SAS® Text Miner

Automate discovery and insights from document collections

What does SAS® Text Miner do?

SAS Text Miner discovers information buried in collections of text. By automatically reading text data and delivering algorithms for rigorous, advanced analyses, the solution makes it possible to grasp future trends and act on new opportunities more precisely and with less risk. It includes advanced linguistic capabilities within the core data mining solution of SAS® Enterprise Miner™ so you can easily extend text insights into structured data mining and predictive analysis.

Why is SAS® Text Miner important?

The software saves money and resources by automating the time-consuming tasks of reading and comprehending electronic text. By consolidating structured (quantitative) data sources with text-based (unstructured) information in a common environment, you gain a more accurate, complete view of your data. Analysis performed using both types of data produces descriptive and predictive models that enable you to spot opportunities and accurately recognize trends, leading to fact-based, prioritized actions.

For whom is SAS® Text Miner intended?

SAS Text Miner is designed primarily for business analysts, modelers and data scientists who must review large volumes of text to extract common topics, ideas and trends. It is applicable across all industries – and for organizations with voluminous collections, it includes high-performance procedures for text mining, to speed the time to results.

Organizations collect huge amounts of text-based information daily, in many languages and dialects. Customer feedback, emails, Web documents, blogs, Twitter feeds, warranty claims, surveys, research studies, client notes, competitive intelligence … the list goes on. No one has time to read it all, much less classify the content into common themes, or make sense of the essential information.

To observe trends, spot new topics, issue alerts about potential problems and flag new business indicators, you must be able to analyze all your data before acting on it. But conversational language is ambiguous, and key messages buried in text data are not easy to discern or process. Most organizations are unable to combine text content with structured data in decision-making contexts.

With SAS Text Miner you can analyze legacy data stored from your system records – and dynamically reach outside to retrieve pertinent, fresh content from the Web. Interactively explore and automatically identify topics from data-driven categories, and find explicit relationships and associations between terms. Use an all-in-one, drag-and-drop interface to mine text data with structured data, then apply derived analytic insights directly into existing scoring systems. Use the solution to match résumés with open positions, predict patient treatment outcomes with different medication regimens, identify factors that motivate buyers to act – and much more.

Key Benefits

- **Reduce decision time through automated processes.** By implementing intelligent algorithms and natural language processing techniques, time-consuming manual activities – such as cluster identification or topic building – can be generated automatically and executed consistently and efficiently.
- **Enhance the discovery process with subject-matter expertise.** Through a unique, data-driven method of identifying key concepts, you can use interactive GUIs to modify relevance scores and guide machine learning results with human insight. Extend text mining efforts beyond start and stop lists – and what the software automatically discovers – using custom entities and active learning.
- **Present a high-level view of data with transparent drill-down capability.** SAS Text Miner offers a visual presentation of the entire data mining process and allows users to drill to relevant details, illustrating and exploring the term connections. An interactive interface lets you investigate derived topics and fine-tune models.
- **Recognize trends and spot business opportunities.** SAS Text Miner structures text into numeric representations that can summarize the collection and become insightful inputs to a full range of predictive and data mining modeling techniques. In turn, you can better understand customer, service and product needs – and predict opportunities for timely exploitation.

Solution Overview

SAS Text Miner provides a rich suite of linguistic and analytical modeling tools specifically developed for discovering and extracting knowledge from collections of text content. Across electronic text snippets, document archives and Web downloads, patterns can be automatically identified as topics and themes, defining explicit associations between terms and phrases. The software provides supervised, unsupervised and semisupervised methods to discover previously unknown patterns in document collections. It structures data in a numeric representation so that it can be included in advanced analytics, such as predictive analysis, data mining and forecasting.

High-performance procedures are included to take advantage of multicore processing hardware that expedites compute-intensive text processing tasks. This version also includes insightful reports describing the results from the rule generator node, providing clarity to model training and validation results.

Automatically create Boolean rules and interactively train models

The text rule builder node automatically generates an ordered set of rules. You can use the software to classify texts based on the detailed terms table and then generate the resultant Boolean logic of the classifications. Use the rules to categorize documents based on rule matches or export them – including the generated AND, OR and NOT operators – for deployment in SAS Enterprise Content Categorization.

The text rule builder node enables semisupervised, active learning. Through this feature, users can interact dynamically with the algorithm. The software automatically learns categories and topics from the collection. Users can then guide the system to an improved solution – enabling interactive model building. The combination of built-in software guidance and human subject-matter expertise results in highly refined models.

Faster processing

Two procedures, HPTMINE and HPTMSCORE, execute multithreaded text processing on an enabled SAS

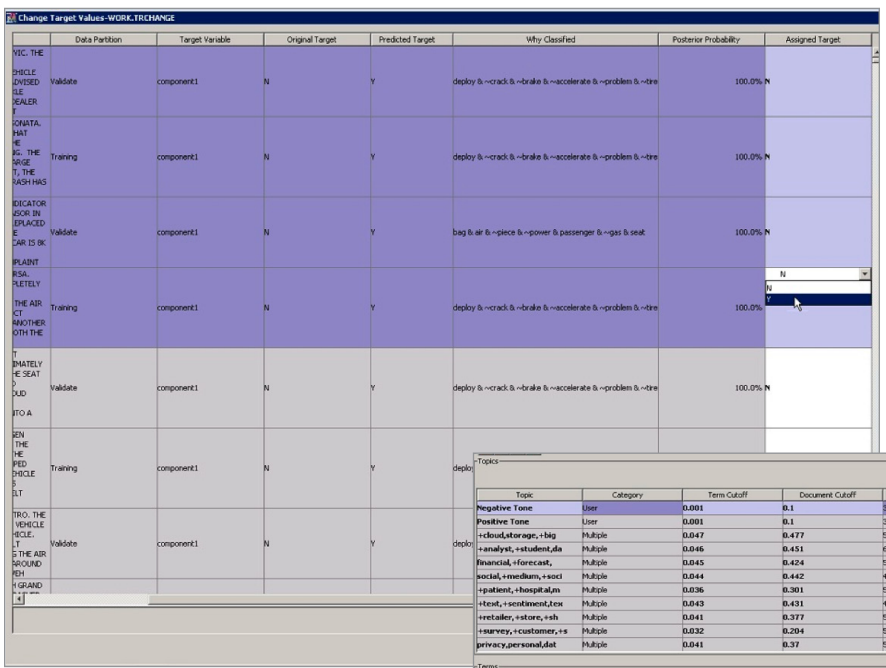

Figure 1: You can generate Boolean rules automatically with the rule generator node and actively teach the model directly by overriding the software-derived results.

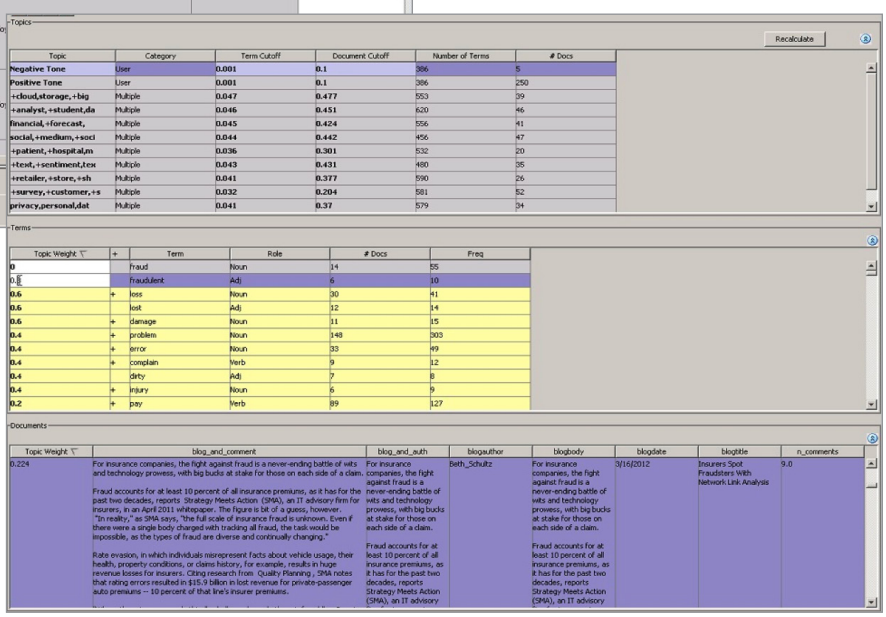

Figure 2: Examine terms driving topic membership in an interactive GUI and, if similar, merge and reassign terms and/or topics to produce desired results.

server. Specific to nondistributed environments, they take advantage of multiple core processors – leading to compute-intensive processing gains in many cases.

Integrated document filtering

Sophisticated dimension reduction techniques enable advanced filtering through weighting, integrated spell checking and transformation of qualitative data into compact formats. Such techniques can structure the unstructured data so that text-based insights are easily included in both predictive and descriptive structured data mining efforts.

Interactive interface for text import defines terms for Web data or text files

With SAS Text Miner, you can use the text import node to create data sets dynamically from files contained in a directory or from the Web. The software reads text stored in a wide variety of document formats, accepting potentially proprietary formats such as Microsoft Word and PDF inputs. The text import node converts the data and also filters or extracts the text from the files – and references the data to a SAS data set. If a URL is specified, this node will also crawl associated websites and retrieve the files, bringing them to the common directory before filtering. This generates a data set that can be used by the text parsing node for the next stage of your analysis. In addition to filtering, the text import node can also identify each document's language and transcode the document to the session encoding format.

More control of table import

Aspects of SAS Text Miner have been further enhanced for node performance and results. Importing table information in dialog boxes is better controlled with Add Table and Replace Table options.

Key Features

Automatic Boolean rule generation makes it easy to classify content

- Lets you describe and predict a target variable based on the detailed terms. Resulting rules can be used to categorize documents based on rule matches.
- Allows rules to also be exported as Boolean rules – and used as a starter rule set for SAS Enterprise Content Categorization.
- Includes output to compare rules between training and validation data.
- Enables active learning by:
  - Providing automated, machine-generated suggestions of categories and topics that can be recharacterized by the user.
  - Modifying the target assigned to the rules, and when rules are regenerated based on these user-defined modifications, the model is updated.

User-friendly, flexible interface

- Merge topics together into one user topic for simplifying similar results.
- Use topic displays to show document terms/all terms, highlighting why a document was assigned to a particular topic.
- Use view mode to illustrate just the terms in a single document or within a topic, or to sort text documents.
- Obtain document-level sentiment insights with an AFFIN sentiment list available as a sample data set with more than 2,000 terms and preassigned polarity weights.
- Modify, save and share process-flow diagrams of text mining analysis.
- Add tables (from previous efforts) to nodes, to extend usability of prior knowledge.
- Extend text nodes further by customizing algorithms or declaring new user-written business rules for predictive modeling, clustering, visualization and reporting – deployable as SAS score code.
- Conforms to accessibility standards for the Windows platform. Accessibility features relate to standards for electronic information technology that were adopted by the US government under Section 508 of the US Rehabilitation Act of 1973.

Integrated document filtering

- Employ sophisticated dimension reduction techniques that enable advanced filtering through weighting, integrated spell checking and transformation of qualitative data into compact formats.
- Create synonym data sets and import previously defined synonyms into the text filter node to improve reusability of existing assets.

Visual analysis of results

- Use the concept link diagram to analyze results visually and to effectively explore the relationships between terms.
- Use interactive diagrams to communicate results to key stakeholders:
  - Employ diagrams that cluster results, derive topic assessments and link associations among terms.
- Use success graph and document rules table to explore generated Boolean rules.

Take advantage of compute power with high-performance processing

- Address compute-intensive text processing tasks, reducing time to results.
- Transform text to structured representations of the collection with SVDs.
- Score large data faster.

Choose predefined entities, define your own, or create custom entities for fact and event extraction

- Define your own multiword terms (phrases such as "drag and drop").
- Choose from one of 18 prespecified entities definitions for address, company, date, phone number, SSN, time and others to ensure extraction from input content.
- Create your own custom entities to be extracted from text inputs, including a list of predefined entities (such as defined districts or product codes) using the SAS Concept Creation for SAS Text Miner add-on.

High-performance text mining

High-performance text mining is enabled to run in a multithreaded mode on a single server using programmable procedures or a node in the interactive GUI. Designed for big text data, these capabilities will grow with your collections, extending to distributed hardware environments (licensed separately). High performance lets you develop models that examine millions and tens of millions of documents, and run the models in minutes or seconds. For more information, visit sas.com/hptextmining.

SAS® Text Miner System Requirements

To learn more about SAS Text Miner system requirements, download white papers, view screenshots and see other related material, please visit sas.com/textminer.

Key Features (continued)

Interactive interface for importing text from the Web or internal file systems

- Lets you dynamically create data sets from files contained in a directory or crawled from the Web.
- Gives access to numerous forms of textual data, including PDFs, Microsoft Word, extended ASCII text, HTML, Microsoft Office formats, spreadsheets, presentations, email and database formats.
- Extracts, transforms and loads textual data into a SAS data set for mining.
- Accepts even potentially proprietary formats, converts the formats, and filters or extracts the text from the files, placing a copy in a plain file and referencing the data to SAS.
- Identifies each document's language and transcodes it to the session encoding format.

Natively supports multiple languages

- Supports Arabic, Chinese, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hebrew, Hungarian, Indonesian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Spanish, Swedish, Thai, Turkish, and Vietnamese. Includes dialects for Simplified and Traditional Chinese, Parisian and Canadian French, Old and New World German, Nynorsk and Bokmål Norwegian, Portuguese for Portugal and Brazil, and Spanish for South America and Spain.

SAS Institute Inc. World Headquarters +1 919 677 8000
To contact your local SAS office, please visit: **sas.com/offices**