



SAS® Enterprise Content Categorization Add-On Modules

Customize your solution for information organization, access and findability

What do SAS® Enterprise Content Categorization add-on modules do?

SAS Enterprise Content Categorization add-on modules, powered by Teragram technology, apply natural language processing (NLP) and advanced linguistic techniques to automate text-processing operations so that additional efficiency gains are rapidly realized by organizations where they are needed most. These unique technologies enable richer processing at the level of words, linguistic relations and word meanings – solving the issues associated with excessive electronic information materials and their exponential growth rate.

Why are SAS® Enterprise Content Categorization add-ons important?

Augmenting content categorization, these add-on capabilities are scalable to huge quantities of information, providing the same processing speed for any amount of data, while encoding that data in a form which is highly compressed. The add-ons enable organizations to customize their content management solution to improve document-centric business process operations, with the ability to add more capabilities as their needs evolve.

For whom are SAS® Enterprise Content Categorization add-ons designed?

They are designed for organizations that want to augment their content categorization solutions to drive faster, efficient information organization and access, and gain real-time awareness of unknown facts and events based on their context.

Unstructured and semistructured content is growing within organizations at an unprecedented rate. All of this information should be maintained, integrated, stored, accessed and distributed efficiently and effectively. Because for content to be useful, people must be able to find the information they need.

However, with today's existing models and processes, it is a challenge to maximize the use of such information. Across most organizations, information is stored and managed in silos based on data type. Content is difficult to locate and use. Organizing and accessing content entails manual updating and after-the-fact indexing. Another common approach is to provide access to a shared content repository with loosely defined naming conventions. This, however, results in small subsets of information being produced that are difficult to navigate and time-consuming to retrieve.

In addition, applying ineffective search techniques to content that lacks a well-defined structure returns large volumes of irrelevant material to queries, and it does not provide the level of specificity needed to be relevant or get the information in the right format, when needed.

SAS extends your business processes that rely on accurate content categorization with several add-on modules, providing effective search and retrieval activities, meaningful summaries of materials, real-time alerts to new content availability and more. These add-ons ensure that you can customize your SAS content categorization solution to meet specific organizational needs.

Key Benefits

- Enable users to find the information they need quickly.** Effective findability retrieves content in context so users can find the information they need whether they know where it is or not. Add-on capabilities include search and indexing to narrow down retrieved information, a high-performance Web crawler that automatically downloads appropriate documents from the Internet for associating with existing taxonomies, a text summarization module that conveys relevant messages within a document in condensed form and a scalable real-time alert notification service that delivers documents to millions of users at individually specified times.
- Drive faster, more efficient information access.** The SAS Search and Indexing add-on module automatically discerns query semantics and enables superior drill-down capabilities to enhance users' investigative techniques. Narrowing down the information to only relevant sources, this add-on applies stemming and automatic spelling correction, enabling richer preprocessing. By applying these linguistic technologies at the preprocessing level, searches become more accurate and relevant.
- Purge content chaos that spans multiple enterprise repositories.** Enterprise repositories often contain many documents that have been duplicated or edited and republished. Extending the categorization of similar content, the SAS Document Duplication Detection add-on helps organizations minimize their content stores, maintaining only those materials that meet the threshold standards of similarity.



SAS® Document Duplication Detection

With SAS Document Duplication Detection, you can overcome a very common problem in document management that stems from the high number of documents that are duplicated or slightly modified and then republished. Within an enterprise repository, the same document may be published in Microsoft Word, plain text format, as HTML or a PDF. It may have been slightly edited over time. In other situations, a generic template may have been used to generate a large number of similar documents. However, this history often is lost and it can be difficult to figure out which documents are actually “near copies” of each other. With this add-on module, SAS Document Duplication Detection is designed to recognize, within a large set, which documents are similar up to a threshold. A configurable similarity threshold allows the system to detect versions of documents that have

been substantially re-edited or to focus on documents that are only small variations of others. In addition, it can return the documents that best meet your set of criteria.

SAS® Text Summarization

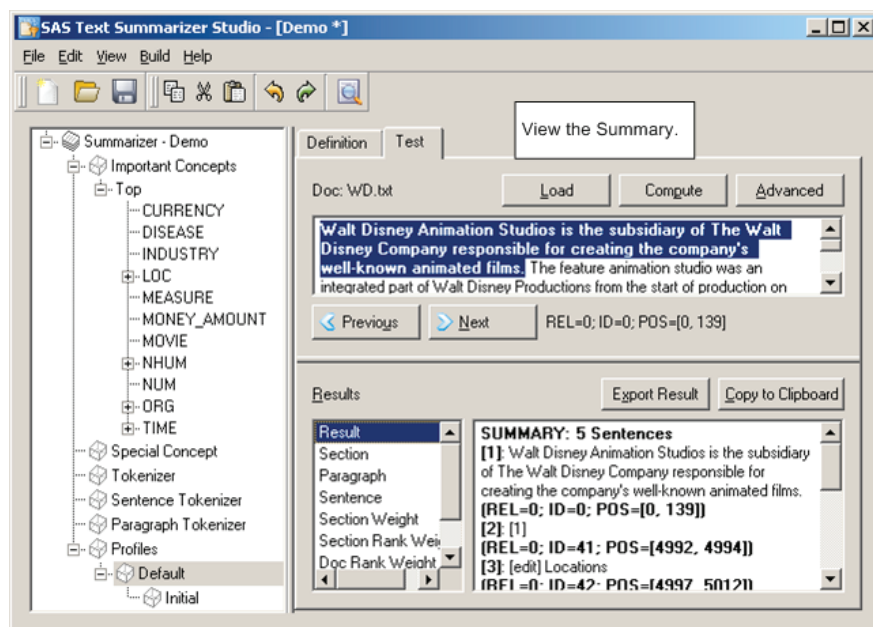
SAS Text Summarization distills documents and creates concise summaries, allowing users to access information faster and more efficiently. It also improves the efficiency at which information is conveyed by allowing the readers to focus on the important concepts by condensing documents in real time as deployed into your daily workflow. Advanced linguistic and parsing technologies are applied to compose editorial-quality abstracts and short summaries of variables based on an organization’s requirements. Documents can be condensed in real time and deployed into existing workflows. They also can be distributed to devices with limited screen space.

SAS® Search and Indexing

SAS Search and Indexing solves the traditional problem where queries return endless, unstructured and often unrelated documents. Instead of looking for information blindly in all sources, you are now able to narrow down the information from relevant predefined categories. Using linguistic technologies that include stemming and automatic spelling correction, words, word meanings and linguistic relations are applied at the preprocessing level. The automatic discernment of query semantics and superior drill-down capabilities enable you to easily investigate categories of multifaceted information. And the ability to process huge quantities of information in real time ensures that queries return useful, relevant responses.

SAS® Web Crawler

SAS Web Crawler is designed for organizations that collect information from the Internet. It provides a powerful Web-crawling application that performs several kinds of downloading operations depending on network bandwidth, specified crawling politeness and information coverage. Starting at a user-specified URL, the crawler follows the hyperlinks on the Web, while repeatedly sending HTTP requests to simultaneously obtain corresponding HTML content and any URLs that exist within that content. High-performance crawling uses a multithreaded mode, and distributed crawling enables you to use many computers to optimize the crawling by spreading out required resources to multiple computers and networks. In addition, there are many enhanced management and configuration features. The returned pages are then analyzed as part of processing – extracting entities or categories for domain-related selections so that only information defined as relevant is saved.



The graphical user interface in SAS Text Summarizer enables you to easily test summary definitions prior to production to ensure relevance thresholds have been defined optimally.

SAS® Content Categorization Information Workbench

The information workbench module, designed for indexers and editors, provides a workflow tool that combines human editorial review with automatic categorization, entity extraction and metadata tagging. This increases productivity and speeds the return on investment, while eliminating the risks of full automation.

SAS® Content Alerts

The real-time alert module provides accurate notification services through a variety of alert media, including e-mails, instant messaging, etc. The alert capability is highly scalable to millions of users with a constant flow of documents.

SAS® Industry Taxonomy Rules

These modules provide prebuilt taxonomies for virtually every industry. Because they already come with an extensive suite of terms, relationships, key entities, attributes and values – as well as some defined rules – the taxonomy kits afford an immediate baseline for analyzing content, helping to identify top priorities for ongoing categorization efforts.

SAS® Text Data Language Pack

SAS Enterprise Content Categorization ships with English and the native language if other than English. Additional languages may be licensed as add-ons.

Key Features for SAS® Enterprise Content Categorization Add-On Modules

SAS® Document Duplication Detection

- Designed to recognize, within a large set, which documents are similar up to a threshold of similarity.
- Configurable similarity threshold allows the system to detect versions of documents that have been substantially re-edited or to focus on documents that are only small variations of others.
- Abstract the documents from their actual format and focus on the content of the document.
- Platforms supported (server only): AIX, HP-UX Itanium, HP-UX PA-RISC, Linux for x86 and x64, Microsoft Windows (x86-32), Microsoft Windows on x64, Solaris on SPARC and Solaris on x64.

SAS® Search and Indexing

- Apply linguistic techniques to search queries and documents at the preprocessing level to provide a more accurate and relevant search.
- Use advanced linguistics technologies such as stemming and automatic spelling correction to provide richer processing at the level of words, linguistic relations and word meanings.
- Organize information into an intuitive hierarchical directory, which encapsulates specific categories into more general categories, allowing for greater flexibility.
- Narrow down search within a category, or browse documents in the category of interest.
- Platforms supported (server only): AIX, HP-UX Itanium, HP-UX PA-RISC, Linux for x86 and x64, Microsoft Windows (x86-32), Microsoft Windows on x64, Solaris on SPARC and Solaris on x64.

SAS® Text Summarization

- Documents are summarized automatically for wide distribution of content.
- Natural order of key sentences describe the essence of text so it is meaningful to readers.
- Define the relative importance of special concepts (i.e., anchor words or word strings) to capture subject-matter expertise.
- Leverage existing concepts and concept taxonomies to define single concepts or relationships and form the basis of definitions that are sought in the identification of key sentences, including Classifier concepts (authority lists), Regex concepts (regular expressions) and Grammar concepts (syntactic patterns).
- Documents written in different languages can be summarized while retaining the inherent meaning within the natural language of the source content. Word tokenization is dependent on the language of the materials being summarized.
- Platforms supported (client): Linux for x86, Microsoft Windows (x86-32 and x64).
- Platforms supported (server): AIX, HP-UX Itanium, HP-UX PA-RISC, Linux for x86 and x64, Microsoft Windows (x86-32), Microsoft Windows on x64, Solaris on SPARC and Solaris on x64.

SAS® Web Crawler

- Starting at a user-specified URL, the crawler follows the hyperlinks in the Web while repeatedly sending HTTP requests to simultaneously obtain corresponding HTML content and any URLs existent within that content.
- High-performance crawling: Used in a multiple-threading mode to allow the configuration of the number of threads.
- Distributed crawling: Distributed running mode to optimize crawling. When multiple crawlers are running simultaneously, each crawler will send the correct set of links to the crawler to which they might belong.
- Incremental crawling: Enables continuous downloads.
- Page quality: Crawl the highest quality pages first, when the quantity of object pages is very large. Duplicates of URLs or page contents are automatically removed.
- Polite downloads prevents complaints or access blocking from crawled sites. Specify the minimum access interval for continuous downloads from each site, maximum parallel connections to each site or domain, or the maximum number of times to retry each failed HTTP request.
- JavaScript parsing: URL extraction from JavaScripts where content is often deeply embedded.
- Logon for cookie-supported and password-protected Web sites.

SAS® Enterprise Content Categorization Add-On Modules Technical Requirements

All add-ons must license SAS Enterprise Content Categorization or the single-user version SAS Content Categorization. Because supported platforms vary for each add-on, please check the features list for specific information.

Client environment

- Linux for x86 (x86-32): RHEL 4, SuSE SLES 9
- Microsoft Windows (x86-32 and x64): Windows XP Professional, Windows Server 2003 family, Windows Vista*

Server environment

- AIX: Versions 5.3 and 6.1 (x64) on POWER architectures
- HP-UX Itanium: HP-UX 11iv2 (11.23), 11iv3 (11.31)
- HP-UX PA-RISC: HP-UX 11iv2 (11.23), 11iv3 (11.31)
- Linux for x86 (x86-32): RHEL 4, SuSE SLES 9
- Linux for x64 (EM64T/AMD64): RHEL 4, SuSE SLES 9
- Microsoft Windows (x86-32): Windows XP Professional, Windows Server 2003, Windows Vista*
- Microsoft Windows on x64 (EM64T/AMD64): Windows XP Professional for x64, Windows Server 2003 for x64, Windows Vista* for x64
- Solaris on SPARC: Versions 9 and 10
- Solaris on x64: Version 10

* NOTE: Windows Vista editions that are supported include Enterprise, Business and Ultimate

SAS® Enterprise Content Categorization Add-On Modules (continued)

- Enhanced management and configuration:
 - Entry points: Specify a list of URLs as seeds to start the crawling and define the number of pages to start from each seed.
 - Portal list: Define URLs to download without extracting new URLs.
 - Link-following restrictions: Define link-following rules with regular expressions to restrict the crawling area – e.g., restrict the crawling in a directory, server or domain.
 - Excluded paths: Provide a list of URL paths that will be excluded in the crawling. Any URL that is not an entry point will not be extracted if it contains an excluded pattern.
- Platforms supported (server only): AIX, HP-UX Itanium, HP-UX PA-RISC, Linux for x86 and x64, Microsoft Windows (x86-32), Microsoft Windows on x64, Solaris on SPARC and Solaris on x64.

SAS® Content Categorization Information Workbench

- Workflow tool incorporating automatic abstracting, categorization and entity extraction that is designed for indexers or editors.
- Combines human editorial review with automatic abstracting, categorization and metadata tagging.
- Increases measurable business value and productivity and dramatically speeds the return on investment while eliminating the risks of full automation.
- Provides a feedback loop to the taxonomy tool for editing the taxonomy based on the use of nodes in the taxonomy.
- Platforms supported (client only): Linux for x86, Microsoft Windows (x86-32 and x64).

SAS® Content Alerts

- Specify HTML, text or XML e-mail alerts.
- Use e-mail, SMS or other means of alerts.
- Multiple alerts to the same user can be combined into a single alert.
- All alerts are encoded in an intermediate XML format for delivery processing.
- Users can specify the time when alerts are sent (time of day or as soon as possible).
- Communicate directly through the SMTP protocol to a send mail server. Automatically check for returned e-mails by accessing a POP server.
- Generate preformatted files for use with existing e-mail programs.
- Highly scalable to millions of users with a constant flow of documents.
- Platforms supported (server only): AIX, HP-UX Itanium, HP-UX PA-RISC, Linux for x86 and x64, Microsoft Windows (x86-32), Microsoft Windows on x64, Solaris on SPARC and Solaris on x64.

SAS® Industry Taxonomy Rules

- Provides an extensive suite of terms, entities and their hierarchical relationships to quick-start categorization efforts, sourced from Wand Inc.
- More comprehensive than other prebuilt taxonomies with attributes, attribute values and SAS predefined rules.
- Available to virtually every industry, and can be readily translated into more than 30 languages.
- Updates are included as part of the licensing agreement.
- Platforms supported (client only): Microsoft Windows (x86-32 and x64).

SAS® Text Data Language Pack

- SAS Enterprise Content Categorization ships with English and a native language if not English.
- Asian, Eastern and Western European, and Middle Eastern languages are available.
- Platforms supported (client only): Microsoft Windows (x86-32 and x64).



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 © 2011, SAS Institute Inc. All rights reserved. 104390_S74470.0511