



## SAS® Text Miner 4.2

Capitalize on the value hidden in textual information

### What does SAS® Text Miner do?

SAS Text Miner incorporates advanced linguistic capabilities, powered by Teragram, within the core data mining solution of SAS® Enterprise Miner™. Consolidating structured (quantitative) data analysis with unstructured (free-form text) provides complete views and meaningful insights within an integrated predictive modeling environment. Automating manual comprehension of the textual data sources, incorporating interactive drill-down reporting and delivering algorithms for rigorous advanced analyses make it possible to grasp future trends and act on new opportunities more efficiently and with less risk.

### Why is SAS® Text Miner important?

SAS Text Miner saves money and resources by automating time-consuming tasks of reading and comprehending text. By consolidating structured data sources with text-based information, you gain a more accurate and complete view of your organization. Analysis performed using both types of data will produce descriptive and predictive models that spot more opportunities and recognize trends more accurately so that actions can be based on better decisions.

### For whom is SAS® Text Miner designed?

SAS Text Miner is designed primarily for business analysts and statisticians who must look through large volumes of text to extract information, ideas and trends. It is applicable across all industries and the public sector, and is especially useful for organizations that are actively building predictive models.

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

To get the most value from collected data, you must be able to analyze it before acting on it. Due to the ambiguity of conversational language, key messages buried in text-based data are not easy to discern, much less process. Most organizations lack the ability to combine text-based information with their structured data in decision-making contexts.

With SAS Text Miner you can classify documents into predefined or data-driven categories and find explicit relationships or associations between topics to analyze textual material along with your structured data. Interactive exploration empowers you to discover patterns in document collections and apply those insights directly to your predictive models, delivering maximum value across all of your information sources.

### Key Benefits

- **Reduces decision time through automated processes.** By implementing intelligent algorithms and vocabulary processing techniques, time-consuming activities previously done manually – such as categorization, tagging or building of topic libraries and document indexes – are generated automatically and executed consistently and efficiently.
- **Enhances the discovery process by uncovering associations and relationships previously undetected.** Why limit your text analytics to searching for terms or querying known items? SAS Text Miner provides a unique data-driven method for identifying new concepts with its rich interactive user interface that highlights paths and links for in-depth document analysis.
- **Visually presents a high-level view of data with the ability to drill down to specific phrases in documents.** SAS Text Miner offers a visual presentation of the entire data-mining process with the ability to drill down to relevant detail illustrating the connections and exploring the links between items in document collections.
- **Enables you to recognize trends and spot business opportunities with a full range of predictive modeling tools.** Analysis of information such as customer letters and call center notes provides valuable information about customer dissatisfaction or insights into service and product needs.



## Product Overview

SAS Text Miner provides a rich suite of linguistic and analytical modeling tools for discovering and extracting knowledge from multiple text documents. After transforming text so that it can be fed into data mining engines, topics and themes are identified as explicit associations so documents can be clustered into related groups for the scoring purposes of predictive modeling. A high-performance search capability, enhanced spell-checking and the processing of multiple topics per document are now provided. Results from SAS Enterprise Content Categorization or the SAS Concept Creation for SAS Text Miner add-on module can be directly integrated into your text mining to complement whatever customized entities you create.

### Access to a variety of document formats and languages

SAS Text Miner can read text stored in a wide variety of document formats; and preprocessing wizards assist in the transformation of files into SAS data sets so they can be input into SAS Text Miner. This enables you to analyze information in a single integrated system from a wide range of sources,

including the Internet and social media networks via Web-crawling capabilities. Customized routines and dictionaries are available for Arabic, Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Spanish and Swedish. Entity extraction is provided for all supported languages. Languages not currently supported may be encoded and analyzed using Unicode UTF-8 encoding.

### User-friendly, flexible interface

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 score code can be exported and deployed as familiar business intelligence clients, including Microsoft Excel, SAS Enterprise Content Categorization, SAS® Enterprise Guide® and/or JMP®.

### Comprehensive text parsing

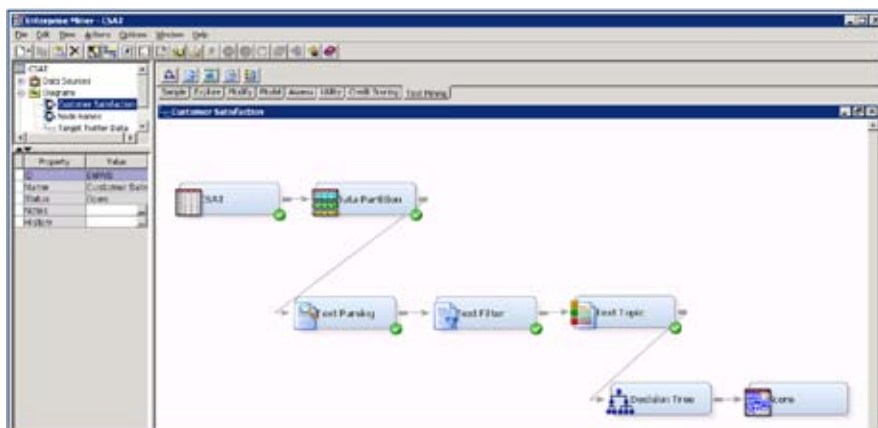
Text parsing decomposes textual data and generates a quantitative representation that is suitable for data mining purposes. The Text Parsing node (new in SAS Text Miner 4.2) decomposes the data into meaningful parts of speech, addresses, phone numbers and company names, including stems or root forms. This enhanced parser lets you choose to ignore words or specify which to treat as synonyms. Negation, multiple-word phrases and user-defined custom entities now compliment previous parsing functionality.

### Dimension reduction

Advanced filtering with weighting, integrated spell-checking and transformation of the qualitative data into compact formats is made possible by SAS Text Miner's sophisticated dimension reduction techniques. Parsed documents can be transformed into a numerical representation using singular value decomposition (SVD), roll-up terms or a combination of the two.

### Text topic identification and clustering

Advanced algorithms automatically group documents into common themes and topics based on content. Instead of requiring that documents belong to a single topic ("hard clustering"), the new Text Topic node in SAS Text Miner 4.2 works from the premise that any particular document could relate to several topics of interest or none at all. These topics can be defined by the user or automatically determined by the tool. The Text Topic node's interactive interface enables users to look at the documents clusters associated with different topics and tweak topic definitions on the fly. Alternatively, when hard clustering is desired, the Text Mining node is used to place



**SAS Text Miner 4.2 includes three new nodes (Text Parsing, Text Filter and Text Topic). Multiple-word phrases, full-text search features and the ability to incorporate user-defined custom entities are just a few of the new capabilities.**

topics into a hierarchy of clusters or a list of clusters. Expectation-maximization clustering processes apply spatial clustering techniques to organize the documents into meaningful groups. Cluster summaries can be displayed in an easy-to-interpret manner within the context of the original text documents. The interactive visualization environment enables analysts to explore concepts and relationships between documents and dynamically make modifications to further tailor their analyses as desired.

### Text filtering

The Text Filter node (new in SAS Text Miner 4.2) provides integrated full-text search capabilities, and automates spell-checking, concept linking and the subsetting of terms and documents. An interactive query will retrieve individual documents matching whatever search parameters you specify. Filters can be based on any characteristic, including presence or absence of terms, and the interactive visualization enables you to probe until you find documents and terms that meet your specifications. Concept maps link terms, phrases and entities in a visual, interactive manner so you can discern previously undetected patterns.

### Directly implement results from other SAS® Analytics software

Seamless integration with SAS premier predictive modeling software or any of the new SAS Text Analytics offerings opens up a full range of mining tools for textual and structured data, as well as data preprocessing, scoring and deployment tools. Organizations with SAS' award-winning analytical software can deploy analytics into their operational environments to identify and solve critical business issues more efficiently.

## Key Features

### Universal data access

- Access to numerous forms of textual data, including PDFs, extended ASCII text, HTML, Microsoft Office formats, spreadsheets, presentations, e-mail and database formats.
- Web-crawling capabilities, including social media discussions such as Twitter and news feeds.
- Ability to extract, transform and load textual data into a SAS data set for mining.

### Support for multiple languages

- Support for Latin-1, Double Byte Character and UTF-8 encodings.
- European languages (Latin-1 encoding): Dutch, English, French, German, Italian, Polish\*\*\*, Portuguese, Spanish and Swedish.
- Eastern languages (Double Byte Character Support): Arabic, Chinese, Japanese, Korean.

### User-friendly, flexible interface

- Text mining is encapsulated into four different nodes corresponding to common tasks. They can be combined in any way depending on the task at hand. These text nodes operate directly with the variety of SAS Enterprise Miner nodes and can be extended further by customizing your own algorithm or declaring a new user-written business rule for predictive modeling, clustering, visualization and reporting, and thus are deployable as SAS score code.
- Process-flow diagrams of text mining analysis can be modified, saved and shared with others.
- Flexible reporting allows results to be published in a concise HTML format.
- The Concept Link diagram displays a visual relationship between terms.

### Text Parsing node \*\*\*

- Default or customized stop lists will remove terms with little or no informational value from your analysis.
- 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.”
- Out-of-the-box support for many different entity types, including person and company names, locations, dates, addresses, measurements, and e-mail and URL addresses. These entities are customized for every language supported.
- User-defined multiword tokens, such as “point and click.”
- User-customized and default synonym lists.
- Comprehensive capabilities include compound word splitting into distinct subterms.

### Dimension reduction techniques

- Roll-up terms automatically identify the  $n$  highest-weighted terms in a document.
- Singular value decomposition (SVD) transforms each document into an  $n$ -dimensional space where the closer two documents are in that space, the more similar they are.

### Text Topic node \*\*\*

- Taxonomy browser displays the default topics automatically generated, as well as the manually created topics defined by the user.
- Documents can be categorized as belonging to zero, one or even many different topics.
- Topics can be customized interactively in an easy-to-comprehend and intuitive visual environment.

### Text clustering algorithms

- Expectation-maximization clustering groups documents into discrete nonoverlapping clusters (also known as hard clustering) using spatial clustering techniques.
- Hierarchical clustering facilitates automatic grouping of documents into taxonomies.
- Profile clusters and topics by incorporating structured data from original documents to enhance overall analysis (e.g., age, purchase propensity, etc.).

## SAS® Text Miner 4.2 Technical Requirements

### Supported platforms

- AIX: Version 5.3 and Version 6.1 on POWER architectures
- HP-UX Itanium: HP-UX 11iv2 (11.23), 11iv3 (11.31)
- Linux for x86 (x86-32): RHEL 4 and 5, SuSE SLES 9 and 10
- Microsoft Windows (x86-32): Windows XP Professional, Windows Vista\*, Windows Server 2003 family
- Microsoft Windows on x64 (EM64T/AMD64): Windows XP Professional for x64, Windows Vista\* for x64, Windows Server 2003 for x64
- Solaris on SPARC: Version 9, 10
- Solaris on x64: Version 10

\* NOTE: Windows Vista Editions that are supported include Enterprise, Business and Ultimate

### Supported Web browsers

- Internet Explorer 6 on Windows XP Pro
- Internet Explorer 7 on Windows XP Pro and Windows Vista\*
- Firefox 2.0 on Windows XP Pro, Windows Vista\* and Linux x86 (SuSE and RHEL)

### Middle tier required/optional software

- SAS client and middle tier require Sun JRE 1.5

### Required software

- SAS Enterprise Miner is required and must be installed on the same machine as SAS Text Miner; or SAS Enterprise Miner for Desktop is required and must be installed on the same machine as SAS Text Miner for Desktop

## Key Features (continued)

### Text Filter node \*\*\*

- Contains a concise view of documents and vocabulary or all terms discovered during parsing.
- Automatically performs spell-checking by mapping misspelled words to the terms they were misspelled from.
- Apply Google-like searches or SQL WHERE clauses to subset analysis (for example, conducting separate warranty analysis for each make or model of automobile).
- Can programmatically and interactively distinguish and filter out unimportant terms, easily map abbreviations and represent other equivalent terms.

### Get a 360-degree view of your data

- Combine textual data with traditional structured data mining to automate, visualize, classify and deploy your predictive modeling results.
- Seamlessly combine quantitative and qualitative data with text analysis to improve predictions.
- Advanced techniques such as neural networks, memory-based reasoning, regression models and decision trees are extensible via the SAS Enterprise Miner Code node which allows more innovation and quicker deployment with less risk.
- Performance assessments of multiple models can be displayed side-by-side to help you select the best one to deploy as score code for categorizing new documents.
- Output from SAS Enterprise Content Categorization can be directly integrated into your text mining analysis. Discovered topics and themes produced by SAS Text Miner are valuable input for SAS Enterprise Content Categorization, especially in situations where taxonomies did not previously exist. \*\*\*

\*\*\* New in SAS Text Miner 4.2 (released December 2009)

TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
sas institute	2391	525	<input checked="" type="checkbox"/>	0.034	COMPANY	Alpha
be	1755	498	<input checked="" type="checkbox"/>	0.044	Verb	Alpha
software	857	485	<input checked="" type="checkbox"/>	0.042	Noun	Alpha
use	694	354	<input checked="" type="checkbox"/>	0.091	Verb	Alpha
status	1201	340	<input checked="" type="checkbox"/>	0.116	Noun	Alpha
system	533	262	<input checked="" type="checkbox"/>	0.151	Noun	Alpha
application	502	217	<input checked="" type="checkbox"/>	0.18	Noun	Alpha
user	343	201	<input checked="" type="checkbox"/>	0.186	Noun	Alpha
have	275	192	<input checked="" type="checkbox"/>	0.184	Verb	Alpha

The powerful search syntax of the Interactive Filter Viewer finds documents depending on words or phrases contained in them – and provides the flexibility to subset your analysis.



SAS Institute Inc. World Headquarters +1 919 677 8000

To contact your local SAS office, please visit: [www.sas.com/offices](http://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 © 2010, SAS Institute Inc. All rights reserved. 101371\_543297.0210