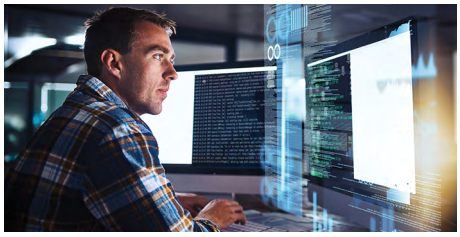
Product Flyer

opentext™



Closing the Gaps in Natural Language Processing

IDOL offers Unified text analytics, speech analytics and video analytics.

"IDOL has helped to automatically search for and extract key concepts from a massive amount of text, video and audio data on a daily basis. This has significantly enhanced user experience and productivity, quality of information and reduced operating costs."

ZHOU QING
Research and Development engineer
Xi'an Panorama Data Co., Ltd

Understanding the complex structure and frequently ambiguous meaning of human language can be very difficult for computers. Common NLP approaches address this challenge by using algorithms that tokenize text into words and tag the words based on their position and function in a sentence—to extract meaning and perform tasks such as summarization, named entity recognition, or sentiment analysis.

OpenText™ IDOL combines sophisticated probabilistic modeling with NLP algorithms to extract concepts and insights from written or spoken language in a fully automated and highly accurate manner.

Extracting Meaning from Human Information

IDOL derives contextual and conceptual insights from data, which allow computers to recognize the relationships that exist within virtually any type of information, structured or unstructured. Similar to NLP, the capability to understand the data makes it possible to automate manual operations in real time by extracting meaning from information and then performing an action. Built upon these principles, IDOL enables you to recognize around 1000 file types with support for 150 languages and connect to over 150 repositories, providing advanced and accurate retrieval of valuable knowledge and business intelligence both inside and outside your enterprise. Unlike NLP technology, which focuses solely on linguistics, IDOL follows a language-independent, statistical approach to understanding human information that is fine-tuned by the use of linguistics. The underlying content analytics technology is built upon the mathematical works of Thomas Bayes and Claude Shannon, and was developed further through innovations covered by over 200 patents. With IDOL, you can identify patterns that naturally occur in human language, whether written or spoken, based on the frequency of terms that correspond to specific concepts.

This approach makes IDOL completely language-independent. In addition, IDOL implements a number of NLP techniques, such as stemming, to fine-tune the analysis in common languages.

Why Use a Probabilistic Approach?

The IDOL statistical approach supports format and language-independence. For this reason, IDOL does not require training on English syntax to determine which concepts are important and which are of little relevance in an English text. The same principle is true for Arabic, Greek, Spanish, and Chinese—or any other language. NLP technologies, on the other hand, require extensive training on structural and syntactical rules to perform segmentation, dictionary lookups, or part-of-speech tagging before ideas can be formed. This training typically requires a large dictionary and a large annotated corpus. In some cases, such corpora are already available, but for less common use cases (e.g., medical data) the vocabulary and syntax are very specific and such corpora

must be built and tagged manually. With IDOL's data analytics capabilities, you can analyze human information automatically, with no additional training. Another challenge of using traditional NLP technology is its difficulty dealing with improper language, such as chat conversations, social media data, or spoken language. Because the information does not follow standard rules of grammar and syntax, NLP has difficulty with the analysis. IDOL overcomes this challenge by treating all words as abstract symbols of meaning that exist in patterns and relationships, regardless of formal linguistic rules.

Leveraging the Benefits of NLP

IDOL leverages NLP algorithms on a second layer of processing to further optimize accuracy and performance on a per-language basis. Stemming algorithms allow IDOL to accurately relate concepts with similar semantic roots. Sentence-breaking libraries, stop-lists, and n-grams are also used to optimize concept separation and proper weighting of terms. On a third layer of analysis, IDOL can perform NLP tasks such as entity extraction or sentiment analysis to facilitate the interaction between the user and the processed information. Once IDOL identifies the important concepts, more than 500 functions may be performed—based on NLP tasks that enable you to interact with information easily and comprehensively.

Figure 1. IDOL Eductaion automaticaly extracts entities such as people, locations, or organizations, from any piece of unstructured data.

What NLP Tasks Can IDOL Perform?

IDOL can help you handle the following types of language-processing tasks:

- **Named entity recognition (NER)**, or entity extraction, locates and classifies elements in text into predefined categories, such as people's names and locations. IDOL Education uses grammar-based techniques to extract entities from any piece of unstructured information. A number of grammar-based techniques are available out of the box (such as proper names, addresses, organizations, phone numbers, or Social Security Numbers). In addition, IDOL Education enables custom entities to be built and deployed to meet specific objectives. Paired with IDOL capabilities to identify patterns and draw relationships between different entities, IDOL Education is a powerful tool for classifying and relating information.
- **Stemming** reduces inflected words to their root and allows IDOL to group together words with similar basic meanings. This enables users to retrieve relevant information even when the specific form of the word is not present in the index. For example, a query for "running" will automatically retrieve information about "running shoes" but also about "runners" or "places to run."
- **Sentence-breaking and character tokenization** are important for languages that use words that are not delimited by spaces. IDOL can be easily configured to break text into sentences and tokenize characters into n-grams of a specified size with great accuracy.
- **Stop words** are extremely common terms of little or no value. Words such as "a," "and," and "the" do not carry any conceptual significance. IDOL can automatically identify such words and exclude them from analysis to increase performance and the accuracy of results.
- **Synonyms** allow users to build conceptual relationships between words and phrases. When a user queries the engine for "college," IDOL recognizes that a college and a university represent the same concept and therefore automatically searches for "university," as well. IDOL can also be configured to treat similar terms as hyponyms or hypernyms.
- **Automatic summarization** creates a brief summary of the contents of a document. IDOL has the ability to create a number of different summaries: a conceptual summary of the most salient concepts in the document as a whole, a contextual summary that relates to the original query, and a simple summary that is comprised of a few sentences from the beginning of a document.
- **Speech recognition** determines the textual representation of natural speech. IDOL can create a transcript of an audio file or a live audio stream, and apply its sophisticated data analytics to detect the main concepts and relationships within the audio. Spoken language differs significantly from written language in that grammatical and syntactical rules are not always followed. In addition, speech recognition is still an imperfect process and therefore automatic transcriptions are prone to errors. Unlike NLP, which understands language via linguistics, the probabilistic model can understand the main concepts of an audio and can deliver insightful information even when the transcription is imperfect.
- **Sentiment analysis** determines the attitude of a writer with respect to the

Connect with Us

document. With the rise of social media, this technology is extremely helpful in determining online opinion, identifying opportunities, and managing reputations. People's comments can be complex and multifaceted, expressing a harsh criticism for one topic and appreciation for another within a single message. While traditional technologies can miss these subtleties, IDOL sentiment analysis capabilities can identify topics in the text and classify the polarity for each topic as positive, negative, or neutral.

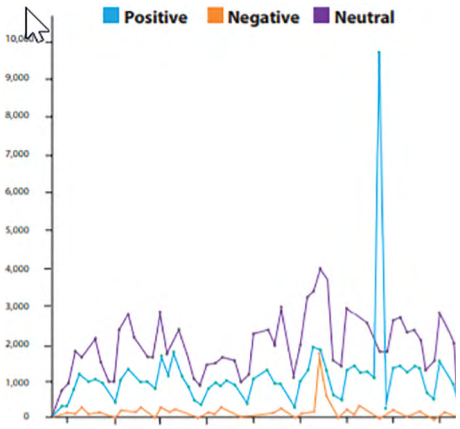


Figure 2. IDOL can analyze sentiment over time for a particular topic and identify points of interest.

The IDOL Advantage

The IDOL probabilistic model is capable of extracting meaning from human information in any language or format. It does not rely on an intimate knowledge of a language's grammatical structure, but rather derives its understanding through the context of the words' occurrence. This is particularly beneficial when analyzing spoken or informal language that does not follow the linguistic rules of pure NLP systems. In addition, the ability to extract information from around 1000 file types—including audio and video—makes this sophisticated technique a very powerful tool that can add great value. Optimized through NLP algorithms, such as stemming or sentence-breaking, IDOL yields incredible accuracy and performance and offers a broad spectrum of functions that can be performed in a fully automated fashion and on a single platform.

Learn more at
www.opentext.com/IDOL

www.opentext.com

opentext™