



SAS® Text Miner 3.1

Capitalize on the value hidden in textual information

What does SAS Text Miner do?

SAS Text Miner provides a rich suite of tools for discovering and extracting intelligence from large document collections. It helps identify trends and business opportunities and generates meaningful insights to key business issues more efficiently and with less risk.

Why is SAS Text Miner important?

SAS Text Miner offers a fully integrated set of text preprocessing tools within a core data mining solution. Both structured (quantitative) data sources as well as unstructured textual information can be consolidated to deliver complete views for improved analyses and decision making.

For whom is SAS Text Miner designed?

SAS Text Miner is designed for anyone who must look through large volumes of text to extract information, ideas and trends. SAS Text Miner uncovers patterns across entire document collections.

Large volumes of text-based information are collected throughout organizations each day. Customer feedback, e-mail, Web documents, blogs, memos, warranty claims, surveys, journal articles, research studies, resumes, client notes, competitive intelligence...the list goes on. No one has time to read all the documents, much less sort and classify bits of essential information from each.

To get the most value from collected data, you must be able to analyze it. But because of the ambiguity and numerous ways to represent similar concepts, information implicit in text-based data is not easy to discern, quantify or analyze. Additionally, most organizations lack the ability to combine text-based information with their structured data.

If textual information cannot be integrated with data held in organizational databases, it is impossible to get a full and accurate view of the enterprise or situation. As a result, important decisions are often made without complete information.

Many organizations today depend on data mining and text mining to gain information and a better understanding of business issues. By exploring and modeling large amounts of data, companies can uncover hidden relationships and patterns that enhance the ability to make accurate predictions and drive competitive advantage.

With SAS Text Miner you can classify documents into predefined or data-driven categories, find explicit relationships or associations between documents and incorporate textual data with structured inputs. The dynamic exploration component helps you find patterns in large document collections, combine those insights with other predictive analytics and gain maximum value from more information sources.

Key benefits

- **Reduces time to decisions, saves money and resources.** There are many tasks that are currently manually performed or completely ignored. With SAS Text Miner, organizational activities are streamlined, resulting in immediate ROI and performance gain.
- **Recognizes trends and predicts business opportunities.** Analysis of information such as blogs, customer feedback and call center notes may provide valuable information about your customer's critical issues, insights into service and product needs. SAS Text Miner combines a variety of information sources, including text and traditional databases, to provide complete views of an organization. By combining structured data and unstructured text and automating the process of analyzing the data, organizations can gain meaningful insights that successfully drive overall business direction.



**THE
POWER
TO KNOW®**

Product overview

SAS Text Miner provides a rich suite of text processing and analysis tools that can uncover underlying themes or concepts across large document collections. Text documents can be clustered automatically into groups, classified into predefined categories and used in conjunction with structured data to build predictive models. Text mining can be described as a three-step process: accessing the unstructured text, parsing the text and turning it into actionable data, and analyzing the newly created data. For each step, SAS Text Miner provides state-of-the-art tools that enable organizations to efficiently extract intelligence from large text collections.

Access to a variety of document formats and languages

SAS Text Miner can read text stored in a variety of document formats, such as PDF, ASCII, HTML, Microsoft Word and WordPerfect. This enables users to analyze information from a wide range of sources, including the Internet via Web crawling capabilities. Customized routines and dictionaries are available for

Danish, Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Norwegian (Bokmal), Portuguese, Spanish, Swedish, Traditional Chinese and Simplified Chinese. Languages not currently supported may be encoded and analyzed using Unicode UTF-8 encoding.

A distinctive, integrated interface

SAS Text Miner's unique, fully integrated graphical user interface significantly reduces text mining time for both business analysts and statisticians. The Java client/SAS server architecture provides informative summary graphics, making it easy to drill down into textual documents to gain deeper insight. With the tiered-server relationships, computational processes can be separated from the user interface. Powerful UNIX and Windows servers can be dedicated to intensive mining while users work from their desktops. This provides unprecedented flexibility for configurations that scale from single users to enterprise solutions. In addition, the interface automatically generates score code as models are built. This SAS score code then can be exported and deployed in

other environments, including through familiar BI clients such as Microsoft Excel, SAS Web Report Studio, SAS Information Delivery Portal and SAS Enterprise Guide.

Text parsing and extraction

Sophisticated text parsing capabilities automatically extract terms and phrases from large text collections while stemming terms to root forms based on parts of speech, as well as finding phrases of interest such as abbreviations, country or organization names and other user-specified entities.

Automatic text cleaning

Unless a document is ready for publishing, it likely contains bad punctuation and spelling errors. Call center documents are a good example of this. An automatic spelling detection capability significantly reduces the manual effort required to create industry- and application-specific ontologies, accurately identifying misspelled words. This leaves more time for true data exploration so you can find the nuggets of gold contained in the text you are analyzing.

Concept linking

User directed concept linking provides flexible navigation to visualize complex hidden relationships between terms, phrases and entities (such as people and place names). Concept linking, now integrated with the Interactive Results Browser for enhanced usability, helps detect patterns in a clear visual fashion that otherwise may not have been observed.

Dimension reduction

Parsed documents can be transformed into a numerical representation using Singular Value Decomposition (SVD), rollup terms, or a combination of the two. SVD is a powerful technique for automatically relating similar terms and documents, eliminating an exhaustive

The screenshot displays the SAS Text Miner Interactive Results Browser interface. The top window shows a document snippet with text about Oracle database procedures and decision support systems. Below this, two tables are visible: 'Terms' and 'Clusters'.

Term	Freq	# Docum...	Weight	Role	All
ans	3166.0	377.0	0.06287560487Prep	Alph	α
n	2486.0	368.0	0.08004323111Prep	Alph	α
gda	2215.0	735.0	0.109704629989Prep	Alph	α
with	1243.0	719.0	0.09681262457Prep	Alph	α
paper	868.0	848.0	0.10284208674Prep	Alph	α
will	1140.0	809.0	0.12222609120Prep	Alph	α
on	865.0	869.0	0.13151313018Prep	Alph	α
can	845.0	820.0	0.12167620111Prep	Alph	α
have	764.0	493.0	0.14821904797Verb	Alph	α
software	887.0	432.0	0.18813992400Prep	Alph	α
system	622.0	379.0	0.18773737979Prep	Alph	α
application	745.0	371.0	0.19664214488Prep	Alph	α
user	896.0	359.0	0.19475197870Prep	Alph	α
provide	495.0	345.0	0.19595957259Prep	Alph	α
data	523.0	334.0	0.21206932508Prep	Alph	α
system	835.0	301.0	0.22818047020Prep	Alph	α
new	454.0	294.0	0.22271662302Adj	Alph	α

#	Descriptive Terms	Freq	Percenta...	RMG
1	+ application, + provide, + will, with, on	61	0.049273021000	1257
2	+ create, html, software, ca, n, on	36	0.109854604200	1070
3	+ table, + will, data, can, on	113	0.091276252010	1188
4	ans/as, + application, + use, software, + will	92	0.074313400720	0997
5	+ new, system, + will, + application, sas	34	0.106239095310	1087
6	+ set, + analysis, data, dat, a, + paper	100	0.080775442630	1114
7	+ program, + macro, + variab, le, can, sas	142	0.114701130895	1041
8	+ system, + analysis, + prov, ide, data, + have	37	0.029886914370	1191

SAS Text Miner's Interactive Results Browser lets users interactively explore concepts and relationships between documents and dynamically make modifications to further tailor analyses.

need to manually generate industry specific ontologies or synonym lists. Rollup terms reduces dimensionality by taking the n highest weighted terms and ignoring the rest. Rollup terms has been shown to be very effective for short documents of around five words.

Text clustering

Text clustering algorithms automatically group documents into common themes and topics based on their content. The Taxonomy Browser automatically creates document taxonomies enabling users to quickly spot key information and drill down into a complete taxonomy of their document collection. Expectation Maximization Clustering groups documents using spatial clustering techniques. Cluster summaries can be quickly generated and easily interpreted in the context of the original text documents. The Interactive Results Browser enables analysts to easily explore concepts and relationships between documents and dynamically make modifications to further tailor their analyses. From the Documents window you can also filter and find similar documents. Documents can be filtered based on any characteristic, including presence or absence of terms, and probed to find documents and terms similar to a target document or term. SAS Text Miner will automatically select and filter clusters to allow closer inspection of specific documents.

Full integration with leading SAS Enterprise Miner™ software

Seamless integration with SAS Enterprise Miner provides a full range of mining tools for text and related structured data, including prediction, classification and clustering, as well as the full range of data access and data preprocessing tools. Organizations can easily make use of SAS' award-winning analytical software to drive sound business decision making.

Key Features

Universal data access

- Access to numerous forms of textual data, including PDF, extended ASCII text, HTML and Microsoft Word.
- Web crawling capabilities.
- Ability to extract, transform and load textual data into a SAS data set for mining.

Support for multiple languages

- Total language list: Danish, Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Norwegian (Bokmal), Portuguese, Spanish, Swedish, Traditional Chinese and Simplified Chinese.
- Support for Latin-1, Double Byte Character and UTF-8 encodings.
- European languages (Latin-1 encoding): Danish, Dutch, English, Finnish, French, German, Italian, Norwegian (Bokmal), Portuguese, Spanish and Swedish.
- Far-Eastern languages (Double Byte Character Support): Japanese, Korean, Simplified Chinese and Traditional Chinese.
- Encoding support for Unicode UTF-8.

Self-documenting interface

- User-friendly interface eliminates manual coding with visual diagrams.
- Process flow diagrams can be modified, saved and shared with others.
- Flexible reporting allows results to be published in a concise HTML format.

Comprehensive text preprocessing capabilities

- Capture and distill the most important underlying information within a document collection.
- Default or customized stop lists for each language to remove terms with little or no informational value.
- Automated spelling correction.
- Stemming to identify root words.
- Part-of-speech tagging based on sentence context.
- Noun group extraction for identifying phrase-level concepts such as 'competitive intelligence.'
- User-defined multiword tokens, such as 'point and click.'
- User-customized and default synonym lists.
- Compound word splitting into distinct sub-terms.

Extensive feature extraction

- Broad customizable data dictionaries can extract particular pieces of information such as names of people, products, organizations, URLs and addresses.
- Extracted entities are then normalized and included in a matrix table.
- Entity extraction is available for English, French, German and Spanish.

Dimension reduction techniques

- Textual data is preprocessed into an information-rich matrix for application of powerful dimension reduction techniques.
- Rollup terms automatically identify the n -highest weighted terms in a document.
- Singular value decomposition (SVD) transforms each document into an n -dimensional subspace.

Text clustering algorithms

- Group documents based on their content.
- Expectation-maximization clustering groups documents using spatial clustering techniques.
- Hierarchical clustering using Ward's agglomerative method facilitates automatic grouping of documents into taxonomies. Documents grouped into hierarchical clusters belong to one leaf cluster as well as its parent clusters.
- Cluster documents downstream in the Process Flow Diagram using K-means or SOM/Kohonen clustering.
- Profile clusters using additional structured data from original documents (age, purchase propensity, etc.).

SAS® Text Miner 3.1 Technical Requirements

Client environment

- Windows (x86-32): Windows 2000 Professional, Windows XP Professional, Windows NT 4 Workstation
- Internet Explorer 5.5+

Server environment

- AIX (64-bit), Release 5.1+
- Solaris (64-bit), Version 8, 9 or 10 on SPARC
- Windows (x86-32): Windows NT 4 Server, Windows 2000 Professional, Windows Server 2003

Required software

SAS Enterprise Miner is required and must be installed on the same machine.

Key Features (continued)

Interactive Results window

- Provides a concise summary of results that includes document, term and cluster tables.
- Sort term table by terms, term frequency, number of documents, weight and term role.
- Expand parent terms to identify child terms and their related statistics.
- Toggle between full and partial text view of the documents.
- Find the n most similar items for the selected document, term or cluster.
- Filter term(s) to show documents that contain them and clusters that contain those documents.
- Filter document(s) to show all terms in the documents, as well as revised cluster counts.
- Filter cluster(s) to show all documents in the filtered clusters, as well as the terms in those documents.
- Modify the keep and drop term lists.
- Treat selected terms as equivalent.
- Re-weight terms using a different algorithm.
- Select the number of SVD dimensions.
- Browse concept links.
- Browse taxonomies.
- View the top n most representative terms for each cluster.
- Re-cluster anytime using a subset of documents or terms.

Document categorization

- Use neural networks, memory-based reasoning, regression and decision trees to assign documents to predefined categories.
- Seamlessly combine quantitative and qualitative data with text analysis to improve predictions.
- Compare performance of multiple models and deploy score code to categorize new documents.

Enhanced performance

- SAS Text Miner 3.1 includes a high-performance, procedure-driven interface, enabling significant performance improvements.



THE
POWER
TO KNOW.

SAS Institute Inc. World Headquarters +1 919 677 8000

To contact your local SAS office, please visit: www.sas.com/offices

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2007, SAS Institute Inc. All rights reserved. 101371_453242.0707