



SAS® Text Miner 4.1

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 Teragram natural language processing 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, 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 process it. But because of the ambiguity and numerous ways to represent similar concepts, information buried 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 predictive models to gain better understanding of business issues and take actions resulting in a competitive advantage. Companies that apply the highly successful, iterative process of data mining by creating predictive and descriptive models are uncovering hidden relationships and patterns in textual data that enhance their ability to make accurate predictions and better decisions.

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 discover patterns in large document collections, combine those insights with your predictive analytics to gain maximum value from all of your information sources.

Key benefits

- **Save money and resources.** There are many tasks that are currently performed manually or completely ignored. With SAS Text Miner, organizational activities are streamlined, resulting in immediate ROI and performance gain.
- **Recognize trends and spot business opportunities.** Analysis of information such as blogs, customer feedback and call center notes may provide valuable information about your customers' critical issues, insights into service and product needs. This helps decision makers gain meaningful insights that successfully drive overall business direction.
- **Process a variety of information sources, including text and traditional databases, to deliver complete views of an organization.** Combining structured data and unstructured data types enables you to automate many of the manual steps required before analysis traditionally begins.



Product overview

SAS Text Miner provides a rich suite of text processing and analysis tools that uncovers underlying themes or concepts across large document collections. Smart default settings allow text documents to be grouped automatically into clusters, or you can set your own predefined categories to be used in conjunction with structured data to build predictive models. Text analytics is a three-step process: accessing the unstructured text, parsing the text and transforming it into usable results.

Access to a variety of document formats and languages

SAS Text Miner can read text stored in a variety of document formats. This enables you to analyze information from a wide range of sources, including the Internet via Web crawling capabilities. Customized routines and dictionaries are available for English, French, German, Italian, Portuguese, Spanish and Chinese (combined Traditional and Simplified Chinese). Entity extraction is provided for all supported languages.

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 saves time for both business analysts and statisticians who face growing amounts of textual inputs. 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 can be exported and deployed as familiar business intelligence clients, including Microsoft Excel, SAS Content Categorization, SAS® Enterprise Guide® and JMP®.

Text parsing and extraction

Sophisticated text parsing capabilities, delivered by the fully integrated Teragram parsing engine, automatically extract terms and phrases from large text collections. Stemming capabilities on root forms are based on parts of speech so that phrases of interest, such as abbreviations, country or organization names and other entities, can be identified accurately.

Automatic text cleaning

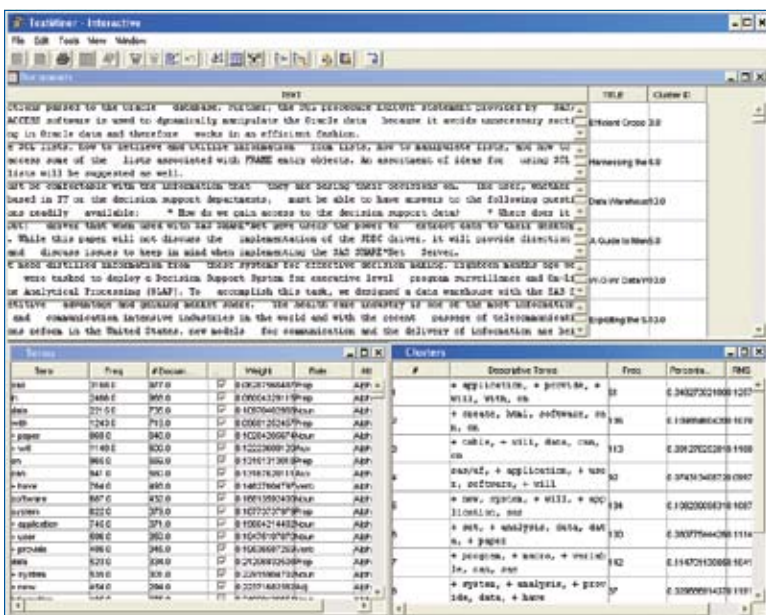
Until a document has been edited and approved for publishing, it likely contains punctuation and spelling errors. Call center documents and e-mails often are full of typos and slang. An automatic spelling detection capability significantly reduces the manual effort required to correct misspelled words, and it also can be applied to create industry- and application-specific ontologies. This leaves more time for true data exploration so you can find information contained in the text you are analyzing.

Concept linking

User-guided concept mapping links terms, phrases and entities (such as people and place names) in a visual, interactive flexible manner. Concept linking, now integrated with interactive training for enhanced navigation, 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), roll-up terms, or a combination of the two. SVD is a powerful technique for automatically relating similar terms and documents, eliminating the need to manually generate industry-specific ontology or synonym lists. Roll-up terms reduce dimensionality by taking the n highest-weighted terms and ignoring



An Interactive Results Browser lets users interactively explore concepts and relationships between documents and dynamically make modifications to further tailor analyses.

the rest. Roll-up terms have shown to be very effective for short documents containing a handful of terms.

Text clustering

Text clustering algorithms automatically group documents into common themes and topics based on 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 training environment enables analysts to explore concepts and relationships between documents and dynamically make modifications to 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 preprocessing, scoring and deployment tools. Organizations can easily make use of SAS' award-winning analytical software to drive sound business decision making and deploy analytics into their operational environments.

Key Features

Universal data access

- Access to numerous forms of textual data, including PDF, extended ASCII text, HTML and Microsoft Office formats.
- 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: English, French, German, Italian, Portuguese, Spanish, and Chinese (combined Traditional and Simplified Chinese).
- Support for Latin-1, Double Byte Character and UTF-8 encodings.
- European languages (Latin-1 encoding): English, French, German, Italian, Portuguese, and Spanish.
- Far-Eastern languages (Double Byte Character Support): Combined 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 all supported languages.

Dimension reduction techniques

- Textual data is preprocessed into an information-rich matrix for application of powerful 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 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 4.1 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
- 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

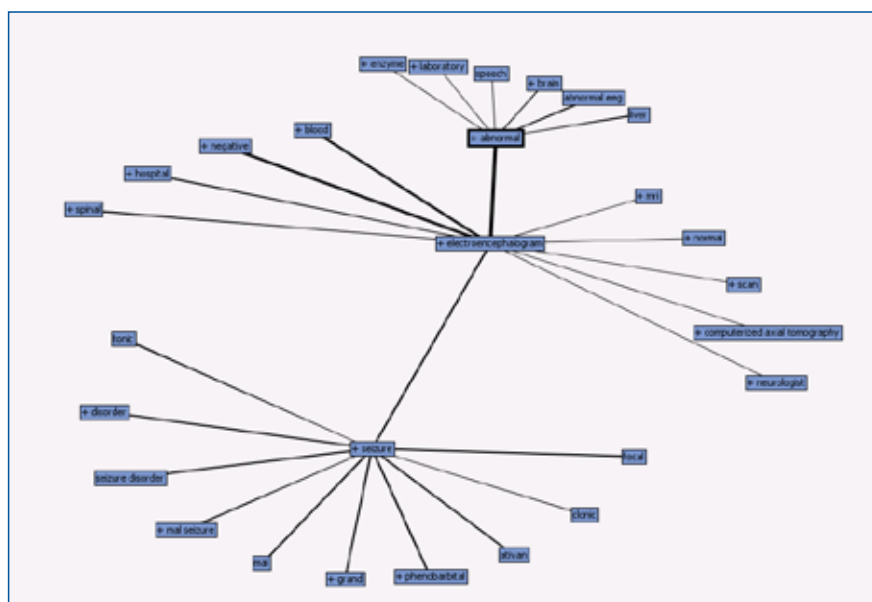
Key Features (continued)

Interactive training

- Provides a concise summary of results that includes document, term and cluster tables.
- Sorts term table by terms, term frequency, number of documents, weight and term role.
- Expands 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.
- Reweight 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.
- Recluster anytime using a subset of documents or terms.

Document categorization

- Advanced techniques such as neural networks, memory-based reasoning, regression models and decision trees will assign documents to predefined categories.
- Seamlessly combine quantitative and qualitative data with text analysis to improve predictions.
- Graphically display performance assessments of multiple models to compare and select the best one to deploy as score code for categorizing new documents.



Interactive text analysis of terms is displayed in this concept linking window. Here you can see a display of terms grouped by clusters according to their association with each other.



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 © 2009, SAS Institute Inc. All rights reserved. 101371_539267.0609