



SAS® Enterprise Content Categorization 5.2

Automated content categorization drives more efficient search and relevant information retrieval

What does SAS® Enterprise Content Categorization do?

SAS Enterprise Content Categorization applies natural language processing (NLP) and advanced linguistic techniques to identify key topics and phrases in electronic text so you can automatically categorize large volumes of multilingual content that is acquired, generated or exists in a repository. It correctly parses, analyzes and extracts content for entities, facts and events to create metadata tags that index documents in a collaborative taxonomy management environment. Based on sophisticated linguistic rules, it can consistently organize, index and trigger dependent informational activities in real time.

Why is SAS® Enterprise Content Categorization important?

The software drives faster, more efficient information organization, access and retrieval, reducing the overhead typically associated with content organization such as manual tagging and indexing. It also improves collaborative knowledge development, retention and sharing.

For whom is SAS® Enterprise Content Categorization intended?

It is designed for organizations that want to improve their information retrieval activities, boost productivity, usability and reuse of document collections, and add document-based intelligence to existing search and content management systems.

Unstructured and semistructured content within organizations is growing at an unprecedented rate – from applications, records and business processes, external documents, scanned images, XML components, Web pages, blogs and forums. Most organizations store and manage this information in silos based on data type or source, tagging the content manually or after the fact using content sampling or content creators.

Common methods for search and information processing are plagued by redundancy, inaccuracy, wasted efforts and missed opportunities. And for newer types of content from sources like blogs and wikis, the old rules no longer apply. Unfortunately, the lack of well-defined taxonomies and indexes based on the materials themselves makes relevant content difficult to locate. Left unmanaged, content loses its relevancy and timeliness.

In the era of “big data”, search and information processing can no longer effectively function as a siloed activity. Managing enterprise content effectively and efficiently as a strategic asset requires a common, underlying organizational structure.

With SAS Enterprise Content Categorization, you can define a hierarchical taxonomy where related topics are readily identified. Then, people across the organization can quickly find the relevant content they need, when they need it, at the level of granularity required.

Key Benefits

- **Find the information you need, when you need it.** With SAS Enterprise Content Categorization, you can find the information you need regardless of whether it's been used before or whether you know its exact location. The flexible, intuitive software provides multiple ways to retrieve content, as it accurately identifies metadata from the content itself and delivers only the most meaningful material related to your inquiry.
- **Improve efficiency and purge content chaos.** The software reduces the overhead of content categorization processes by applying sophisticated linguistic rules to identify and extract terms, then automatically applying the defined intelligence to the content. Use it with large repositories to determine which documents are similar, contain only small variations or have been substantially modified.
- **Extend existing investments.** The software transforms corporate textual data into a reusable asset, extending the value of existing investments by building upon existing indexes and integrating with content management systems like Documentum and Microsoft SharePoint, and with search technologies like Endeca and FAST ESP.



Solution Overview

SAS Enterprise Content Categorization is an easy-to-use application that lets organizations automatically create better relevance, enhanced usability and quality content from search and retrieval activities. By applying natural language processing and advanced linguistic techniques, it correctly parses, analyzes and extracts content entities, facts and events, creating the necessary metadata to accurately represent a document's keywords and phrases.

The metadata is applied to incoming documents before other systems index the content. This added layer of intelligence, which can include definitions for identical or duplicate matches – and which is defined by the documents themselves – provides the information that search and content management systems must have in order to pinpoint only the relevant information.

SAS Enterprise Content Categorization drives faster, more efficient information organization and document retrieval – and drastically reduces the time spent searching for documents and the overhead associated with manual tagging, retrospective indexing, and search and content management system replacement.

Entity, fact and event extraction

SAS Enterprise Content Categorization automatically detects and extracts entities, facts and events from text. For example, entities and concepts such as people and company names, publicly traded businesses, titles and positions, and geographical locations can be automatically extracted based on definitions that you develop. Concepts can be simple (Jim Goodnight); relational (Jim Goodnight, CEO, SAS); or, sophisticated (for example, with prebuilt co-reference operators to resolve pronouns or with case-insensitive operators for greater rule-matching precision). The rules then mimic the way the human mind processes information.

Contextual extraction

SAS Enterprise Content Categorization can locate and return related pieces of data that form a fact or event, and there is no requirement for precompiled dictionaries to identify unknown information. Only the facts and events with the highest relevance, or those with the longest matches, are returned. Matching criteria can be customized using contextual markers, parts of speech, identifiers for uppercase or lowercase words, and Boolean operators.

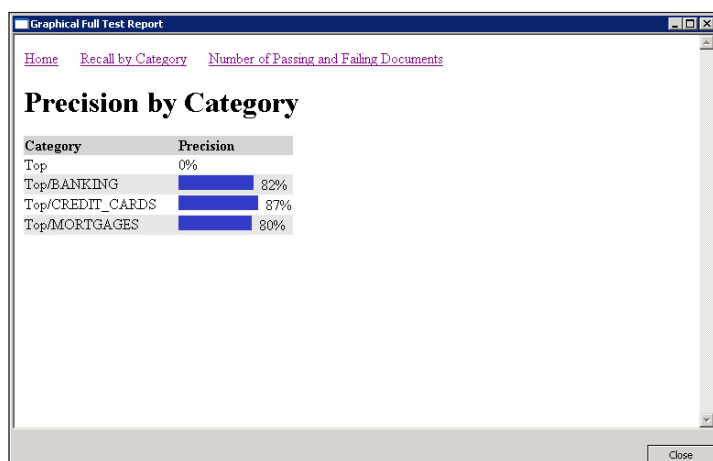
Category classification

The software automatically applies advanced linguistic technologies to identify and classify key information your business requires. You can easily define category rules to classify documents that match this rule, while excluding those that don't. The software can automatically categorize new documents based on examples of documents evaluated. Using extensive linguistic dictionaries covering a host of languages, it automatically extracts the relevant linguistic features from documents and associates them with their given categories. You can control the number of terms that an automatic category rule generates for each category; identify relationships between categories and concepts; and use syntax checking and duplicate rule elimination for classifier concepts.

Collaborative taxonomy management

Using an intuitive GUI, you can create and manage taxonomies in a secured, controlled and audited environment. Develop a hierarchical taxonomy where related topics are grouped together, or create a flat taxonomy where there is no relationship between nodes. Automatically generate categories from Wikipedia if there are no pre-existing categories. The prebuilt graphical reports can help identify taxonomy rule refinement needs with statistics for category matches, such as the precision, recall and numbers of passing and failing documents.

Multiple taxonomists and developers, working individually or in teams, can access projects based on set permission levels. This enables teams of specialists to collaboratively define and refine taxonomy development. Once developed, this collective work is implemented to process vast quantities of text in real time, increasing the efficiency and accuracy of information retrieval and deployment across the enterprise.



Explicit reports examining precision and recall help guide the model development process.

Support for more than 27 languages

With SAS Enterprise Content Categorization, you can build projects using one or more of the 27-plus Asian, Eastern and Western European, and Middle Eastern languages. Advanced linguistic technologies let you leverage part-of-speech recognition, tagging and case sensitivity, and Wikipedia category generation in English and most European and Asian languages. SAS Enterprise Content Categorization ships with English and the native language, if other than English. Additional languages may be licensed as add-ons.

Add-on modules available

A variety of add-on modules are available for SAS Enterprise Content Categorization to suit your organization's unique needs. Choose from a variety of prebuilt industry-specific taxonomy starter kits to help jump-start your document classification initiatives. You can also use add-ons for search and indexing, Web crawling, text summarization, document duplication detection and more.

Key Features

Entity, fact and event extraction

- Locate and return related pieces of data that form a fact or an event based on their context in real time (e.g., a person in relation to a company, the merger of two companies, etc.).
- Identify unknown information without a requirement for precompiled dictionaries.
- Customize matching criteria using options such as contextual markers, parts of speech tags and Boolean operators.
- Use prebuilt co-reference operators in an intuitive GUI to help resolve pronouns more easily, addressing syntactical functions such as:
 - Linking to a matched string with its canonical form.
 - Co-referencing classifier definitions.
 - Restricting forward and preceding co-reference matches.
 - Assigning a new concept name for a match on a specific term.
- XML fields can be limited for specified matches.
- Case-insensitive operators allow for greater rule-matching precision, including:
 - Prebuilt stemming can match on all word forms or only on noun or verb forms.
 - Sentence and paragraph operator defines the number of word, noun or verb matches within a paragraph.
- Write more than one rule to extract all of the possible permutations of the data you seek.
- Disambiguate facts and events by excluding certain matches.

Category classification

- Control the number of terms that an automatic category rule generates for each category.
- Adjust relevance schemes (frequency-based or zone-based) for concepts used in category rules.
- Use linguistic rules, which are unique identifying terms, or add Boolean operators to your unique terms for added specificity in determining category membership.
- Develop a list of unique identifying terms for each category rule.
- Weight selective terms or the categories themselves, creating more exclusive membership requirements.
- Use the testing facilities to validate application of rules and definitions to batch, entire or content components.
- Includes enhanced XML handling along with better syntax checking and duplicate rule elimination for classifier concepts.
- Automatically apply the rules and definitions to incoming texts using the client APIs in C, C++, C#, .NET, Java, Perl or Python.

Collaborative and flexible taxonomy management

- Define user-permission levels, including read, write, category rules and concept definitions.
- Use an unlimited number of taxonomy nodes, and apply categories and concepts generated to large volumes of input documents.
- Develop a hierarchical taxonomy where related topics are grouped together, or create a flat taxonomy where there is no relationship between any of the nodes in the taxonomy tree.
- Automatically generate categories from Wikipedia data to jump-start your taxonomy development.
- Benefit from improved PDF conversion.

Out-of-the-box integration

- Configure APIs to automatically tag content from Microsoft Office SharePoint, Endeca, FAST ESP and Documentum.
- Documents are tagged prior to indexing to speed up processing time.
- Extends capabilities of existing search tools and content management systems, increasing the relevance of retrieved materials.

Continued on reverse

SAS® Enterprise Content Categorization 5.2 System Requirements

To learn more about SAS Enterprise Content Categorization system requirements, download white papers, view screenshots and see other related material, please visit www.sas.com/categorization.

Key Features (continued)

Support for more than 27 languages

- Language tools: NLP/advanced linguistic technologies that leverage:
 - Part-of-speech recognition and tagging: Recognizes nouns, verbs, adjectives, etc.
 - Stemming: Locates the various forms of an input noun or verb.
 - Case sensitivity: Specifies uppercase and/or lowercase recognition for concepts.
- Use the following two options with Germanic and Asian languages:
 - Compound recognition and compound decomposition: Break apart the recognized compound words.
 - Segmentation for Asian languages.
- The product ships with English and the native language if other than English. Additional languages are licensed as add-ons.

Add-on modules available

- Multiple, prebuilt industry-specific taxonomy starter kits are available as add-on modules for SAS Enterprise Content Categorization:
 - Provide immediate ROI by classifying industry-specific content to help jump-start document classification initiatives.
 - Include detailed concepts and attribute values with predefined rules, helpful for initiating taxonomy project development.
- Search and indexing add-ons automatically discern query semantics and enable superior drill-down and investigative capabilities by categorizing multifaceted information:
 - Include an easy-to-use interface for search and document processing.
 - Can extract entities, concepts and facts within a unified document processor.
 - Support multiple document schemas with multiple instantiations, and also retain the original URL in any split documents.
- Web crawling add-ons automatically download documents from the Internet by performing several different kinds of downloading operations based on network bandwidth, crawling politeness and information coverage:
 - Have an easy-to-use interface for defining and managing Web and internal file system crawls.
 - Set individual quotas for a specified URL, and define project quotas.
- Other add-on modules include text summarization, document duplication detection, content alerts and additional languages.