opentext™

WHITE PAPER

OpenText Intelligent Classification helps users gain insight from unstructured content

How to uncover insights and information that optimizes your content

Contents

OpenText™ Intelligent Classification helps users gain insight from unstructured content. It enables enterprises to take control of their knowledge assets to manage and grow their business efficiently. Using thoughtfully selected text analytics techniques, such as metadata federation or crawlers to access data from multiple repositories, this tool can extract from content the meaningful pieces of information and help users connect with the content most relevant to them.

This white paper focuses on how OpenText Intelligent Classification streamlines and speeds up a key task of content analytics, the semantic annotation of content, to make documents more "findable" and usable for uses ranging from indexing and content curation to claim form processing and creating new, value-added content products. It takes on the task of tagging content with semantic metadata, traditionally done manually, and frees up workers from many hours of repetitive labor to exert more judgment in content management.

But what is semantic metadata? How is it found by a machine and how does the machine know what to look for? Once semantic metadata is found, how can it be used? This white paper explores the OpenText approach to deriving this valuable information. It will discuss the methodologies used in OpenText Intelligent Classification and the various approaches to identifying patterns and trends and explore the larger issue of how artificial intelligence (AI) and machine learning can play a role in an organization's business processes.

Introduction: Mining the text for meaning

Unstructured data includes emails, social media posts and business documents—essentially anything that contains textual content (i.e. natural language in written form). This kind of data has historically been harder to manage and extract useful insights from than structured data (i.e. figures and entries that fit in databases), because it is more complex and ambiguous, requiring human understanding of language in order to analyze its meaning.

Now, any information-rich organization can use OpenText Intelligent Classification to analyze unstructured data and other content, highlight its value and enhance comprehension on what is going on within the organization. With machine learning (ML) techniques that let it quickly and flexibly annotate massive amounts of unstructured data with semantic metadata, OpenText Intelligent Classification improves content discovery, lowering operating costs and helping users reach new insights more easily.

The following discussion of how OpenText Intelligent Classification works uses terms that are common in the field of content analytics but may not be familiar to all readers. For example, "text mining" and "text analytics" are sometimes used interchangeably. Although they are both methods of extracting valuable information from written text, they are not quite the same thing. Other definitions are in the box below.

Name	Definition
Text mining	The process of deriving high-quality information and relationships from textual content. The information and relationships derived are often expressed in the form of metadata and relationships between metadata. Text mining seeks to deduce and extract a variety of information from textual content, such as important concepts, named entities (people, companies, geographical locations) and topics. It operates at the level of a single document or other piece of text.
Text analytics	In contrast with text mining, this term applies to a collection of texts, drawing insights and finding patterns that exist both at the individual document level and at the multi-document level.
Machine learning	A branch of AI concerned with the design and development of algorithms that allow computers to improve their performance on specific tasks, learning by examples from real data, such as sensor data or database information. ML algorithms use input data and statistical analysis to predict an output value, based on previous experience or data. ML techniques can be grouped under two types: supervised and unsupervised learning.
Supervised learning	Includes approaches where the machine learning classification models or algorithms are trained on known data prior to being used on target content. This approach is best leveraged when the results of learning are known. Supervised methods allow careful supervision of the learning process by selecting the teaching materials—choosing examples that "teach" or train the ML model. For example, internal emails that are about a given topic even if they do not contain pre-identified keywords. After the first supervised learning cycle, if users feel the algorithms are not identifying the targets accurately enough, they can provide more training targets. They can also complement these evolving classification models with explicit rules created by domain experts. Then they can repeat the cycle until they are satisfied with the results and finally apply the models to real-world data.
Unsupervised learning	Refers to ML approaches where users know what data is given to the algorithm, but not what results to expect.
Clustering and association algorithms	Techniques in unsupervised learning that can detect duplicates or similar/replicated material in a document collection. Clustering and association can be helpful in use cases such as e-discovery or marking redundant documents for deletion in content migration.
Metadata	Data that provides information about other data, such as a file's size, type, author, main topic, names of people mentioned and date it was created or edited. Metadata comes in two basic types: Semantic (or editorial) and system metadata.
Semantic (editorial) metadata	A major subset of metadata that conveys meaning from somewhere within the text, such as identifying names, dates, places and topics. The metadata can then be searched or automatically matched to assure that users connect with content that is most relevant to them.

Name	Definition
System metadata	The other major subset, system metadata covers features outside the meaning of the text, such as file size, type and the person who last edited it.
Named entities	Items in the larger world that are identifiable by proper names: People, places, dates, products, organizations and titles such as CEO. For example, the noun "city" is not a named entity, but "Waterloo" is a named entity.
Personally identifiable information (PII)	Information such as full names, addresses, ID numbers and health conditions that could identify an individual person. PII is a type of named entity. From a practical standpoint, it is an especially important type because laws in many countries, including the US, Canada, the UK and the European Union protect PII from unauthorized sharing. This means organizations need to know where it is in their stores of information.
Token	A series of characters separated by spaces that comprise a word, phrase, sentence, symbol or similar element of meaning. Two or more tokens can be combined into concepts, such as the term "Human Resources."

OpenText Intelligent Classification allows the user to easily call upon complex natural language processing methods and algorithms, combining statistical methods, linguistics insights and machine learning, to yield value from unstructured content. (In the second half of this paper, we will describe the functionalities within each module of OpenText Intelligent Classification and how they contribute to the overall results.) The broader OpenText ecosystem of AI-enhanced analytics further allows the user to combine the resulting metadata with structured data to answer deeper and broader analytic queries.

OpenText Intelligent Classification can help organizations through the automation and semi-automation of a number of semantic metadata:

- **Concepts:** Key terms that are considered the most relevant for a given document
- **Named Entities:** Any term or sequence of terms that can be referred to (personal names, companies, brands, PII, etc.)
- **Classifications/Topics:** The different topics (such as "finance") represented in a document; also classifications (such as "annual financial reports") to which the document belongs
- **Sentiment and Emotion:** The overall tone of the document and each sentence within (i.e. positive, negative, neutral), whether it expresses facts or opinions referred to individually as "subjectivity", and the extent to which it expresses specific emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust)
- **Summary:** A summary of a document's content
- **Languages:** The language or languages in the document
- **Named Entity and Relationships:** Any term, or sequence of terms, that can be referred to and related to each other

Creating and using semantic metadata

Assigning semantic metadata throughout a text or group of texts is one of the fundamental techniques OpenText Intelligent Classification uses to derive useful insights and help users connect with the content that's most relevant to them. For easier reading, we will simply say "metadata" here, keeping in mind that non-semantic types of metadata are also relevant from a broader analytics perspective. The metadata can be classified and searched, quickly highlighting aspects of the text that humans could only find by reading carefully. Once metadata is added to a document, useful relationships can be built. For example, a user can create the list of all published documents for a given author and, in turn, retrieve all documents by that author.

Users routinely create metadata manually during the process of authoring, editing and sharing content. A basic example is when an author creates a document and gives it a title. Or a user can manually add key terms associated with a document in an Enterprise Content Management system, perhaps for search engine optimization (SEO) or easier retrieval. However, manually creating metadata is very time-consuming and not always on-target, so organizations seek text analysis tools that can automate the process to make it faster and smarter.

OpenText Intelligent Classification can assign metadata to a document either semi-automatically or completely automatically, depending on which mode the user prefers.

Automated metadata assignment

In completely automated mode, OpenText Intelligent Classification can automatically process textual data, storing the metadata so it can be used as is, without necessary revision.

For instance, it could verify that documents or emails do not contain specific personally identifiable information (PII), such as full legal names and credit card purchase histories.

Combining different semantic metadata is an excellent way to improve search efficiency. For example, a user may only be interested in content that was authored by Mary J. Jones, contains a credit card number and mentions OpenText. Having all this metadata ahead of time reduces the number of documents searched, delivering the information that users need much faster.

Semi-automated metadata assignment

In semi-automatic assignment, OpenText Intelligent Classification can spot and recommend important keywords for each document, which can then be reviewed by the author or a content manager. For example, it could suggest what key phrases constitute variations on entity names, such as New York City, the Big Apple, Manhattan or New York. Then, a reviewer can select what is or is not a relevant organization to extract.

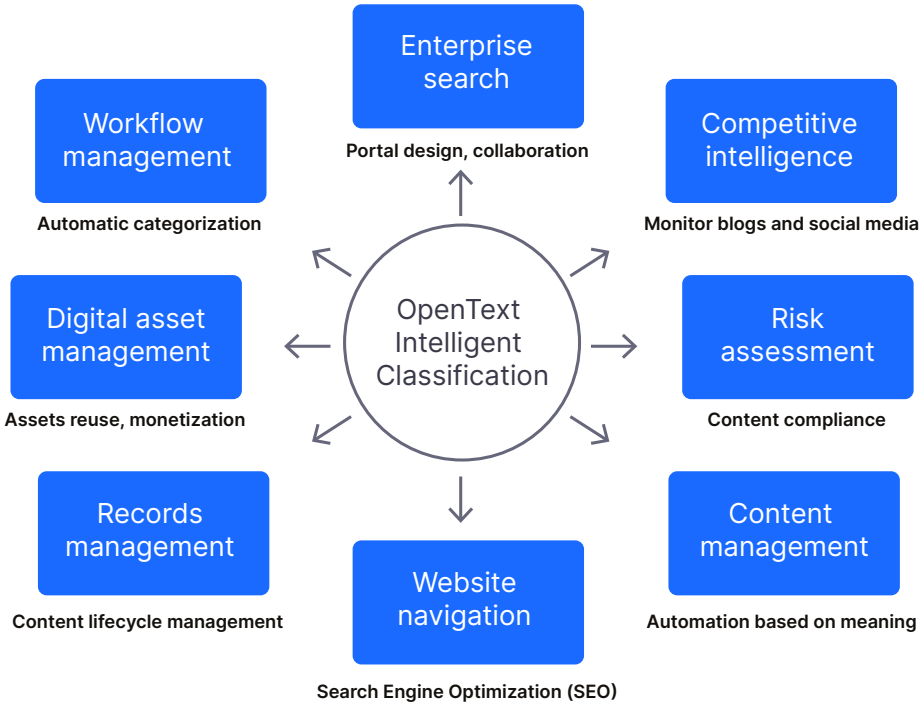
During the installation of OpenText Intelligent Classification, an organization can configure its preferences for fully or semi-automated metadata assignment depending on the use case. Fully automated metadata assignment is better for tasks that require highly reliable retrieval of all responsive documents even if that also brings up some non-responsive ones, such as complying with a consumer's request under the GDPR to hand over all their PII. In information retrieval terms, this is known as high recall.

By contrast, semi-automated assignment is desirable where the use case requires identifying and retrieving the most relevant documents, for example, in replying to a customer support ticket, where the organization wants to make sure the client gets the most helpful answer as soon as possible. This is known as high precision. It also allows the user to tweak and refine the search criteria, giving rounds of feedback that make the OpenText Intelligent Classification instance more accurate.

By offering a flexible choice of either semi-automated or fully automated metadata assignment, OpenText makes it easier to create an effective metadata management strategy. This helps:

- Increase the value of content through the creation of new points of access or re-purposing.
- Create new products and/or services (e.g. for tailored market segments, which couple content with customer usage across different brand silos, etc.).
- Retrieve information more quickly, reliably and cost-effectively, speeding time-to-value.

- Give organizations better insight into the overall number, range and contents of their various files.
- Reduce the amount of redundant, outdated, trivial and/or risky information being retained.
- Improve the customer/user experience.
- Track and protect personal or sensitive information more effectively.
- Drive traffic and business to websites.
- Find and access information from many locations (e.g. search engines, new portals, content aggregators, etc.)



OpenText Intelligent Classification streamlines a wide range of work processes

Methodology

OpenText Intelligent Classification uses a variety of carefully chosen methodologies that complement each other, so that it can identify and extract the most relevant semantic metadata quickly and accurately.

Linguistic patterns

Based on statistical patterns of the frequency and position of low-level textual features such as tokens, the solution looks for linguistic patterns to identify words and phrases (structured sequences of words) that are likely to be concepts and entities.

This is where OpenText Intelligent Classification transforms the tokens and n-grams from “sequences of items” to real lexical units (parts of speech such as nouns, verbs, adjectives, etc.) that are found within their syntactic environment. The solution recognizes lexical units, groups them, filters out functional words, such as conjunctions or prepositions, and finally exposes the most conceptually loaded terms of text, namely the syntactic categories Nouns and Noun Phrases.

This is how OpenText Intelligent Classification can identify the patterns of value that contribute to having the machine understand what the text is about, exactly as people do. For example, when someone reads an article about the last Toronto Raptors-Golden State Warriors game, how do they know it is about basketball? They recognize specific words or words sequences used in the article: basketball, two teams, five players, backboard, defender’s hoop, free throws, NBA. Likewise, the tools in OpenText Intelligent Classification extract only what is relevant for textual comprehension.

Machine learning

OpenText uses some of the most powerful algorithms in machine learning to extract relevant insight from text. Some of the approaches include:

Decision trees

OpenText Intelligent Classification uses decision trees to enhance machine learning processes and resolve certain language ambiguities, increasing the accuracy of extractions. From a computing perspective, decision trees are a decision support tool commonly used in business processes, specifically in decision analysis, to help identify a strategy most likely to reach a certain goal.

Post-processing algorithms

Users can configure OpenText Intelligent Classification to apply post-processing algorithms that select relevant semantic metadata extractions. “Post-processing” refers to actions a reviewer takes after the automated or semi-automated metadata assignment process described above. This is important because it allows the users reviewing a collection of text to distinguish what is truly meaningful from what is not. These post-processing algorithms can expose the metadata according to different factors, such as relevancy, frequency, engine confidence and position.

For example, they could be set to look for articles that use the word “tomatoes” often enough to confirm that they are about growing that particular crop, rather than a quick roundup of plants recommended for a summer vegetable garden. Specifically, this would look like:

MTM = Cat: <tomato> Weight: 82.5

Post-Processing: Expose only categories > 85.0 Result: Document not shown to user

They can also filter out “stop words,” such as “and,” “the,” or “who,” which are too common to be valuable search terms.

Authority files

- Organization names
- Person names
- Geopolitical locations
- Geophysical names
- Trademark
- Events
- Life sciences
- Features
- Currency, date and time
- Computer and internet
- Twitter
- Personally identifiable information (PII)

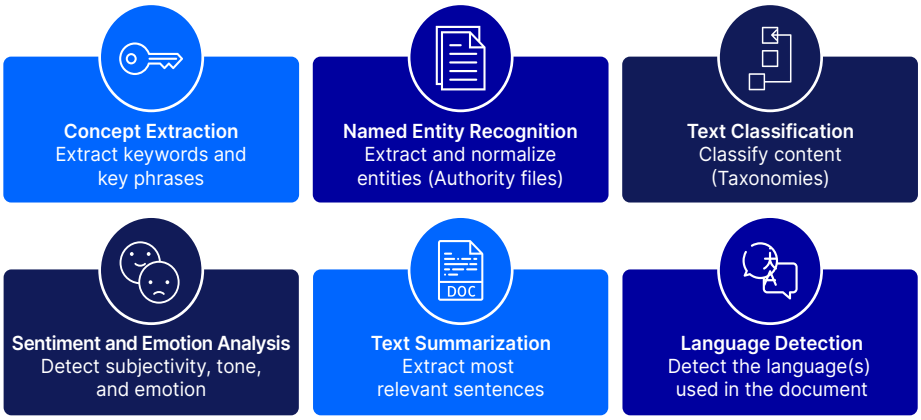
Knowledge engineering

It is important to separate ML from knowledge engineering. A number of OpenText Intelligent Classification's modules use ML to improve their accuracy by predicting extraction results based on statistical observations, without being explicit definitions. The knowledge engineering functionalities embedded in OpenText Intelligent Classification empower users to create and manage their own taxonomy and authority files and their own classification or extraction rules, which can call upon different condition types.

Modules and architecture

OpenText Intelligent Classification comprises six modules, each with its own configurations and providing different types of semantic metadata.

These six modules can be leveraged independently or together to extract meaning from various types of content (user-generated content, product reviews, legal contracts, HR documents, etc.) reaching across multiple industries, domains and levels of formality.



Content analytic modules

Concept Extraction

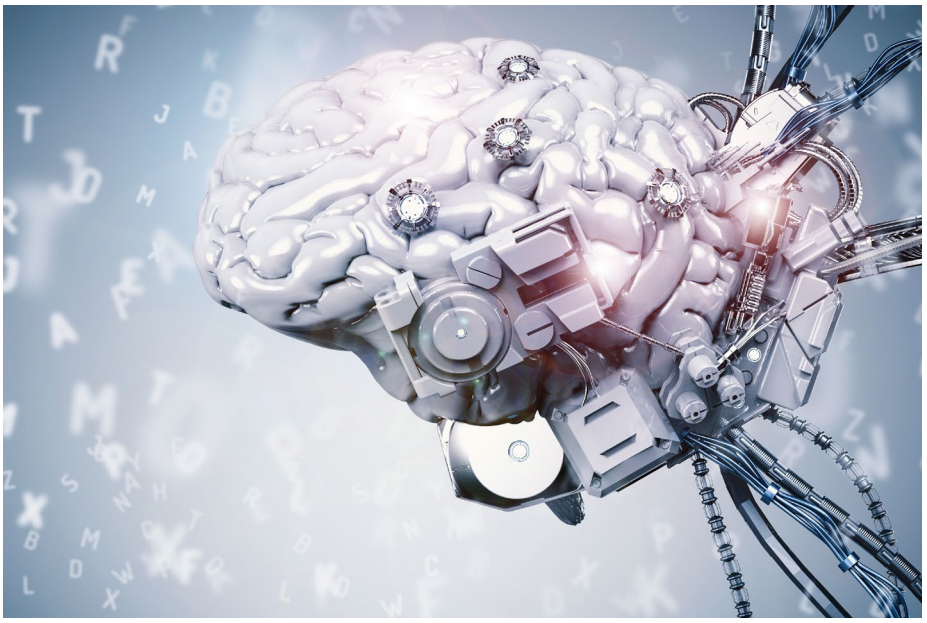
Concept Extraction identifies meaningful information from documents using special grammatical, pattern-based algorithms to extract key concepts.

Concepts: A keyword (simple concept) or key phrase (complex concept) found within text. For example, 'financial statement' would be a key phrase, while 'goal' would be a keyword. OpenText Intelligent Classification Concept Extraction uses a linguistic approach to concept extraction. Conceptually, it operates in three steps:

1. **Word cropping:** A first analysis is performed distinguishing keywords from numbers, abbreviations, delimiters, etc.
2. **Part of speech tagging:** For each keyword, a tag for its grammatical role is assigned (e.g. verb, noun, pronoun, adjective, etc.)
3. **Extraction of part of speech patterns:** Relevant grammatical patterns (e.g. noun-noun, adjective-noun, etc.) are located and exposed.

This approach allows configuration for specific uses (e.g. news/magazine stories, photo captions, legal content, business memos, etc.). For example, fully grammatical sentences in legal content or forms might require a different configuration than one for user-generated content in social media photo captions.

Typical uses include the creation of on-the-fly document clusters, automatic population of a document's keywords metadata field and increased search engine optimization (SEO) rankings, by automatically populating metadata within HTML headers, URLs, etc.



Named Entity Recognition

Named Entity Recognition (NER) locates and extracts places, people, organizations and anything else with a name.

Named entities: Concepts that belong to specific, pre-defined subject categories, typically people's names, geographic locations, organization names and trademarks/products. For example, "Open Text Corporation" would be an organization entity; "North America" is a geographic entity and "Albert Einstein" is a person name entity. OpenText's entity extractor includes normalization algorithms to ensure standardized entity assignments across all content.

The software can be taught to recognize variations of different names for the same entity. For example, a document containing an entity labeled "RedDot" (now a subsidiary of OpenText) or an entity labeled "OpenText Systems Inc." can be indexed with the standard Open Text Corporation version of the entity, allowing consistent and relevant search results across all content. The same applies to examples like "John Smith" and "J. Smith" or "United Nations," "UN" and "U.N." These algorithms rely on advanced machine learning techniques and ISO/NISO standard authority files, which can be managed using OpenText Intelligent Classification Annotation Studio.

Knowledge bases

- Business and finance
- General business
- International Press Telecommunications Council
- Industry Classification Benchmark
- Library of Congress Thesaurus for Graphic Materials
- Energy Technology Data Exchange and International Nuclear Information System
- Generally accepted accounting principles (GAAP)
- Records management
- Résumés
- Retention

Client-specific authority files can also be imported in NER and managed using OpenText Intelligent Classification Annotation Studio. More details, including a list of authority file topics included with the software, appear below. Further, users can create their own new authority files in the Annotation Studio.

Entities highlight the who, where and what of documents. Typical uses include inline tagging, creation of live content indexes and lists of "The most talked about..." companies, persons, locations or products. Uses also include increased SEO rankings, by automatically populating metadata within HTML headers, URLs, etc.

Text Classification

Text Classification indexes and sorts documents by classification and provides relevancy rankings. It creates document profiles by analyzing concepts and querying them against an extensible knowledge base.

Classifications: A specific division in a system of classification. The system can be a taxonomy, a thesaurus or a controlled vocabulary. For example, "football" is a classification referring to the sport, while "biology" would refer to the science.

Classifications express the topic of a document. For example, if the biology classification is assigned to a document, it means this document is about biology, even if the keyword biology is not mentioned in it. That means when someone is looking for a document about biology, using the classification metadata, they can find relevant documents. By contrast, a full-text search would only yield results where the keyword is mentioned, which can lead to irrelevant results.

The Text Classification module can manage an unlimited number of taxonomies at the same time. As first mentioned in the "Creating Semantic Metadata" section on page 6, this can be set up as a completely automated service where precision is typically favored (e.g. batch processing documents overnight or in real time with little to no manual review) or a semi-automated service, where recall is typically favored (e.g. a manual review of all or some of the recommendations is performed). The Text Classification module can also be configured to auto-assess the quality of its classifications by providing a confidence score. This score can be used to highlight the documents that should be reviewed. This enables the setup of a cost-efficient production workflow, where the review team can focus on less accurate results.

OpenText Intelligent Classification offers many out-of-the-box knowledge bases covering important business and scientific topics, to speed up automated categorization of your content.

Further, it allows users to train their own organization-specific knowledge bases in the Annotation Studio. This is extremely useful for custom classification of documents in a repository or as part of document workflow.

Sentiment and Emotion Analysis

Sentiment analysis detects the tone of documents and can determine whether a document, or even a sentence, expresses an overall mood (negative, positive or neutral) and whether that statement is subjective (opinion) or objective. This

Supported languages

Arabic, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hebrew, Hungarian, Icelandic, Irish, Italian, Japanese, Latvian, Lithuanian, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Slovak, Slovene, Spanish, Swedish, Turkish, Ukrainian and Vietnamese

Level of language support varies per language. For more information, contact us.

feature can also identify the overall sentiment attached throughout the text to a named entity, such as a product or organization. Emotion analysis determines the emotions expressed within the text (anticipation, anger, disgust, fear, joy, sadness, surprise, trust), along with their classifications based upon the weight for each.

Typical uses include the automatic clustering and/or moderation of social media, user-generated content (UGC), including surveys, emails, forums, blogs, chats and tweets, Voice of the Customer tracking, measuring public opinion and other business intelligence applications.

Text Summarization

Text Summarization identifies key sentences in a document and uses them to create a summary.

Summary: A shortened version of the original document, highlighting the most important sentences.

Typical uses include the creation of document teasers to fuel content memberships and document summaries to enable the quick review/sorting/selection of important assets. Uses also include increased SEO rankings by allowing the organic creation of optimized HTML code, increasing the code to content ratios or by automatically populating metadata within HTML headers, etc.

Language Detection

Language Detection finds the language of text or paragraphs within text.

This module is statistically trained on n-grams coming from a variety of languages. Language scores are provided per document or per paragraph. This module can automatically classify documents based on their language; the metadata can also be used in SEO use cases. OpenText Intelligent Classification supports Arabic, Chinese (simplified), Dutch, English, French, German, Hebrew, Italian, Japanese, Portuguese, and Spanish with full, or most, functionality, and dedicated natural language processing. It also supports basic concept and entity extraction in other languages that use Latin alphabets, such as Polish, Romanian and Vietnamese and non-Latin alphabets, including Greek, Russian and Ukranian—even non-alphabetic languages.

Additional components

These components round out the functionality of the complete OpenText Intelligent Classification platform.

Core Engine: The Core Engine is the binary package that contains configuration files and Java libraries.

Data packs: OpenText Intelligent Classification offers a range of knowledge bases in key subjects, with industry-specific terminology, providing a handy head start for new content analytics projects. The data packs include foundational metadata specific to each language your text is in. Users can further expand these data packs with specific configurations aligned to business needs.

Resources

Join the conversation ›

Keep up to date ›

Learn more ›

Professional services available
AI & Analytics Services ›

OpenText Intelligent Classification Annotation Studio: Annotation Studio allows users to:

- Build and manage taxonomies and vocabularies.
- Train a text classification against a given taxonomy.
- Have an on-the-fly interaction with the content tagging process.
- Maintain customized controlled vocabularies as they change over time or over evolving content strategies.
- Search through data structures.
- Automate data structures.

REST APIs: Most of the functionality of OpenText Intelligent Classification is exposed via a REST API. This widely used interface enables an organization to embed text mining capabilities into other enterprise applications. For example, as part of a loan processing workflow, an organization might want to call OpenText Intelligent Classification to classify incoming documents.

Conclusion

This paper has provided an overview of the various machine learning algorithms and content analytic techniques on which OpenText Intelligent Classification is based, as well as the specific functions of each module. This analytics package offers a wide range of techniques for enriching unstructured text with structured metadata, including extracting concepts, named entities, categories, sentiments, and emotions to improve its “findability” and usability.

OpenText Intelligent Classification includes a unique synthesis of statistical and grammatical analysis, a choice of more than 30 languages, and the flexibility to combine fully automated extraction processes with human-guided metadata and configure search parameters. It is also easily embeddable into other applications, making text analytics simpler and more approachable for many industries and use cases.

OpenText Intelligent Classification can transform the way an organization identifies meaning across a wide range of text sources and collections, unlocking the power of information to fuel data-driven business decisions and more efficient business processes.

To learn more about how OpenText Intelligent Classification can unlock new value and insights from enterprise content, consider the Semantic Strategy Workshop. Participants will work with an OpenText computational linguist on site to get an overview of how OpenText works and explore various content challenges that can be addressed. Email PortfolioAnalyticsPS@opentext.com for further details.

opentext™