheken und erweiterten Ressourcen, die darauf abgestimmt sind, wichtige Konzepte und Muster zu extrahieren. Diese Extraktionskonzepte 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 Kategorienset ausgewählt wurde, kann IBM SPSS Modeler Text Analytics Ihre Datensätze extrahieren und kategorisieren.

Hinweis: TAPs können austauschbar zwischen IBM SPSS Text Analytics for Surveys und IBM SPSS Modeler Text Analytics erstellt und verwendet werden.

Erstellung von Text Analysis Packages

Wenn Sie über eine Sitzung mit mindestens einer Kategorie und einigen Ressourcen verfügen, können Sie aus dem Inhalt der geöffneten interaktiven Workbenchsitzung ein Text Analysis Package (TAP) erstellen. Das Set von 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" des Vorlageneditors oder des Ressourceneditors festgelegt.

So erstellen Sie ein Text Analysis Package

1. Wählen Sie in den Menüs die Optionsfolge **Datei > Text Analysis Packages > Paket erstellen** aus. Das Dialogfeld "Paket erstellen" wird angezeigt.
2. Wechseln Sie zu dem Verzeichnis, in dem Sie das TAP speichern möchten. Standardmäßig werden TAPs im Unterverzeichnis `\TAP` des Produktinstallationsverzeichnisses gespeichert.
3. Geben Sie im Feld **Dateiname** einen Namen für das TAP ein.
4. Geben Sie im Feld **Paketbeschriftung** eine Beschriftung ein. Wenn Sie einen Dateinamen eingeben, wird dieser Name automatisch als Beschriftung angezeigt, die Sie aber ändern können.
5. Um ein Kategorienset aus dem TAP auszuschließen, inaktivieren 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.
6. Benennen Sie Kategoriensets um. Die Spalte **Neues Kategorienset** enthält standardmäßig generische Namen, die durch Hinzufügen des Präfixes `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 geöffneten Sitzung wird nicht geändert.

7. Ändern Sie bei Bedarf die Reihenfolge der Kategoriensets mithilfe der Pfeiltasten rechts von der Tabelle "Kategorienset".
8. Klicken Sie auf **Speichern**, um das Text Analysis Package zu erstellen. Das Dialogfeld wird geschlossen.

Laden von Text Analysis Packages

Bei der Konfiguration eines Textmining-Modellierungsknotens müssen Sie angeben, welche Ressourcen für die Extraktion 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 Workbenchsitzung gestartet und dieses Kategorienset im Bereich "Kategorien" angezeigt. So können Sie Ihre Dokumente und Datensätze unmittelbar mithilfe dieser Kategorien scoren und diese Kategorien anschließend optimieren, erstellen und erweitern, bis sie Ihre Anforderungen erfüllen. Weitere Informationen finden Sie im Thema „Methoden und Strategien zur Erstellung von Kategorien" auf Seite 110.

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.

So laden Sie ein Text Analysis Package

1. Bearbeiten Sie den Textmining-Modellierungsknoten.
2. Wählen Sie auf der Registerkarte "Modell" im Abschnitt **Ressourcen kopieren von** die Option *Text Analysis Package* aus.
3. Klicken Sie auf **Laden**. Das Dialogfeld "Text Analysis Package laden" wird geöffnet.
4. Wechseln Sie an die Position des TAP mit den Ressourcen und dem Kategorienset, die Sie in diesen Knoten kopieren möchten. Standardmäßig werden TAPs im Unterverzeichnis `\TAP` des Produktinstallationsverzeichnisses gespeichert.
5. Geben Sie im Feld **Dateiname** einen Namen für das TAP ein. Die Beschriftung wird automatisch angezeigt.
6. Wählen Sie das Kategorienset aus, das Sie verwenden möchten. Dieses Kategorienset wird in der interaktiven Workbenchsitzung angezeigt. Sie können diese Kategorien anschließend manuell oder über die Optionen zum Erstellen und Erweitern von Kategorien optimieren und verbessern.
7. 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 Kategorienset 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. Dazu müssen Sie die Sitzung geöffnet haben, die die Informationen enthält, die Sie in das TAP aufnehmen möchten. Beim Aktualisieren können Sie auswählen, ob Sie Kategoriensets anfügen, Ressourcen ersetzen, die Paketbeschriftung ändern oder Kategoriensets umbenennen/neu ordnen möchten.

So aktualisieren Sie ein Text Analysis Package

1. Wählen Sie in den Menüs die Optionsfolge **Datei > Text Analysis Packages > Paket aktualisieren** aus. Das Dialogfeld "Paket aktualisieren" wird angezeigt.

2. Wechseln Sie in das Verzeichnis, das das Text Analysis Package enthält, das Sie aktualisieren möchten.
3. Geben Sie im Feld **Dateiname** einen Namen für das TAP ein.
4. Um die linguistischen Ressourcen im TAP durch die Ressourcen in der aktuellen Sitzung zu ersetzen, wählen Sie die Option **Die Ressourcen in diesem Paket durch die in der offenen Sitzung ersetzen** aus. Es ist im Allgemeinen sinnvoll, 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.
5. 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.
6. Um das neue Kategorienset aus der geöffneten 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.
7. Um Kategoriensets aus dem TAP zu entfernen, inaktivieren Sie das jeweilige Kontrollkästchen **Einschließen**. Unter Umständen möchten Sie ein Kategorienset, das bereits im TAP vorhanden war, entfernen, da Sie ein verbessertes Set hinzufügen. Inaktivieren 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.
8. 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 geöffneten Sitzung wird nicht geändert. Wenn zwei Kategoriensets denselben Namen haben, werden ihre Namen rot angezeigt, bis Sie das Duplikat korrigieren.
9. 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.
10. Klicken Sie auf **Aktualisieren**, um die Änderungen, die Sie am ausgewählten TAP vorgenommen haben, zu speichern.

So speichern Sie ein Text Analysis Package

1. Wechseln Sie zu dem Verzeichnis, in dem Sie die TAP-Datei speichern möchten. Standardmäßig werden TAP-Dateien im Unterverzeichnis "TAP" des Installationsverzeichnisses gespeichert.
2. Geben Sie im Feld "Dateiname" einen Namen für die TAP-Datei ein.
3. Geben Sie im Feld "Paketbeschriftung" eine Beschriftung ein. Wenn Sie einen Dateinamen eingeben, wird dieser Name automatisch als Beschriftung verwendet. Sie können diese Beschriftung jedoch umbenennen. Sie müssen eine Beschriftung haben.
4. Klicken Sie auf **Speichern**, um das neue Paket zu erstellen.

Bearbeiten und Optimieren von Kategorien

Nachdem Sie einige Kategorien erstellt haben, sollten Sie diese untersuchen und Anpassungen vornehmen. Neben der Optimierung 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 vorneh-

men wollen. Nach der Verwendung mindestens eines Verfahrens 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. Weitere Informationen finden Sie im Thema „Erläuterung von Kategorien” auf Seite 115.

Hier finden Sie einige Optionen zur Optimierung Ihrer Kategorien, von denen die meisten auf den folgenden Seiten beschrieben werden:

Hinzufügen von Deskriptoren zu Kategorien

Nach der Verwendung von automatisierten Verfahren liegen Ihnen mit großer Wahrscheinlichkeit noch Extraktionsergebnisse vor, die in keiner der Kategoriedefinitionen verwendet wurden. Sie sollten die betreffende Liste im Bereich "Extraktionsergebnisse" 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

1. Wählen Sie im Bereich "Extraktionsergebnisse" und im Datenbereich die Elemente aus, die Sie einer neuen oder einer vorhandenen Kategorie hinzufügen möchten.
2. Wählen Sie in den Menüs **Kategorien > Zu Kategorie hinzufügen** aus. Im Dialogfeld "Alle Kategorien" wird das Set von Kategorien 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.

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

1. Wählen Sie die zu bearbeitende Kategorie im Fensterbereich "Kategorien" aus.
2. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Kategoriedefinitionen** aus. Das Dialogfeld "Kategoriedefinitionen" wird geöffnet.
3. Wählen Sie den zu bearbeitenden Deskriptor aus und klicken Sie auf die entsprechende Schaltfläche in der Symbolleiste.

In der folgenden Tabelle sind die einzelnen Symbolleistenschaltflächen beschrieben, mit denen Sie die Kategoriedefinitionen bearbeiten können.

Tabelle 31. 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. Weitere Informationen finden Sie im Thema „Verwenden von Kategorieregeln” auf Seite 134.
	Verschiebt die einzelnen ausgewählten Deskriptoren jeweils als eigene neue Kategorie

Tabelle 31. Symbolleistenschaltflächen und Beschreibungen (Forts.).

Symbole	Beschreibung
 **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

1. Wählen Sie im Fensterbereich "Kategorien" die Kategorien aus, die Sie in eine andere Kategorie verschieben möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Kategorien > In Kategorie verschieben** aus. 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.

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.

So glätten Sie eine Kategorie

1. Wählen Sie im Kategoriebereich die Kategorie (oberste Ebene oder Unterkategorie) aus, die Sie glätten möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Kategorien > Kategorien glätten** aus. Die Unterkategorien werden entfernt und die Deskriptoren werden in der ausgewählten Kategorie zusammengeführt.

Zusammenführen bzw. Kombinieren von Kategorien

Um mindestens zwei vorhandene Kategorien in einer neuen Kategorie zu kombinieren, können Sie sie zusammenführen. Wenn Sie Kategorien zusammenführen, wird eine neue Kategorie mit einem generischen Namen erstellt. Alle Konzepte, Typen 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

1. Wählen Sie im Fensterbereich "Kategorien" die Elemente aus, die zusammengeführt werden sollen.
2. Wählen Sie in den Menüs die Optionsfolge **Kategorien > Kategorien zusammenführen** aus. Das Dialogfeld "Kategorieeigenschaften" wird angezeigt und Sie können einen Namen für die neu erstellte Kategorie eingeben. Die ausgewählten Kategorien werden in der neuen Kategorie als Unterkategorien zusammengeführt.

Löschen von Kategorien

Wenn Sie eine Kategorie nicht mehr benötigen, können Sie sie löschen.

So löschen Sie eine Kategorie

1. Wählen Sie im Fensterbereich "Kategorien" die zu löschende(n) Kategorie(n) aus.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Löschen** aus.

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-/Datensatzset 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 Datasets kann diese Technik erheblich längere Verarbeitungszeiten nach sich ziehen.

Hinweis: Verwenden Sie die Option **Maximale Anzahl an Dokumenten für die Berechnung von Clustern** im Dialogfeld "Cluster erstellen", um Cluster mit nur einem Subset aller Dokumente oder Datensätze zu erstellen.

Das Clustering ist ein Prozess, der mit der Analyse eines Sets 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. Weitere Informationen finden Sie im Thema „Berechnen von Werten für Ähnlichkeitszusammenhänge” auf Seite 158.

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).

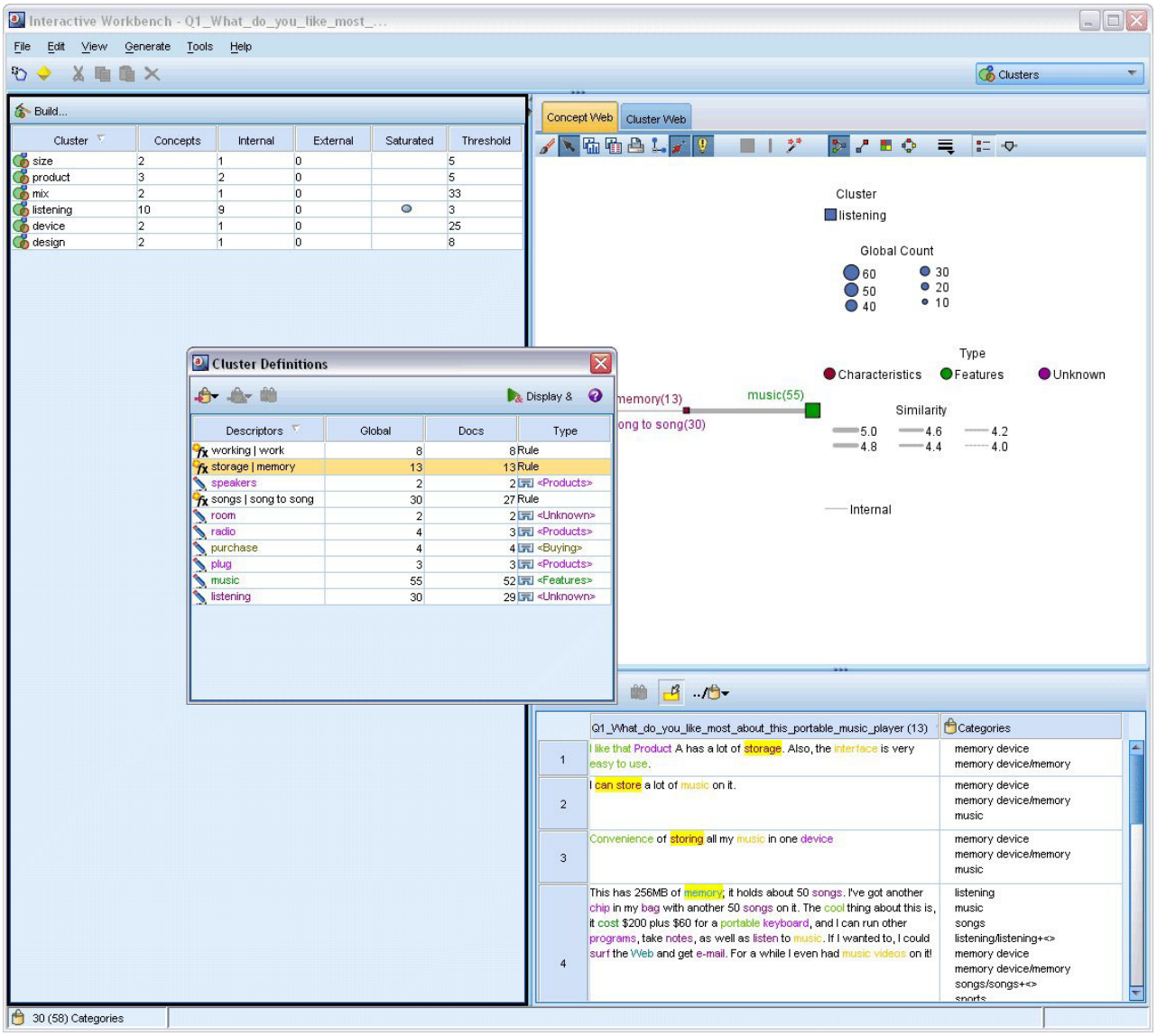


Abbildung 30. Clusteransicht

Die Clusteransicht ist in drei Bereiche unterteilt, die jeweils durch Auswahl des entsprechenden Namens in der Clusteransicht aus- oder eingeblendet werden können.

- **Clusterbereich.** In diesem Bereich erstellen und verwalten Sie Ihre Cluster. Weitere Informationen finden Sie im Thema „Untersuchen von Clustern” auf Seite 159.
- **Visualisierungsbereich.** In diesem Bereich können Sie Ihre Cluster und deren Interaktionen visuell untersuchen. Weitere Informationen finden Sie im Thema „Clusterdiagramme” auf Seite 171.
- **Datenbereich.** Hier können Sie Text, der in den Dokumenten und Datensätzen enthalten ist, die im Dialogfeld "Clusterdefinitionen" ausgewählt wurden, untersuchen und prüfen. Weitere Informationen finden Sie im Thema „Clusterdefinitionen” auf Seite 160.

Erstellen von Clustern

Wenn Sie die Clusteransicht zum ersten Mal aufrufen, werden keine Cluster angezeigt. Sie können Cluster über die Menüs (**Tools > Cluster erstellen**) erstellen oder indem Sie in der Symbolleiste auf die Schaltfläche **Erstellen...** klicken. Daraufhin wird das Dialogfeld "Cluster erstellen" geöffnet, in dem Sie die Einstellungen und Grenzwerte für das Erstellen Ihrer Cluster definieren können.

Hinweis: Wenn die Extraktionsergebnisse nicht mehr mit den Ressourcen übereinstimmen, wird dieser Bereich, ebenso wie der Bereich "Extraktionsergebnisse", gelb dargestellt. Sie können dann eine erneute Extraktion durchführen, um die neusten Extraktionsergebnisse zu erhalten. Der Bereich wird anschließend nicht mehr gelb angezeigt. Bei jeder Extraktion wird jedoch der Clusterbereich gelöscht. Anschließend müssen Sie Ihre Cluster neu erstellen. Cluster werden nicht von einer Sitzung zur nächsten gespeichert.

Die folgenden Bereiche und Felder sind im Dialogfeld "Cluster erstellen" verfügbar:

Eingaben

Tabelle **Eingaben**. Cluster werden aus Deskriptoren erstellt, die aus bestimmten Typen abgeleitet wurden. Sie können in der Tabelle die Typen auswählen, die in den Erstellungsprozess eingeschlossen werden sollen. Standardmäßig sind die Typen ausgewählt, die die meisten Datensätze oder Dokumente erfassen.

Konzepte für das Clustering: Wählen Sie die Auswahlmethode 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 obersten Konzepten, einem Prozentsatz von obersten Konzepten oder mit allen Konzepten durchführen:

- **Zahl basierend auf Dokumentanzahl.** Wenn Sie **Obere Anzahl von 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 Dokumentanzahl.** 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.

Maximale Anzahl an Dokumenten für die Berechnung 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.

Ausgabegrenzwerte

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. Weitere Informationen finden Sie im Thema „Berechnen von Werten für Ähnlichkeitszusammenhänge".

Paarbildung spezifischer Konzepte verhindern. Wählen Sie dieses Kontrollkästchen aus, um den Vorgang der Gruppierung oder Paarbildung von zwei Konzepten in der Ausgabe zu unterbinden. Klicken Sie zum Erstellen oder Verwalten von Konzeptpaaren auf **Paare verwalten**. Weitere Informationen finden Sie im Thema „Verwalten von Linkausnahmepaaren" auf Seite 123.

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 der Dokumentenanzahl 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.

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

Abbildung 31. Formel des Ähnlichkeitskoeffizienten

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 Dokumentenset 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 Szenarios, die verdeutlichen, wie der Koeffizient und der Zusammenhangswert berechnet werden.

Tabelle 32. 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

Tabelle 32. Beispiel für Konzepthäufigkeiten (Forts.)

Konzept/Paar	Szenario A	Szenario B
Konzeptpaar: IJ	Kommt in 20 Dokumenten gemeinsam vor	Kommt in 20 Dokumenten gemeinsam vor
Ähnlichkeitskoeffizient	1	0,22222
Ähnlichkeitszusammenhangswert	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. Weitere Informationen finden Sie im Thema „Clusterdefinitionen" auf Seite 160.
- **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. Weitere Informationen finden Sie im Thema „Erstellen von Clustern" auf Seite 156.
- **Schwellenwert.** Für alle im Cluster enthaltenen gemeinsam vorkommenden Konzeptpaare ist dies der niedrigste Ähnlichkeitszusammenhangswert von allen im Cluster vorhandenen. Weitere Informationen finden Sie im Thema „Berechnen von Werten für Ähnlichkeitszusammenhänge" auf Seite 158. 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. Weitere Informationen finden Sie im Thema „Clusterdiagramme" auf Seite 171. Sie können den Inhalt der Tabelle auch ausschneiden und in eine andere Anwendung einfügen.

Wenn die Extraktionsergebnisse nicht mehr mit den Ressourcen übereinstimmen, wird der Bereich, ebenso wie der Bereich "Extraktionsergebnisse", gelb dargestellt. Sie können dann eine erneute Extraktion durchführen, um die neusten Extraktionsergebnisse zu erhalten. Der Bereich wird anschließend nicht

mehr gelb angezeigt. Bei jeder Extraktion 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**).

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 in „Datenbereich" auf Seite 116.

Wenn Sie in diesem Dialogfeld Konzepte auswählen, ändert dies auch das Netzdiagramm des Konzepts. Weitere Informationen finden Sie im Thema „Clusterdiagramme" auf Seite 171. Wenn Sie im Dialogfeld "Clusterdefinitionen" mindestens ein Konzept auswählen, werden für dieses Konzept im Visualisierungsbereich die externen und internen Zusammenhänge angezeigt.

Spaltenbeschreibungen

Es werden Symbole angezeigt, über die Sie die einzelnen Deskriptoren problemlos identifizieren können.

Tabelle 33. Symbole für Spalten und Deskriptoren

Spalten	Beschreibung
Deskriptoren	Der Name des Konzepts.
 **Globalwert**	Zeigt, wie oft dieser Deskriptor im gesamten Dataset 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 als Kategorieregel hinzuzufügen. Weitere Informationen finden Sie im Thema „Kookkurrenzregeln" auf Seite 127. Über die Schaltflächen in der Symbolleiste können Sie die Konzepte Kategorien hinzufügen.

Tabelle 34. Schaltflächen in 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 einer neuen oder einer vorhandenen Kategorie hinzu. Weitere Informationen finden Sie im Thema „Verwenden von Kategorieregeln" auf Seite 134.
	Fügt alle ausgewählten Konzepte als eigene neue Kategorie hinzu

Tabelle 34. Schaltflächen in der Symbolleiste zum Hinzufügen von Konzepten zu Kategorien (Forts.)

Symbole	Beschreibung
	Aktualisiert die Anzeige des Daten- und des Visualisierungsbereichs in Funktion der ausgewählten Deskriptoren

Hinweis: Sie können Konzepte auch mithilfe der Kontextmenüs einem Typ hinzufügen, als Synonyme oder als Ausschlusselemente.

Kapitel 12. Untersuchen von Textlinkanalyse

In der Ansicht "Textlinkanalyse" (TLA) können Sie Musterergebnisse für die Textlinkanalyse untersuchen. Die Textlinkanalyse ist ein Verfahren zum Musterabgleich, mit der Sie Musterregeln definieren und mit tatsächlich extrahierten Konzepten und Beziehungen, die in Ihrem Text gefunden werden, vergleichen können.

Beispiel: Die bloße Extraktion von allgemeinen Daten zu einem Unternehmen ist für Sie wenig aussagekräftig. Mit der TLA erkennen Sie gegebenenfalls 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-Musterergebnisse extrahiert haben, können Sie diese in den Typ- und Konzeptmusterbereichen der Ansicht "Textlinkanalyse" überprüfen. Weitere Informationen finden Sie im Thema „Typ- und Konzeptmuster" auf Seite 165. Außerdem können Sie sie im Daten- oder Visualisierungsbereich 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 **Musterextraktion für Textlinkanalyse aktivieren** im Dialogfeld "Extraktionseinstellungen" auswählen. Weitere Informationen finden Sie im Thema „Extrahieren von TLA-Musterergebnissen" auf Seite 164.

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 im Lieferumfang von IBM SPSS Modeler Text Analytics enthalten sind. 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 Textsprachen *außer* Japanisch definieren. Muster bestehen aus Makros, Wortlisten sowie Wortlücken und bilden eine boolesche Abfrage (Regel), die mit dem Eingangstext abgeglichen wird. Weitere Informationen finden Sie im Thema Kapitel 19, „Textlinkregeln", auf Seite 227.

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 Ansicht "Textlinkanalyse" sichtbar. Jeder Bereich kann aus- oder eingeblendet werden, indem Sie seinen Namen aus dem Menü "Ansicht" auswählen:

- **Typ- und Konzeptmusterbereiche.** In diesen beiden Bereichen können Sie Ihre Muster erstellen und untersuchen. Weitere Informationen finden Sie im Thema „Typ- und Konzeptmuster" auf Seite 165.
- **Visualisierungsbereich.** In diesem Bereich können Sie die Interaktionen der Konzepte und Typen in Ihren Mustern visuell untersuchen. Weitere Informationen finden Sie im Thema „Textlinkanalyse-Diagramme" auf Seite 172.
- **Datenbereich.** Sie können Text, der in Dokumenten und Datensätzen enthalten ist, die mit der Auswahl in einem anderen Bereich übereinstimmen, untersuchen und überprüfen. Weitere Informationen finden Sie im Thema „Datenbereich" auf Seite 167.

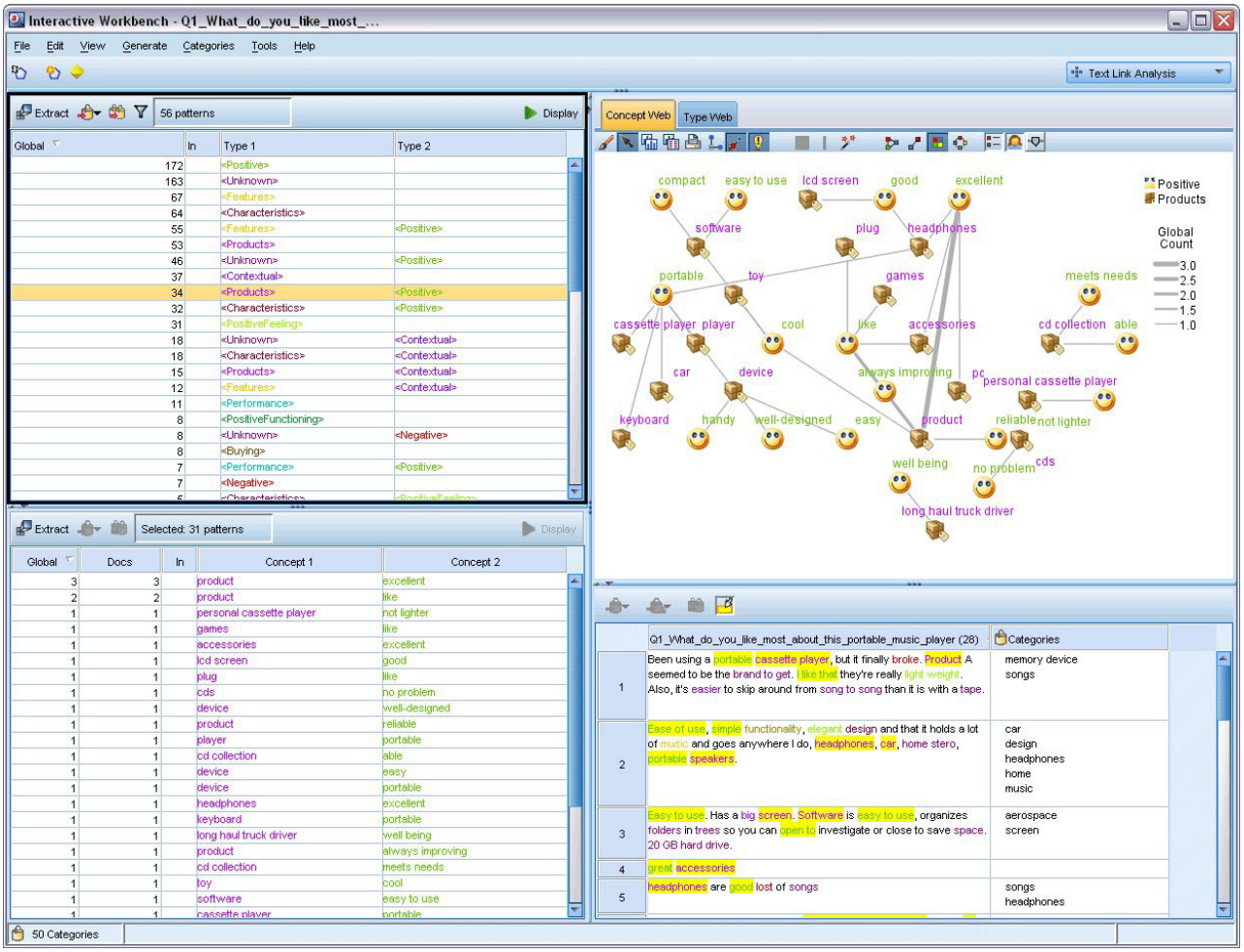


Abbildung 32. Ansicht "Textlinkanalyse"

Extrahieren von TLA-Musterergebnissen

Der Extraktionsprozess liefert eine Reihe von Konzepten und Typen und, sofern aktiviert, Textlinkanalysemuster (TLA-Muster). Extrahierte TLA-Muster können Sie in der Ansicht "Textlinkanalyse" sehen. Wenn die Extraktionsergebnisse nicht mit den Ressourcen übereinstimmen, färben sich die Musterbereiche gelb und signalisieren damit, dass eine erneute Extraktion andere Ergebnisse liefern würde.

Sie müssen die Extraktion dieser Muster in der Knoteneinstellung oder im Dialogfeld "Extrahieren" über die Option **Musterextraktion für Textlinkanalyse aktivieren** auswählen. Weitere Informationen finden Sie im Thema „Extrahieren von Daten” auf Seite 94.

Hinweis: Die für den Extraktionsprozess benötigte Zeit steht in direkter Beziehung zur Größe Ihres Datasets. 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

1. Wählen Sie in den Menüs die Optionsfolge **Tools > Extrahieren** aus. Alternativ können Sie auf die Symbolleistenschaltfläche **Extrahieren** klicken.

2. Ändern Sie nach Bedarf die Optionen, die Sie verwenden möchten. Beachten Sie, dass die Option **Musterextraktion 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. Weitere Informationen finden Sie im Thema „Extrahieren von Daten" auf Seite 94.
3. Klicken Sie auf **Extrahieren**, um die Extraktion zu starten.

Sobald die Extraktion beginnt, öffnet sich die Statusanzeige. Wenn Sie die Extraktion abbrechen möchten, klicken Sie auf **Abbrechen**. Beim Abschluss der Extraktion wird das Dialogfeld geschlossen und die Ergebnisse werden im Bereich angezeigt. Weitere Informationen finden Sie im Thema „Typ- und Konzeptmuster".

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. So könnte es beispielsweise sein, dass es Ihnen nicht ausreicht, 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.

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. Weitere Informationen finden Sie im Thema Kapitel 19, „Textlinkregeln", auf Seite 227.

Bei der Extraktion 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 als `<Organization>` + `<Location>` + `<Positive>` angezeigt, 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 Extraktionsergebnissen in der Ansicht "Kategorien und Konzepte" können Sie hier die Ergebnisse überprüfen. Wenn Sie die Typen und Konzepte, aus denen diese Muster bestehen, optimieren möchten, können Sie diese Optimierung im Bereich "Extraktionsergebnisse" in der Ansicht "Kategorien und Konzepte" oder direkt im Ressourceneditor vornehmen und Ihre Muster erneut extrahieren. Wenn ein

Konzept, ein Typ oder Muster in einer Kategoriedefinition als solches oder als Teil einer Regel verwendet wird, wird ein Kategorie- oder Regelsymbol in der Spalte **In** in der Muster- oder Extraktionsergebnistabelle angezeigt.

Filtern von TLA-Ergebnissen

Wenn Sie mit sehr großen Datasets arbeiten, kann der Extraktionsprozess 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.

In der TLA-Ansicht enthält das Dialogfeld "Filter" die folgenden Bereiche und Felder.

Nach Häufigkeit filtern. Mit diesem Filter werden nur Ergebnisse mit einem bestimmten globalen Häufigkeitswert 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 gefunden wird, und wird in der Spalte **Dokumente** angezeigt.

Beispiel: Wenn ein Muster 300-mal in 500 Datensätzen gefunden 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 35. 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.
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 (angenommen, Sie verwenden eine englische Version der Software).

Abbildung 33. Filterergebnisse, Beispiel 1

In diesem Beispiel wird in der Symbolleiste angezeigt, dass die Anzahl der zurückgegebenen Muster aufgrund des im Filter angegebenen 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 34. Filterergebnisse, Beispiel 2

In diesem Beispiel 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

1. Wählen Sie in den Menüs die Optionsfolge **Tools > Filter** aus. Das Filterdialogfeld wird geöffnet.
2. Wählen und optimieren Sie die Filter, die Sie verwenden möchten.
3. Klicken Sie auf **OK**, um die Filter anzuwenden und die neuen Ergebnisse anzuzeigen.

Datenbereich

Bei der Extraktion und Exploration von Textlinkanalyse-Mustern sollen manche Daten, mit denen Sie arbeiten, überprüft werden. Beispielsweise sollen die tatsächlichen Datensätze, in denen eine Gruppe von Mustern gefunden wurde, angezeigt werden. Sie können Datensätze oder Dokumente im Datenbereich überprüfen, der sich unten rechts befindet. Wird dieser nicht standardmäßig angezeigt, wählen Sie die Optionsfolge **Ansicht > Fenster > Daten** in den Menüs aus.

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. Weitere Informationen finden Sie im Thema „Optionen: Registerkarte "Sitzung"" auf Seite 88.

Datenbereich anzeigen und aktualisieren

Die Anzeige des Datenbereichs wird nicht automatisch aktualisiert, da die automatische Datenaktualisierung bei umfangreicheren Datasets sehr zeitaufwendig sein könnte. Daher können Sie bei der Auswahl von Typ- oder Konzeptmustern in dieser Ansicht auf **Anzeigen** 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 Datasets 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.

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 Farbcodierung entspricht den Typen, die den Konzepten zugewiesen sind. Alternativ können Sie die Maus über farbcodierte 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 gegebenenfalls 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 Datenbereich 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**. Weitere Informationen finden Sie im Thema „Optionen: Registerkarte "Sitzung"” auf Seite 88.
- **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. Weitere Informationen finden Sie im Thema „Kategorierelevanz” auf Seite 117.
- **Kategorieanzahl.** Führt die Anzahl der Kategorien auf, denen ein Datensatz angehört.

Kapitel 13. Visualisierung von Diagrammen

Die Kategorie- und Konzeptansicht, die Clusteransicht und die Textlinkanalyseansicht 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.

- **Ansicht "Kategorien und Konzepte".** Diese Ansicht enthält drei Diagramme: *Kategoriebalken*, *Kategorienetzdiagramm* und *Tabelle für Kategorienetzdiagramm*. In dieser Ansicht werden die Diagramme nur aktualisiert, wenn Sie auf **Anzeigen** klicken. Weitere Informationen finden Sie im Thema „Kategoriendiagramme und Grafiken".
- **Ansicht "Cluster".** Diese Ansicht enthält zwei Netzdiagramme: *Konzeptnetzdiagramm* und *Clusternetzdiagramm*. Weitere Informationen finden Sie im Thema „Clusterdiagramme" auf Seite 171.
- **Ansicht "Textlinkanalyse".** Diese Ansicht enthält zwei Netzdiagramme: *Konzeptnetzdiagramm* und *Typnetzdiagramm*. Weitere Informationen finden Sie im Thema „Textlinkanalyse-Diagramme" auf Seite 172.

Weitere Informationen zu allen zum Bearbeiten von Diagrammen verwendeten allgemeinen Symbolleisten und Paletten finden Sie im Abschnitt über das Bearbeiten von Diagrammen in der Onlinehilfe oder in der Datei *modeler_nodes_general_book.pdf*, die sich im Ordner *\Documentation\en* auf der IBM SPSS Modeler-DVD 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 in der Ansicht "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 Optimierung 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 "Extraktionsergebnisse" oder "Kategorien" bzw. 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 in der Ansicht "Kategorien und Konzepte" bietet die folgenden Diagramme:

- **Kategoriebalkendiagramm.** In einer Tabelle und einem Balkendiagramm werden gemäß Ihrer Auswahl und den zugeordneten Kategorien die Überschneidungen zwischen den Dokumenten/Datensätzen dargestellt. Im Balkendiagramm wird auch das jeweilige Verhältnis der Dokumente/Datensätze in Kategorien zu der Gesamtzahl der Dokumente/Datensätze angezeigt. Weitere Informationen finden Sie im Thema „Kategoriebalkendiagramm" auf Seite 170.

- **Kategorienetzdiagramm.** In diesem Diagramm wird die Überschneidung der Dokumente/Datensätze der Kategorien angezeigt, denen die Dokumente/Datensätze gemäß der Auswahl in den anderen Bereichen zugeordnet sind. Weitere Informationen finden Sie im Thema „Kategorienetzdiagramm".
- **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. Weitere Informationen finden Sie im Thema „Tabelle für Kategorienetzdiagramm".

Weitere Informationen finden Sie im Thema Kapitel 10, „Kategorisieren von Textdaten", auf Seite 107.

Kategoriebalkendiagramm

Auf dieser Registerkarte werden eine Tabelle und ein Balkendiagramm angezeigt, die gemäß Ihrer Auswahl und den zugeordneten Kategorien die Überschneidungen zwischen den Dokumenten/Datensätzen anzeigen. Im Balkendiagramm wird auch das jeweilige Verhältnis der Dokumente/Datensätze in Kategorien zu der Gesamtzahl der Dokumente oder Datensätze angezeigt. 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.** In dieser Spalte wird auf visuelle Art das jeweilige Verhältnis der Dokumente oder Datensätze in einer Kategorie zu der Gesamtzahl der Dokumente oder Datensätze angezeigt.
- **Auswahl %.** In dieser Spalte wird ein 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 angezeigt.
- **Dokumente.** In dieser Spalte wird die Anzahl der Dokumente oder Datensätze in einer Auswahl für die jeweilige Kategorie angezeigt.

Kategorienetzdiagramm

Auf dieser Registerkarte wird ein Kategorienetzdiagramm angezeigt. Im Netzdiagramm wird die Überschneidung der Dokumente oder Datensätze der Kategorien angezeigt, denen die Dokumente bzw. Datensätze gemäß der Auswahl in den anderen Bereichen zugeordnet sind. Vorhandene Kategoriebeschriftungen werden im Diagramm angezeigt. Sie können über die Symbolleistenschaltflächen in diesem Bereich das Layout des Diagramms auswählen (Netz-, Kreis-, Rasterlayout oder gerichtetes Layout).

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 einem Knoten platzieren, zeigt eine QuickInfo den Namen (oder die Beschriftung) 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. Weitere Informationen finden Sie im Thema „Verwenden von Diagrammsymbolleisten und Paletten" auf Seite 173.

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.** In dieser Spalte wird der Name der ersten Kategorie angezeigt, gefolgt von der - in Klammern eingeschlossenen - Gesamtzahl der darin enthaltenen Dokumente oder Datensätze.
- **Kategorie 2.** In dieser Spalte wird der Name der zweiten Kategorie angezeigt, gefolgt von der - in Klammern eingeschlossenen - Gesamtzahl der darin enthaltenen Dokumente oder Datensätze.

Clusterdiagramme

Nach dem Erstellen der Cluster können Sie diese visuell in den Netzdiagrammen im Visualisierungsbereich untersuchen. Im Visualisierungsbereich sind zwei Clustering-Perspektiven 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.

Hinweis: 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. Weitere Informationen finden Sie im Thema „Verwenden von Diagrammsymbolleisten und Paletten” auf Seite 173.

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. Weitere Informationen finden Sie im Thema „Konzeptnetzdiagramm”.
- **Clusternetzdiagramm**. Dieses Diagramm zeigt den oder die ausgewählten Cluster und alle externen Links zwischen den ausgewählten Clustern als gepunktete Linien an. Weitere Informationen finden Sie im Thema „Clusternetzdiagramm” auf Seite 172.

Weitere Informationen finden Sie im Thema Kapitel 11, „Analyse von Clustern”, auf Seite 155.

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 farbcodiert ist. Weitere Informationen finden Sie im Thema „Erstellen von Typen” auf Seite 203.

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.

Hinweis: 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. Weitere Informationen finden Sie im Thema „Verwenden von Diagrammsymbolleisten und Paletten” auf Seite 173.

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.

Hinweis: 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. Weitere Informationen finden Sie im Thema „Verwenden von Diagrammsymbolleisten und Paletten” auf Seite 173.

Textlinkanalyse-Diagramme

Nach dem Extrahieren Ihrer Textlinkanalyse-Muster (TLA-Muster) können Sie diese visuell in den Netzdiagrammen im Visualisierungsbereich untersuchen. Der Visualisierungsbereich bietet zwei Perspektiven der TLA-Muster: ein Konzeptnetzdiagramm und ein Typnetzdiagramm. 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.

Hinweis: 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. Weitere Informationen finden Sie im Thema „Verwenden von Diagrammsymbolleisten und Paletten” auf Seite 173.

In der Textlinkanalyseansicht 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 Vorkommen in der ausgewählten Tabelle an. Weitere Informationen finden Sie im Thema „Konzeptnetzdiagramm” auf Seite 173.
- **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 Vorkommen in der ausgewählten Tabelle an. Knoten werden entweder durch eine Typfarbe oder durch ein Symbol dargestellt. Weitere Informationen finden Sie im Thema „Typnetzdiagramm” auf Seite 173.

Weitere Informationen finden Sie im Thema Kapitel 12, „Untersuchen von Textlinkanalyse”, auf Seite 163.

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. Weitere Informationen finden Sie im Thema „Verwenden von Diagrammsymbolleisten und Paletten".

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 gefunden 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. Weitere Informationen finden Sie im Thema „Verwenden von Diagrammsymbolleisten und Paletten".

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 Farben so anpassen, dass sie den in Ihrem Unternehmen geltenden Stilrichtlinien entsprechen. In diesen Modus wechseln Sie, indem Sie in den Menüs die Optionsfolge **Ansicht > Fensterbereich 'Visualisierung' > Bearbeitungsmodus** auswählen (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 feststellen, dass Sie nicht alle davon verwenden, können Sie die nicht benötigten Symbolleisten ausblenden, um den Platz in dem Dialogfeld zu vergrößern, in dem das Diagramm angezeigt wird. Sie wählen Symbolleisten aus bzw. heben deren Auswahl auf, indem Sie im Menü "Ansicht" auf den entsprechenden Symbolleisten- oder Palettennamen klicken.

Weitere Informationen zu allen zum Bearbeiten von Diagrammen verwendeten allgemeinen Symbolleisten und Paletten finden Sie im Abschnitt über das Bearbeiten von Diagrammen in der Onlinehilfe oder in der Datei *modeler_nodes_general_book.pdf*, die sich im Ordner *\Documentation\en* auf der IBM SPSS Modeler-DVD befindet.

Tabelle 36. Schaltflächen in 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 Beschriftungen 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.

Tabelle 36. Schaltflächen in der Symbolleiste "Textanalyse" (Forts.).

Schaltfläche/Liste	Beschreibung
	Wählen Sie die Art der Netzanzeige für die Diagramme in der Ansicht "Kategorien und Konzepte" und in der Ansicht "Textlinkanalyse" 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. • **Rasterlayout.** 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 aus, 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 Textlinkanalyseansicht.
	Eine Umschaltfläche, mit der die Links-Schiebeleiste 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 für deren Unterkategorien an.
	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 Extraktionsprozess beruht erheblich darauf, dass durch die linguistischen Ressourcen festgelegt wird, wie Informationen aus Textdaten extrahiert werden. Standardmäßig stammen diese Ressourcen aus Ressourcenvorlagen.

IBM SPSS Modeler Text Analytics wird mit einer Reihe von spezialisierten **Ressourcenvorlagen** geliefert, die in Form von Bibliotheken und erweiterten Ressourcen ein Set von linguistischen und nicht linguistischen Ressourcen enthalten. Diese unterstützen Sie bei der Definition der Art und Weise, wie Ihre Daten gehandhabt und extrahiert werden sollen. Weitere Informationen finden Sie im Thema Kapitel 15, „Vorlagen und Ressourcen", auf Seite 179.

Im Knotendialogfeld können Sie eine Kopie der Ressourcen der Vorlage in den Knoten laden. Sobald Sie eine interaktive Workbenchsitzung gestartet haben, können Sie diese Ressourcen bei Bedarf an die Daten dieses Knotens anpassen. Während einer interaktiven Workbenchsitzung können Sie in der Ressourceneditoransicht mit Ihren Ressourcen arbeiten. Wenn Sie eine interaktive Sitzung starten, erfolgt eine Extraktion mit den Ressourcen, die im Knotendialogfeld geladen wurden, sofern Sie die Daten und Extraktionsergebnisse im Knoten nicht zwischengespeichert haben.

Bearbeiten von Ressourcen im Ressourceneditor

Der Ressourceneditor ermöglicht den Zugriff auf ein Set von Ressourcen, mit denen die Extraktionsergebnisse (Konzepte, Typen und Muster) für eine interaktive Workbenchsitzung erzeugt werden. Dieser Editor ist dem Vorlageneditor sehr ähnlich, im Ressourceneditor 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 einer nachfolgenden interaktiven Workbenchsitzung wiederhergestellt werden. Weitere Informationen finden Sie im Thema „Aktualisieren von Modellierungsknoten und Speichern" auf Seite 90.

Falls Sie die Vorlagen direkt bearbeiten möchten, die zum Laden von Ressourcen in Knoten verwendet werden, wird die Verwendung des Vorlageneditors empfohlen. Viele der Aufgaben, die Sie im Ressourceneditor durchführen können, werden auf dieselbe Weise durchgeführt wie im Vorlageneditor. Dazu zählen die folgenden Aufgaben:

- **Arbeiten mit Bibliotheken**. Weitere Informationen finden Sie im Thema Kapitel 16, „Arbeiten mit Bibliotheken", auf Seite 191.
- **Erstellen von Typwörterbüchern**. Weitere Informationen finden Sie im Thema „Erstellen von Typen" auf Seite 203.
- **Hinzufügen von Termen zu Wörterbüchern**. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen" auf Seite 204.
- **Erstellen von Synonymen**. Weitere Informationen finden Sie im Thema „Definieren von Synonymen" auf Seite 209.
- **Importieren und Exportieren von Vorlagen**. Weitere Informationen finden Sie im Thema „Import und Export von Vorlagen" auf Seite 187.
- **Veröffentlichen von Bibliotheken**. Weitere Informationen finden Sie im Thema „Veröffentlichen von Bibliotheken" auf Seite 197.

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

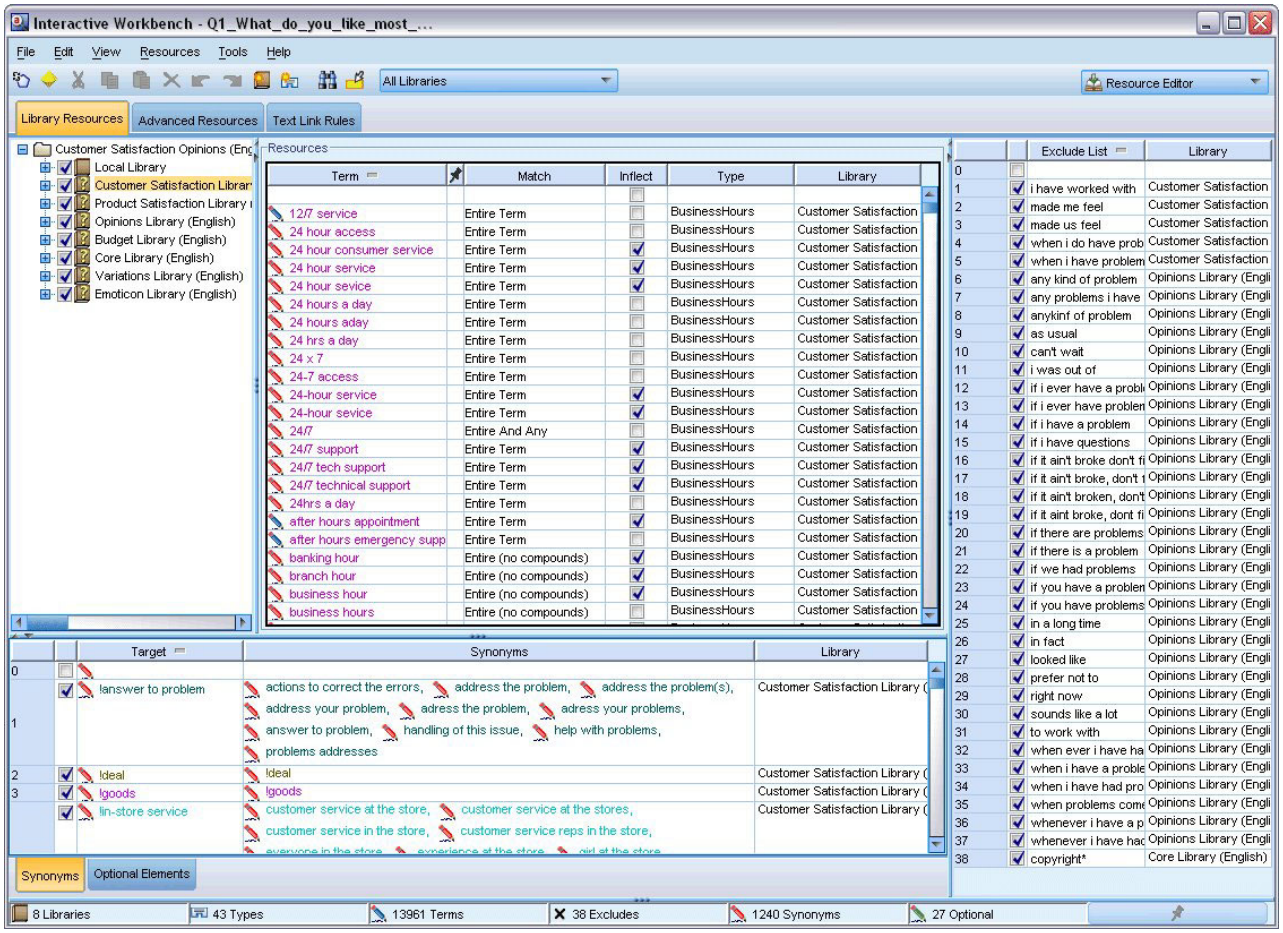


Abbildung 35. Ressourceneditoransicht für andere Sprachen als Japanisch

Für japanischen Text

Die Editorschnittstelle für die japanische Textsprache unterscheidet sich von der Schnittstelle für andere Textsprachen.

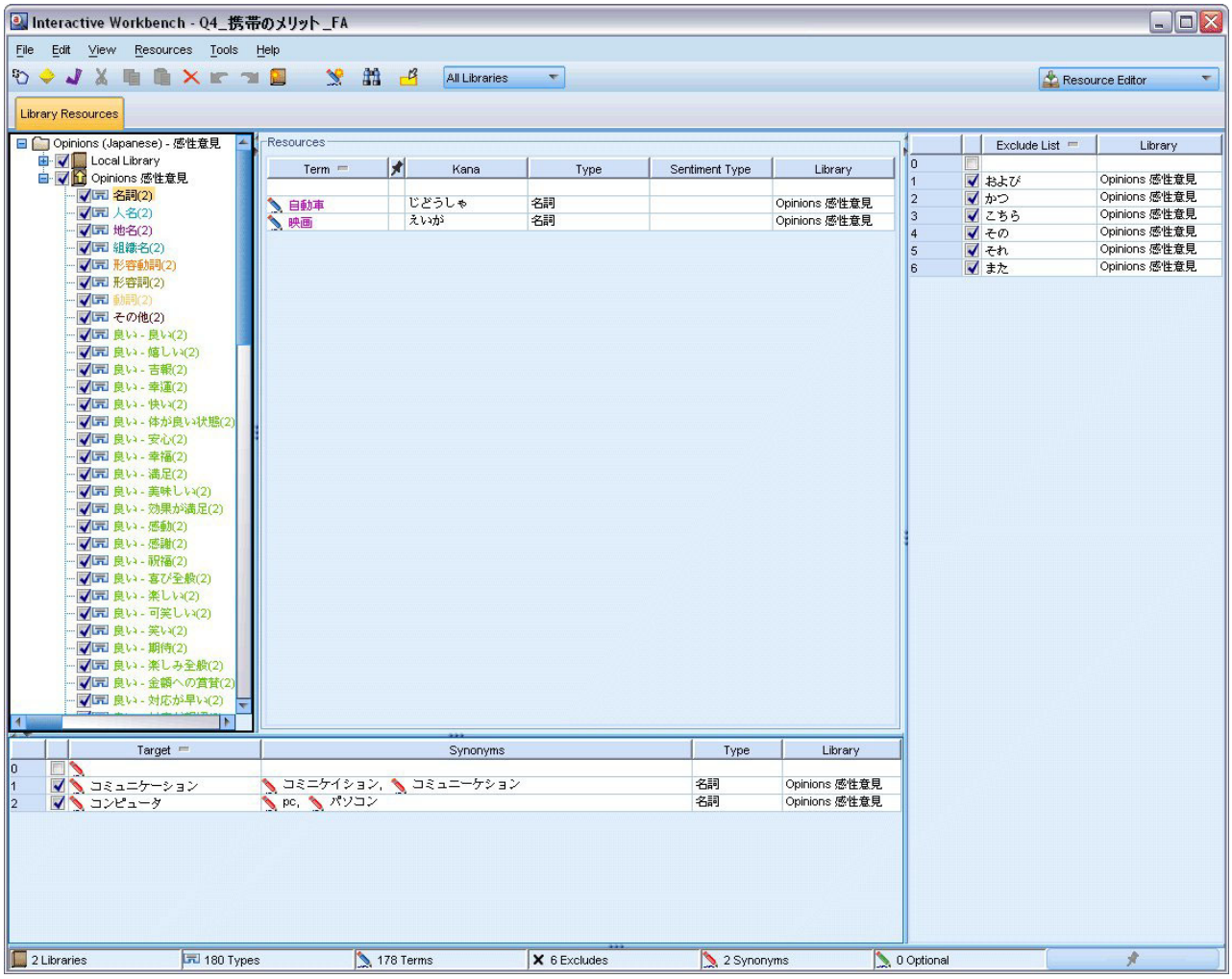


Abbildung 36. Ressourceneditoransicht für japanischen Text

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. Weitere Informationen finden Sie im Thema „Kopieren von Ressourcen aus TAPs und Vorlagen" auf Seite 27.

Hinweis: Sie können Ihre Bibliotheken auch veröffentlichen und mit anderen Benutzern gemeinsam nutzen. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von Bibliotheken" auf Seite 196.

So erstellen (oder aktualisieren) Sie eine Vorlage

1. Wählen Sie in den Menüs der Ansicht "Ressourceneditor" die Optionsfolge **Ressourcen > Ressourcenvorlage erstellen** aus. Das Dialogfeld "Ressourcenvorlage erstellen" wird geöffnet.
2. 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.
3. 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. Weitere Informationen finden Sie im Thema „Aktualisieren von Knotenressourcen nach dem Laden" auf Seite 185.

Wechseln von Ressourcenvorlagen

Wenn Sie die derzeit in der Sitzung geladenen Ressourcen durch eine Kopie der Ressourcen einer anderen Vorlage ersetzen möchten, können Sie zu diesen Ressourcen wechseln. Dadurch werden die geladenen Ressourcen überschrieben, die in dieser Sitzung vorhanden sind. Wenn Sie die Ressourcen wechseln, um vordefinierte Textlinkanalyse-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.

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 Ressourceneditor 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 Workbenchsitzung beibehalten möchten.

Hinweis: 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. Weitere Informationen finden Sie im Thema „Aktualisieren von Modellierungsknoten und Speichern" auf Seite 90.

So wechseln Sie Ressourcen

1. Wählen Sie in den Menüs der Ansicht "Ressourceneditor" die Optionsfolge **Ressourcen > Ressourcenvorlagen wechseln** aus. Das Dialogfeld "Ressourcen wechseln" wird geöffnet.
2. Wählen Sie die gewünschte Vorlage aus der Tabelle aus.
3. 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. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von Bibliotheken" auf Seite 196.

Kapitel 15. Vorlagen und Ressourcen

IBM SPSS Modeler Text Analytics erfasst und extrahiert Schlüsselkonzepte aus Textdaten schnell und genau. Dieser Extraktionsprozess beruht erheblich darauf, dass durch die linguistischen Ressourcen festgelegt wird, wie Informationen aus Textdaten extrahiert werden. Weitere Informationen finden Sie im Thema „Funktionsweise der Extraktion" auf Seite 5. Sie können die Optimierung für diese Ressourcen in der Ressourceneditoransicht 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 Optimierungen für spezifische Sprachen und Anwendungen vorzunehmen. Da die mitgelieferten Ressourcen gegebenenfalls nicht perfekt an den Kontext Ihrer Daten angepasst sind, können Sie diese Ressourcenvorlagen bearbeiten oder sogar benutzerdefinierte Bibliotheken erstellen und verwenden, die speziell auf die Daten Ihrer Organisation abgestimmt sind. Diese Ressourcen liegen in verschiedenen Formaten vor und jede davon kann in Ihrer Sitzung 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 Ressourcenset bilden, das auf eine bestimmte Domäne oder einen bestimmten Kontext (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 mindestens ein spezialisiertes Kategorienset, das mithilfe jener Ressourcen generiert wurde, sodass die Kategorien und die Ressourcen gemeinsam gespeichert werden und wiederverwendbar sind. Weitere Informationen finden Sie im Thema „Verwendung von Text Analysis Packages" auf Seite 148.
- **Bibliotheken**. Bibliotheken werden als Bausteine für TAPs und Vorlagen verwendet. Sie 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. Weitere Informationen finden Sie im Thema Kapitel 16, „Arbeiten mit Bibliotheken", auf Seite 191.

Hinweis: Im Laufe der Extraktion werden auch einige kompilierte interne Ressourcen verwendet. Diese kompilierten Ressourcen enthalten eine große Anzahl von Definitionen, die die Typen in der Kernbibliothek ergänzen. Diese kompilierten Ressourcen können nicht bearbeitet werden.

Der Ressourceneditor ermöglicht den Zugriff auf ein Set von Ressourcen, mit denen die Extraktionsergebnisse (Konzepte, Typen und Muster) erzeugt werden. Im Ressourceneditor können Sie eine Vielzahl von Aufgaben durchführen, darunter:

- **Arbeiten mit Bibliotheken**. Weitere Informationen finden Sie im Thema Kapitel 16, „Arbeiten mit Bibliotheken", auf Seite 191.
- **Erstellen von Typwörterbüchern**. Weitere Informationen finden Sie im Thema „Erstellen von Typen" auf Seite 203.
- **Hinzufügen von Termen zu Wörterbüchern**. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen" auf Seite 204.
- **Erstellen von Synonymen**. Weitere Informationen finden Sie im Thema „Definieren von Synonymen" auf Seite 209.
- **Aktualisierung der Ressourcen in TAPs**. Weitere Informationen finden Sie im Thema „Aktualisierung von Text Analysis Packages" auf Seite 150.
- **Erstellen von Vorlagen**. Weitere Informationen finden Sie im Thema „Erstellen und Aktualisieren von Vorlagen" auf Seite 177.

- **Importieren und Exportieren von Vorlagen**. Weitere Informationen finden Sie im Thema „Import und Export von Vorlagen" auf Seite 187.
- **Veröffentlichen von Bibliotheken**. Weitere Informationen finden Sie im Thema „Veröffentlichen von Bibliotheken" auf Seite 197.

Vorlageneditor im Vergleich zum Ressourceneditor

Es gibt zwei Methoden zum Arbeiten mit und Bearbeiten von Vorlagen, Bibliotheken und deren Ressourcen. Sie können im Vorlageneditor oder im Ressourceneditor mit linguistischen Ressourcen arbeiten.

Vorlageneditor

Im Vorlageneditor können Sie Ressourcenvorlagen ohne eine interaktive Workbenchsitzung und unabhängig von einem bestimmten Knoten oder Stream erstellen und bearbeiten. Mithilfe dieses Editors können Sie Ressourcenvorlagen vor dem Laden in den Textlinkanalyse- und Textmining-Modellierungsknoten erstellen bzw. bearbeiten.

Der Zugriff auf den Vorlageneditor erfolgt über die IBM SPSS Modeler-Hauptsymbolleiste oder über das Menü **Tools > Text Analytics-Vorlageneditor**.

Ressourceneditor

Der Ressourceneditor, auf den in einer interaktiven Workbenchsitzung zugegriffen werden kann, bietet die Möglichkeit, mit den Ressourcen innerhalb des Kontexts eines bestimmten Knotens und Datasets zu arbeiten. Beim Hinzufügen eines Textmining-Modellierungsknotens zu einem Stream können Sie eine Kopie des Inhalts einer Ressourcenvorlage oder eines Text Analysis Package (Kategoriensets *und* Ressourcen) laden, um zu steuern, wie Text für Textmining extrahiert wird. Beim Starten einer interaktiven Workbenchsitzung können Sie neben dem Erstellen von Kategorien, Extrahieren von Textlinkanalysemustern und Erstellen von Kategoriemodellen auch die Ressourcen für die Daten dieser Sitzung in der integrierten Ressourceneditoransicht optimieren. Weitere Informationen finden Sie im Thema „Bearbeiten von Ressourcen im Ressourceneditor" auf Seite 175.

Beim Arbeiten mit Ressourcen in einer interaktiven Workbenchsitzung 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. Weitere Informationen finden Sie im Thema „Aktualisieren von Modellierungsknoten und Speichern" auf Seite 90.

Falls Sie Ihre Änderungen unter der ursprünglichen Vorlage speichern möchten, deren Inhalt in den Modellierungsknoten kopiert wurde, um diese aktualisierte Vorlage in anderen Knoten laden zu können, können Sie aus den Ressourcen eine Vorlage erstellen. Weitere Informationen finden Sie im Thema „Erstellen und Aktualisieren von Vorlagen" auf Seite 177.

Editorschnittstelle

Bei den im Vorlageneditor oder im Ressourceneditor durchgeführten Operationen geht es um die Verwaltung und Optimierung der linguistischen Ressourcen. Diese Ressourcen werden in Form von Vorlagen und Bibliotheken gespeichert. Weitere Informationen finden Sie im Thema „Typwörterbücher" auf Seite 201.

Registerkarte "Bibliotheksressourcen"

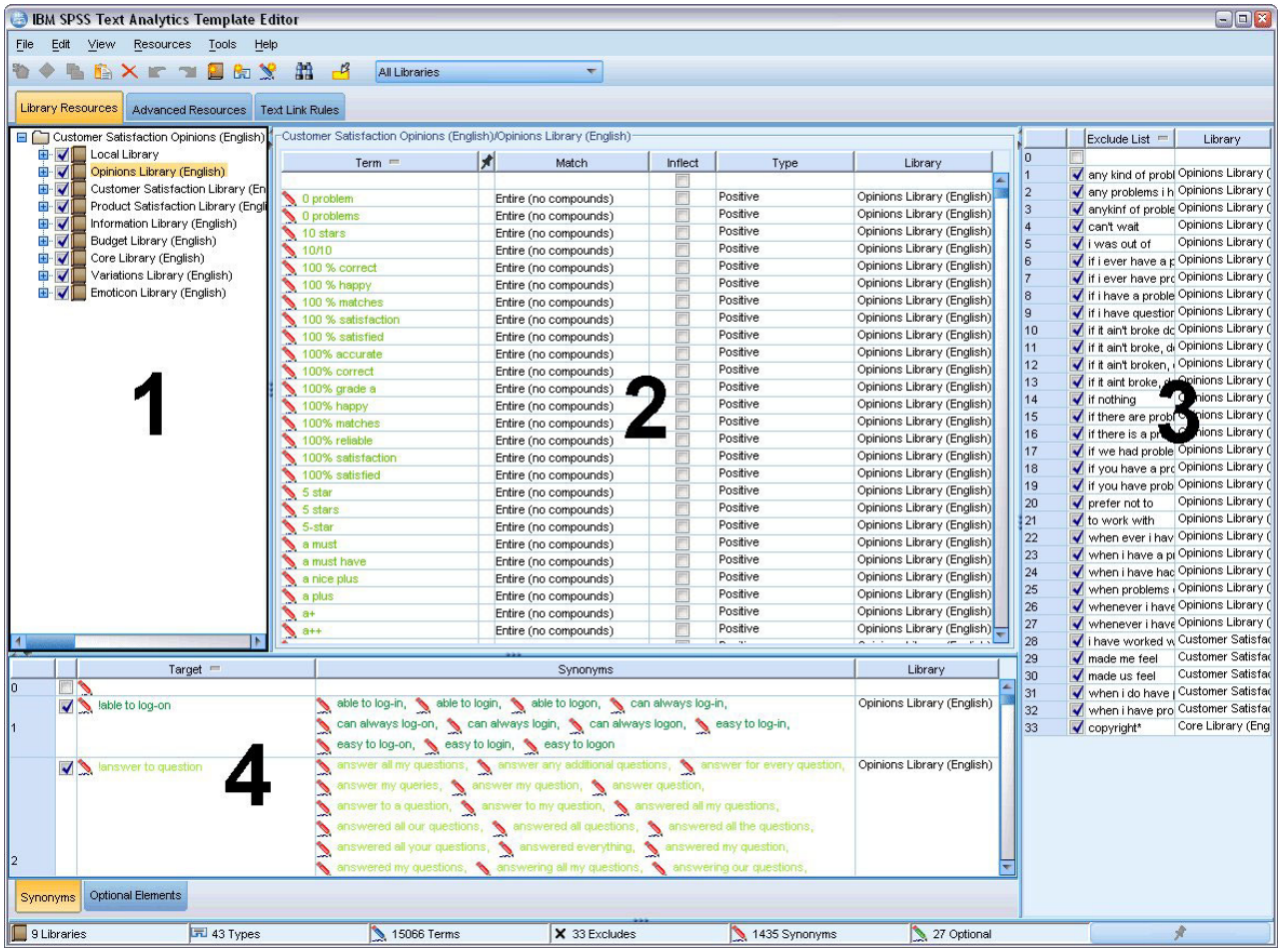


Abbildung 37. Textmining-Vorlageneditor

Die Schnittstelle gliedert sich in die vier folgenden Bereiche:

1. Fensterbereich "Bibliotheksbaum". Dieser Plan befindet sich links oben und zeigt eine Baumstruktur mit den Bibliotheken an. Sie können die Bibliotheken in diesem Baum aktivieren bzw. inaktivieren 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 Typenset angezeigt. Sie können diese Liste auch über das Menü **Ansicht** filtern, wenn Sie sich nur auf eine bestimmte Bibliothek konzentrieren möchten.

2. Bereich für Termlisten von Typwörterbüchern. Dieser Fensterbereich befindet sich rechts neben dem Bibliotheksbaum und zeigt die Termlisten der Typwörterbücher für die im Baum ausgewählten Bibliotheken an. Ein **Typwörterbuch** ist eine Zusammenstellung von Termen, die unter einer Beschriftung (bzw. einem Typ oder einem Namen) gruppiert werden sollen. Beim Lesen der Textdaten vergleicht die Extraktionsengine die im Text gefundenen Wörter mit den Termen in den Typwörterbüchern. Wenn ein extrahiertes Konzept als Term in einem Typwörterbuch vorkommt, wird der entsprechende Typname zugewiesen. Sie können das Typwörterbuch als eigenes Wörterbuch mit Termen betrachten, die etwas gemeinsam haben. Der Typ `<Location>` (Position) in der Kernbibliothek beispielsweise enthält Konzepte wie `new orleans`, `great britain` und `new york`. Diese Terme stehen jeweils für geografische Orte. Eine Bibliothek kann ein oder mehrere Typwörterbücher enthalten. Weitere Informationen finden Sie im Thema „Typwörterbücher” auf Seite 201.

3. Fensterbereich "Ausschlusswörterbuch". Dieser Fensterbereich befindet sich auf der rechten Seite und zeigt die Sammlung der Terme an, die aus den endgültigen Extraktionsergebnissen ausgeschlossen werden. Die Terme in diesem Ausschlusswörterbuch werden nicht im Bereich "Extraktionsergebnisse" angezeigt. Ausgeschlossene Terme werden in der Bibliothek Ihrer Wahl gespeichert. Jedoch zeigt der Bereich "Ausschlusswörterbuch" alle ausgeschlossenen Terme für alle Bibliotheken an, die im Bibliotheksbaum sichtbar sind. Weitere Informationen finden Sie im Thema „Ausschlusswörterbücher” auf Seite 212.

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 Terme unter einem Haupt- oder Zielkonzept in den endgültigen Extraktionsergebnissen. 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 Austauschwörterbuch enthalten. Weitere Informationen finden Sie im Thema „Substitutions-/Synonymwörterbücher” auf Seite 208. Bitte beachten Sie, dass die Registerkarte "Optionale Elemente" nicht für Ressourcen gilt, deren Textsprache Japanisch ist.

Hinweise:

- 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. Weitere Informationen finden Sie im Thema „Anzeigen von Bibliotheken” auf Seite 194.
- Die Editorschnittstelle für die japanische Textsprache unterscheidet sich von der Schnittstelle für andere Textsprachen.

Registerkarte "Erweiterte Ressourcen"

Die erweiterten Ressourcen stehen in der zweiten Registerkarte der Editoransicht zur Verfügung. Sie können die erweiterten Ressourcen auf dieser Registerkarte überprüfen und bearbeiten. Weitere Informationen finden Sie im Thema Kapitel 18, „Informationen zu erweiterten Ressourcen”, auf Seite 215.

Wichtig! Diese Registerkarte ist nicht für Ressourcen verfügbar, die für japanischen Text optimiert sind.

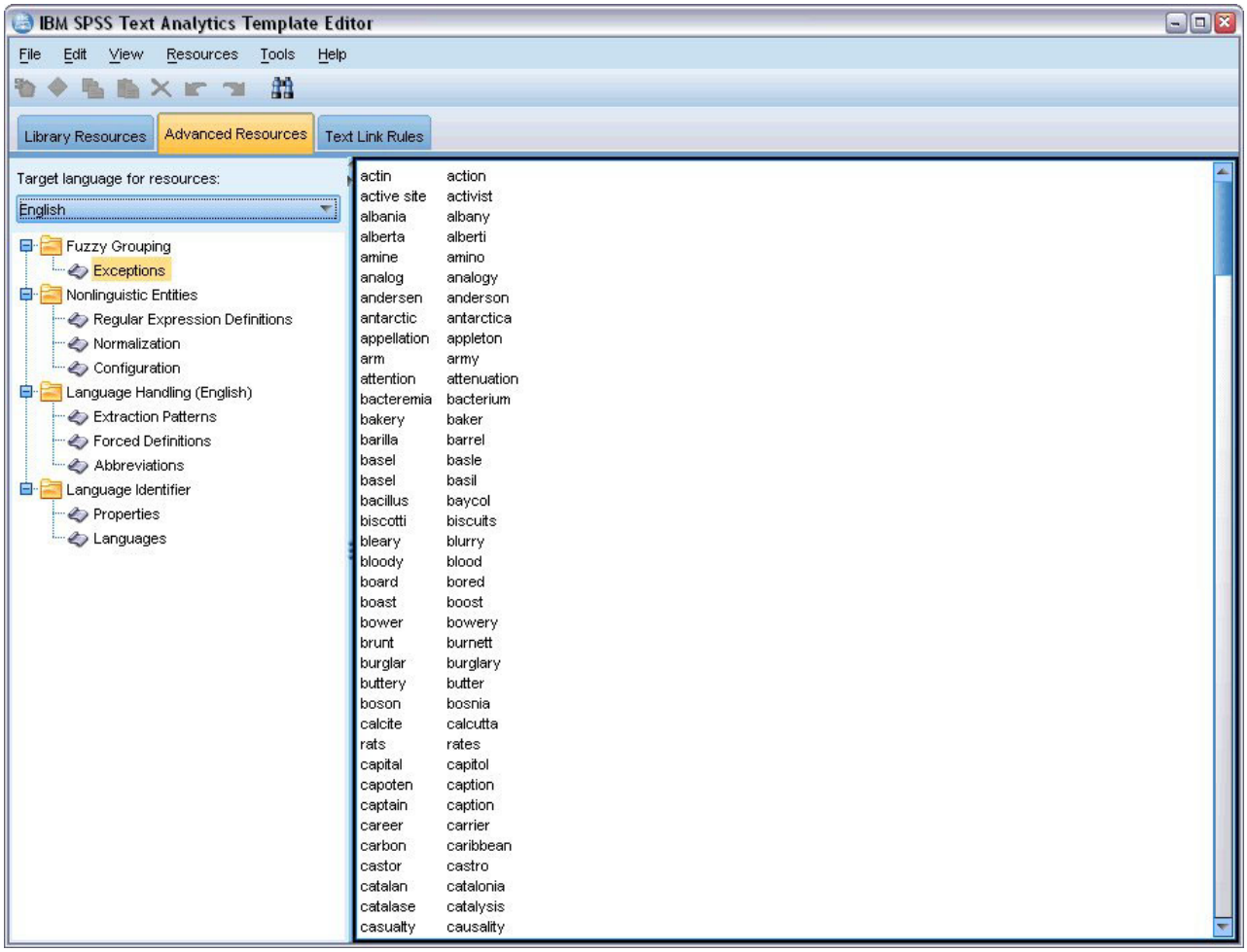


Abbildung 38. Textmining-Vorlageneditor - Registerkarte "Erweiterte Ressourcen"

Registerkarte "Textlinkregeln"

Ab Version 14 sind die Textlinkanalyseregeln auf ihrer eigenen Registerkarte der Editoransicht bearbeitbar. 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. Weitere Informationen finden Sie im Thema Kapitel 19, „Textlinkregeln", auf Seite 227.

Wichtig! Diese Registerkarte ist nicht für Ressourcen verfügbar, die für japanischen Text optimiert sind.

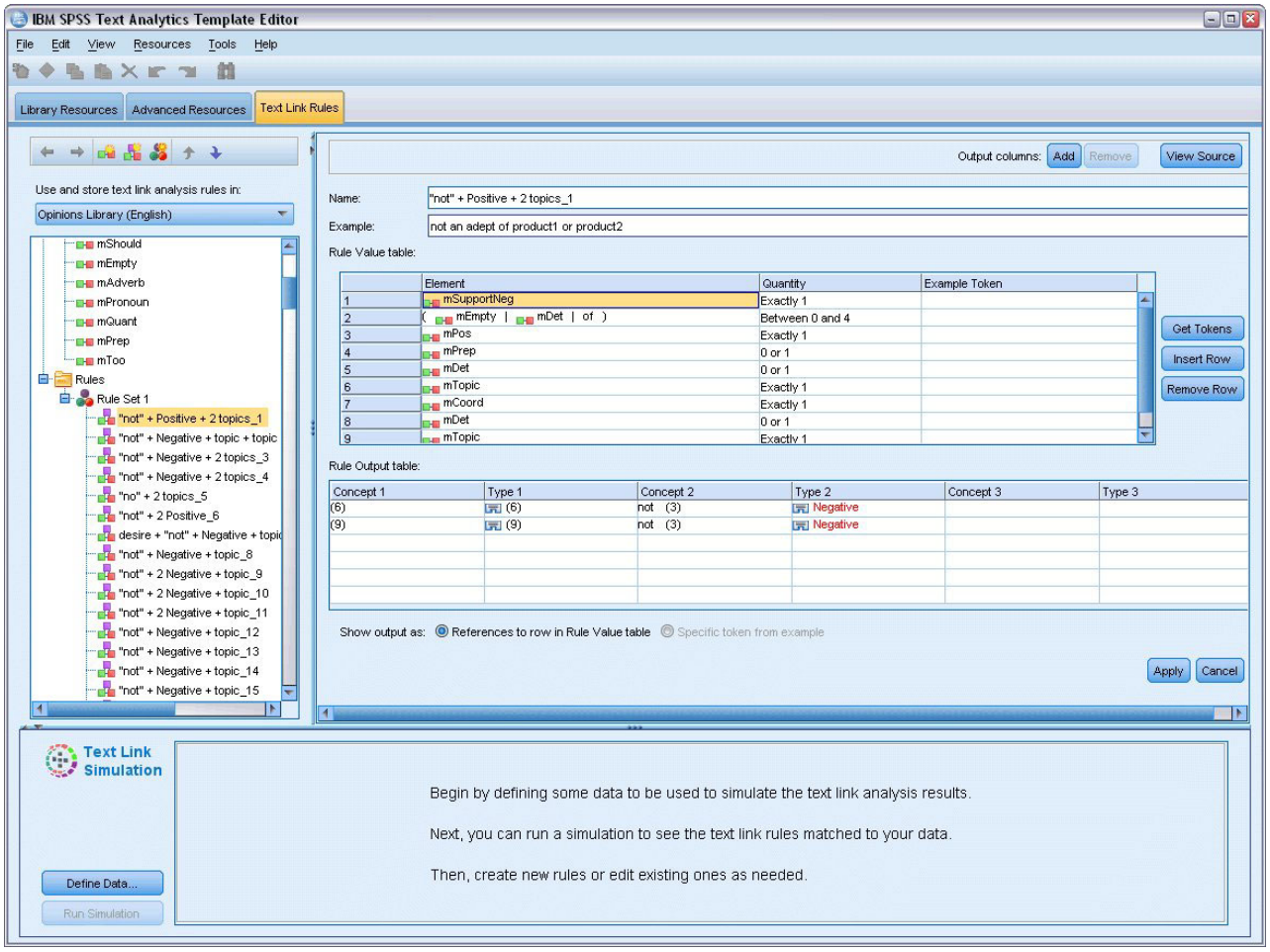


Abbildung 39. Textmining-Vorlageneditor - Registerkarte "Textlinkregeln"

Öffnen von Vorlagen

Beim Starten des Vorlageneditors 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 "Vorlage öffnen" verwenden. Weitere Informationen finden Sie im Thema „Import und Export von Vorlagen” auf Seite 187.

So öffnen Sie eine Vorlage

1. Wählen Sie in den Menüs im Vorlageneditor die Optionsfolge **Datei > Ressourcenvorlage öffnen** aus. Das Dialogfeld "Ressourcenvorlage öffnen" wird geöffnet.
2. Wählen Sie die gewünschte Vorlage aus der Tabelle aus.

3. 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, die Sie hier ausgewählt haben, wird angezeigt. Wenn Sie Änderungen an Ihren Ressourcen vorgenommen haben und Ihre Bibliotheken für eine künftige Verwendung speichern möchten, können Sie sie vor dem Öffnen einer weiteren veröffentlichen, aktualisieren und für die gemeinsame Nutzung freigeben. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von Bibliotheken" auf Seite 196.

Speichern von Vorlagen

Im Vorlageneditor 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. Weitere Informationen finden Sie im Thema „Kopieren von Ressourcen aus TAPs und Vorlagen" auf Seite 27.

Alternativ müssen Sie bei Verwendung der Option **Gespeicherte interaktive Sitzung verwenden** auf der Registerkarte "Modell" des Textminingknotens (das heißt, Sie verwenden Ressourcen aus einer vorherigen interaktiven Workbenchsitzung) innerhalb der interaktiven Workbenchsitzung zu den Ressourcen dieser Vorlage wechseln. Weitere Informationen finden Sie im Thema „Wechseln von Ressourcenvorlagen" auf Seite 178.

Hinweis: Sie können Ihre Bibliotheken auch veröffentlichen und mit anderen Benutzern gemeinsam nutzen. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von Bibliotheken" auf Seite 196.

So speichern Sie eine Vorlage

1. Wählen Sie in den Menüs im Vorlageneditor die Optionsfolge **Datei > Ressourcenvorlage speichern** aus. Das Dialogfeld "Ressourcenvorlage speichern" wird geöffnet.
2. 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.
3. Geben Sie bei Bedarf eine Beschreibung ein, um in der Tabelle einen Kommentar oder eine Anmerkung anzuzeigen.
4. 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. Weitere Informationen finden Sie im Thema „Aktualisieren von Knotenressourcen nach dem Laden".

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 dieser Ressourcen 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 durch eine aktuellere 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. Weitere Informationen finden Sie im Thema „Kopieren von Ressourcen aus TAPs und Vorlagen" auf Seite 27.

Wenn Sie jedoch mit interaktiven Sitzungsdaten in einem Textmining-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** inaktiviert ist. Die Schaltfläche ist inaktiviert, da Sie zu einem bestimmten Zeitpunkt während der interaktiven Workbenchsitzung die Option **Modellierungsknoten aktualisieren** ausgewählt 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 Ressourceneditor gewechselt werden.

Methode 2: Wechseln von Ressourcen im Ressourceneditor

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 Textmining-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 Workbenchsitzung durch Ausführen des Streams und Wechseln der Ressourcen im Ressourceneditor starten. Weitere Informationen finden Sie im Thema „Wechseln von Ressourcenvorlagen" auf Seite 178.

Um die Arbeit der Sitzung einschließlich Ressourcen für nachfolgende Sitzungen beizubehalten, müssen Sie den Modellierungsknoten innerhalb der interaktiven Workbenchsitzung aktualisieren, sodass die Ressourcen (und weitere Daten) unter dem Knoten gespeichert werden. Weitere Informationen finden Sie im Thema „Aktualisieren von Modellierungsknoten und Speichern" auf Seite 90.

Hinweis: 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.

Verwalten von Vorlagen

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. Weitere Informationen finden Sie im Thema „Import und Export von Vorlagen" auf Seite 187.

Hinweis: Mit diesem Produkt installierte (oder gelieferte) 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 löschen, wird sie auf die ursprünglich installierte Version zurückgesetzt.

So benennen Sie eine Vorlage um

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Ressourcenvorlagen verwalten** aus. Das Dialogfeld "Vorlagen verwalten" wird geöffnet.
2. Wählen Sie die Vorlage aus, die Sie umbenennen möchten, und klicken Sie auf **Umbenennen**. Das Namensfeld wird in der Tabelle editierbar.
3. Geben Sie einen neuen Namen ein und drücken Sie die Eingabetaste. Es wird ein Bestätigungs-Dialogfeld geöffnet.
4. Wenn Sie die Namensänderung anwenden möchten, klicken Sie auf **Ja**. Wenn Sie sie verwerfen möchten, klicken Sie auf **Nein**.

So löschen Sie eine Vorlage

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Ressourcenvorlagen verwalten** aus. Das Dialogfeld "Vorlagen verwalten" wird geöffnet.
2. Wählen Sie im Dialogfeld "Vorlagen verwalten" die zu löschende Vorlage aus.
3. Klicken Sie auf **Löschen**. Es wird ein Bestätigungs-Dialogfeld geöffnet.
4. Klicken Sie auf **Ja**, um zu löschen, oder auf **Nein**, um den Anforderungsvorgang abzubrechen. Wenn Sie auf **Ja** klicken, wird die Vorlage gelöscht.

Import und Export 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 in verschiedenen Dialogfeldern zur Verfügung.

- Dialogfeld "Vorlage öffnen" im Vorlageneditor
- Dialogfeld "Ressourcen laden" im Textmining-Modellierungsknoten und Textlinkanalyseknoten
- Dialogfeld "Vorlagen verwalten" im Vorlageneditor und im Ressourceneditor

So importieren Sie eine Vorlage

1. Klicken Sie im Dialogfeld auf **Importieren**. Das Dialogfeld "Vorlage importieren" wird geöffnet.
2. 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

1. Wählen Sie im Dialogfeld die zu exportierende Vorlage aus und klicken Sie auf **Exportieren**. Das Dialogfeld "Verzeichnis auswählen" wird geöffnet.
2. 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 Vorlageneditors

Wenn Sie die Arbeit im Vorlageneditor beendet haben, können Sie Ihre Arbeit speichern und den Editor beenden.

So beenden Sie den Vorlageneditor

1. Wählen Sie in den Menüs die Optionsfolge **Datei > Schließen** aus. Das Dialogfeld "Speichern und schließen" wird geöffnet.
2. Wählen Sie die Option **Änderungen in der Vorlage speichern** aus, um die geöffnete Vorlage vor dem Schließen des Editors zu speichern.
3. 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. Weitere Informationen finden Sie im Thema „Veröffentlichen von Bibliotheken" auf Seite 197.

Sichern von Ressourcen

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.

Hinweis: Eine Sicherung und Wiederherstellung ist nur für eine übereinstimmende Hauptversion Ihrer Software möglich. Wenn Sie beispielsweise eine Sicherung der Version 15 erstellen, können Sie diese Sicherung nicht in Version 16 wiederherstellen.

So sichern Sie die Ressourcen

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Sicherungstools > Ressourcen sichern** aus. Das Dialogfeld "Sicherung" wird geöffnet.
2. Geben Sie einen Namen für Ihre Sicherungsdatei ein und klicken Sie auf **Speichern**. Das Dialogfeld wird geschlossen und die Sicherungsdatei wird erstellt.

So stellen Sie die Ressourcen wieder her

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Sicherungstools > Ressourcen wiederherstellen** aus. Ein Alert informiert Sie darüber, dass bei einer Wiederherstellung der Inhalt Ihrer Datenbank überschrieben wird.
2. Klicken Sie auf **Ja**, um fortzufahren. Das Dialogfeld wird geöffnet.
3. 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.

Import von Ressourcendateien

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 codiert sein. Außerdem können Sie für Japanisch keine Ausschlusslisten importieren.

Jede importierte Datei darf nur einen Eintrag pro Zeile enthalten und wenn die Inhalte folgendermaßen strukturiert sind, gilt Folgendes:

- Als Liste mit Wörtern oder Wortfolgen (ein Wort bzw. eine Wortfolge pro Zeile). Die Datei wird als Termliste für ein Typwörterbuch importiert. Dabei übernimmt das Typwörterbuch den Namen der Datei ohne die Erweiterung.
- Als Liste von Einträgen wie *Term1* `<TAB>` *Term2*. In diesem Fall wird die Datei als Liste von Synonymen importiert. Dabei ist *Term1* das Set des zugrunde liegenden Terms und *Term2* ist der Zielterm.

So importieren Sie eine einzelne Ressourcendatei

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Dateien importieren > Einzelne Datei importieren** aus. Das Dialogfeld "Datei importieren" wird geöffnet.
2. 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

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Dateien importieren > Gesamtes Verzeichnis importieren** aus. Das Dialogfeld "Verzeichnis importieren" wird geöffnet.
2. 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.
3. Wählen Sie das Verzeichnis aus, aus dem die Dateien importiert werden sollen. Es werden keine Unterverzeichnisse gelesen.
4. 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. Arbeiten mit Bibliotheken

Die Ressourcen, die von der Extraktionsengine zum Extrahieren und Gruppieren von Termen verwendet werden, enthalten immer mindestens eine Bibliothek. Die Bibliotheken werden im Bibliotheksbaum im oberen linken Bereich des Vorlageneditors und des Ressourceneditors angezeigt. Sie bestehen aus drei Arten von Wörterbüchern: Typwörterbücher, Substitutionswörterbücher und Ausschlusswörterbücher. Weitere Informationen finden Sie im Thema Kapitel 17, „Informationen zu Bibliothekwörterbüchern", auf Seite 201.

Die ausgewählte Ressourcenvorlage oder die ausgewählten Ressourcen aus dem TAP enthalten verschiedene Bibliotheken, die Ihnen die unverzügliche Extraktion von Konzepten aus Ihren Textdaten ermöglichen. Sie können jedoch auch eigene Bibliotheken erstellen und zur Wiederverwendung auch veröffentlichen. Weitere Informationen finden Sie im Thema „Veröffentlichen von Bibliotheken" auf Seite 197.

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 Vorlageneditor können Sie eine neue Vorlage und darin eine Bibliothek zum Extrahieren und Gruppieren von Termen aus der Automobilbranche 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 Terme aus verschiedenen Unterbranchen, z. B. zu elektronischen Geräten, Motoren, Kühlsystemen, oder sogar Terme eines bestimmten Herstellers oder Markts gruppieren. Sie können für jede Gruppe eine Bibliothek erstellen und sie dann so veröffentlichen, dass sie mit mehreren Textdatasets verwendet werden kann. So können Sie Bibliotheken hinzufügen, die dem Kontext Ihrer Textdaten am besten entsprechen.

Hinweis: Zusätzliche Ressourcen können auf der Registerkarte "Erweiterte Ressourcen konfiguriert und verwaltet werden. Einige beziehen sich auf alle Bibliotheken und verwalten nicht linguistische Entitäten, Ausnahmen bei der Fuzzy-Gruppierung usw. Zudem können Sie die bibliotheksspezifischen Musterregeln für Textlinkanalysen auf der Registerkarte "Textlinkregeln" bearbeiten. Weitere Informationen finden Sie im Thema Kapitel 18, „Informationen zu erweiterten Ressourcen", auf Seite 215.

Mitgelieferte Bibliotheken

Es werden standardmäßig mehrere Bibliotheken mit IBM SPSS Modeler Text Analytics installiert. Mit diesen vorformatierten Bibliotheken können Sie auf Tausende vordefinierter Terme und Synonyme sowie auf viele verschiedene Typen zugreifen. Diese mitgelieferten Bibliotheken sind auf verschiedene Domänen abgestimmt und in verschiedenen Sprachen verfügbar.

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 Typwörterbuch. Diese ist besonders beim Vornehmen von Änderungen oder bei Optimierungen nützlich, die direktvon der Ansicht "Kategorien und Konzepte", der Ansicht "Cluster" und der Ansicht "Textlinkanalyse" aus vorgenommen werden (z. B. Hinzufügen eines Worts zu einem Typ). In diesem Fall werden die Änderungen und Optimierungen automatisch in der ersten Bibliothek des Bibliotheksbaum im Ressourceneditor gespeichert; standardmäßig ist dies die *lokale Bibliothek*. Diese Bibliothek können Sie nicht veröffentlichen, weil sie für die Sitzungsdaten spezifisch ist. Um ihren Inhalt zu veröffentlichen, müssen Sie die Bibliothek zunächst umbenennen.

- **Kernbibliothek.** 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 Terme in einem ihrer Typwörterbücher aufgeführt werden, sind die in der Kernbibliothek dargestellten Typen tatsächlich Ergänzungen zu den leistungsfähigen Typen in den internen, kompilierten Ressourcen, die im Lieferumfang Ihres Textmining-Produkts enthalten sind. Diese internen, kompilierten Ressourcen enthalten Tausende von Termen für jeden Typ. Daher kann ein Term dennoch extrahiert und mit einem Kerntyp versehen werden, auch wenn Sie ihn nicht in der Termliste des Typwörterbuchs sehen. Dies erklärt, wie Namen, z. B. *George*, extrahiert werden und den Typ `<Person>` erhalten können, wenn nur *John* in dem Typwörterbuch `<Person>` in der Kernbibliothek aufgeführt wird. Wenn Sie die Kernbibliothek nicht einschließen, werden diese Typen gegebenenfalls trotzdem in den Extraktionsergebnissen aufgeführt, da die kompilierten Ressourcen, die diese Typen enthalten, weiterhin von der Extraktionsengine verwendet werden.
- **Opinions Library.** Am häufigsten zum Extrahieren von Meinungen und Stimmungen aus Textdaten verwendet. Diese Bibliothek beinhaltet Tausende von Wörtern für Einstellungen, Qualifikationsmerkmale und Präferenzen, die - wenn sie zusammen mit anderen Termen 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 auf der Registerkarte "Textlinkregeln" angegeben werden. Weitere Informationen finden Sie im Thema Kapitel 19, „Textlinkregeln”, auf Seite 227.
- **Budget Library.** Wird zum Extrahieren von Termen zum Thema "Kosten" verwendet. Diese Bibliothek enthält zahlreiche Wörter und Wortfolgen, die Adjektive, Vermerke und Entscheidungen zu den Themen "Preis" oder "Qualität" darstellen.
- **Variantenbibliothek.** 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, es mit einer speziellen Vorlage für die Art der verwendeten Textdaten zu versuchen und Ihre Änderungen an diesen Ressourcen auszuführen, anstatt einer allgemeineren Vorlage nur individuelle Bibliotheken hinzuzufügen.

Es sind außerdem kompilierte Ressourcen im Lieferumfang von IBM SPSS Modeler Text Analytics enthalten. Sie werden bei jedem Extraktionsprozess 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 Terms, der einen Typ durch diese kompilierten Ressourcen erhalten hat, in ein anderes Wörterbuch erzwingen. Weitere Informationen finden Sie im Thema „Erzwingen von Termen” auf Seite 207.

Erstellen von Bibliotheken

Sie können beliebig viele Bibliotheken erstellen. Nach dem Erstellen einer neuen Bibliothek können Sie Typwörterbücher in dieser Bibliothek erstellen und Terme, Synonyme und Ausschlüsse eingeben.

So erstellen Sie eine Bibliothek

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Neue Bibliothek** aus. Das Dialogfeld Bibliothekseigenschaften wird geöffnet.
2. Geben Sie im Textfeld "Name" einen Namen für die Bibliothek ein.
3. Bei Bedarf geben Sie einen Kommentar im Textfeld "Anmerkungen" ein.
4. Klicken Sie auf **Veröffentlichen**, wenn Sie diese Bibliothek nun veröffentlichen möchten, bevor Sie Einträge in die Bibliothek vornehmen. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von Bibliotheken” auf Seite 196. Sie können sie auch später zu einem beliebigen Zeitpunkt veröffentlichen.

5. 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 Typwörterbuch angezeigt, das automatisch in die Bibliothek aufgenommen worden ist. In dieses können Sie sofort Terme aufnehmen. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen" auf Seite 204.

Hinzufügen öffentlicher Bibliotheken

Um eine Bibliothek aus anderen Sitzungsdaten wiederzuverwenden, fügen Sie sie Ihren aktuellen Ressourcen hinzu, vorausgesetzt, es handelt sich um eine öffentliche Bibliothek. Eine **öffentliche Bibliothek** ist eine Bibliothek, die veröffentlicht worden ist. Weitere Informationen finden Sie im Thema „Veröffentlichen von Bibliotheken" auf Seite 197.

Wichtig! Es ist nicht möglich, nicht japanischen Ressourcen eine japanische Bibliothek hinzuzufügen oder umgekehrt.

Wenn Sie eine öffentliche Bibliothek 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 kann das Dialogfeld "Konflikte auflösen" angezeigt werden, wenn Konflikte zwischen den Termen und Typen in einer Bibliothek und den anderen lokalen Bibliotheken festgestellt werden. Lösen Sie diese Konflikte auf oder akzeptieren Sie die vorgeschlagenen Lösungen, um den Vorgang abzuschließen. Weitere Informationen finden Sie im Thema „Auflösen von Konflikten" auf Seite 198.

Hinweis: Wenn Sie Ihre Bibliotheken jedes Mal aktualisieren, wenn Sie eine interaktive Workbenchsitzung starten, oder veröffentlichen, wenn Sie eine Sitzung schließen , ist die Gefahr geringer, dass Bibliotheken asynchron werden. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von Bibliotheken" auf Seite 196.

So fügen Sie eine Bibliothek hinzu

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Bibliothek hinzufügen** aus. Das Dialogfeld "Bibliotheken hinzufügen" wird geöffnet.
2. Wählen Sie die Bibliotheken aus der Liste aus.
3. 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. Weitere Informationen finden Sie im Thema „Auflösen von Konflikten" auf Seite 198.

Suchen von Termen und Typen

Sie können mit der Suchfunktion in den verschiedenen Bereichen im Editor suchen. Im Editor können Sie die Optionsfolge **Bearbeiten > Suchen** in den Menüs auswählen, woraufhin die Symbolleiste "Suchen" angezeigt wird. Mit dieser Symbolleiste können Sie jeweils ein Vorkommen suchen. Durch erneutes Klicken auf **Suchen** können Sie nachfolgende Vorkommen Ihres Suchbegriffs 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 auswählen, ob die Suche Groß-/Kleinschreibung unterscheiden soll.

So finden Sie Zeichenfolgen in der Ansicht

1. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Suchen** aus. Die Symbolleiste "Suchen" wird angezeigt.
2. Geben Sie die zu suchende Zeichenfolge ein.
3. Klicken Sie auf die Schaltfläche **Suchen**, um die Suche zu starten. Das nächste Vorkommen des Terms oder Typs wird hervorgehoben.
4. Klicken Sie erneut auf die Schaltfläche, um zu den nachfolgenden Vorkommen zu navigieren.

Anzeigen von Bibliotheken

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 auf der Registerkarte "Bibliotheksressourcen", inaktiviert aber keine Bibliotheken bei der Extraktion. Weitere Informationen finden Sie im Thema „Inaktivieren lokaler Bibliotheken" auf Seite 195.

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 in 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 Extraktion gelesen.

So ändern Sie die Bibliothekenansicht

1. Wählen Sie in den Menüs der Registerkarte "Bibliotheksressourcen" die Optionsfolge **Ansicht > Bibliotheken** aus. Es wird ein Menü mit allen lokalen Bibliotheken geöffnet.
2. 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.

Verwalten lokaler Bibliotheken

Lokale Bibliotheken sind (im Gegensatz zu öffentlichen Bibliotheken) die Bibliotheken innerhalb Ihrer interaktiven Workbenchsitzung oder innerhalb einer Vorlage. Weitere Informationen finden Sie im Thema „Verwalten öffentlicher Bibliotheken" auf Seite 195. Sie sollten außerdem einige grundlegende Maßnahmen zur Verwaltung der lokalen Bibliotheken durchführen, u. a.: Umbenennen, Inaktivieren oder Löschen einer lokalen Bibliothek.

Umbenennen lokaler Bibliotheken

Sie können lokale Bibliotheken umbenennen. Wenn eine lokale Bibliothek umbenannt wird, wird sie von der öffentlichen Version getrennt, falls eine solche vorhanden ist. In diesem Fall können nachfolgende Änderungen nicht mehr in der öffentlichen Version gemeinsam genutzt werden. Sie können diese lokale Bibliothek unter ihrem 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.

Hinweis: Eine öffentliche Bibliothek kann nicht umbenannt werden.

1. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Bibliothekseigenschaften** aus. Das Dialogfeld "Bibliotheken-Eigenschaften" wird geöffnet.

So benennen Sie ein lokale Bibliothek um

1. Wählen Sie in der Baumansicht die Bibliothek aus, die Sie umbenennen möchten.
2. Geben Sie im Textfeld "Name" einen neuen Namen für die Bibliothek ein.
3. 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.

Inaktivieren lokaler Bibliotheken

Um eine Bibliothek vorübergehend aus dem Extraktionsprozess 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 Extraktion ignoriert werden soll.

So inaktivieren Sie eine Bibliothek

1. Wählen Sie im Bibliotheksbaum die Bibliothek aus, die Sie inaktivieren möchten.
2. Drücken Sie die Leertaste. Das Kontrollkästchen links neben dem Namen wird inaktiviert.

Löschen lokaler Bibliotheken

Sie können eine Bibliothek entfernen, ohne die öffentliche Version der Bibliothek zu löschen, und umgekehrt. Durch 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. Weitere Informationen finden Sie im Thema „Verwalten öffentlicher Bibliotheken".

So löschen Sie eine lokale Bibliothek

1. Wählen Sie in der Baumansicht die zu löschende Bibliothek aus.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Löschen** aus, um die Bibliothek zu löschen. Die Bibliothek wird entfernt.
3. 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.

Hinweis: Es muss immer eine Bibliothek bestehen bleiben.

Verwalten öffentlicher Bibliotheken

Um lokale Bibliotheken wiederzuverwenden, können Sie sie veröffentlichen und dann über das Dialogfeld "Bibliotheken verwalten" (**Ressourcen > Bibliotheken verwalten**) anzeigen und so mit ihnen arbeiten. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von Bibliotheken" auf Seite 196. Zu den grundlegenden Maßnahmen zur Verwaltung von öffentlichen Bibliotheken gehören das Importieren, Exportieren oder Löschen einer öffentlichen Bibliothek. Eine öffentliche Bibliothek kann nicht umbenannt werden.

Öffentliche Bibliotheken importieren

1. Klicken Sie im Dialogfeld "Bibliotheken verwalten" auf **Importieren...**. Das Dialogfeld "Bibliothek importieren" wird geöffnet.
2. Wählen Sie die zu importierende Bibliothekendatei (**.lib*) und wählen Sie **Bibliothek zu aktuellem Projekt hinzufügen**, wenn Sie diese Bibliothek außerdem lokal hinzufügen möchten.
3. Klicken Sie auf **Importieren**. Das Dialogfeld wird geschlossen. Wenn bereits eine öffentliche Bibliothek mit demselben Namen vorhanden ist, werden Sie aufgefordert, die zu importierende Bibliothek umzubenennen oder die aktuelle öffentliche Bibliothek zu überschreiben.

Öffentliche Bibliotheken exportieren

Sie können öffentliche Bibliotheken in das *.lib*-Format exportieren, sodass Sie sie mit anderen Benutzern gemeinsam nutzen können.

1. Wählen Sie im Dialogfeld "Bibliotheken verwalten" die Bibliothek, die in die Liste exportiert werden soll.
2. Klicken Sie auf **Exportieren**. Das Dialogfeld "Verzeichnis auswählen" wird geöffnet.

3. 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.

Öffentliche Bibliotheken 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.

1. Wählen Sie im Dialogfeld "Bibliotheken verwalten" die zu löschende Bibliothek aus. Sortieren Sie die Liste, indem Sie auf den entsprechenden Titel klicken.
2. 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.

Gemeinsame Nutzung von Bibliotheken

Mithilfe von Bibliotheken können Sie Ressourcen leicht in mehreren interaktiven Workbenchsitzungen gemeinsam nutzen. Bibliotheken können in zwei Status oder Versionen vorliegen. Bibliotheken, die im Editor bearbeitet werden können und Teil einer interaktiven Workbenchsitzung sind, werden als **lokale Bibliotheken** bezeichnet. Während der Arbeit an einer interaktiven Workbenchsitzung können Sie zahlreiche Änderungen, z. B. in der *Pflanzen*-Bibliothek, vornehmen. Wenn Ihre Änderungen auch für andere Daten nützlich sein können, können Sie diese Ressourcen durch Erstellen einer **öffentlichen Bibliotheken**-Version der *Pflanzen*-Bibliothek verfügbar machen. Wie der Name andeutet, ist eine öffentliche Bibliothek für alle anderen Ressourcen in allen interaktiven Workbenchsitzungen verfügbar.

Die öffentlichen Bibliotheken 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 öffentliche Bibliotheken. Sie können die Ressourcen in diesen Bibliotheken bearbeiten und dann eine neue öffentliche Version erstellen. Die neuen Versionen sind dann in anderen interaktiven Workbenchsitzungen 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 Workbenchsitzung heraus aktualisiert wurde. Wenn Ihre Bibliotheken asynchron werden, können Sie sie neu synchronisieren. Zum Synchronisieren der Versionen einer Bibliothek werden lokale Bibliotheken erneut veröffentlicht und/oder aktualisiert.

Wenn Sie eine interaktive Workbenchsitzung starten oder schließen , werden Sie aufgefordert, alle Bibliotheken zu synchronisieren, die aktualisiert oder erneut veröffentlicht werden sollten. Sie können außerdem den Synchronisierungsstatus ihrer lokalen Bibliothek leicht an dem Symbol erkennen, das neben dem Namen der Bibliothek in der Baumansicht angezeigt wird, oder indem Sie das Dialogfeld "Bibliotheken-Eigenschaften" 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 37. Status der Synchronisierung lokaler Bibliotheken.

Symbol	Statusbeschreibung lokaler Bibliotheken
	Unveröffentlicht - Die lokale Bibliothek ist nie veröffentlicht worden.
	Synchronisiert - Die lokale und die öffentliche Version der Bibliothek stimmen überein. Dies gilt auch für die *lokale Bibliothek*, 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.
	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 lokale als auch die öffentliche Bibliothek enthalten Änderungen, die in der anderen nicht enthalten sind. Sie müssen die lokale Bibliothek 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.

Hinweis: Wenn Sie Ihre Bibliotheken jedes Mal aktualisieren, wenn Sie eine interaktive Workbenchsitzung 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 eine beliebige Anzahl öffentlicher Bibliotheken Ihren Ressourcen hinzufügen.

Wenn eine öffentliche Version einer Bibliothek gemeinsam genutzt wird, ist die Wahrscheinlichkeit größer, dass Unterschiede zwischen der lokalen und der öffentlichen Version entstehen. Beim Starten oder Schließen und Veröffentlichen über eine interaktive Workbenchsitzung bzw. beim Öffnen oder Schließen einer Vorlage über den Vorlageneditor, wird eine Nachricht angezeigt, damit Sie Bibliotheken veröffentlichen und/oder aktualisieren können, die nicht mehr synchron mit den Versionen derjenigen 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 lokalen Bibliothek zusammenführen möchten.

Veröffentlichen von Bibliotheken

Wenn eine bestimmte Bibliothek noch nie veröffentlicht worden ist, gehört zur Veröffentlichung die Erstellung einer öffentlichen Kopie Ihrer lokalen Bibliothek in der Datenbank. Bei der erneuten Veröffentlichung einer Bibliothek wird der Inhalt der öffentlichen Version durch den Inhalt der lokalen Bibliothek ersetzt. Nach der erneuten Veröffentlichung können Sie diese Bibliothek in allen anderen Streamsitzungen 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 lokalen Bibliothek 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.

So veröffentlichen Sie lokale Bibliotheken in der Datenbank

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Bibliotheken veröffentlichen** aus. Das Dialogfeld "Bibliotheken veröffentlichen" wird geöffnet, wobei alle Bibliotheken, die veröffentlicht werden sollten, standardmäßig ausgewählt sind.
2. Wählen Sie das Kontrollkästchen links von jeder Bibliothek, die Sie veröffentlichen oder erneut veröffentlichen möchten.
3. Klicken Sie auf **Veröffentlichen**, um die Bibliotheken in der Datenbank "Bibliotheken verwalten" zu veröffentlichen.

Aktualisieren von Bibliotheken

Wenn Sie eine interaktive Workbenchsitzung 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.

Hinweis: Wenn Sie Ihre Bibliotheken jedes Mal aktualisieren, wenn Sie eine interaktive Workbenchsitzung starten, oder veröffentlichen, wenn Sie eine Sitzung schließen, ist die Gefahr geringer, dass Bibliotheken asynchron werden. Weitere Informationen finden Sie im Thema „Gemeinsame Nutzung von Bibliotheken” auf Seite 196.

So aktualisieren Sie lokale Bibliotheken

1. Wählen Sie in den Menüs die Optionsfolge **Ressourcen > Bibliotheken aktualisieren** aus. Das Dialogfeld "Bibliotheken aktualisieren" wird geöffnet, wobei alle Bibliotheken, die veröffentlicht werden sollten, standardmäßig ausgewählt sind.
2. Wählen Sie das Kontrollkästchen links von jeder Bibliothek, die Sie veröffentlichen oder erneut veröffentlichen möchten.
3. Klicken Sie zum Aktualisieren der lokalen Bibliotheken auf **Aktualisieren**.

Auflösen von Konflikten

Konflikte zwischen lokaler und öffentlicher Bibliothek

Wenn Sie eine Streamsitzung starten, führt IBM SPSS Modeler Text Analytics einen Vergleich der lokalen Bibliotheken mit den im Dialogfeld "Bibliotheken verwalten" aufgeführten Bibliotheken durch. Wenn lokale Bibliotheken in Ihrer Sitzung 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 Projektdatei als lokal eingestuft werden.** Mit dieser Option werden alle lokalen Bibliotheken unverändert beibehalten. Sie können Sie später erneut veröffentlichen oder aktualisieren.
- **Alle auf diesem Computer veröffentlichten Bibliotheken.** Diese Option ersetzt die angezeigten lokalen Bibliotheken durch die Versionen in der Datenbank.
- **Alle aktuelleren Bibliotheken.** Diese Option ersetzt die älteren lokalen Bibliotheken 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 Terme

Durch Hinzufügen einer öffentlichen Bibliothek oder Aktualisierung einer lokalen Bibliothek können Konflikte oder doppelte Einträge unter den Termen und Typen in diesem Wörterbuch und den Termen 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 Terme bearbeiten" zu verifizieren oder zu ändern. Weitere Informationen finden Sie im Thema „Erzwingen von Termen” auf Seite 207.

Das Dialogfeld "Erzwungene Terme bearbeiten" enthält alle Paare der in Konflikt stehenden Terme oder Typen. Die Konfliktpaare werden durch abwechselnde Hintergrundfarben optisch unterschieden. Diese Farben können im Dialogfeld "Optionen" geändert werden. Weitere Informationen finden Sie im Thema „Optionen: Registerkarte "Anzeigen"” auf Seite 88. Das Dialogfeld "Erzwungene Terme bearbeiten" enthält zwei Registerkarten:

- **Duplikate.** Diese Registerkarte enthält die in den Bibliotheken ermittelten doppelten Einträge. Wenn ein Reißzwecken-Symbol hinter einem Term angezeigt wird, ist das Vorkommen dieses Terms erzwungen worden. Wenn ein schwarzes X-Symbol angezeigt wird, wird das Vorkommen dieses Terms bei der Extraktion ignoriert, weil er an anderer Stelle erzwungen worden ist.
- **Benutzerdefiniert.** Diese Registerkarte enthält eine Liste der Terme, die manuell im Termbereich des Typwörterbuchs und nicht durch Konflikte erzwungen worden sind.

Hinweis: Das Dialogfeld "Erzwungene Terme bearbeiten" wird geöffnet, nachdem Sie Einträge 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

1. Wählen Sie im Dialogfeld "Erzwungene Terme bearbeiten" das Optionsfeld in der Spalte "Verwenden" für den zu erzwingenden Term.
2. Nach Abschluss der Auswahl klicken Sie auf **OK**, um die erzwungenen Terme 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 **Typwörterbuch** enthält eine Zusammenstellung von Termen, die unter einer Beschriftung oder einem Typnamen gruppiert sind. Beim Lesen der Textdaten vergleicht die Extraktionsengine die im Text gefundenen Wörter mit den Termen, die in den Typwörterbüchern definiert sind. Bei der Extraktion werden die gebeugten Formen der Terme und Synonyme eines Typs unter einem Zielterm gruppiert, der als Konzept bezeichnet wird. Extrahierte Konzepte werden dem Typwörterbuch zugewiesen, in dem sie als Terme angezeigt werden. Sie können Ihre Typwörterbücher in den oben links und in der Mitte gelegenen Bereichen des Editors (im Bibliotheksbaum und im Termbereich) verwalten. Weitere Informationen finden Sie im Thema „Typwörterbücher".
- Das **Substitutionswörterbuch** enthält eine Zusammenstellung von Wörtern, die als Synonyme oder optionale Elemente definiert sind und zur Gruppierung ähnlicher Terme unter einem Zielterm verwendet werden, dem sogenannten "Konzept" in den Extraktionsendergebnissen. Sie können Ihre Substitutionswörterbücher im unten links gelegenen Bereich des Editors über die Registerkarte "Synonyme" und die Registerkarte "Optional" verwalten. Weitere Informationen finden Sie im Thema „Substitutions-/Synonymwörterbücher" auf Seite 208.
- Das **Ausschlusswörterbuch** enthält eine Sammlung von Termen und Typen, die aus den endgültigen Extraktionsergebnissen entfernt werden. Sie können Ihre Ausschlusswörterbücher im ganz rechts gelegenen Bereich des Editors verwalten. Weitere Informationen finden Sie im Thema „Ausschlusswörterbücher" auf Seite 212.

Weitere Informationen finden Sie im Thema Kapitel 16, „Arbeiten mit Bibliotheken", auf Seite 191.

Typwörterbücher

Ein **Typwörterbuch** besteht aus dem Typnamen bzw. der Beschriftung und einer Liste von Termen. Typwörterbücher 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 Workbenchsitzung befinden. Andernfalls können Sie Wörterbücher für eine bestimmte Vorlage im Vorlageneditor bearbeiten.

Wenn die Extraktionsengine Ihre Textdaten liest, vergleicht sie die im Text gefundenen Wörter mit den in Ihren Typwörterbüchern definierten Termen. Terme sind Wörter oder Wortfolgen in den Typwörterbüchern in Ihren linguistischen Ressourcen.

Wenn ein Wort mit einem Term übereinstimmt, wird es dem Typnamen für diesen Term zugeordnet. Wenn die Ressourcen während des Extraktionsprozesses gelesen werden, werden die im Text gefundenen Terme einigen Verarbeitungsschritten unterzogen, bevor sie Konzepte im Bereich "Extraktionsergebnisse" werden. Wenn mehrere Terme, die zu demselben Typwörterbuch gehören, von der Extraktionsengine als synonym eingestuft werden, werden sie unter dem am häufigsten auftretenden Term zusammengefasst und im Bereich "Extraktionsergebnisse" als **Konzept** zusammengefasst. So könnten beispielsweise die Terme `Frage` und `Abfrage` letzten Endes unter dem Konzeptnamen `Frage` zusammengefasst werden.

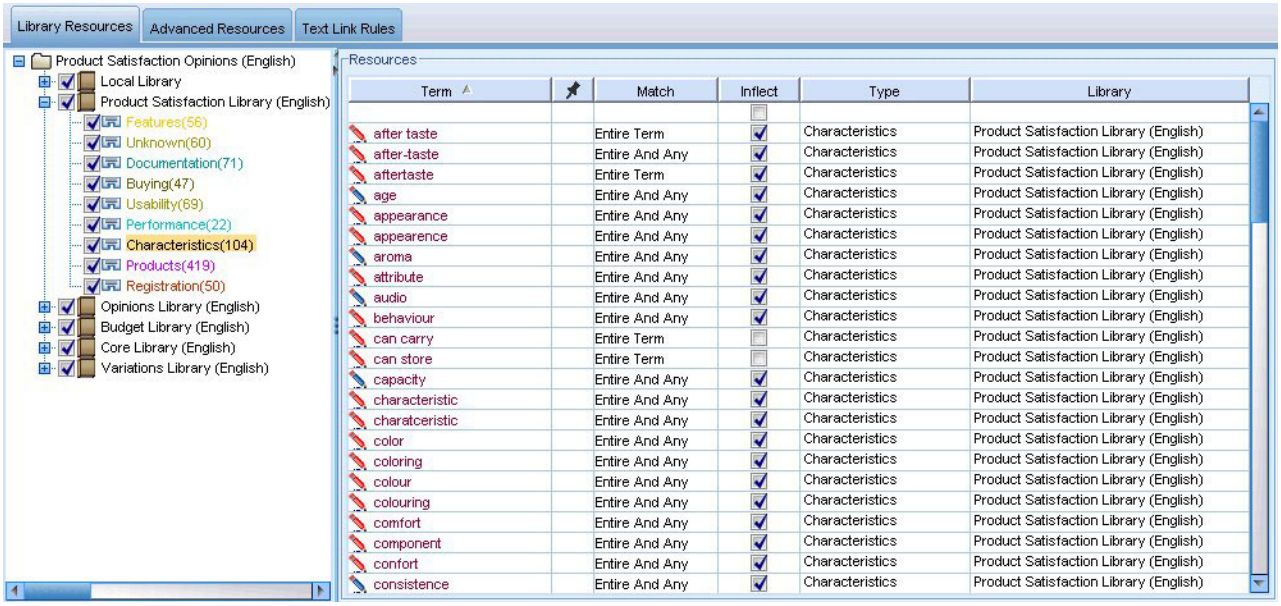


Abbildung 40. Bibliotheksbaum und Termbereich

Die Liste der Typwörterbücher wird links im Bereich des Wörterbuchbaums angezeigt. Der Inhalt der einzelnen Typwörterbücher wird im mittleren Bereich angezeigt. Typwörterbücher enthalten mehr als nur eine Liste mit Termen. Die festgelegte Abgleichsoption bestimmt, wie die Übereinstimmung der in Ihren Textdaten enthaltenen Wörter und Wortfolgen mit den in den Typwörterbüchern definierten Termen ermittelt wird. Eine **Abgleichsoption** legt fest, wie ein Term hinsichtlich eines in den Textdaten enthaltenen möglichen Worts oder einer möglichen Wortfolge verankert ist. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen" auf Seite 204.

Hinweis: Für japanischen Text gelten nicht alle Optionen, z. B. die Abgleichsoption und gebeugte Formen.

Darüber hinaus können Sie die Terme in Ihrem Wörterbuch ausweiten, indem Sie angeben, ob Sie im Wörterbuch automatisch gebeugte Formen der Terme generieren und hinzufügen möchten. Wenn Sie gebeugte Formen generieren, werden dem Typwörterbuch automatisch Pluralformen für im Singular angegebene Terme, Singularformen für im Plural angegebene Terme und Adjektive hinzugefügt. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen" auf Seite 204.

Hinweis: Für die meisten Sprachen gilt: Konzepte , die in keinem Typwörterbuch gefunden, aber aus dem Text extrahiert werden, erhalten automatisch den Typ `<Unknown>`

Integrierte Typen

IBM SPSS Modeler Text Analytics wird mit einem Set linguistischer Ressourcen geliefert, die als Bibliotheken und kompilierte Ressourcen vorliegen. Die mitgelieferten Bibliotheken enthalten ein Set integrierter Typwörterbücher, z. B. `<Location>`, `<Organization>`, `<Person>` und `<Product>`

Hinweis: Das Set der integrierten Standardtypen für japanischen Text setzt sich anders zusammen.

Diese Typwörterbücher werden von der Extraktionsengine verwendet, um den extrahierten Konzepten Typen zuzuweisen, z. B. den Typ `<Location>` zum Konzept `Paris`. Obwohl in den integrierten Wörterbüchern eine Vielzahl von Termen definiert ist, decken diese nicht alle Möglichkeiten ab. Sie können daher Terme hinzufügen oder eigene Wörterbücher anlegen. Eine Beschreibung der Inhalte eines der mitgelieferten Typwörterbücher 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 aus.

Hinweis: Neben den mitgelieferten Bibliotheken enthalten die kompilierten Ressourcen (die ebenfalls von der Extraktionsengine 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 Term, dessen Typ durch die kompilierten Wörterbücher bestimmt wurde, in ein anderes Wörterbuch übertragen wird. Weitere Informationen finden Sie im Thema „Erzwingen von Termen” auf Seite 207.

Erstellen von Typen

Sie können Typwörterbücher erstellen, mit denen ähnliche Terme gruppiert werden können. Wenn während des Extraktionsprozesses in diesen Wörterbüchern vorhandene Terme 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 Termen beginnen.

Wichtig!: Für japanische Ressourcen können Sie keine neuen Typen erstellen.

Wenn Sie Text über Nahrungsmittel analysieren und Terme zum Thema Gemüse gruppieren möchten, können Sie Ihr eigenes Typwörterbuch `<Gemüse>` erstellen. Dort können Sie dann Terme wie `Karotte`, `Brokkoli` und `Spinat` hinzufügen, sofern Sie der Meinung sind, dass es sich um wichtige Terme handelt, die im Text vorkommen. Wenn dann während der Extraktion einer dieser Terme 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 Terme stattdessen generieren lassen. Wenn Sie diese Option auswählen, erkennt die Extraktionsengine automatisch die Singular- bzw. Pluralformen der Terme sowie andere zu diesem Typ 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.

Das Dialogfeld "Typeigenschaften" enthält die folgenden Felder.

Name. Der Name des Typwörterbuchs, das sie gerade erstellen. Es wird empfohlen, keine Leerzeichen in Typnamen zu verwenden, insbesondere wenn mindestens zwei Typnamen mit demselben Wort beginnen.

Hinweis: Es gibt einige Beschränkungen für Typnamen und die Verwendung von Symbolen. Verwenden Sie beispielsweise keine Symbole wie "@" oder "!" im Namen.

Standardabgleich. Das Attribut "Standardabgleich" legt fest, wie die Extraktionsengine die Übereinstimmung dieses Terms mit den Textdaten ermittelt. Jedes Mal, wenn Sie diesem Typwörterbuch einen Term hinzufügen, wird diesem automatisch dieses Abgleichsattribut zugewiesen. Die Abgleichseinstellung können Sie in der Termliste jederzeit manuell ändern. Zu den Optionen gehören: **Gesamter Term**, **Start**, **End**, **Alle**, **Start oder Ende**, **Gesamt und Start**, **Gesamt und Ende**, **Gesamt und (Start oder Ende)** und **Gesamt (keine Zusammensetzungen)**. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen” auf Seite 204. Diese Option gilt nicht für japanische Ressourcen.

Hinzufügen zu. Dieses Feld gibt die Bibliothek an, in der Sie Ihr neues Typwörterbuch erstellen.

Gebeugte Formen standardmäßig generieren. Diese Option weist die Extraktionsengine an, die grammatikalische Morphologie zu verwenden, um ähnliche Formen der Terme, die Sie diesem Wörterbuch hinzufügen, zu erfassen und gruppieren (z. B. Singular- oder Pluralformen des Terms). Diese Option ist besonders hilfreich, wenn Ihr Typ überwiegend Substantive enthält. Wenn Sie diese Option auswählen, besitzen alle diesem Typ hinzugefügten neuen Terme 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 Benutzerschnittstelle dargestellten Typen festlegen. Wenn Sie die Option **Übergeordnete**

Farbe verwenden auswählen, wird für dieses Typwörterbuch die Standardtypfarbe verwendet. Die Standardfarbe wird im Dialogfeld "Optionen" festgelegt. Weitere Informationen finden Sie im Thema „Optionen: Registerkarte "Anzeigen"" auf Seite 88. Wenn Sie **Benutzerdefiniert** auswählen, wählen Sie aus der Dropdown-Liste eine Farbe aus.

Anmerkung. Dieses Feld ist optional und kann für Kommentare und Beschreibungen verwendet werden.

So erstellen Sie ein Typwörterbuch

1. Wählen Sie eine Bibliothek aus, in der Sie ein neues Typwörterbuch erstellen möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Tools > Neuer Typ** aus. Das Dialogfeld "Typ-Eigenschaften" wird geöffnet.
3. Geben Sie den Namen für Ihr Typwörterbuch in das Textfeld **Name** ein und wählen Sie die gewünschten Optionen aus.
4. Klicken Sie auf **OK**, um das Typwörterbuch zu erstellen. Der neue Typ wird im Bibliotheksbaum und im mittleren Bereich angezeigt. Sie können sofort beginnen, Terme hinzuzufügen. Weitere Informationen finden Sie in „Hinzufügen von Termen".

Hinweis: Diese Anweisungen zeigen auf, wie Sie in der Ressourceneditoransicht oder im Vorlageneditor Änderungen vornehmen können. Beachten Sie, dass Sie solche Optimierungen auch direkt im Bereich "Extraktionsergebnisse", im Datenbereich, im Kategoriebereich oder im Dialogfeld "Clusterdefinitionen" in den anderen Ansichten vornehmen können. Weitere Informationen finden Sie im Thema „Optimieren von Extraktionsergebnissen" auf Seite 101.

Hinzufügen von Termen

Im Bereich des Bibliotheksbaums werden die Bibliotheken angezeigt. Diese können erweitert werden, um die darin enthaltenen Typwörterbücher anzuzeigen. Im mittleren Bereich zeigt eine Liste je nach Auswahl im Baum die in der ausgewählten Bibliothek oder im ausgewählten Typwörterbuch enthaltenen Terme an.

Wichtig! Terme werden für japanische Ressourcen abweichend definiert.

Im Ressourceneditor können Sie Terme einem Typwörterbuch direkt im Termbereich oder über das Dialogfeld "Neue Terme hinzufügen" hinzufügen. Bei den hinzugefügten Termen 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 auf, wie Sie in der Ressourceneditoransicht oder im Vorlageneditor Änderungen vornehmen können. Beachten Sie, dass Sie solche Optimierungen auch direkt im Bereich "Extraktionsergebnisse", im Datenbereich, im Kategoriebereich oder im Dialogfeld "Clusterdefinitionen" in den anderen Ansichten vornehmen können. Weitere Informationen finden Sie im Thema „Optimieren von Extraktionsergebnissen" auf Seite 101.

Termspalte

Geben Sie in dieser Spalte ein Wort oder eine Wortfolge in die Zelle ein. In welcher Farbe der Term angezeigt wird, hängt von der Farbe des Typs ab, in dem der Term gespeichert wurde. Die Typfarben können Sie im Dialogfeld "Typeigenschaften" ändern. Weitere Informationen finden Sie im Thema „Erstellen von Typen" auf Seite 203.

Erzwingungsspalte

Wenn Sie in dieser Spalte in dieser Zelle ein Reißzweckensymbol hinzufügen, weisen Sie die Extraktionsengine an, das Vorkommen desselben Terms in allen anderen Bibliotheken zu ignorieren. Weitere Informationen finden Sie im Thema „Erzwingen von Termen" auf Seite 207.

Übereinstimmungsspalte

Wählen Sie in dieser Spalte eine Übereinstimmungsoption aus, um festzulegen, wie die Extraktionsengine die Übereinstimmung dieses Terms mit den Textdaten ermittelt. Beispiele finden Sie in der Tabelle. Sie können den Standardwert ändern, indem Sie die Typeigenschaften bearbeiten. Weitere Informationen finden Sie im Thema „Erstellen von Typen" auf Seite 203. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Abgleich ändern** aus. Folgende Beispiele sind grundlegende Abgleichsoptionen, da Kombinationen aus ihnen ebenfalls möglich sind:

- **Start**. Dieser Typ wird zugewiesen, wenn der im Wörterbuch gefundene Term mit dem ersten Wort eines aus dem Text extrahierten Konzepts übereinstimmt. Wenn Sie beispielsweise `Apfel` eingeben, gilt `Apfelkuchen` als Übereinstimmung.
- **Ende**. Dieser Typ wird zugewiesen, wenn der im Wörterbuch gefundene Term mit dem letzten Wort eines aus dem Text extrahierten Konzepts übereinstimmt. Wenn Sie beispielsweise `Apfel` eingeben, gilt `Klarapfel` als Übereinstimmung.
- **Alle**. Dieser Typ wird zugewiesen, wenn der im Wörterbuch gefundene Term mit irgendeinem Wort eines aus dem Text extrahierten Konzepts übereinstimmt. Wenn Sie z. B. `Apfel` eingeben, sorgt die Option **Alle** dafür, dass `Apfelkuchen`, `Klarapfel` und `Klarapfelkuchen` demselben Typ zugeordnet werden.
- **Gesamter Term**. Dieser Typ wird zugewiesen, wenn das gesamte aus dem Text extrahierte Konzept exakt mit dem im Wörterbuch vorhandenen Term übereinstimmt. Das Hinzufügen eines Terms als **Gesamter Term**, **Gesamt und Start**, **Gesamt und Ende**, **Gesamt und Beliebig** oder **Gesamt (keine Zusammensetzungen)** erzwingt die Extraktion eines Terms.

  Da der Typ `<Person>` außerdem nur zweiteilige Namen wie *edith piaf* oder *mohandas gandhi* extrahiert, sollten Sie gegebenenfalls explizit die Vornamen in dieses Typwörterbuch aufnehmen, wenn Sie versuchen, einen Vornamen zu extrahieren, wenn kein Nachname angegeben ist. Wenn Sie beispielsweise alle Instanzen von *edith* als Namen erfassen möchten, sollten Sie `edith` dem Typ `<Person>` über **Gesamter Term** oder **Gesamt und Start** hinzufügen.
- **Gesamt (keine Zusammensetzungen)**. Wenn das gesamte aus dem Text extrahierte Konzept exakt mit dem Term im Wörterbuch übereinstimmt, wird dieser Typ zugewiesen und die Extraktion wird beendet, um zu verhindern, dass der Term bei der Extraktion mit einer längeren Zusammensetzung abgeglichen wird. Wenn Sie beispielsweise `Apfel` eingeben, wird dem Term `Apfel` als Typ **Gesamt (keine Zusammensetzung)** zugeordnet, die Zusammensetzung `Grüner Apfel` wird jedoch nicht extrahiert, sofern dies nicht an anderer Stelle erzwungen wird.

In der folgenden Tabelle wird davon ausgegangen, dass der Term `Apfel` in einem Typwörterbuch enthalten ist. Je nach Abgleichsoption wird in dieser Tabelle angezeigt, welche Konzepte extrahiert und mit Typen versehen würden, wenn sie im Text gefunden würden.

Tabelle 38. Abgleichsbeispiele.

Abgleichsoptionen für den Term:  **Apfel**	**Extrahierte Konzepte**			
	Apfel	**Apfel***kuchen*	*reifer* **Apfel**	*selbst gemachter* **Apfel***kuchen*
Gesamter Term	✔			
Start		✔		
Ende			✔	
Start oder Ende		✔	✔	

Tabelle 38. Abgleichsbeispiele (Forts.).

Abgleichsoptionen für den Term:  **Apfel**	**Extrahierte Konzepte**			
	Apfel	**Apfel***kuchen*	*reifer* **Apfel**	*selbst gemachter* **Apfel***kuchen*
Gesamt und Start	✔	✔		
Gesamt und Ende	✔		✔	
Gesamt und (Start oder Ende)	✔	✔	✔	
Alle		✔	✔	✔
Gesamt und Beliebig	✔	✔	✔	✔
Gesamt (keine Zusammensetzungen)	✔	*nie extrahiert*	*nie extrahiert*	*nie extrahiert*

Flektierungsspalte

Wählen Sie in dieser Spalte aus, ob die Extraktionsengine während der Extraktion gebeugte Formen dieses Terms generieren soll, sodass 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 in den Menüs die Optionsfolge **Bearbeiten > Beugung ändern** aus.

Typspalte

Wählen Sie in dieser Spalte ein Typwörterbuch 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 in den Menüs die Optionsfolge **Bearbeiten > Typ ändern** aus.

Bibliotheksspalte

In dieser Spalte wird die Bibliothek angezeigt, in der Ihr Term gespeichert wird. Sie können einen Term mit der Maus auf einen anderen Typ im Bibliotheksbaum ziehen, um seine Bibliothek zu ändern.

So fügen Sie einen einzelnen Term einem Typwörterbuch hinzu

1. Wählen Sie im Bibliotheksbaum das Typwörterbuch aus, zu dem Sie den Term hinzufügen möchten.
2. Geben Sie in der mittleren Liste der Terme Ihren Term in die erste verfügbare leere Zelle ein und wählen Sie die gewünschten Optionen für diesen Term aus.

So fügen Sie mehrere Terme einem Typwörterbuch hinzu

1. Wählen Sie im Bibliotheksbaumbereich das Typwörterbuch aus, dem Sie die Terme hinzufügen möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Tools > Neue Terme** aus. Das Dialogfeld "Neue Terme hinzufügen" wird geöffnet.

3. Geben Sie die Terme ein, die Sie zum ausgewählten Typwörterbuch hinzufügen möchten. Sie können die Terme eingeben oder mehrere Terme kopieren und einfügen. Wenn Sie mehrere Terme eingeben, müssen Sie diese mit dem Trennzeichen, das im Dialogfeld "Optionen" definiert ist, trennen oder jeden Term in einer neuen Zeile eingeben. Weitere Informationen finden Sie im Thema „Festlegen von Optionen” auf Seite 88.
4. Klicken Sie auf **OK**, um die Terme zum Wörterbuch hinzuzufügen. Als Abgleichsoption wird automatisch die Standardoption für dieses Typwörterbuch festgelegt. Das Dialogfeld wird geschlossen und die neuen Terme erscheinen im Wörterbuch.

Erzwingen von Termen

Wenn Sie möchten, dass ein Term einem bestimmten Typ zugewiesen wird, können Sie ihn zum entsprechenden Typwörterbuch hinzufügen. Wenn jedoch mehrere Terme mit demselben Namen vorliegen, muss die Extraktionsengine wissen, welcher Typ verwendet werden soll. Sie werden daher aufgefordert, den zu verwendenden Typ auszuwählen. Dies wird als **Erzwingen** der Zuordnung eines Terms zu einem Typ bezeichnet. Diese Option ist vor allem beim Überschreiben der Typzuordnung aus einem kompilierten (internen, nicht bearbeitbaren) Wörterbuch hilfreich. Generell wird empfohlen, doppelte Terme von vornherein zu vermeiden.

Durch das Erzwingen werden die anderen Stellen, an denen der Term vorkommt, nicht *entfernt*. Diese werden stattdessen von der Extraktionsengine ignoriert. Sie können später ändern, welche Fundstelle verwendet wird, indem Sie das Erzwingen eines Terms festlegen oder aufheben. Es kann außerdem erforderlich sein, dass Sie erzwingen, dass ein Term in ein Typwörterbuch eingefügt wird, wenn Sie eine öffentliche Bibliothek hinzufügen oder aktualisieren.

In der Erzwingungsspalte (2. Spalte im Termbereich) sehen Sie, welche Terme erzwungen oder ignoriert werden. Wenn ein Reißzweckensymbol angezeigt wird, bedeutet dies, dass dieses Vorkommen des Terms erzwungen wurde. Wenn ein schwarzes X-Symbol angezeigt wird, bedeutet dies, dass dieses Vorkommen des Terms während des Extrahierens ignoriert wird, da er anderweitig erzwungen wurde. Wenn Sie einen Term erzwingen, wird er außerdem in der Farbe des Typs angezeigt, für den er erzwungen wurde. Dies bedeutet, dass ein Term, 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 Term an einer anderen Stelle vorkommt, wird das Dialogfeld "Konflikte auflösen" geöffnet, in dem Sie auswählen können, welche Fundstelle verwendet wird.

Umbenennen von Typen

Sie können ein Typwörterbuch umbenennen oder andere Wörterbucheinstellungen ändern, indem Sie die Typeigenschaften bearbeiten.

Wichtig! Es wird empfohlen, keine Leerzeichen in Typnamen zu verwenden, insbesondere wenn mindestens zwei Typnamen mit demselben Wort beginnen. Es ist außerdem empfehlenswert, die Typen in der Kernbibliothek und Opinions Library nicht umzubenennen oder ihre Standardübereinstimmungsattribute zu ändern.

So benennen Sie einen Typ um

1. Wählen Sie im Bereich des Bibliotheksbaums das Typwörterbuch aus, das Sie umbenennen möchten.
2. Klicken Sie mit der rechten Maustaste und wählen Sie im Kontextmenü **Typeigenschaften** aus. Das Dialogfeld "Typ-Eigenschaften" wird geöffnet.
3. Geben Sie den neuen Namen für Ihr Typwörterbuch in das Textfeld "Name" ein.
4. Klicken Sie auf **OK**, um den neuen Namen zu übernehmen. Der neue Name wird im Bibliotheksbaum angezeigt.

Verschieben von Typen

Sie können ein Typwörterbuch 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

1. Wählen Sie im Bereich des Bibliotheksbaums das Typwörterbuch aus, das Sie verschieben möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Nach oben verschieben** aus, um das Typwörterbuch 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

1. Wählen Sie im Bereich des Bibliotheksbaums das Typwörterbuch aus, das Sie verschieben möchten.
2. 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.)
3. Wählen Sie im Feld "Hinzufügen zu" die Bibliothek aus, in die Sie das Typwörterbuch verschieben möchten.
4. Klicken Sie auf **OK**. Das Dialogfeld wird geschlossen und der Typ befindet sich in der von Ihnen ausgewählten Bibliothek.

Inaktivieren und Löschen von Typen

Wenn Sie ein Typwörterbuch vorübergehend entfernen möchten, können Sie im Bibliotheksbaum das Kontrollkästchen links neben dem Namen des Wörterbuchs inaktivieren. Dies bewirkt, dass das Wörterbuch in Ihrer Bibliothek bleibt, dass die Inhalte bei der Prüfung von Konflikten und während des Extraktionsvorgangs aber ignoriert werden.

Sie können Typwörterbücher auch dauerhaft aus einer Bibliothek entfernen.

So inaktivieren Sie ein Typwörterbuch

1. Wählen Sie im Bereich des Bibliotheksbaums das Typwörterbuch aus, das Sie inaktivieren möchten.
2. Drücken Sie die Leertaste. Das Kontrollkästchen links neben dem Typnamen wird inaktiviert.

So löschen Sie ein Typwörterbuch

1. Wählen Sie im Bereich des Bibliotheksbaums das Typwörterbuch aus, das Sie löschen möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Löschen** aus, um das Typwörterbuch zu löschen.

Substitutions-/Synonymwörterbücher

Ein **Substitutionswörterbuch** ist eine Sammlung von Termen, mit deren Hilfe ähnliche Terme unter einem Zielterm 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 Workbenchsitzung befinden. Andernfalls können Sie Wörterbücher für eine bestimmte Vorlage im Vorlageneditor bearbeiten.

In diesem Wörterbuch 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 Extraktion 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 Extraktionsengine diese einem einzigen Zielterm zuordnet.

Die Substitution der Verwendung von Synonymen und optionalen Elementen verringert die Anzahl an Konzepten im Bereich "Extraktionsergebnisse", indem sie in sinnvollere, aussagekräftigere Konzepte mit einer höheren Häufigkeit (Dokumentenanzahl) zusammengefasst werden.

Hinweis: Für japanische Ressourcen gelten optionale Elemente nicht und sind daher nicht verfügbar. Zusätzlich werden Synonyme für japanischen Text etwas abweichend behandelt.

Synonyme

Synonyme verknüpfen mindestens zwei Wörter mit derselben Bedeutung. Mithilfe von Synonymen können Sie außerdem Terme mit ihren Abkürzungen gruppieren oder auch falsch geschriebene Wörter mit dem Term 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 **Ziel**term, d. h. der Term, unter dem die Extraktionsengine alle Synonyme zusammenfassen soll. Wenn dieser Zielterm nicht als Synonym eines anderen Zielterms verwendet wird oder er ausgeschlossen wird, ist es wahrscheinlich, dass dieser Term das Konzept wird, das im Bereich "Extraktionsergebnisse" angezeigt wird. Der zweite Teil besteht aus einer Liste von Synonymen, die unter dem Zielterm zusammengefasst werden.

Wenn Sie beispielsweise `Automobil` durch `Fahrzeug` ersetzen möchten, dann ist `Automobil` das Synonym und `Fahrzeug` der Zielterm.

Sie können jedes beliebige Wort in die Spalte **Synonym** setzen, wenn das Wort jedoch bei der Extraktion nicht gefunden wird und der Term eine Abgleichsoption mit `Gesamt` hatte, kann keine Substitution vorgenommen werden. Der Zielterm muss jedoch nicht extrahiert werden, damit die Synonyme unter diesem Term zusammengefasst werden.

Optionale Elemente

Optionale Elemente kennzeichnen optionale Wörter in einem zusammengesetzten Term bzw. einer Wortfolge, die während der Extraktion ignoriert werden können, um ähnliche Terme auch dann zusammenzuhalten, wenn sie im Text leicht unterschiedlich auftreten. Optionale Elemente sind einzelne Wörter, die, wenn sie aus einer Zusammensetzung entfernt werden, eine Übereinstimmung mit einem anderen Term darstellen können. Diese einzelnen Wörter können an einer beliebigen Stelle innerhalb des zusammengesetzten Worts vorkommen - am Anfang, in der Mitte oder am Ende. Sie können optionale Elemente auf der Registerkarte "Optional" definieren.

Um z. B. die Terme `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 Extraktion sowohl `internet access speed` als auch `internet speed` gefunden werden, dann werden diese unter dem Term gruppiert, der am häufigsten vorkommt.

Hinweis: Für japanische Textressourcen steht die Registerkarte "Optionale Elemente" nicht zur Verfügung, da optionale Elemente nicht gelten.

Definieren von Synonymen

Auf der Registerkarte "Synonyme" können Sie in die Leerzeile am Anfang der Tabelle eine Synonymdefinition eingeben. Beginnen Sie, indem Sie den Zielterm und seine Synonyme definieren. Sie können auch die Bibliothek auswählen, in der diese Definition gespeichert werden soll. Während der Extraktion werden alle Fundstellen der Synonyme für die endgültige Extraktion unter dem Zielterm gruppiert. Weitere Informationen finden Sie im Thema „Hinzufügen von Termen” auf Seite 204.

Wenn Ihre Textdaten viele Fachausdrücke aus der Telekommunikation enthalten, liegen beispielsweise folgende Terme 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 Term wie wireless phone behandelt und zusammen in der Liste der Terme ausgegeben.

Wenn Sie Ihre Typwörterbücher erstellen, können Sie einen Term eingeben, für den Ihnen drei oder vier Synonyme einfallen. In diesem Fall könnten Sie alle Terme und anschließend Ihren Zielterm in das Substitutionswörterbuch eingeben und dann die Synonyme durch Ziehen übertragen.

Hinweis: Bei japanischen Texten werden Synonyme etwas anders gehandhabt.

Die Substitution von Synonymen wird auch auf gebeugte Formen (wie Pluralformen) der Synonyme angewendet. Je nach Kontext sollten Sie Einschränkungen für die Substitution der Terme 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 Zielterm ersetzt werden. Ein Ausrufezeichen direkt vor dem Zielterm (!Zielterm) bedeutet hingegen, dass kein weiterer Bestandteil dieses zusammengesetzten Zielterms oder Varianten davon ersetzt werden sollen.
- **Stern (*).** Ein direkt nach dem Synonym stehender Stern (*) (Synonym*) bedeutet, dass dieses Wort durch den Zielterm ersetzt werden soll. Wenn Sie beispielsweise den Quellenterm manage* als Synonym und management als Zielterm definiert haben, dann wird associate managers durch den Zielterm associate management ersetzt. Sie können nach dem Wort auch ein Leerzeichen und einen Stern (*) einfügen (Synonym *); Beispiel: internet *. Wenn Sie den Zielterm 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 Stern (*) als Platzhalterzeichen beginnen.
- **Winkelzeichen (^).** Ein vor einem Synonym stehendes Winkelzeichen und ein Leerzeichen (^ Synonym) bedeuten, dass die Synonymgruppierung nur dann durchgeführt wird, wenn der Term mit dem Synonym beginnt. Wenn Sie beispielsweise ^ wage als das Synonym und income als den Zielterm definiert haben und beide Terme extrahiert werden, dann werden beide unter dem Term 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 Term mit dem Synonym endet. Wenn Sie beispielsweise cash $ als das Synonym und money als den Zielterm definiert haben und beide Terme extrahiert werden, dann werden beide unter dem Term 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 (z. B. ^ Synonym $), stimmt ein Term nur bei exakter Übereinstimmung mit dem Synonym überein. Dies bedeutet, dass im extrahierten Term vor oder nach dem Synonym keine Wörter stehen dürfen, damit die Synonymgruppierung stattfindet. Beispiel: Sie definieren ^ van $ als Synonym und truck als Ziel, 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 Quellentexts auftritt, wird das Synonym außerdem automatisch extrahiert.

Hinweis: Diese Sonderzeichen und Platzhalter werden für japanischen Text nicht unterstützt.

So fügen Sie einen Synonymeintrag hinzu

1. Klicken Sie im Substitutionsbereich oben links auf die Registerkarte **Synonyme**.
2. Geben Sie in der leeren Zeile am Anfang der Tabelle Ihren Zielterm in die Zielspalte ein. Der von Ihnen eingegebene Zielterm wird farbig angezeigt. Die Farbe stellt den Typ dar, als der der Term er-

scheint oder in den er erzwungen wird, sofern dies der Fall ist. Wenn der Term schwarz dargestellt wird, bedeutet dies, dass er in keinem der Typwörterbücher vorkommt.

3. Klicken Sie auf die zweite Zelle rechts neben dem Zielterm und geben Sie die Synonyme ein. Trennen Sie die einzelnen Einträge mithilfe des globalen Trennzeichens, das im Dialogfeld "Optionen" definiert ist. Weitere Informationen finden Sie im Thema „Festlegen von Optionen” auf Seite 88. Die von Ihnen eingegebenen Terme werden farbig angezeigt. Die Farbe stellt den Typ dar, in der der Term angezeigt wird. Wenn der Term schwarz dargestellt wird, bedeutet dies, dass er in keinem der Typwörterbücher vorkommt.
4. Klicken Sie auf die letzte Zelle, um die Bibliothek auszuwählen, in der die Synonymdefinition gespeichert werden soll.

Hinweis: Diese Anweisungen zeigen auf, wie Sie in der Ressourceneditoransicht oder im Vorlageneditor Änderungen vornehmen können. Beachten Sie, dass Sie solche Optimierungen auch direkt im Bereich "Extraktionsergebnisse", im Datenbereich, im Kategoriebereich oder im Dialogfeld "Clusterdefinitionen" in den anderen Ansichten vornehmen können. Weitere Informationen finden Sie im Thema „Optimieren von Extraktionsergebnissen” auf Seite 101.

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 Extraktionsengine gleicht Einträge mit im Text in Groß- oder Kleinbuchstaben vorkommenden Wörtern ab.

Hinweis: Für japanische Ressourcen gelten optionale Elemente nicht und sind daher nicht verfügbar.

Hinweis: Terme werden mithilfe des Trennzeichens getrennt, das im Dialogfeld "Optionen" definiert ist. Weitere Informationen finden Sie im Thema „Festlegen von Optionen” auf Seite 88. Wenn das von Ihnen eingegebene optionale Element das Trennzeichen als Teil des Terms enthält, müssen Sie diesem einen umgekehrten Schrägstrich (\) voranstellen.

So fügen Sie einen Eintrag hinzu

1. Klicken Sie im Editor im Substitutionsbereich unten links auf die Registerkarte "Optional".
2. Klicken Sie in der Spalte "Optionale Elemente" auf die Zelle der Bibliothek, zu der Sie diesen Eintrag hinzufügen möchten.
3. Geben Sie das optionale Element ein. Trennen Sie die einzelnen Einträge mithilfe des globalen Trennzeichens, das im Dialogfeld "Optionen" definiert ist. Weitere Informationen finden Sie im Thema „Festlegen von Optionen” auf Seite 88.

Inaktivieren und Löschen von Substitutionen

Sie können einen Eintrag vorübergehend entfernen, indem Sie ihn in Ihrem Wörterbuch inaktivieren. Inaktivierte Einträge werden bei der Extraktion ignoriert.

Sie können zudem veraltete Einträge aus Ihrem Substitutionswörterbuch löschen.

So inaktivieren Sie einen Eintrag

1. Wählen Sie in Ihrem Wörterbuch den Eintrag aus, den Sie inaktivieren möchten.
2. Drücken Sie die Leertaste. Das Kontrollkästchen links neben dem Eintrag wird inaktiviert.

Hinweis: Sie können auch links neben dem Eintrag auf das Kontrollkästchen klicken, um es zu inaktivieren.

So löschen Sie einen Synonymeintrag

1. Wählen Sie in Ihrem Wörterbuch den Eintrag aus, den Sie löschen möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Löschen** aus oder drücken Sie die Löschtaste auf Ihrer Tastatur. Der Eintrag wird aus dem Wörterbuch entfernt.

So löschen Sie einen Eintrag für ein optionales Element

1. Doppelklicken Sie in Ihrem Wörterbuch auf den Eintrag, den Sie löschen möchten.
2. Löschen Sie diesen Term manuell.
3. Drücken Sie die Eingabetaste, damit die Änderung wirksam wird.

Ausschlusswörterbücher

Ein **Ausschlusswörterbuch** ist eine Liste von Wörtern, Wortfolgen oder Teilzeichenfolgen. Terme, die mit einem Eintrag im Ausschlusswörterbuch übereinstimmen, werden ignoriert oder aus der Extraktion ausgeschlossen. Ausschlusswörterbücher werden im rechten Bereich des Editors verwaltet. In der Regel handelt es sich bei den zu dieser Liste hinzugefügten Termen um Füllwörter oder -wortfolgen, die im Text verwendet werden, aber keine wichtigen Informationen enthalten und daher die Extraktionsergebnisse 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 Workbenchsitzung befinden. Andernfalls können Sie Wörterbücher für eine bestimmte Vorlage im Vorlageneditor bearbeiten.

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 den Stern als Platzhalter verwenden. Die im Ausschlusswörterbuch erfassten Einträge werden verwendet, um Konzepte von der Extraktion auszuschließen. Falls ein Eintrag auch an anderer Stelle deklariert ist, beispielsweise in einem Typwörterbuch, wird er in den anderen Wörterbüchern durchgestrichen dargestellt, was anzeigt, dass er zurzeit ausgeschlossen ist. Diese Zeichenfolge muss nicht in den Textdaten vorkommen oder als Teil eines Typwörterbuchs deklariert sein, um angewendet zu werden.

Hinweis: Wenn Sie dem Ausschlusswörterbuch ein Konzept hinzufügen, das in einem Synonymeintrag als Ziel verwendet wird, dann werden auch das Ziel und alle Synonyme ausgeschlossen. Weitere Informationen finden Sie im Thema „Definieren von Synonymen” auf Seite 209.

Verwenden von Platzhaltern (*)

Für alle Textsprachen außer Japanisch können Sie den Stern als Platzhalter verwenden, wenn der Ausschlusseintrag als Teilzeichenfolge behandelt werden soll. Alle Terme, die von der Extraktionsengine gefunden werden und ein Wort enthalten, das mit einer im Ausschlusswörterbuch angegebenen Zeichenfolge beginnt oder endet, werden von der endgültigen Extraktion ausgeschlossen. Es gibt jedoch zwei Fälle, in denen ein Platzhalterzeichen nicht zulässig ist:

- Bindestrich (-), vor dem ein Stern (*) als Platzhalterzeichen steht (*-)
- Apostroph ('), vor dem ein Stern (*) als Platzhalterzeichen steht (*'*s*)

Tabelle 39. Beispiele für Ausschlusseinträge.

Aufnahme	Beispiel	Ergebnisse
Wort	*weiter*	Es werden keine Konzepte (bzw. die darin enthaltenen Terme) extrahiert, die das Wort *weiter* enthalten.

Tabelle 39. Beispiele für Ausschlusseinträge (Forts.).

Aufnahme	Beispiel	Ergebnisse
Wortfolge	*zum Beispiel*	Es werden keine Konzepte (bzw. die darin enthaltenen Terme) extrahiert, die die Wortfolge *zum Beispiel* enthalten.
Teilweise	*Recht**	Schließt alle Konzepte (oder die zugehörigen Terme) aus, die Variationen des Worts *Rechtsschutz* enthalten, wie `recht haben`, `rechtens`, `Rechte` oder `rechts`.
Teilweise	**ware*	Schließt alle Konzepte (bzw. die darin enthaltenen Terme) aus, die mit Variationen des Worts *ware* übereinstimmen bzw. diese enthalten, wie `Freeware`, `Shareware`, `Software`, `Hardware`, `Handelsware` oder `Einkaufsware`.

So fügen Sie Einträge hinzu

1. Geben Sie in die Leerzeilen am Anfang der Tabelle einen Term ein. Der von Ihnen eingegebene Term wird farbig angezeigt. Die Farbe stellt den Typ dar, in der der Term angezeigt wird. Wenn der Term schwarz dargestellt wird, bedeutet dies, dass er in keinem der Typwörterbücher vorkommt.

So inaktivieren Sie Einträge

Sie können einen Eintrag vorübergehend entfernen, indem Sie ihn in Ihrem Ausschlusswörterbuch inaktivieren. Inaktivierte Einträge werden bei der Extraktion ignoriert.

1. Wählen Sie in Ihrem Ausschlusswörterbuch den Eintrag aus, den Sie inaktivieren möchten.
2. Drücken Sie die Leertaste. Das Kontrollkästchen links neben dem Eintrag wird inaktiviert.

Hinweis: Sie können auch links neben dem Eintrag auf das Kontrollkästchen klicken, um es zu inaktivieren.

So löschen Sie Einträge

Sie können nicht mehr benötigte Einträge aus Ihrem Ausschlusswörterbuch löschen.

1. Wählen Sie in Ihrem Ausschlusswörterbuch den Eintrag aus, den Sie löschen möchten.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Löschen** aus. 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 Fuzzy-Gruppierung und nicht linguistische Typdefinitionen. Sie können auf der Registerkarte "Erweiterte Ressourcen" im Vorlageneditor oder Ressourceneditor mit diesen Ressourcen arbeiten.

Wichtig! Diese Registerkarte ist nicht für Ressourcen verfügbar, die für japanischen Text optimiert 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 optimiert werden. Weitere Informationen finden Sie im Thema „Zielsprache für Ressourcen” auf Seite 217.
- **Fuzzy-Gruppierung (Ausnahmen).** Hiermit werden Wortpaare aus dem Algorithmus für Fuzzy-Gruppierung ausgeschlossen (Rechtschreibfehlerkorrektur). Weitere Informationen finden Sie im Thema „Fuzzy-Gruppierung” auf Seite 218.
- **Nicht linguistische Entitäten.** Hiermit wird aktiviert bzw. inaktiviert, welche linguistischen Elemente extrahiert werden und welche regulären Ausdrücke und Normalisierungsregeln bei ihrer Extraktion angewendet werden. Weitere Informationen finden Sie im Thema „Nicht linguistische Entitäten” auf Seite 219.
- **Sprachbehandlung.** Darüber wird festgelegt, wie Sätze strukturiert werden (Extraktionsmuster und erzwungene Definitionen) und wie Abkürzungen für die ausgewählte Sprache verwendet werden. Weitere Informationen finden Sie im Thema „Sprachbehandlung” auf Seite 224.
- **Sprachenkennung.** Hiermit wird die automatische Sprachenkennung konfiguriert, die aufgerufen wird, wenn die Sprache auf **Alle** festgelegt ist. Weitere Informationen finden Sie im Thema „Language Identifier” auf Seite 225.

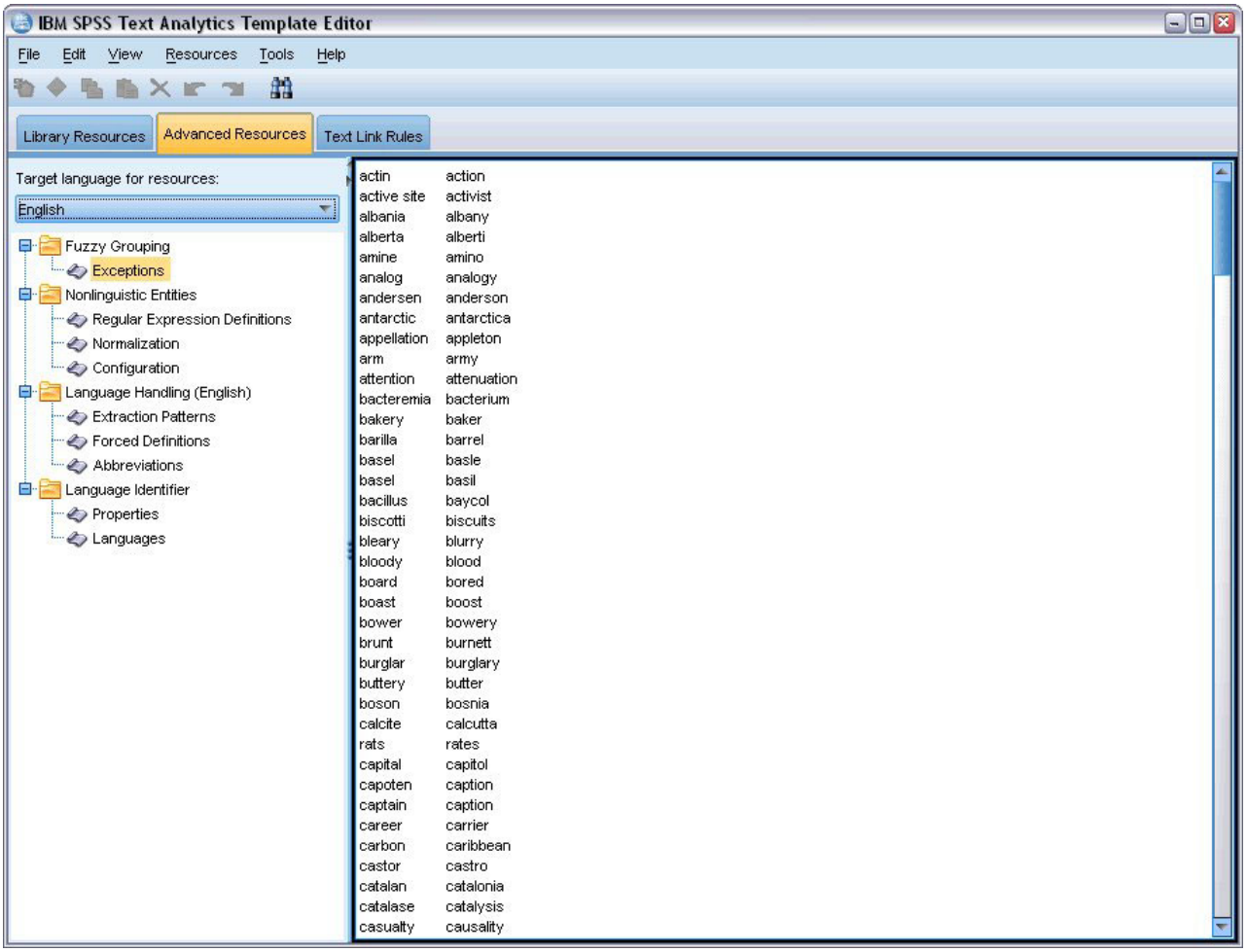


Abbildung 41. Textmining-Vorlageneditor - Registerkarte "Erweiterte Ressourcen"

Hinweis: Mit der Symbolleiste zum Suchen und Ersetzen können Sie Informationen schnell finden oder in einem Abschnitt identische Änderungen durchführen. Weitere Informationen finden Sie im Thema „Ersetzen" auf Seite 217.

So bearbeiten Sie erweiterte Ressourcen

1. Suchen Sie den Ressourcenabschnitt, den Sie bearbeiten möchten, und wählen Sie ihn aus. Der Inhalt wird im rechten Fensterbereich angezeigt.
2. Über die Menübefehle oder die Schaltflächen in der Symbolleiste können Sie Inhalte ausschneiden, kopieren oder einfügen.
3. 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 Textlinkanalyse durchführen, haben Sie eventuell Hunderte von Makros und Musterdefinitionen. Mit der Suchfunktion können Sie eine bestimmte Regel schnell finden. Für die Suche nach Informationen in einem Abschnitt können Sie die Symbolleiste "Suchen" verwenden.

So verwenden Sie die Suchfunktion

1. Suchen Sie den Ressourcenabschnitt, den Sie durchsuchen möchten, und wählen Sie ihn aus. Der Inhalt wird im rechten Bereich des Editors angezeigt.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Suchen** aus. Oben rechts im Dialogfeld "Erweiterte Ressourcen bearbeiten" wird die Symbolleiste "Suchen" angezeigt.
3. Geben Sie die Wortfolge, nach der Sie suchen möchten, in das Textfeld ein. Mit den Schaltflächen in 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.
4. Klicken Sie auf **Suchen**, um die Suche zu starten. Wenn eine Übereinstimmung gefunden wird, wird der Text im Fenster markiert.
5. Klicken Sie erneut auf **Suchen**, um nach der nächsten Übereinstimmung zu suchen.

Hinweis: 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"

1. 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.
2. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Ersetzen** aus. Das Dialogfeld "Ersetzen" wird geöffnet.
3. Geben Sie in das Textfeld **Suchen nach** das Wort ein, nach dem Sie suchen möchten.
4. Geben Sie in das Textfeld **Ersetzen durch** die Zeichenfolge ein, die Sie anstelle des gefundenen Texts verwenden möchten.
5. Wählen Sie die Option **Nur ganzes Wort** aus, wenn Sie nur vollständige Wörter suchen und ersetzen möchten.
6. Wählen Sie **Groß-/Kleinschreibung beachten**, wenn Sie nur Wörter suchen oder ersetzen möchten, deren Schreibweise exakt übereinstimmt.
7. 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.
8. Klicken Sie auf **Ersetzen**, um den ausgewählten Treffer zu ersetzen.
9. Klicken Sie auf **Ersetzen**, um alle im Abschnitt gefundenen Übereinstimmungen zu ersetzen. Anschließend wird eine Nachricht mit der Anzahl der durchgeführten Ersetzungen angezeigt.
10. Wenn Sie Ihre Ersetzungen durchgeführt haben, klicken Sie auf **Schließen**. Das Dialogfeld wird geschlossen.

Hinweis: Wenn Sie beim Ersetzen einen Fehler gemacht haben, können Sie die Ersetzung rückgängig machen, indem Sie das Dialogfeld schließen und in den Menüs die Optionsfolge **Bearbeiten > Rückgängig** auswählen. Dies müssen Sie für jede Änderung, die Sie rückgängig machen möchten, wiederholen.

Zielsprache für Ressourcen

Ressourcen werden für eine bestimmte Textsprache erstellt. Die Sprache, für die diese Ressourcen optimiert werden, wird auf der Registerkarte "Erweiterte Ressourcen" definiert. Sie können, falls gewünscht, eine andere Sprache wählen, indem Sie die jeweilige Sprache im Kombinationsfeld **Zielsprache für Ressourcen** auswählen. Außerdem wird die hier aufgelistete Sprache als Sprache für Text Analysis Packages angezeigt, 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 Extraktionssprache übereinstimmen. Auch wenn die Sprache nur selten geändert wird, können Sie dies tun, wenn Sie die Sprachenoption `ALLE` während der Extraktion verwenden wollten, weil Sie Text in mehr als einer Sprache erwarteten. Durch Änderung der Sprache können Sie beispielsweise auf die Sprachverwendungsressourcen für Extraktionsmuster und Abkürzungen zugreifen und Definitionen für die gewünschte Sekundärsprache erzwingen. Denken Sie jedoch daran, vor der Veröffentlichung oder Speicherung der vorgenommenen Ressourcenänderungen oder vor dem Ausführen einer weiteren Extraktion die Sprache wieder auf die Primärsprache umzustellen, die Sie extrahieren möchten.

Fuzzy-Gruppierung

Wenn Sie im Textminingknoten unter "Extraktionseinstellungen" **Rechtschreibung korrigieren für eine minimale Anzahl an Stammzeichen von** auswählen, haben Sie den Algorithmus für Fuzzy-Gruppierung aktiviert.

Mit der Fuzzy-Gruppierung können Sie Wörter einfacher gruppieren, die häufig falsch oder ähnlich geschrieben werden, indem Sie vorübergehend alle Vokale (außer dem ersten) und doppelt/dreifach auftretende Konsonanten aus den extrahierten Wörtern entfernen und anschließend einen Vergleich dieser Wörter durchführen. Die Funktion für die Fuzzy-Gruppierung wird während des Extraktionsprozesses auf die extrahierten Terme angewendet und die Ergebnisse werden dann verglichen, um eventuelle Übereinstimmungen zu ermitteln. Wenn dies der Fall ist, werden die ursprünglichen Terme in der endgültigen Extraktionsliste gruppiert. Die Gruppierung erfolgt unter dem Term, der am häufigsten in den Daten vorkommt.

Hinweis: Wenn die zwei miteinander verglichenen Terme unterschiedlichen Typen zugeordnet werden (ausschließlich des Typs `<Unknown>`), wird das Fuzzy-Gruppierungsverfahren auf dieses Paar nicht angewendet. Die Terme müssen also zu demselben Typ oder zum Typ `<Unknown>` gehören, damit dieses Verfahren 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 Fuzzy-Gruppierung ausschließen. Hierzu geben Sie die fälschlicherweise als übereinstimmend eingestuften Wortpaare auf der Registerkarte "Erweiterte Ressourcen" im Abschnitt "Ausnahmen" ein. Weitere Informationen finden Sie im Thema Kapitel 18, „Informationen zu erweiterten Ressourcen”, auf Seite 215.

Das folgende Beispiel veranschaulicht, wie eine Fuzzy-Gruppierung durchgeführt wird. Wenn die Fuzzy-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 Fuzzy-Gruppierung zwei Wörter zu einem Paar verbunden, 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. Weitere Informationen finden Sie im Thema „Definieren von Synonymen” auf Seite 209.

Formatierungsregeln für Ausnahmen bei der Fuzzy-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 Entitäten

Bei der Arbeit mit bestimmten Datenarten sind Datumsangaben, Sozialversicherungsnummern, Prozentsätze und andere nicht linguistische Entitäten von Interesse. Diese Elemente werden explizit in der Konfigurationsdatei deklariert, in der Sie auch die Elemente aktivieren bzw. inaktivieren können. Weitere Informationen finden Sie im Thema „Konfiguration" auf Seite 222. Um die Ausgabe der Extraktionsengine zu optimieren, wird die Eingabe aus der nicht linguistischen Verarbeitung so normalisiert, dass ähnliche Elemente gemäß den vordefinierten Formaten gruppiert werden. Weitere Informationen finden Sie im Thema „Normalisierung" auf Seite 222.

Hinweis: Sie können die Extraktion nicht linguistischer Elemente in den Extraktionseinstellungen aktivieren und inaktivieren.

Verfügbare nicht linguistische Entitäten

Die in der folgenden Tabelle aufgeführten nicht linguistischen Entitäten können extrahiert werden. Der Typname wird in Klammern angegeben.

Tabelle 40. Nicht linguistische Entitäten, die extrahiert werden können

Adressen	(`<Address>`)
Aminosäuren	(`<Aminoacid>`)
Währungen	(`<Currency>`)
Datumsangaben	(`<Date>`)
Verzögerung	(`<Delay>`)
Ziffern	(`<Digit>`)
E-Mail-Adressen	(`<email>`)
HTTP/URL-Adressen	(`<url>`)
IP-Adresse	(`<IP>`
Organisationen	(`<Organization>`)
Prozentwerte	(`<Percent>`)
Produkte	(`<Product>`)
Proteine	(`<Gene>`)
Telefonnummern	(`<PhoneNumber>`)
Zeitangaben	(`<Time>`)
Sozialversicherungsnummern (USA)	(`<SocialSecurityNumber>`)
Gewichte und Maßangaben	(`<Weights-Measures>`)

Bereinigen von Text für die Verarbeitung

Bevor die Extraktion nicht linguistischer Entitäten beginnt, wird der Eingabetext bereinigt. Bei diesem Schritt werden die folgenden temporären Änderungen durchgeführt, damit nicht linguistische Entitäten identifiziert und als solche extrahiert werden können:

- Jede Folge von mindestens zwei Leerzeichen wird durch ein einzelnes Leerzeichen ersetzt.

- Tabulatorzeichen werden durch Leerzeichen ersetzt.
- Einzelne Zeilenendezeichen oder Sequenzzeichen werden durch ein Leerzeichen ersetzt, während mehrere Zeilenendezeichen als Absatzende markiert werden. Das Zeilenende kann durch Rücklauf (Carriage Return, CR) und/oder Zeilenvorschub (Line Feed, LF) gekennzeichnet werden.
- HTML- und XML-Tags werden temporär unterdrückt und ignoriert.

Definitionen regulärer Ausdrücke

Beim Extrahieren nicht linguistischer Entitäten sollen gegebenenfalls 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". Weitere Informationen finden Sie im Thema Kapitel 18, „Informationen zu erweiterten Ressourcen”, auf Seite 215.

Die Datei ist in mehrere Abschnitte gegliedert. Der erste Abschnitt trägt die Bezeichnung `[macros]`. Neben diesem Abschnitt kann jeweils ein zusätzlicher Abschnitt für jede nicht linguistische Entität vorliegen. Sie können weitere Abschnitte in diese Datei aufnehmen. Die Regeln in den einzelnen Abschnitten sind jeweils nummeriert (*regexp1*, *regexp2* usw.). Diese 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 eine Entität sprachabhängig. Eine Entität gilt dann als sprachabhängig, wenn es für den Sprachparameter in der Konfigurationsdatei einen Wert ungleich 0 besitzt. Weitere Informationen finden Sie im Thema „Konfiguration” auf Seite 222. Ist eine Entität sprachabhängig, muss die Sprache dem Abschnittsnamen vorangestellt werden, z. B. `[english/PhoneNumber]`. Wenn Sie der Entität `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 Extraktionsengine 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.

Sonderzeichen . [] {} () \ * + ? | ^ $

Alle Zeichen entsprechen sich selbst, mit Ausnahme der folgenden Sonderzeichen, die für einen bestimmten Zweck in Ausdrücken verwendet werden: `.[{()\*+?|^$` Um diese Zeichen als solche zu verwenden, muss ihnen in der Definition ein umgekehrter Schrägstrich (\) vorangestellt werden.

Beispiel: Wenn Sie versuchen, Internetadressen 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 -quantoren ? + * {}

Damit flexiblere Definitionen möglich sind, können Sie mehrere Platzhalterzeichen verwenden, die für reguläre Ausdrücke Standard sind. Dabei handelt es sich um `* ? +`

- Ein *Stern* * gibt an, dass *null oder mehr* Instanzen der vorangehenden Zeichenfolge vorhanden sind. Übereinstimmungen für `ab*c` sind beispielsweise "*ac*", "*abc*", "*abbbc*" usw.
- Ein *Pluszeichen* + gibt an, dass *mindestens eine* Instanz der vorangehenden Zeichenfolge vorhanden ist. Übereinstimmungen für `ab+c` sind beispielsweise "*abc*", "*abbc*", "*abbbc*", aber nicht "*ac*".
- Ein *Fragezeichen ?* gibt an, dass *null oder eine* Instanz der vorangehenden Zeichenfolge vorhanden ist. Übereinstimmungen für `modell?ing` sind beispielsweise sowohl *modelling*" als auch "*modeling*".
- *Beschränken von Wiederholungen mit Klammern {}* gibt die Grenzen der Wiederholung an. Beispiel:

`[0-9]{n}` entspricht einer Ziffer, die exakt *n*-mal wiederholt wird. Eine Übereinstimmung für `[0-9]{4}` ist beispielsweise "*1998*", aber weder "*33*" noch "*19983*".

`[0-9]{n}` entspricht einer Ziffer, die *n-mal oder öfter* wiederholt wird. Eine Übereinstimmung für `[0-9]{3,}` ist beispielsweise "*199*" oder "*1998*", aber nicht "*19*".

`[0-9]{n,m}` entspricht einer Ziffer, die zwischen *n- und m-mal (inklusive)* wiederholt wird. Eine Übereinstimmung für `[0-9]{3,5}` ist beispielsweise "*199*", "*1998*" oder "*19983*", aber weder "*19*" noch "*199835*".

Optionale Leerzeichen und Bindestriche

In manchen Fällen ist es erforderlich, ein optionales Leerzeichen in eine Definition einzufügen. Wenn Sie beispielsweise Währungen wie "*uruguayan pesos*", "*uruguayan peso*", "*uruguay pesos*", "*uruguay peso*", "*pesos*" oder "*peso*" extrahieren möchten, 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 keine Übereinstimmung mit "*pesos*" oder "*peso*" geben, 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.). Diese 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 inaktivieren, 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 Entitäten werden die gefundenen Entitäten so normalisiert, dass ähnliche Entitäten 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. Weitere Informationen finden Sie im Thema Kapitel 18, „Informationen zu erweiterten Ressourcen", auf Seite 215. 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.

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 inaktivieren, geben Sie am Anfang der entsprechenden Zeile eine Raute (#) ein. Soll ein Eintrag aktiviert werden, entfernen Sie das Nummernzeichen (#) am Anfang der zugehörigen Zeile.

Englische Daten in Normalisierung

Standardmäßig werden Datumsangaben in einer englischen Vorlage im amerikanischen Datumsformat erkannt, d. h. Monat, Datum, Jahr. Wenn Sie dieses Format in das Format "Tag, Monat, Jahr" ändern müssen, inaktivieren Sie die Zeile "format:US" (indem Sie das Nummernzeichen (#) am Anfang der Zeile hinzufügen) und aktivieren Sie "format:UK" (indem Sie das Nummernzeichen (#) aus dieser Zeile entfernen).

Konfiguration

Sie können die nicht linguistischen Entitätstypen, die Sie extrahieren möchten, in der Konfigurationsdatei der nicht linguistischen Entitäten aktivieren bzw. inaktivieren. Wenn Sie die nicht benötigten Entitäten inaktivieren, wird die Verarbeitung beschleunigt. Dies erfolgt im Abschnitt **Konfiguration** auf der Registerkarte "Erweiterte Ressourcen". Weitere Informationen finden Sie im Thema Kapitel 18, „Informationen zu erweiterten Ressourcen", auf Seite 215. Wenn die Extraktion nicht linguistischer Entitäten aktiviert ist, liest die Extraktionsengine diese Konfigurationsdatei während des Extraktionsprozesses, um zu ermitteln, welche Typen nicht linguistischer Entitäten extrahiert werden sollen.

Für diese Datei gilt die folgende Syntax:

```
#Name<TAB>Sprache<TAB>Code
```

Tabelle 41. Syntax für Konfigurationsdatei.

Spaltenbeschriftung	Beschreibung
`#Name`	Wortlaut, mit dem die anderen beiden Dateien, die für die Extraktion nicht linguistischer Entitäten erforderlich sind, auf diese nicht linguistischen Entitäten 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 verwendet werden kann, z. B. IP/URL/E-Mail-Adressen; `1` = Französisch; `2` = Englisch; `4` = Deutsch; `5` = Spanisch; `6` = Niederländisch; `8` = Portugiesisch; `10` = Italienisch.
`Code`	Code für die Wortart. Die meisten Entitäten 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 Entitäten extrahiert und dann wird die Rolle dieser Entitäten mithilfe der Extraktionsmuster in einem größeren Zusammenhang ermittelt. Prozentsätze erhalten beispielsweise den Wert "a". Angenommen, 30 % wird als nicht linguistische Entität extrahiert. Dieses Element wird als Adjektiv erkannt. Wenn der Text dann die Wortgruppe "30% salary increase" enthält, erfüllt die nicht linguistische Entität "30%" das Wortartmuster "ann" (Adjektiv-Nomen-Nomen).

Reihenfolge bei der Definition von Entitäten

Die Reihenfolge, in der Sie die Entitäten in dieser Datei deklarieren, ist von Bedeutung und wirkt sich auf die Extraktion aus. Die Einträge werden in der angegebenen Reihenfolge angewendet. Wenn Sie die Reihenfolge ändern, ändern sich auch die Ergebnisse. Die spezifischsten nicht linguistischen Entitäten müssen vor den allgemeineren definiert werden.

Beispielsweise wird die nicht linguistische Entität "`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 ist die nicht linguistische Entität "`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 inaktivieren, geben Sie am Anfang der entsprechenden Zeile eine Raute (#) ein. Soll eine Entität 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
- Erzwungene Definitionen
- Abkürzungen

Extraktionsmuster

Beim Extrahieren von Informationen aus den Dokumenten wendet die Extraktionsengine ein Set fest codierter Wortartmuster (POS-Muster, POS - Part of Speech) auf einen "Stapel" von Wörtern im Text an, um so infrage kommende Terme (Wörter und Wortfolgen) für die Extraktion 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 Wortart-Extraktionsmuster. In IBM-Textminingprodukten ist jede Wortart mit einem einzelnen Buchstaben gekennzeichnet, sodass 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 einem Set 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 inaktivieren, 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 Extraktionsengine den Text und erkennt dabei die Wortart (Part of Speech - POS) für jedes gefundene Wort. In einigen Fällen kann ein Wort verschiedene POS-Rollen annehmen, je nach Kontext. Im Bereich **Erzwungene Definitionen** der Registerkarte "Erweiterte Ressourcen" können Sie eine bestimmte Wortartrolle erzwingen oder dieses Wort komplett aus der Verarbeitung ausschließen. Weitere Informationen finden Sie im Thema Kapitel 18, „Informationen zu erweiterten Ressourcen”, auf Seite 215.

Um eine Wortartrolle für ein bestimmtes Wort zu erzwingen, müssen Sie in diesem Abschnitt eine Zeile mit der folgenden Syntax hinzufügen:

```
Term:Code
```

Tabelle 42. Beschreibung der Syntax.

Aufnahme	Beschreibung
`Term`	Der Name eines Terms.
`Code`	Ein einstelliger Code für die Wortartregel. Sie können bis zu sechs verschiedene Wortartcodes pro Uniterm angeben. Mit dem Code `s` (Kleinbuchstabe "s") können Sie außerdem die Extraktion eines bestimmten Worts in zusammengesetzte Wörter bzw. Wortfolgen unterbinden, z. B. `additional:s`.

Formatierungsregeln für erzwungene Definitionen

- Eine Zeile pro Wort.
- Doppelpunkte sind in Termen 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 Wortartcodes pro Zeile ein. Unterstützte Wortartcodes werden im Abschnitt "Extraktionsmuster" angezeigt. Weitere Informationen finden Sie im Thema „Extraktionsmuster” auf Seite 224.
- Der Stern (*) 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 Term noch als Bestandteil eines zusammengesetzten Worts extrahiert. Ist eine Wortübereinstimmung jedoch explizit als Term 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 Extraktionsengine 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 in einer Synonymdefinition angezeigt wird oder in einem Typwörterbuch als Term 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 Sprachenoption **Alle** greift auf Language Identifier zurück, eine Engine für die automatische Spracherkennung. Language Identifier scannt die Dokumente und ermittelt dabei diejenigen, die in einer unterstützten Sprache verfasst sind. Bei der Extraktion werden anschließend automatisch 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. In der folgenden Tabelle werden die Parameter beschrieben, die Sie im Abschnitt **Sprachenkennung - Eigenschaften** auf der Registerkarte "Erweiterte Ressourcen" festlegen können. Weitere Informationen finden Sie im Thema Kapitel 18, „Informationen zu erweiterten Ressourcen”, auf Seite 215.

Tabelle 43. Parameterbeschreibungen

Parameter	Beschreibung
`NUM_CHARS`	Gibt die Anzahl der Zeichen an, die die Extraktionsengine lesen soll, um die Sprache des Texts zu ermitteln. Je niedriger der Wert, desto schneller wird die Sprache ermittelt. Je höher der Wert, mit desto größerer Genauigkeit wird die Sprache ermittelt. Beim Wert 0 wird der gesamte Text im Dokument gelesen.

Tabelle 43. Parameterbeschreibungen (Forts.)

Parameter	Beschreibung
USE_FIRST_SUPPORTED _LANGUAGE	Gibt an, ob die Extraktionsengine 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 **Sprachenkennung - Sprachen** auf der Registerkarte "Erweiterte Ressourcen" bearbeiten.

Je mehr Sprachen vorhanden sind, desto wahrscheinlicher können diese falsch positive Ergebnisse hervorrufen und die Verarbeitungsgeschwindigkeit senken. Entfernen Sie also gegebenenfalls die weniger häufig gebrauchten 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 "Extraktion" aktiviert ist, werden die Textdaten mit diesen Regeln abgeglichen. Wird eine Übereinstimmung gefunden, wird das Textlinkanalyse-Muster extrahiert und präsentiert. Diese Regeln werden auf der Registerkarte "Textlinkregeln" definiert.

Konzepte 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 zu den Zusammenhängen 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-Regeln (TLA-Regeln) enthalten. Wenn Sie Vorlagen auswählen, wird über ein Symbol in der Spalte "TLA" angezeigt, ob die Vorlagen TLA-Regeln besitzen.

Textlinkanalysemuster werden während der Musterabgleichsphase des Extraktionsprozesses 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".

Hinweis: Die Unterstützung von Variablen wurde in Version 13 eingestellt. Verwenden Sie stattdessen Makros. Weitere Informationen finden Sie im Thema „Arbeiten mit Makros” auf Seite 232.

Bearbeiten von Textlinkregeln

Sie können Regeln direkt auf der Registerkarte "Textlinkregeln" in der Ansicht "Vorlageneditor" oder "Ressourceneditor" 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 Extraktion 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. Navigieren Sie im Vorlageneditor oder im Ressourceneditor zur 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, alle Ihre Regeln in einer Bibliothek zu speichern, es sei denn, dies wird aus einem bestimmten Grund nicht gewünscht.

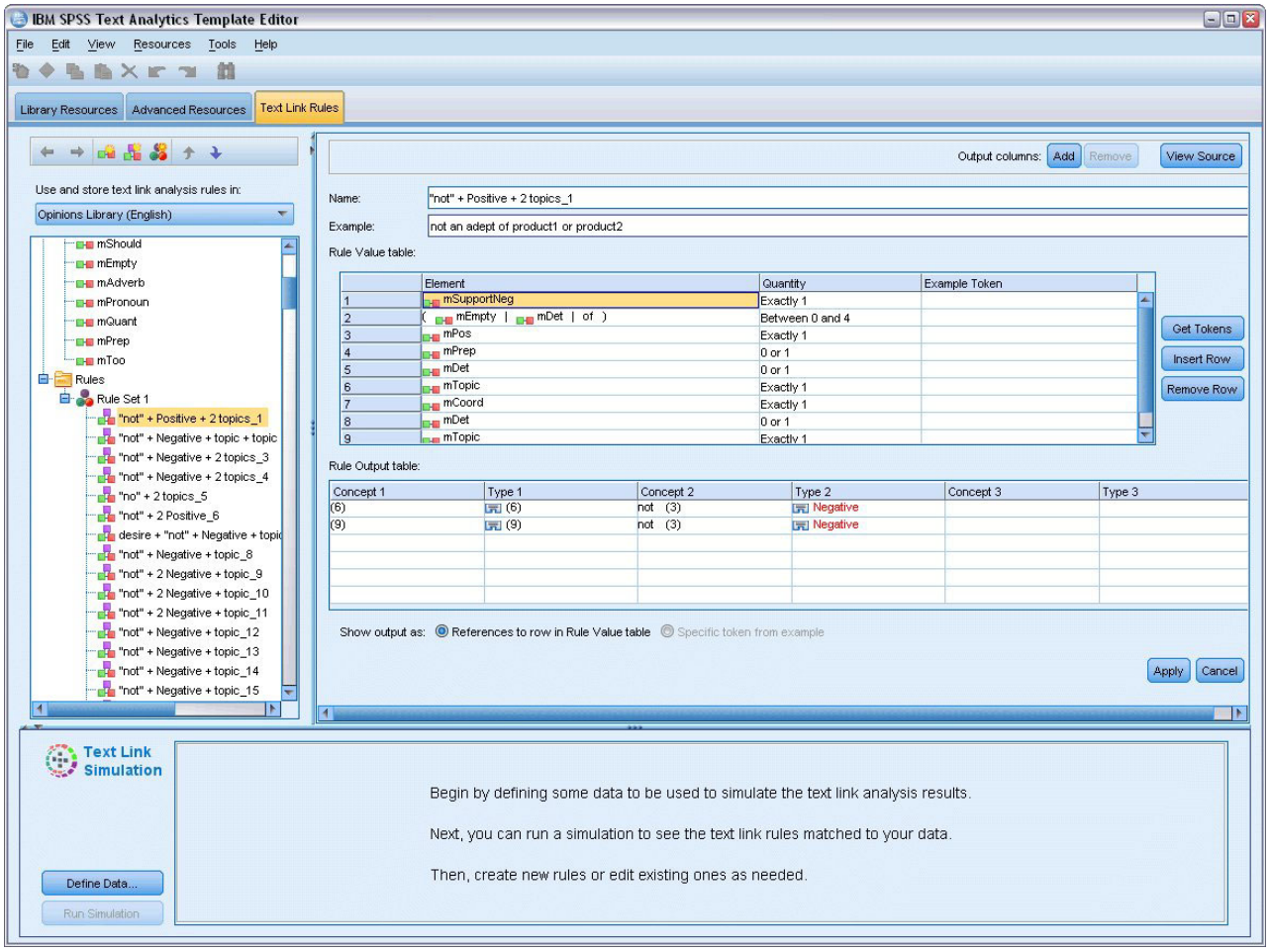


Abbildung 42. Registerkarte "Textlinkregeln"

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 Ergebnisse mit einem Beispieltext und erstellen oder bearbeiten Sie übereinstimmende Regeln basierend auf den Kriterien, wie die aktuellen Regeln Muster aus den Simulationsdaten extrahieren.
- Erstellen Sie eine völlig neue Regel oder bearbeiten Sie eine bestehende Regel.
- Arbeiten Sie direkt in der Quellenansicht.

Wann Regeln erstellt oder bearbeitet werden sollten

Obwohl die in jeder Vorlage enthaltenen Textlinkanalyseregeln oft für die Extraktion 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. Weitere Informationen finden Sie im Thema „Spezielle Makros: mTopic, mNonLingEntities, SEP" auf Seite 235.

- 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 <Pharmaceuticals>, <Car Manufacturing>, <Finance> 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 Text: Max Mustermann ruft Martina Mustermann an. Sie möchten jedoch, dass diese Regel nicht nur Anrufe, sondern auch E-Mail-Kommunikation erfasst. Sie könnten der Regel einen nicht linguistischen Entitätstyp 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 Extraktionseinstellungen eine Extraktion 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 Teilsatz je nach Kontext) werden in einer Simulationsausgabe die gesammelten Tokens und TLA-Regeln angezeigt, die ein Muster in diesem Text aufgedeckt haben. Ein **Token** wird als beliebiges Wort oder beliebige Wortfolge definiert, das/die während des Extraktionsprozesses 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. Navigieren Sie im Vorlageneditor oder im Ressourceneditor zur 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, alle Ihre Regeln in einer Bibliothek zu speichern, es sei denn, dies wird aus einem bestimmten Grund nicht gewünscht.

Wichtig! Es wird dringend empfohlen, bei der Verwendung einer Datendatei sicherzustellen, dass der enthaltene Text kurz ist, um die Verarbeitungszeit zu minimieren. Das Ziel einer Simulation ist es, 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-Extraktion aus, um Ergebnisse für ein vollständigeres Dataset 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

1. Klicken Sie im Simulationsbereich unten auf der Registerkarte **Textlinkregeln** auf **Daten definieren**. Wählen Sie alternativ, sofern zuvor keine Daten definiert wurden, die Optionsfolge **Tools > Simulation ausführen** in den Menüs aus. Der Assistent "Simulationsdaten" wird geöffnet.
2. 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.

- **Dateidatenquelle bestimmen**. 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 die gesamte Arbeitsmappe 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 Textminingknoten verbunden hätten.

Wichtig! Es wird dringend empfohlen, bei der Verwendung einer Datendatei sicherzustellen, dass der enthaltene Text kurz ist, um die Verarbeitungszeit zu minimieren. Das Ziel einer Simulation ist es, 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-Extraktion aus, um Ergebnisse für ein vollständigeres Dataset zu erhalten. Diese Simulation dient nur zu Testzwecken und zum Verfassen von Regeln.

3. Klicken Sie auf **Simulation ausführen**, um den Simulationsprozess zu starten. Es wird ein Fortschrittsdialog angezeigt. Wenn Sie in einer interaktiven Sitzung arbeiten, sind die Extraktionseinstellungen für die Simulation diejenigen, die derzeit in der interaktiven Sitzung ausgewählt sind (siehe **Tools > Extraktionseinstellungen** in der Ansicht "Konzepte und Kategorien"). Wenn Sie im Vorlageneditor arbeiten, werden als Extraktionseinstellungen für die Simulation die Standardextraktionseinstellungen verwendet, die mit denen identisch sind, die auf der Registerkarte "Experten" eines Textlinkanalyseknotens angezeigt werden. Weitere Informationen finden Sie im Thema „Informationen zu den Simulationsergebnissen”.

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 Extraktions- und Simulationsprozesses werden die Ergebnisse der Simulation angezeigt.

Für jeden während der Extraktion 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 nachdem, wie der Extraktor den Text in lesbare Abschnitte unterteilt hat.

Ein **Token** wird als beliebiges Wort oder beliebige Wortfolge definiert, das/die während des Extraktionsprozesses identifiziert wurde. Zum Beispiel können in dem Satz *Mein Onkel lebt in New York* die folgenden Tokens während der Extraktion 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, Literalzeichenfolgen 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.

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 auf dieser Registerkarte), werden die Ergebnisse als ohne Übereinstimmung betrachtet und die Schaltflächen **Nächste Nichtübereinstimmung** und **Vorherige Nichtübereinstimmung** 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 Extraktionseinstellungen geändert haben, möchten Sie vielleicht eine erneute Simulation durchführen. Um eine Simulation erneut auszuführen, klicken Sie im Simulationsbereich auf **Simulation ausführen** 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 Extraktionsprozess aus den Simulationsdaten identifiziert wurde, die Sie im Assistenten definiert haben. Unter einem Satz verstehen wir entweder ein Wort, einen Satz oder einen Teilsatz, je nachdem, wie der Extraktor den Text in lesbare Abschnitte unterteilt hat.

Systemansicht. Eine Sammlung von Tokens, die durch den Extraktionsprozess identifiziert wurden.

- **Eingabetexttoken.** Jedes im Eingabetext gefundene Token. Tokens wurden bereits weiter oben in diesem Thema definiert.
- **Typisiert als.** Wenn ein Token als Konzept identifiziert und einem Typ zugeordnet wurde, wird der zugehörige Typname (z. B. `<Unknown>`, `<Person>`, `<Location>`) in dieser Spalte angezeigt.
- **Übereinstimmendes Makro.** 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 **Regelwerttabelle** 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 "*I like pizza*" während der Simulation dem Typ `<Unknown>` zugeordnet und mit dem Makro `mTopic` abgeglichen worden sein, wenn Sie die Vorlage "Grundlegende Ressourcen (Englisch)" verwendet haben. In diesem Fall wird `mTopic` als das Element in der generierten Regel verwendet. Weitere Informationen finden Sie im Thema „Arbeiten mit Textlinkregeln” auf Seite 236.

Navigation durch Regeln und Makros im Baum

Wenn die Textlinkanalyse während der Extraktion durchgeführt wird, werden die Textlinkregeln verwendet, die in der ausgewählten Bibliothek auf der Registerkarte **Textlinkregeln** 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. Navigieren Sie im Vorlageneditor oder im Ressourceneditor zur Registerkarte **Textlinkregeln**. Auf dieser Registerkarte können Sie die Bibliothek in Ihrer Vorlage angeben, die die TLA-Regeln enthält, die Sie verwenden oder bearbeiten möch-

ten. Aus diesem Grund wird dringend empfohlen, alle Ihre Regeln in einer Bibliothek zu speichern, es sei denn, dies wird aus einem wichtigen oder speziellen Grund nicht gewünscht.

Sie können auf der Registerkarte "Textlinkregeln" festlegen, in welcher Bibliothek Sie arbeiten möchten, indem Sie diese Bibliothek in der Dropdown-Liste **Regeln für Textlinkanalyse verwenden und speichern in:** auf dieser Registerkarte auswählen. Wenn die Textlinkanalyse während der Extraktion durchgeführt wird, werden die Textlinkregeln verwendet, die in der ausgewählten Bibliothek auf der Registerkarte **Textlinkregeln** 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, alle Ihre Regeln in einer Bibliothek zu speichern, es sei denn, dies wird aus einem bestimmten 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, wird ein Kontextmenü geöffnet, über das Sie weitere Aufgaben durchführen können, zum Beispiel:

- Ein neues Makro im Baum erstellen und rechts im Editor öffnen.
- Eine neue Regel im Baum erstellen und 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 inaktivieren, 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 "Tokens 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 Nachricht lautet in etwa: **Warnung: Unbekannter Typ oder unbekanntes Makro**. Damit werden Sie darauf hingewiesen, dass ein Element, das durch `$etwas` in der Quellenansicht definiert würde, beispielsweise `$myType`, kein traditioneller 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 `OR` (`|`) 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 Ih-

ren Textlinkanalyseregeln vornehmen zu müssen. Die meisten mitgelieferten TLA-Regeln enthalten vordefinierte Makros. Makros werden im oberen Bereich des Baums ganz links auf der Registerkarte "Textlinkregeln" angezeigt.

Es werden folgende Felder und Tabellen in den Simulationsergebnissen angezeigt:

Name: Ein eindeutiger Name, der dieses Makro identifiziert. Es wird empfohlen, vor den Makronamen den kleingeschriebenen Buchstaben "m" zu 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 Zeichenpräfix $ verwenden, damit der Extraktionsprozess 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 muss die Groß-/Kleinschreibung beachtet werden.
- Diese Werte können eine Kombination aus Typen, Literalzeichenfolgen, Wortlücken oder Makros oder eines dieser Elemente enthalten. Weitere Informationen finden Sie im Thema „Unterstützte Elemente für Regeln und Makros” auf Seite 242.
- Um in einem Makro einen Wert für ein Element einzugeben, doppelklicken Sie auf die Zeile, in der Sie arbeiten möchten. Es wird ein bearbeitbares Textfeld angezeigt, in das Sie eine Typenreferenz, eine Makroreferenz, eine Literalzeichenfolge oder eine Wortlücke eingeben können. Klicken Sie alternativ mit der rechten Maustaste in die Zelle, um ein Kontextmenü anzuzeigen, in dem übliche Makros, Typnamen und nicht linguistische Typnamen aufgelistet werden. Um auf einen Typ oder ein Makro zu verweisen, müssen Sie dem Makro- oder Typnamen das Zeichen "$" voranstellen, wie beispielsweise in `$mTopic` für das Makro `mTopic`. Um Argumente zu kombinieren, müssen Sie diese mithilfe von runden Klammern `( )` zu einer Gruppe zusammenfassen, und das Zeichen `|` verwenden, um den booleschen Operator `OR` zu kennzeichnen.
- Sie können Zeilen in der Tabelle "Makrowert" mithilfe der Schaltflächen im rechten Bereich hinzufügen oder entfernen.
- Geben Sie jedes Element in einer eigenen Zeile ein. Wenn Sie zum Beispiel ein Makro erstellen möchten, das eine von drei Literalzeichenfolgen wie `bin OR war OR ist` darstellt, würden Sie jede Literalzeichenfolge 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 Anleitungen und Beschreibungen für den Makroeditor. Weitere Informationen finden Sie im Thema „Arbeiten mit Makros” auf Seite 232.

Erstellen neuer Makros

1. Wählen Sie in den Menüs die Optionsfolge **Tools > Neues Makro** aus. Klicken Sie alternativ auf das Symbol "Neues Makro" in der Baumsymbolleiste, um ein neues Makro im Editor zu öffnen.
2. Geben Sie einen eindeutigen Namen ein und definieren Sie die Makrowertelemente.
3. Klicken Sie, wenn Sie fertig sind, auf **Anwenden**, um auf Fehler zu prüfen.

Bearbeiten von Makros

1. Klicken Sie auf den Makronamen im Baum. Das Makro wird rechts im Editorbereich geöffnet.
2. Nehmen Sie die gewünschten Änderungen vor.
3. Klicken Sie, wenn Sie fertig sind, auf **Anwenden**, um auf Fehler zu prüfen.

Inaktivieren und Löschen von Makros

Inaktivieren von Makros

Wenn Sie möchten, dass ein Makro während der Verarbeitung ignoriert wird, können Sie es inaktivieren. Dies kann in Regeln, die noch auf dieses inaktivierte Makro verweisen, zu Warnungen oder Fehlernachrichten führen. Seien Sie beim Löschen und Inaktivieren von Makros vorsichtig.

1. Klicken Sie auf den Makronamen im Baum. Das Makro wird rechts im Editorbereich geöffnet.
2. Klicken Sie mit der rechten Maustaste auf den Namen.
3. Wählen Sie in den Kontextmenüs **Inaktivieren** aus. 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 Fehlernachrichten führen. Seien Sie beim Löschen und Inaktivieren von Makros vorsichtig.

1. Klicken Sie auf den Makronamen im Baum. Das Makro wird rechts im Editorbereich geöffnet.
2. Klicken Sie mit der rechten Maustaste auf den Namen.
3. Wählen Sie in den Kontextmenüs **Löschen** aus. Das Makro wird aus der Liste entfernt.

Fehlersuche, Speichern und Abbrechen

Anwenden von Makroänderungen

Wenn Sie auf eine Stelle außerhalb des Makroeditors oder auf **Anwenden** klicken, wird das Makro automatisch nach Fehlern gescannt. Wird ein Fehler gefunden, müssen Sie diesen beheben, bevor Sie in einen anderen Bereich der Anwendung wechseln.

Wenn jedoch weniger schwerwiegende Fehler gefunden werden, wird nur eine Warnung angezeigt. Wenn Ihr Makro zum Beispiel unvollständige Definitionen oder Definitionen ohne Verweis auf Typen oder andere Makros enthält, wird eine Warnung angezeigt. Wenn Sie auf **Anwenden** klicken, wird bei allen nicht behobenen Warnungen ein Warnsymbol links neben dem Makronamen im Regel- und Makrobaum im linken Bereich angezeigt.

Durch das Anwenden eines Makros wird Ihr Makro nicht permanent gespeichert. Durch das Anwenden wird im Validierungsprozess nach Fehlern und Warnungen gesucht.

Speichern von Ressourcen in einer interaktiven Workbenchsitzung

1. So speichern Sie die Änderungen an Ihren Ressourcen während einer interaktiven Workbenchsitzung, 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 Ressourcen zurückgreifen können. Weitere Informationen finden Sie im Thema „Aktualisieren von Modellierungsknoten und Speichern" auf Seite 90. Speichern Sie anschließend Ihren Stream. Speichern Sie Ihren Stream im Hauptbereich von IBM SPSS Modeler nach der Aktualisierung des Modellierungsknotens.
2. So speichern Sie die Änderungen an Ihren Ressourcen während einer interaktiven Workbenchsitzung, damit Sie sie in anderen Streams verwenden können:

- Aktualisieren Sie die verwendete Vorlage oder erstellen Sie eine neue. Weitere Informationen finden Sie im Thema „Erstellen und Aktualisieren von Vorlagen" auf Seite 177. Dadurch werden die Änderungen für den aktuellen Knoten nicht übernommen (siehe vorheriger Schritt).
- Alternativ können Sie auch das verwendete TAP aktualisieren. Weitere Informationen finden Sie im Thema „Aktualisierung von Text Analysis Packages" auf Seite 150.

Speichern von Ressourcen im Vorlageneditor

1. Veröffentlichen Sie zunächst die Bibliothek. Weitere Informationen finden Sie im Thema „Veröffentlichen von Bibliotheken” auf Seite 197.
2. Speichern Sie anschließend die Vorlage über die Optionsfolge **Datei > Ressourcenvorlage speichern** in den Menüs.

Abbrechen von Makroänderungen

1. Wenn Sie die Änderungen verwerfen möchten, klicken Sie auf **Abbrechen**.

Spezielle Makros: mTopic, mNonLingEntities, SEP

Die Vorlage "Meinungen" (und ähnliche Vorlagen) sowie die Vorlagen "Grundlegende Ressourcen" 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. `<Negative>` oder `<Positive>`) oder ein Typ ist, der in den erweiterten Ressourcen als nicht linguistische Entität 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 nicht linguistische Entitäten auf der Registerkarte "Erweiterte Ressourcen" angegeben.

Angenommen, Sie haben über eine Meinungsvorlage neue Typen in den Ressourcen erstellt: `<Gemüse>` und `<Obst>`. Ohne dass Änderungen nötig sind, werden Ihre neuen Typen als `mTopic`-Typen behandelt, sodass Sie automatisch die positiven, negativen, neutralen und kontextbezogenen Meinungen über Ihre neuen Typen erfassen können. Bei der Extraktion 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 mindestens einer Regel 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 Entitäten in den Abschnitt **Nicht linguistische Entitäten** der Registerkarte "Erweiterte Ressourcen" einfügen, werden sie ebenfalls automatisch als `mNonLingEntities` verarbeitet, wenn nicht anders angegeben. Weitere Informationen finden Sie im Thema „Nicht linguistische Entitäten” auf Seite 219.

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 boolesche Abfrage, die für den Abgleich eines Satzes eingesetzt wird. Textlinkanalyseregeln enthalten mindestens eines der folgenden Argumente: Typen, Makros, Literalzeichenfolgen oder Wortlücken. Sie müssen mindestens eine Textlinkanalyseregel definiert haben, um TLA-Ergebnisse extrahieren zu können.

Die folgenden Bereiche und Felder werden auf der Registerkarte "Textlinkregeln" im Regeleditor angezeigt:

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 generieren, um zu sehen, wie dieser mit der Regel übereinstimmt und wie er ausgegeben wird. Ein **Token** wird als beliebiges Wort oder beliebige Wortfolge definiert, das/die während des Extraktionsprozesses identifiziert wurde. Zum Beispiel können in dem Satz *Mein Onkel lebt in New York* die folgenden Tokens während der Extraktion 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, Literalzeichenfolgen und Wortlücken sein. Nur einem Typ zugeordnete Wörter oder Wortfolgen können Konzepte sein.

Regelwerttabelle. 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 Kombination von Typen, Literalzeichenfolgen, Wortlücken (`<Beliebiges Token>`) oder Makros oder als eines dieser Elemente ein. Weitere Informationen finden Sie im Thema „Unterstützte Elemente für Regeln und Makros” auf Seite 242. Doppelklicken Sie auf die Elementzelle, um die Informationen direkt einzugeben. Klicken Sie alternativ mit der rechten Maustaste in die Zelle, um ein Kontextmenü anzuzeigen, in dem übliche Makros, Typnamen und nicht linguistische Typnamen aufgelistet werden. 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 diese mithilfe von runden Klammern `( )` zu einer Gruppe zusammenfassen, und das Zeichen `|` verwenden, um den booleschen Operator `OR` zu kennzeichnen. Beachten Sie, dass bei den Werten die Groß-/Kleinschreibung beachtet werden muss.
- 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 **Beispieltoken**. Wenn Sie auf **Tokens abrufen** klicken, teilt das Programm den **Beispiel**text 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.

Regelausgabetabelle. Jede Zeile in dieser Tabelle legt fest, wie die TLA-Musterausgabe in den Ergebnissen angezeigt wird. 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 Wortfolgen 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 Wortfolgen 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 auszuwählen, das Sie in der Ausgabe sehen möchten. Alternativ können Sie auch Elemente aus der Regelwerttabelle ziehen und in der Ausgabe ablegen. Wenn Sie beispielsweise eine Regel definiert haben, die eine Referenz zu dem Makro `mTopic` in Zeile 2 der Regelwerttabelle 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 Regelausgabetabelle 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 Regelwerttabelle definiert wurde, dann ziehen Sie diesen Typ aus der Regelwerttabelle 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 die 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 Regelwerttabelle 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 Regelwerttabelle definierte Element zu verweisen.

Weiterhin können Sie sogar Methoden kombinieren. Angenommen, in Ihrer Regelwerttabelle 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 Ausgabespalte 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 Regelausgabetabelle.

Wichtig! Denken Sie daran, dass andere linguistische Operationen bei der Extraktion 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 (Synonyme und andere Gruppierungen).

- **Ausgabe anzeigen als.** Standardmäßig ist die Option **Zeilenreferenzen in Regelwerttabelle** 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 "Beispieltokens" in der Regelwerttabelle vorhanden sind, können Sie die Ausgabe für diese jeweiligen Tokens ansehen, indem Sie die Option auswählen.

Hinweis: 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 derzeit drei Paare angezeigt werden und Sie auf "Hinzufügen" klicken, werden der Tabelle zwei weitere Spalten (Konzept 4 und Typ 4) 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 Extraktion von TLA-Ergebnissen aktiviert:

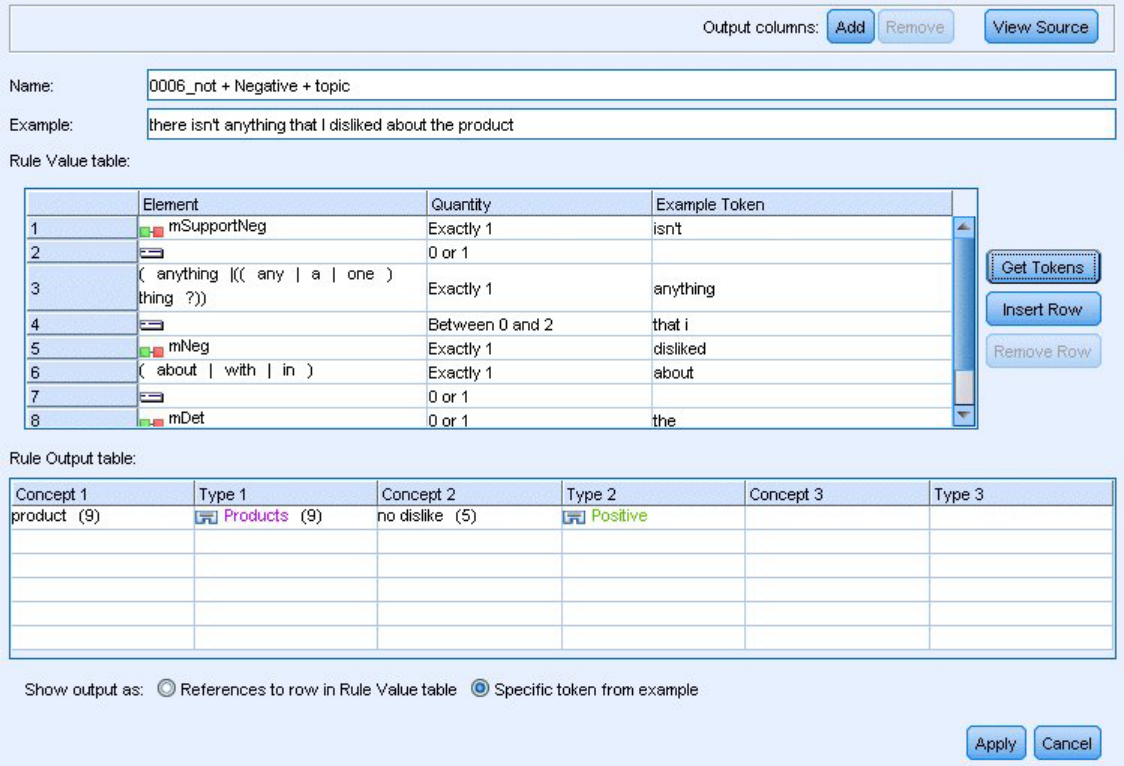


Abbildung 43. Registerkarte "Textlinkregeln": Regeleditor

Bei der Extraktion liest die Extraktionsengine jeden Satz und versucht, die folgende Sequenz abzugleichen:

Tabelle 44. Beispiel für Extraktionssequenz

Element (Zeile)	Beschreibung der Argumente
1	Das Konzept aus einem der durch die Makros `mPos` oder `mNeg` dargestellten Typen oder aus dem Typ `<Uncertain>`.
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 Regelwerttabelle definiert wurde, + ein beliebiges Konzept oder ein beliebiger Typ, der `mPos`, `mNeg` oder `<Uncertain>` entspricht, wie in Zeile 1 der Regelwerttabelle 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. Weitere Informationen finden Sie im Thema „Arbeiten mit Textlinkregeln" auf Seite 236.

Erstellen neuer Regeln

1. Wählen Sie in den Menüs die Optionsfolge **Tools > Neue Regel** aus. Klicken Sie alternativ auf das Symbol "Neue Regel" in der Baumsymbolleiste, um eine neue Regel im Editor zu öffnen.
2. Geben Sie einen eindeutigen Namen ein und definieren Sie die Regelwertelemente.
3. Klicken Sie, wenn Sie fertig sind, auf **Anwenden**, um auf Fehler zu prüfen.

Bearbeiten von Regeln

1. Klicken Sie auf den Regelnamen im Baum. Die Regel wird rechts im Editorbereich geöffnet.
2. Nehmen Sie die gewünschten Änderungen vor.
3. Klicken Sie, wenn Sie fertig sind, auf **Anwenden**, um auf Fehler zu prüfen.

Inaktivieren und Löschen von Regeln

Inaktivieren von Regeln

Wenn Sie möchten, dass eine Regel während der Verarbeitung ignoriert wird, können Sie sie inaktivieren. Lassen Sie beim Löschen und Inaktivieren von Regeln Vorsicht walten.

1. Klicken Sie auf den Regelnamen im Baum. Die Regel wird rechts im Editorbereich geöffnet.
2. Klicken Sie mit der rechten Maustaste auf den Namen.
3. Wählen Sie in den Kontextmenüs **Inaktivieren** aus. 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 Inaktivieren von Regeln Vorsicht walten.

1. Klicken Sie auf den Regelnamen im Baum. Die Regel wird rechts im Editorbereich geöffnet.
2. Klicken Sie mit der rechten Maustaste auf den Namen.
3. Wählen Sie in den Kontextmenüs **Löschen** aus. Die Regel wird aus der Liste entfernt.

Fehlersuche, Speichern und Abbrechen

Anwenden von Regeländerungen

Wenn Sie auf eine Stelle außerhalb des Regeleditors oder auf **Anwenden** 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, wird nur eine Warnung angezeigt. Wenn Ihre Regel zum Beispiel unvollständige Definitionen oder Definitionen ohne Verweis auf Typen oder Makros enthält, wird eine Warnung angezeigt. Wenn Sie auf **Anwenden** klicken, wird bei allen nicht behobenen Warnungen ein Warnsymbol links neben dem Regelnamen links im Baum angezeigt.

Durch das Anwenden einer Regel wird Ihre Regel nicht permanent gespeichert. Durch das Anwenden wird im Validierungsprozess nach Fehlern und Warnungen gesucht.

Speichern von Ressourcen in einer interaktiven Workbenchsitzung

1. So speichern Sie die Änderungen an Ihren Ressourcen während einer interaktiven Workbenchsitzung, 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 Ressourcen zurückgreifen können. Weitere Informationen finden Sie im Thema „Aktualisieren von Modellierungsknoten und Speichern" auf Seite 90. Speichern Sie anschließend Ihren Stream. Speichern Sie Ihren Stream im Hauptbereich von IBM SPSS Modeler nach der Aktualisierung des Modellierungsknotens.
2. So speichern Sie die Änderungen an Ihren Ressourcen während einer interaktiven Workbenchsitzung, damit Sie sie in anderen Streams verwenden können:
   - Aktualisieren Sie die verwendete Vorlage oder erstellen Sie eine neue. Weitere Informationen finden Sie im Thema „Erstellen und Aktualisieren von Vorlagen" auf Seite 177. Dadurch werden die Änderungen für den aktuellen Knoten nicht übernommen (siehe vorheriger Schritt).
   - Alternativ können Sie auch das verwendete TAP aktualisieren. Weitere Informationen finden Sie im Thema „Aktualisierung von Text Analysis Packages" auf Seite 150.

Speichern von Ressourcen im Vorlageneditor

1. Veröffentlichen Sie zunächst die Bibliothek. Weitere Informationen finden Sie im Thema „Veröffentlichen von Bibliotheken" auf Seite 197.
2. Speichern Sie anschließend die Vorlage über die Optionsfolge **Datei > Ressourcenvorlage speichern** in den Menüs.

Abbrechen von Regeländerungen

1. Wenn Sie die Änderungen verwerfen möchten, klicken Sie im Editorbereich auf **Abbrechen**.

Verarbeitungsreihenfolge für Regeln

Wenn während der Extraktion eine Textlinkanalyse ausgeführt wird, wird ein "Satz" (Teilsatz, Wort, Wortfolge) 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, Ihre Regeln absteigend von der genauesten bis zur allgemeinsten Regel zu sortieren. Die genauesten Regeln sollten sich im oberen Bereich des Baums befinden. Um die Reihenfolge einer genauen Regel oder eines Regelsets zu ändern, wählen Sie aus dem Kontextmenü des Regel- und Makrobaums **Nach oben** oder **Nach unten** aus oder verwenden Sie die Schaltflächen mit dem Aufwärtspfeil und dem Abwärtspfeil 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 angezeigt wird, 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 16 können Sie die Verarbeitungsreihenfolge nur bestimmen, indem Sie eine Regel im Baum nach oben oder unten verschieben, oder durch ihre Position in der Quellenansicht.

Nehmen wir beispielsweise an, Ihr Text enthält die folgenden beiden Sätze:

Ich mag Sardellen

Ich mag Sardellen und grüne Paprika

Nehmen wir zudem an, dass zwei Textlinkanalyseregeln mit den folgenden Werten vorhanden sind:

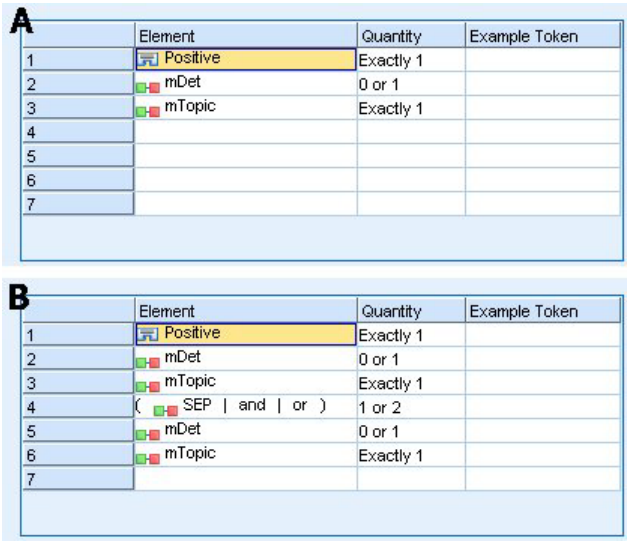


Abbildung 44. 2 Beispielregeln

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 Durchläufe)

Ein Regelset ist eine hilfreiche Möglichkeit, zusammengehörige Regeln im Regel- und Makrobaum zusammenzufassen und mehrere Durchläufe 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 Verknüpfungen 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, ein besonderes Set 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 aus. Sie können anschließend direkt unter einem Regelsetknoten im Baum neue Regeln erstellen oder vorhandene Regeln in ein Regelset verschieben.

Wenn Sie eine Extraktion mithilfe von Ressourcen durchführen, in denen die Regeln in Regelsets zusammengefasst sind, ist die Extraktionsengine gezwungen, mehrere Durchläufe durch den Text durchzuführen, um in den einzelnen Durchläufen 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.

Hinweis: Sie können bis zu 512 Regeln in ein Regelset aufnehmen.

Erstellen neuer Regelsets

1. Wählen Sie in den Menüs die Optionsfolge **Tools > Neues Regelset** aus. Klicken Sie alternativ auf das Symbol "Neues Regelset" in der Symbolleiste des Baums. Es wird ein Regelset im Regelbaum angezeigt.
2. Fügen Sie diesem Regelset neue Regeln hinzu oder verschieben Sie bestehende Regeln in das Set.

Inaktivieren von Regelsets

1. Klicken Sie mit der rechten Maustaste auf den Regelsetnamen im Baum.
2. Wählen Sie in den Kontextmenüs **Inaktivieren** aus. Das Regelsetsymbol wird grau und alle Regeln in diesem Regelset werden ebenfalls inaktiviert und während der Verarbeitung ignoriert.

Löschen von Regelsets

1. Klicken Sie mit der rechten Maustaste auf den Regelsetnamen im Baum.
2. Wählen Sie in den Kontextmenüs **Löschen** aus. 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 anderen 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 Dollarzeichen ($) voranstellen, z. B. `$mTopic`. Bei dem Makronamen muss die Groß-/Kleinschreibung beachtet werden. Wenn Sie Makros über die Kontextmenüs auswählen, können Sie aus allen Makros auswählen, die auf der aktuellen Registerkarte "Textlinkregeln" definiert wurden.

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 Dollarzeichen ($) voranstellen, z. B. `$Person`. Bei dem Typnamen muss die Groß-/Kleinschreibung beachtet werden. Wenn Sie die Kontextmenüs verwenden, können Sie aus allen Typen aus dem aktuellen Set von Ressourcen auswählen.

Wenn Sie auf einen nicht erkannten Typ verweisen, erhalten Sie eine Warnung und neben der Regel steht ein Warnsymbol im Regel- und Makrobaum, bis Sie den Fehler beheben.

Literalzeichenfolgen

Um Informationen einzubeziehen, die nie extrahiert wurden, können Sie eine Literalzeichenfolge definieren, nach der die Extraktionsengine suchen soll. Allen extrahierten Wörtern oder Wortfolgen wurde ein Typ zugewiesen. Daher können sie nicht in Literalzeichenfolgen verwendet werden. Wenn Sie ein Wort verwenden, das extrahiert wurde, wird es ignoriert, selbst wenn der Typ `<Unknown>` ist.

Eine Literalzeichenfolge kann aus einem Wort oder mehreren Wörtern bestehen. Folgende Regeln müssen bei der Definition einer Liste von Literalzeichenfolgen eingehalten werden:

- Schließen Sie die Liste der Zeichenfolgen in Klammern ein, z. B. (sein). Falls eine Auswahl von Literalzeichenfolgen vorhanden ist, muss jede Zeichenfolge 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 booleschen OR.
- 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 Literalzeichenfolgen wiederzuverwenden, definieren Sie sie als Makro und setzen Sie dieses Makro dann in anderen 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 Buchstaben z. B. als Literalzeichenfolge 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 Regelwerttabelle oder der Makrowerttabelle) oder in der Quellenansicht hinzufügen. Wenn Sie Ihrer Textlinkanalyseregel beispielsweise ($mTopic @{0,2} !($Positive) $Budget) hinzufügen, suchen Sie nach einem Text, der (1) einen Term enthält, der einem der Typen im Makro mTopic zugewiesen ist, (2) eine Wortlücke enthält, die von null bis zu zwei Wörtern lang ist, (3) keine Instanzen eines Terms enthält, der dem Typ <Positive> zugewiesen ist, und (4) einen Term enthält, der dem Typ<Budget> zugewiesen 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 die Elementzelle eingeben, indem Sie auf die Zelle doppelklicken.

Wortlücken (<Beliebiges Token>)

Eine Wortlücke, auch als <Beliebiges Token> bezeichnet, definiert einen numerischen Bereich von Tokens, die zwischen zwei Elementen vorliegen können. Wortlücken sind sehr hilfreich, wenn sehr ähnliche Wortfolgen abgeglichen werden, die sich aufgrund zusätzlicher Determinatoren, Präpositionalphrasen, Adjektive oder ähnlicher Wörter nur wenig voneinander unterscheiden.

Tabelle 45. Beispiel für die Elemente in einer Regelwerttabelle ohne Wortlücke

#	Element
1	 Unknown
2	 mBeHave
3	 Positive

Hinweis: In der Quellenansicht wird dieser Wert folgendermaßen definiert: $Unknown $mBeHave $Positive

Dieser Wert hat Sätze wie "*Das Hotelpersonal war nett*" als Übereinstimmung, wobei *Hotelpersonal* zum Typ <Unknown> gehört, *war* sich unter dem Makro mBeHave befindet und *nett* zum Typ <Positive> gehört. Der Satz "*Das Hotelpersonal war sehr nett*" hingegen stellt keine Übereinstimmung dar.

Tabelle 46. Beispiel für die Elemente in einer Regelwerttabelle mit einer Wortlücke (<Beliebiges Token>)

#	Element
1	 `Unknown`
2	 `mBeHave`
3	
4	 `Positive`

Hinweis: In der Quellenansicht wird dieser Wert folgendermaßen definiert: `$Unknown $mBeHave @{0,1} $Positive`

Wenn Sie Ihrem Regelwert eine Wortlücke hinzufügen, hat diese sowohl "*Das Hotelpersonal war nett*" als auch "*Das Hotelpersonal war sehr nett*" als Übereinstimmung.

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.

Anzeigen und Arbeiten im Quellenmodus

Der TLA-Editor generiert für jede Regel und jedes Makro den zugrunde liegenden Quellcode, den der Extraktor zum Abgleichen und Erzeugen einer TLA-Ausgabe verwendet. Falls Sie lieber mit dem Code arbeiten, können Sie diesen Quellcode anzeigen und ihn direkt bearbeiten, indem Sie auf die Schaltfläche "Quelle anzeigen" im oberen Bereich des Editors klicken. Die Quellenansicht springt zur/zum derzeit ausgewählten Regel oder Makro. Es wird jedoch empfohlen, die Editorbereiche zu verwenden, um die Gefahr von Fehlern zu minimieren.

Klicken Sie auf **Quelle verlassen**, wenn Sie mit dem Anzeigen 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, wird dringend empfohlen, immer nur eine Regel oder ein Makro gleichzeitig zu bearbeiten. Validieren Sie die Ergebnisse nach der Bearbeitung eines Makros durch Extraktion. Wenn Sie mit dem Ergebnis zufrieden sind, wird empfohlen, 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 = Makroname
value = ([Typname|Makroname|Literalzeichenfolge|Wortlücke])
```

Tabelle 47. Makroeinträge.

`[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 mindestens einem Typ, einer Literalzeichenfolge, einer Wortlücke oder einem Makro. Weitere Informationen finden Sie im Thema „Unterstützte Elemente für Regeln und Makros" auf Seite 242. Um Argumente zu kombinieren, müssen Sie Klammern ( ) verwenden, um die Argumente zu einer Gruppe zusammenzufassen, und das Zeichen | verwenden, um den booleschen Operator OR zu kennzeichnen.

Zusätzlich zu den Richtlinien und der Syntax, die im Abschnitt über Makros behandelt wurden, gibt es in der Quellenansicht einige Aspekte 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 inaktiviert 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 Terms, der mit *einem* der folgenden Typen übereinstimmt: `<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 = Musternamе
value = [$Typname|Makroname|Wortlücken|Literalzeichenfolgen]
output = $Zahl[\t]#Zahl[\t]$Zahl[\t]#Zahl[\t]$Zahl[\t]#Zahl[\t]
```

Tabelle 48. Regeleinträge.

`[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. Weitere Informationen finden Sie im Thema „Unterstützte Elemente für Regeln und Makros" auf Seite 242.
`output`	Das Ausgabeformat für die resultierenden, übereinstimmenden Muster, die im Text festgestellt wurden. Die Ausgabe entspricht nicht immer der exakten, ursprünglichen Position der Elemente im Quellentext. 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 Tabulatorcode `\t` (z. B. `$1\t#1\t$3\t#3`). • `$` und eine Zahl ruft den Term auf, der mit dem Argument übereinstimmt, das im Wertparameter an dieser Position definiert ist. $1 bezieht sich also auf den Term, der mit dem ersten für den Wert definierten Argument übereinstimmt. • `#` und eine Zahl ruft den Typnamen des Elements an dieser Position auf. Wenn ein Objekt eine Liste von Literalzeichenfolgen ist, wird der Typ `<Unknown>` zugewiesen. • Der Wert `Null\tNull` erzeugt keine Ausgabe.

Zusätzlich zu den Richtlinien und der Syntax, die im Abschnitt über Regeln behandelt wurden, gibt es in der Quellenansicht einige Aspekte 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 = $mTopic?` oder `value = @{0,1}` als erstes Element wäre beispielsweise nicht zulässig.
- Es ist möglich, einem Token eine Menge (oder Anzahl an Instanzen) zuzuordnen. Dies ist insbesondere dann nützlich, wenn Sie eine Regel für alle Fälle schreiben, anstatt je eine Regel für die einzelnen Fälle zu notieren. Mit der Zeichenfolge `($SEP|and)` können Sie beispielsweise eine Übereinstimmung mit `,` (Komma) oder mit dem Wort `and` ermitteln. Wenn Sie diese Regel mit einer Menge ergänzen (die Literalzeichenfolge wird zu `($SEP|and){1,2}`), wird in allen nachstehenden Fällen eine Übereinstimmung erkannt: `","` `"and"` `", and"`.
- Leerzeichen werden zwischen dem Makronamen und dem Dollar- bzw. Fragezeichen ($ und ?) in der Textlinkanalyseregel `value` nicht unterstützt.
- Leerzeichen werden in der Textlinkanalyseregel `output` nicht unterstützt.
- Soll ein Element inaktiviert werden, geben Sie eine Raute (#) am Anfang der entsprechenden Zeilen ein.

Beispiel. Angenommen, Ihre Ressourcen enthalten die folgende TLA-Textlinkanalyseregel und Sie haben die Extraktion 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 Extraktion liest die Extraktionsengine jeden Satz und versucht, die folgende Sequenz abzugleichen:

Tabelle 49. Beispiel für Extraktionssequenz.

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 (`$Function`)
5	Eine der folgenden Zeichenfolgen: `"of"`, `"with"`, `"for"`, `"in"`, `"to"` oder `"at"`,
6	0 oder 1 Wort (`@{0,1}`)
7	Der Name einer Organisation (`$Organization`)
8	0, 1 oder 2 Wörter (`@{0,2}`)
9	Der Name eines Orts (`$Location`)

Dieses Beispiel einer Textlinkanalyseregel würde folgende Sätze oder Wortfolgen abgleichen:

Jean Doe, the HR director of IBM in France

Jean Doe was the former HR director of IBM in France

IBM appointed Jean Doe as the HR director of IBM in France

Dieses Beispiel einer Textlinkanalyseregel würde folgende Ausgabe erzeugen:

```
jean doe <Person> hr director <Function> ibm <Organization> france <Location>
```

Erläuterung:

- `jean doe` ist der Term, der `$1` (dem ersten Element in der Textlinkanalyseregel) entspricht, und `<Person>` ist der Typ für `jean doe` (`#1`),
- `hr director` ist der Term, der `$4` (dem vierten Element in der Textlinkanalyseregel) entspricht, und `<Function>` ist der Typ für `hr director` (`#4`),
- `ibm` ist der Term, der `$7` (dem siebten Element in der Textlinkanalyseregel) entspricht, und `<Organization>` ist der Typ für `ibm` (`#7`),
- `france` ist der Term, der `$9` (dem neunten Element in der Textlinkanalyseregel) entspricht, und `<Location>` ist der Typ für `france` (`#9`)

Regelsets in der Quellenansicht

```
[set(<ID>)]
```

Dabei gibt `[set (<ID>)]` 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 zu 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.
```

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 Extraktion von TLA-Mustern stattfinden. In diesem Fall wird `merger` während der Synonymgruppierungsphase des Extraktionsprozesses unter `merges with` gruppiert. Und da `merges with` zum Typ `<ActiveVerb>` gehört, wird dieser Typname in der endgültigen TLA-Musterausgabe angezeigt. 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 (Synonyme und andere Gruppierungen).

Anstatt komplexe Regeln wie die vorherige zu verfassen, kann es einfacher sein, mit zwei Regeln zu arbeiten. Die 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>`

Die 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>`

Bemerkungen

Diese Informationen wurden für weltweit angebotene Produkte und Dienstleistungen erarbeitet.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Alle in diesem Dokument enthaltenen Leistungsdaten stammen aus einer kontrollierten Umgebung. Die Ergebnisse, die in anderen Betriebsumgebungen erzielt werden, können daher erheblich von den hier erzielten Ergebnissen abweichen. Einige Daten stammen möglicherweise von Systemen, deren Entwicklung noch nicht abgeschlossen ist. Eine Gewährleistung, dass diese Daten auch in allgemein verfügbaren Systemen erzielt werden, kann nicht gegeben werden. Darüber hinaus wurden einige Daten unter Umständen durch Extrapolation berechnet. Die tatsächlichen Ergebnisse können davon abweichen. Benutzer dieses Dokuments sollten die entsprechenden Daten in ihrer spezifischen Umgebung prüfen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corp in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite „Copyright and trademark information" unter www.ibm.com/legal/copytrade.shtml.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Markten oder eingetragene Marken der Intel Corporation oder der Tochtergesellschaften des Unternehmens in den USA und anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA, anderen Ländern oder beidem.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein.

Index

Sonderzeichen

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z