SAS® Contextual Analysis

Structure your unstructured text to gain more insight into your organization

§sas
THE POWER TO KNOW®

What does SAS® Contextual Analysis do?

With SAS Contextual Analysis, you can quickly derive insight from your text data by categorizing documents, extracting facts and understanding document sentiment – all from a single interface. You can also define rules without creating a training corpus.

Why is SAS® Contextual Analysis important?

SAS Contextual Analysis makes it easy to identify emerging issues and upcoming trends from your text data – information you can use to better understand your business and your customers. It can even help you stay in compliance with industry regulations.

For whom is SAS® Contextual Analysis designed?

Organizations can easily get buried in too many documents and don't have the time or resources to manually review them. Because SAS Contextual Analysis is easy to get up and running – and it's customizable – it's ideal for both IT and business users.

Documents, customer communications and other types of unstructured data can quickly pile up, becoming an insurmountable mass of material that's impossible to manually organize or review. But the right technology can turn that pile of information into a gold mine.

SAS Contextual Analysis is an advanced linguistic solution that helps you derive key business insight from unstructured text. It uses a combination of machine learning and subject-matter expertise to uncover trends and themes in your data. These insights would've otherwise stayed hidden, but can give you an in-depth understanding of your business.

And while SAS Contextual Analysis is an advanced technology, it's not complicated to use. You don't have to create a training corpus for taxonomies – the software discovers them for you, automatically identifying topics that serve as a starting point. And it's flexible. You can refine the machine-generated category rules to make them even more specific, delivering the exact insight you need for even more business value.

Benefits

- **Easily manage your analysis through a single, integrated system**. With SAS Contextual Analysis, you can get up and running quickly and customize rules when necessary – all from a single, integrated environment. The discovery work is performed automatically right from the start; the software also allows you to set your own definitions using natural language processing techniques.

- **Gain more control with a guided, hybrid approach**. SAS Contextual Analysis enables you to customize your text analysis through a hybrid approach by adding your subject-matter expertise to the automated machine-learning insight. The technology syntactically identifies common themes, category rules and

document sentiment based on the data. At any time you can review and modify the machine-learned results to meet your specific needs. You have more flexibility, power and control so you get faster results.

- **Prepare your data for even broader reach**. Not only does SAS Contextual Analysis help you gain value from unstructured data, it prepares your data for other technologies so you can visually interact with the results. Implement visual analytics or create new metadata to enrich other applications.

Overview

Unstructured data can be unwieldy, but it's also rife with potential value when you have the right technology. Neglecting to examine your unstructured information could lead to missing important business insights, such as customer concerns, emerging trends or even early signs of potential problems. And to stay competitive, you need this type of intelligence.

But not all organizations have the time or resources to comb through endless documents. Or if they did, the work would still be susceptible to human error. Today's data-rich work environments demand technology that's not only powerful, but also intuitive, capable and customizable. And that's what you'll get with SAS Contextual Analysis, a technology built to give you important insights, which lead to a competitive edge.

Single system for guided text model development and deployment

Having just one environment that leads you through the process to develop, refine and deploy linguistic models helps simplify your workload and document categorization. You'll gain more control over the system's parameters because they're easily manageable from a single interface.

Hybrid approach to categorizing documents

Machine learning syntactically discovers common themes, category rules and document sentiment, while your subject-matter expertise enables you to customize your analysis. With both, you can uncover even more contextual meaning – and get better results faster.

Natural language processing

Manually reviewing documents is time-consuming and can leave you open to errors. Data-driven natural language processing eliminates the need for tedious manual definitions. Once you identify the input data source, the solution automatically identifies term frequency counts, parts of speech, synonyms and stems.

Discovery of topics for initial taxonomy development

Pinpointing central themes across multiple documents is a difficult, if not impossible, manual task. SAS Contextual Analysis does this automatically, using machine learning and statistical methods to identify the core themes in the collection.

Generation of configurable rules and improved linguistic context

SAS Contextual Analysis simplifies the task of Boolean rule definition and taxonomy development by automatically creating categorization rules using a patent-pending algorithm. Once the rules are established, they can be semantically enhanced and refined with defined concepts.

Direct integration with SAS®

Discovering key business insights through text analysis is valuable on its own, but it's also a step toward deeper analysis. SAS Contextual Analysis provides rich text-based results that prepare your data for further exploration, such as with SAS Visual Analytics.
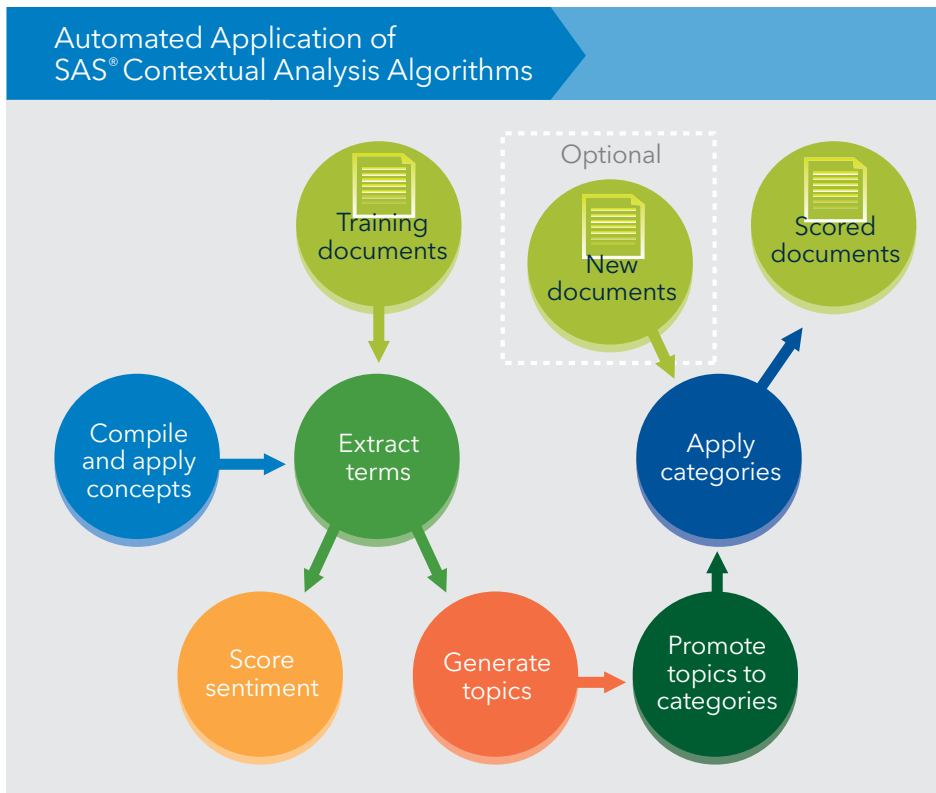

Figure 1: Machine-generated results are open to the analyst for further refinement.

Key Features

Single system for guided text model development and deployment

- Single, point-and-click GUI interface for natural language processing and:
  - Terms extraction.
  - Topic discovery.
  - Category definition.
  - Concept specification.
  - Document-level sentiment identification.
- Guided text model development for the investigation, refinement and deployment of text models.
- Ability to import existing SAS Content Categorization taxonomies for project initiation.
- Ability to create, modify and enable (or disable) custom concepts and test linguistic rule definitions with validation checks within the same interactive interface.
- Prebuilt project creation wizard for initial project definition.
- In-line help to guide analysts through the text model development process.
- Properties panel detailing the status of each step of text model creation, processing status and message dialogue to help diagnose model development issues (such as an insufficient number of documents in the collection to generate topics).
- Project management options that permit multiple text models to be simultaneously developed and run, with associated descriptions indicating each model's development status.
- Centralized metadata management for all project properties.

Hybrid approach to categorizing documents

- Automated natural language processing for term identification (modifiable with start and stop lists), synonym detection, spell checking and stemming.
- Drag-and-drop customization to easily change kept or dropped parent terms between windows.

- Topic discovery based on SAS Text Miner for machine-generated taxonomy initialization.
- Option to use a data source later – to build taxonomies first and then apply to a corpus of documents.
- Ability to visually modify term inclusion in topics and merge similar topics to refine generated results.
- Ability to promote topics to categories with rule generation for initial data-driven category rule definition.
- Ability to edit system-derived categories with an extensive array of prebuilt linguistic operators to further refine rules.
- Ability to use 18 predefined, configurable concepts for common entity types such as location and phone number, with the flexibility to define custom concepts and facts to increase the context sensitivity.
- Ability to enable or disable concepts for specific evaluations.
- Validation of rule syntax for some or all defined concept rules.
- Automated, data-driven, document-level sentiment for a document generated as part of topic discovery or as selected for a category.

Natural language processing

- Natural language processing provided by the software includes automated parsing, tokenization, part-of-speech tagging, synonym detection, spell checking and stemming.
- Output includes lists of terms to drop/keep and term frequency counts.
- Drag and drop between "keep" and "drop" terms windows (parent terms only).
- Additional text processing includes the ability to apply start and stop lists (for terms to include/exclude from processing). Includes a preliminary stop list for common terms typically not desired in analysis (like articles and short conjunctions). Analysts can

edit the predefined stop list (English only at this time).
- User-defined start/stop lists can be used by the software.
- Custom-defined categories can be included by defining the corresponding column in the input SAS data.
- Custom concepts can also be added in the custom concepts panel, written with predefined operator syntax.
- A text analyst has the option to develop a taxonomy without input documents, and later run the taxonomy against the corpus.

Automatic discovery of topics for initial taxonomy development

- Automated machine discovery identifies the core themes in the input document collection with associated relevance scores.
- Term relationships within topics can be interrogated and explored with term clouds (with configurable thresholds), interactive term maps and by drilling into topics to evaluate relevancy and refine discovered topics.
- Topic-level sentiment is visually depicted for the documents associated with that theme.
- Control the number of topics generated by splitting topics (splitting one topic into two, repeatedly if necessary), and merging similar topics into one topic.
- Desired topics are promoted to categories for classification rule definitions and are further enhanced with concept definitions.

Generation of configurable categorization rules

- Automated initial category rule definition based on user-refined generated topics.
- Easy-to-understand Boolean rule definitions create the categorization model (i.e., taxonomy).

Key Features (continued)

- System-generated rule robustness is detailed in true positive, false positive and false negative visual diagnostics.
- Visual depiction of category sentiment is produced in the document view.
- Rules can be edited, enhanced, removed or defined from scratch as custom categories.
- An extensive list of prebuilt rule operators is available to the analyst for detailed rule-model specification.
- Analysts can include prebuilt categorization hierarchies (i.e., taxonomies) from a predefined field in an input data set or from an existing SAS Content Categorization Studio project.
- Rules can be tested on an input data set prior to deployment.

Improved context with concepts

- Concepts provide contextual specificity to categories to refine models for more exact meaning extraction.
- Predefined concepts are available – no rule writing is required. These address common entity definitions for date, location, time, etc.
- Pre-existing concepts can be imported from existing SAS Content Categorization Studio projects.
- Custom concepts can be written using a suite of predefined operators.
- Both predefined and pre-existing concept rules can be edited and refined as desired.
- Concept rules are validated at the click of a button with diagnostic error messages detailing any identified syntax issues.
- Canonical or resolved form for concepts can be returned during scoring.

Natively supports multiple languages

- Supports processing of documents in Chinese (simplified and traditional) , Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish and Turkish.
- Provides native language interface for English, French, Dutch, Chinese and Korean.

Scalable text analytics solution

- Scalable document category and sentiment discovery using distributed processing across multiple cores of a single server.
- Scalable scoring of new documents with automatically created score code for distributed processing.
- Use SAS In-Database Code Accelerator for Hadoop to perform scoring in Hadoop, eliminating the need to move data.
- Faster decisions with data compression at project run time and scoring.

Direct integration with SAS®

- Integration with the SAS Metadata Server (now directly inputs SAS data sets into text analysis).
- For interactive exploration of results, there is direct integration with the SAS Visual Analytics Hub.
- Model scoring code download feature applies contextual models in any other SAS application. Includes viewing and downloading of sentiment and concept score code, as well as category score code for your use in scoring other documents.

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

§sas

THE POWER TO KNOW®