



## SAS® Text Miner

*Capitalize on the value hidden in textual information*

Think of the huge volumes of text documents that flood your organization each day: Web pages, e-mail messages, articles, memos, customer feedback, warranty claims, patent information, surveys, research studies, resumes, client notes, competitive intelligence and more. There isn't time to read everything, much less sort and classify the essential bits of information from each document.

Likewise, most businesses have neither the time nor resources to capitalize fully on the value of the large amounts of textual data they generate in daily operations. Natural language text is not designed for analytical processing. Because of its ambiguity and the numerous ways to represent similar concepts, information implicit in textual data is not easy to discern, quantify or exploit.

And even if you were able to tap into this wealth of information, what next? How could you possibly integrate it with traditional structured data such as age, job classification and income information, to gain a better understanding of the big picture?

### **Integrating text mining into the data mining process**

Many organizations today depend on data mining to help them better understand their customers and their business. By exploring and modeling large amounts of structured data, companies can uncover hidden relationships and patterns of information and enhance the ability of decision makers to make accurate predictions that drive competitive advantage.

However, the bulk of real-world data is unstructured. To get the most value from unstructured text-based data, an organization must be able to analyze it automatically – just as with more structured data.

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.

For example, companies implementing analytical customer relationship management (CRM) programs can use SAS Text Miner to manage large volumes of inbound customer e-mails, categorizing them for faster routing and follow-up. Free-form text notes collected by call center representatives can be categorized into meaningful clusters so that intelligent decisions can be made about which products are most appropriate for individual customers.

Human resource professionals can use SAS Text Miner to sift through thousands of resumes and applications to properly match qualified applicants with job openings. They can also combine free-form survey responses with other responses to identify trends and create results they can act on.

With the ability to evaluate and analyze text-based notes about patients, pharmaceutical companies can enhance the clinical trials process. Search queries to large document databases can be refined to focus on articles of interest.



With SAS Text Miner, anyone who must sift through large volumes of text to extract information, ideas and trends will be able to transform these massive, largely untapped stores of data into insightful and valuable knowledge.

## SAS Text Miner Features

### Universal data access

With access to numerous forms of textual data, including Adobe Portable Document Format (PDF), extended ASCII Text, HTML and Microsoft Word, users can extract, transform, and load their textual data into a SAS data set for text mining.

### Support for multiple languages

Advanced text parsing is provided for English, French and German, or a combination of those languages through automatic language identification. Basic parsing is also provided for many other languages with words that are delimited by spaces and/or punctuation.

### Numerous text preprocessing methods

Once textual data is read into a SAS data set, Text Miner provides a comprehensive set of text preprocessing capabilities to capture and distill the most important underlying information within the document collection. These capabilities include:

- Default or customized stop lists for each language to remove terms with little or no information value.
- Stemming to identify root words, such as *running* and *run*, *bills* and *bill*, and *excluding* and *exclude*.
- Part-of-speech tagging based on sentence context. For example, tagging recognizes that “Bill took it to the bank” uses bank as a noun while “Bill can bank on it” uses it as a verb.
- Noun group extraction identifies phrase-level concepts such as *data mining* and *competitive intelligence*.

- User-defined multiword tokens, such as *cup holder* or *point and click*.
- User-customized and default synonym lists.
- Splitting of compound words into distinct subterms. (This is particularly important in languages such as German, where compound words are created by concatenating several simpler words.)

### Extensive feature extraction capabilities

Broad, customizable data dictionaries can extract particular pieces of information, such as names of people, products, organizations, URLs, and addresses. The extracted entities are then normalized and included in a matrix representation.

### Dimension reduction techniques

With the textual data preprocessed into an information-rich matrix, powerful dimension reduction techniques can be applied.

- Rollup terms provide a standard method of reduction that chooses the  $n$ -highest weighted terms to represent a document.
- Singular value decomposition (SVD) projects each document into an  $n$ -dimensional subspace that best fits the document collection. In this reduced space, similar documents tend to be placed near one another.

### Unique clustering algorithms

After applying a dimension reduction technique, the Text Mining node provides two clustering techniques to group documents based on their content:

- Expectation-maximization clustering groups documents so that they belong to each cluster with an assigned probability.
- Hierarchical clustering facilitates grouping documents into taxonomies. Documents grouped into hierarchical clusters belong to one leaf cluster as well as its parent clusters.

Both techniques assist users with cluster profiling by providing a list of the most diagnostic terms for each cluster.

Because Text Miner is fully integrated with SAS® Enterprise Miner®, the Clustering and Self-Organizing Maps nodes of Enterprise Miner can be used to cluster documents downstream in the Process Flow Diagram. The clusters also can be profiled using additional structured data (such as age, purchase propensity, etc.) that may have been collected with the original documents.

### Document categorization

Once the text has been preprocessed and transformed into a numerical representation of the documents, Enterprise Miner tools such as neural networks, memory-based reasoning, regression and decision trees can be used to assign documents to predefined categories. Unlike other document categorization tools, SAS Text Miner can seamlessly combine additional quantitative and qualitative data with the text analysis data to improve predictions. Finally, users can compare the performance of multiple models in the Assessment node and deploy the score code to categorize new documents.

### Interactive results viewer

The Text Miner results viewer provides a concise summary of results that includes document, term and cluster tables (Figure 1). Interactive features allow users to:

- Sort the term table by terms, term frequency, number of documents, weight and term role.
- 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.

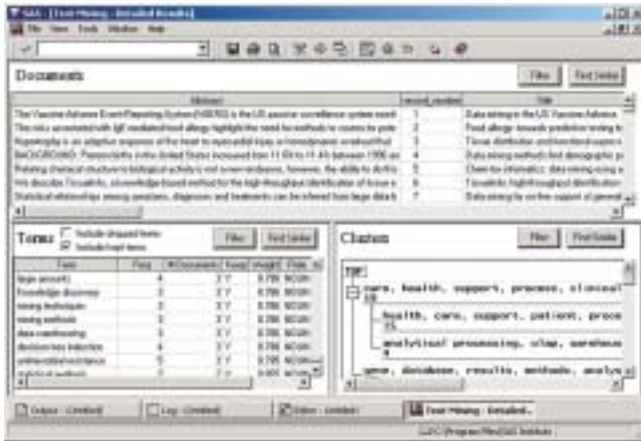


Figure 1: The interactive results viewer provides a concise summary of the text mining results.

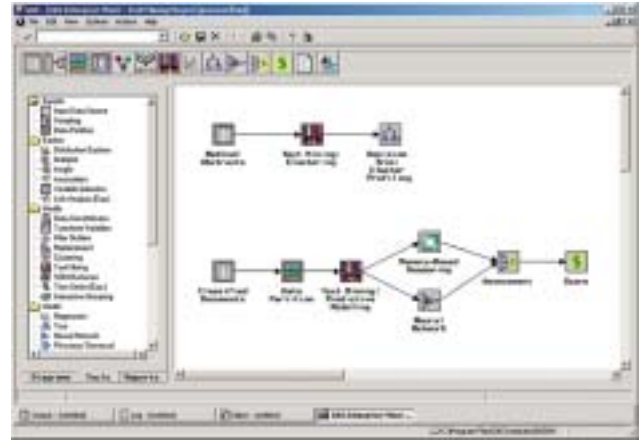


Figure 2: SAS Text Miner runs within the intuitive point-and-click process flow environment of Enterprise Miner enabling you to seamlessly incorporate textual data into the mining process.

- Filter document(s) to show all terms in the documents, as well as revised cluster counts.
- Filter clusters 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.
- Reweight terms using a different algorithm.
- Select the number of SVD dimensions.
- View the top *n* most representative terms for each cluster.
- Recluster at any time using a subset of documents or terms. Often the initial clustering is performed in the results viewer rather than at node run time.

**Easy-to-use self-documenting interface**

The graphical user interface, designed around Enterprise Miner's exclusive Process Flow Diagram, eliminates manual coding and significantly reduces text-mining time for both

business analysts and statisticians. The process flow can be modified, saved and shared with other analysts.

**Flexible reporting capabilities**

The results from a text mining process flow diagram can be published in a concise HTML report.

**The SAS® Intelligence Advantage**

With the integration of text mining into SAS' proven Enterprise Miner solution for data mining, SAS is the first software vendor to offer a complete data mining solution for analyzing both unstructured and structured data (Figure 2).

SAS is the market leader in providing a new generation of business intelligence software and services that create true enterprise intelligence. SAS solutions are used at more than 40,000 sites – including 90 percent of the Fortune 500 – to develop more profitable relationships with customers and suppliers; to enable better, more accurate and informed decisions; and to drive organizations forward. SAS is the only vendor

that completely integrates leading data warehousing, analytics and traditional BI applications to create intelligence from massive amounts of data. For more than 25 years, SAS has been giving customers around the world *The Power to Know*®. Visit us at [www.sas.com](http://www.sas.com).



World Headquarters  
and SAS Americas  
SAS Campus Drive  
Cary, NC 27513 USA  
Tel: (919) 677 8000  
Fax: (919) 677 4444  
U.S. & Canada sales:  
(800) 727 0025

SAS International  
PO Box 10 53 40  
Neuenheimer Landstr. 28-30  
D-69043 Heidelberg, Germany  
Tel: (49) 6221 4160  
Fax: (49) 6221 474850  
**[www.sas.com](http://www.sas.com)**

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 © 2003, SAS Institute Inc. All rights reserved. 49150US\_253220.0903

Contains LinguistX® from Inxight Software, Inc. Copyright © 1996-2002. All rights reserved. [www.inxight.com](http://www.inxight.com).  
Contains Thing Finder™ Server from Inxight Software, Inc. Copyright © 1996-2002. All rights reserved. [www.inxight.com](http://www.inxight.com).