



SAS® Text Miner

*Texte haben viel zu bieten:
Entdecken Sie den Wert der darin enthaltenen Informationen!*

Der Großteil der Daten in der realen Welt liegt in unstrukturierter Form vor. Um den größtmöglichen Nutzen aus unstrukturierten Textdaten zu ziehen, muss ein Unternehmen in der Lage sein, diese Daten automatisch zu analysieren – ebenso wie es bei strukturierten Daten der Fall ist. Denken Sie nur an die riesigen Mengen an Textdokumenten, mit denen Ihr Unternehmen Tag für Tag überschwemmt wird.

Die meisten Unternehmen verfügen im Tagesgeschäft weder über die Zeit noch die Ressourcen, um den Wert der enormen Mengen an Textdaten voll nutzen zu können. Natürlichsprachlicher Text eignet sich nicht zur analytischen Weiterverarbeitung. Aufgrund ihrer Ambiguität und der unzähligen verschiedenen Möglichkeiten, ähnliche Konzepte darzustellen, sind die in Textdaten implizit enthaltenen Informationen nicht leicht zu erkennen, zu quantifizieren oder auszuwerten.

Und selbst wenn es einem Unternehmen gelingen würde, diese Informationsquelle anzuzapfen – was dann? Wie könnte man diese Informationen mit traditionellen strukturierten Daten wie Alters-, Berufs- und Einkommensangaben zusammenführen, um sich ein Bild von der Gesamtsituation zu machen?

Nutzen

- Analyse und Klassifizierung von Texten
- Umwandlung von unstrukturierten Texten in strukturierte Daten
- Bewältigung einer großen Anzahl von Text-Dokumenten
- Integration in den ETL- und Data Mining-Prozess

Integration von Text Mining in den Data Mining-Prozess

Wann immer es darum geht, große Textmengen durchzusehen, um Informationen, Ideen und Trends zu extrahieren – mit dem SAS Text Miner können diese riesigen und weitestgehend ungenutzten Datenspeicher in aussagekräftiges und wertvolles Wissen verwandelt werden. Beispiele hierfür sind:

- Fachartikel, Prospekte, Quartalsberichte,
- Wettbewerbsinformationen, Handelsregistereintragen
- Ausschreibungen, Schadenmeldungen, Garantieforderungen
- Kundenfeedback, Kundengesprächsnotizen, Memos
- Umfragen, Forschungsberichte, Patentinformationen
- Bewerbungen, Lebensläufe
- Websites, E-Mails





Abb. 2: Der SAS Text Miner läuft innerhalb der intuitiven Point-and-Click-Prozessflussumgebung des Enterprise Miner. So können Sie Textdaten nahtlos in Mining-Prozesse einbinden.

SAS Text Miner bietet eine breite Palette an Textverarbeitungs- und Analysetools, mit deren Hilfe den Texten zugrunde liegende Themen oder Konzepte auch in umfangreichen Dokumentensammlungen offen gelegt werden können. Textdokumente können automatisch in Gruppen zusammengefasst, in vordefinierte Kategorien klassifiziert und in Verbindung mit strukturierten Daten zur Erstellung von Prognosemodellen verwendet werden.

- Customer Relationship Management

Beispielsweise können Unternehmen, die mit Programmen für analytisches CRM (Customer Relationship Management) arbeiten, mit Hilfe des SAS Text Miner große Mengen eintreffender Kunden-E-Mails verwalten und kategorisieren und auf diesem Weg den Weiterleitungs- und Bearbeitungsvorgang deutlich effizienter gestalten.

- Call Center – Produktinformationen

Freie Textnotizen, die von Mitarbeitern des Call Centers gesammelt werden, können sinnvoll in Gruppen zusammengefasst werden. So sind fundierte Entscheidungen darüber möglich, welche Produkte für den jeweiligen Kunden am besten geeignet sind.

- Umfragen – Freitexte auswerten

Der SAS Text Miner kann dazu genutzt werden, Umfrageergebnisse in Freiform mit anderen Antworten bzw. Daten zu

kombinieren, um Trends zu identifizieren und Ergebnisse als Grundlage für Handlungsentscheidungen zu erhalten. Es können außerdem Umfrageergebnisse in Freiform mit anderen Antworten kombiniert werden, um Trends zu identifizieren und Ergebnisse zu erhalten, auf deren Grundlage Handlungsentscheidungen getroffen werden können.

- Pharma – klinische Forschung

Durch die Fähigkeit, textbasierte Notizen auszuwerten und zu analysieren, können Pharmaunternehmen ihre klinischen Testprozesse optimieren. Suchanfragen für große Dokumentendatenbanken können so eingeeengt werden, dass nur noch die Artikel gefunden werden, die tatsächlich von Interesse sind.

Leistungsmerkmale von SAS Text Miner

- Universeller Datenzugriff

Die Anwender haben Zugriff auf zahlreiche Formen von Textdaten, z. Bsp. Adobe Portable Document Format (PDF), erweiterter ASCII-Text, HTML und Microsoft Word, und können so Textdaten zum Zwecke des Text Mining extrahieren, umwandeln und in eine SAS Datei laden.

- Unterstützung mehrerer Sprachen

Erweitertes Text-Parsing ist mittels automatischer Spracherkennung für Deutsch oder eine Kombination aus Deutsch und weiteren Sprachen (Englisch, Französisch etc.) möglich.

Elementare Parsing-Funktionen stehen außerdem für viele andere Sprachen zur Verfügung, sofern die einzelnen Wörter durch Leerzeichen bzw. Interpunktion voneinander getrennt sind.

- Methoden zur Textvorverarbeitung

Sobald die Textdaten in eine SAS Datei eingelesen worden sind, stehen die umfassenden Textvorverarbeitungsfunktionen von Text Miner zur Verfügung. So können die wichtigsten Informationen in der Dokumentensammlung erfasst und herausgefiltert werden. Text Miner bietet u. a. folgende Funktionen:

- Voreingestellte oder anwenderdefinierte Sperrlisten für jede Sprache zur Entfernung von Begriffen mit wenig oder keinem Informationswert
- Wortarterkennung, basierend auf dem Satzkontext. Beispielsweise wird erkannt, dass der Begriff „Leben“ in „Das Leben wird immer teurer“ als Substantiv und in „Leben lässt es sich hier gut“ als Verb verwendet wird
- Herauslösen von Nominalkonstruktionen zur Identifikation von Konzepten auf Phrasenebene
- Vom Anwender zu definierende Einheiten, bestehend aus mehreren Wörtern, wie z.B. Data Warehouse oder Point and Click
- Benutzerdefinierte und voreingestellte Synonymlisten

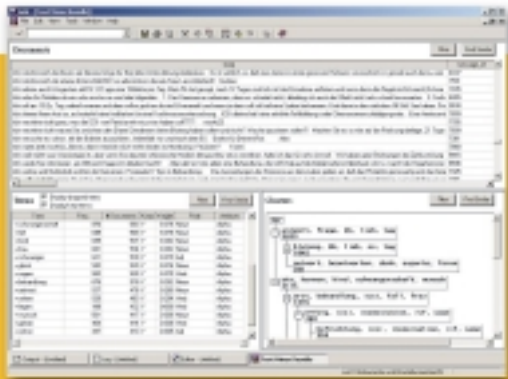


Abb. 1: Der interaktive Results Viewer liefert eine kurze Zusammenfassung der Text Mining-Ergebnisse.

- Aufspaltung von Komposita in die einzelnen Teilbegriffe. Dies ist insbesondere in Sprachen wie dem Deutschen wichtig, wo Komposita durch die direkte Aneinanderreihung mehrerer Einzelbegriffe gebildet werden.

- Komplexe Funktionen zur Merkmalsextraktion

Umfangreiche anwenderspezifische Datenlexika können einzelne Informationen extrahieren, wie z.B. die Namen von Personen, Produkten, Unternehmen oder auch URLs und Adressen. Die extrahierten Einheiten werden anschließend normiert und in eine Matrixdarstellung aufgenommen.

- Techniken zur Dimensionsreduktion

Sobald die vorverarbeiteten Textdaten in eine Matrix überführt worden sind, können leistungsstarke Techniken zur Dimensionsreduktion angewandt werden.

- Rollup-Begriffe stellen eine Standardmethode zur Reduktion dar, bei der die am höchsten gewichteten Begriffe für die Darstellung eines Dokuments ausgewählt werden
- Durch die sog. Singular Value Decomposition (SVD) wird jedes Dokument in einen n-dimensionalen Unterraum projiziert, der der jeweiligen Dokumentensammlung am besten entspricht. In diesem

reduzierten Raum werden ähnliche Dokumente in der Regel nebeneinander abgelegt

- Eindeutige Clustering-Algorithmen

Nach der Anwendung einer Technik zur Dimensionsreduktion stellt der Text Mining-Knoten zwei Clustering-Techniken zur Verfügung, um die Dokumente ihrem Inhalt entsprechend zu gruppieren.

- Beim Erwartungsmaximierungs-Clustering werden Dokumente so gruppiert, dass sie jedem Cluster mit einer bestimmten Wahrscheinlichkeit angehören
- Hierarchisches Clustering erleichtert die Gruppierung der Dokumente

Beide Techniken liefern dem Anwender bei der Erstellung von Cluster-Profilen eine Liste der aussagekräftigsten Begriffe für jeden Cluster. Der Text Miner wird vollständig in den SAS® Enterprise Miner™ eingebunden. Deshalb können die Knoten für Clustering und Self-Organizing Maps im Enterprise Miner zur anschließenden Gruppierung von Dokumenten im Prozessflussdiagramm benutzt werden. Mit Hilfe zusätzlicher strukturierter Daten (wie z. B. Alter, Kaufneigung), die eventuell mit den ursprünglichen Dokumenten erhoben wurden, können außerdem Profile für diese Cluster erstellt werden.

Benefits

- Intelligente, automatische Methoden zur Verarbeitung textueller Information zur Erschließung von Wissen
- Schnelle und ressourcenschonende Nutzung von unstrukturierten Texten als Datenquelle
- Strukturierte Informationserfassung ohne Verlust des Kontextes der Dokumente
- analytische Weiterverarbeitung von Texten in Kombination mit Data Mining-Verfahren
- Regelwerke für Deutsch und weitere Sprachen auch in Kombination nutzbar
- Erschließung des Wissens in nachgelagerten Prozessen

- Kategorisierung von Dokumenten

Sobald der Text vorverarbeitet und in eine numerische Darstellung der Dokumente umgewandelt worden ist, kann mit Hilfe von Enterprise Miner-Tools wie neuronalen Netzen, Memory-based Reasoning, Regression und Entscheidungsbäumen eine Zuordnung der Dokumente zu vordefinierten Kategorien erfolgen.

Im Gegensatz zu herkömmlichen Tools für die Kategorisierung von Dokumenten kann der SAS Text Miner nahtlos zusätzliche quantitative und qualitative Daten mit den Textanalysedaten kombinieren und so die Vorhersagegenauigkeit erhöhen. Schließlich können die Anwender auch die Leistung verschiedener Modelle im Assessment-Knoten miteinander vergleichen und einen Rangschlüssel für die Kategorisierung neuer Dokumente verwenden.

- Interaktiver Results Viewer

Der Results Viewer des Text Miner bietet eine knappe Zusammenfassung der Ergebnisse, u. a. mit Tabellen der Dokumente, Begriffe und Cluster (Abb. 1). Mit Hilfe der interaktiven Funktionen kann der Anwender

- die Begriffstabelle nach Begriffen, Begriffshäufigkeit, Anzahl der Dokumente, Gewichtung und Rolle des Begriffs sortieren;
- zwischen Voll- und Teilanzeige der Texte in den Dokumenten hin- und herschalten;

- nach den n-ähnlichsten Größen für das ausgewählte Dokument bzw. den ausgewählten Begriff oder Cluster suchen lassen;
- Begriffe filtern, um die Dokumente in denen sie enthalten sind, sowie die Cluster, die ihrerseits diese Dokumente enthalten anzuzeigen;
- Dokumente filtern, um alle Begriffe in den Dokumenten sowie die revidierte Clusteranzahl anzuzeigen;
- Cluster filtern, um alle Dokumente in den gefilterten Clustern sowie die Begriffe in diesen Dokumenten anzuzeigen;
- die Liste der beizubehaltenden bzw. zu vernachlässigenden Begriffe ändern;
- ausgewählte Begriffe als äquivalent behandeln;
- Begriffe mit Hilfe eines anderen Algorithmus neu gewichten;
- die Anzahl der SVD-Dimensionen auswählen;
- die n-repräsentativsten Begriffe für jeden Cluster anzeigen lassen sowie;
- zu jedem beliebigen Zeitpunkt mit einer Teilmenge von Dokumenten oder Begriffen eine Umgruppierung durchführen. Oft wird die ursprüngliche Gruppierung im Results Viewer und nicht während der Knotenlaufzeit durchgeführt.

- Anwenderfreundliche selbstdokumentierende Schnittstelle

Die grafische Benutzeroberfläche, die um das ausgefeilte Prozessflussdiagramm des Enterprise Miner herum

geschaffen wurde, macht die manuelle Kodierung überflüssig und reduziert den zeitlichen Text Mining-Aufwand sowohl für Geschäftsprozessanalytiker als auch für Statistiker erheblich. Der Prozessfluss kann modifiziert, abgespeichert und anderen Analytikern zur Verfügung gestellt werden.

- Flexible Berichtsfunktionen

Die Ergebnisse eines Prozessflussdiagramms im Bereich Text Mining können in einem kurzen HTML-Bericht veröffentlicht werden.

Vorteile durch SAS® Intelligence

Die Integration von Text Mining in die bewährte Data Mining-Lösung von SAS, den Enterprise Miner, macht SAS zum ersten Softwarehersteller, der eine umfassende Data Mining-Lösung zur Analyse sowohl unstrukturierter als auch strukturierter Daten anbietet (Abb. 2).



SAS Institute GmbH
In der Neckarhelle 162
D-69118 Heidelberg
Tel: 06221/415-123
Fax: 06221/415-145

www.sas.de

SAS World Headquarters
SAS Campus Drive
Cary, NC 27513 USA
Tel: (919) 677 8000
Fax: (919) 677 4444
Web: www.sas.com