Best Practices in Text Analytics and Natural Language Processing

Produced by:

KMWorld Magazine
Specialty Publishing Group

For information on participating in the next white paper in the "Best Practices" series, contact:

Stephen Faig
Group Sales Director
908.795.3702
sfaig@infotoday.com

LaShawn Fugate
Account Executive
859.361.0667
lashawn@infotoday.com

KMWorld
KMWorld.com

Access Innovations, Inc.
Innovative • Effective • Customizable

Access Innovations, Inc.
6301 Indian School Road NE, Suite 400
Albuquerque, NM, 87110
PH: 505.998.0800
FAX: 505.256.1080
Contact: info@accessinn.com
Web: www.accessinn.com

searchblox

SearchBlox Software, Inc.
4870 Sadler Road, Suite 300
Glen Allen, VA 23060
PH: 866.933.3626
Contact: sales@searchblox.com
Web: www.searchblox.com

Speed Reading on Steroids

By Marydee Ojala, Conference Program Director, Information Today, Inc.

At one point in my life, I was entranced by the possibilities of speed reading. Just think, being able to get through the assigned chapters in my textbooks in minutes instead of hours, with the added benefit of complete comprehension of difficult concepts. It was more realistic than the idea of putting the textbook under your pillow and hoping the words would get into your brain by osmosis as you slept. I could point to several bad exam results to prove this doesn't work. But speed reading held such great promise!

I found it intriguing that President John F. Kennedy could, purportedly, read 1,200 words per minute. Another president, Jimmy Carter, took speed reading courses while in the White House. If such prominent people could embrace speed reading, why not me? My fascination with speed reading probably happened during the heyday of the Evelyn Wood Speed Reading Dynamics courses. The company has gone through numerous owners, some of whom advertised heavily, making the concept of speed reading very visible; other times it faded from view. Although I found the notion of being able to read quickly and still master complex information concepts extremely appealing, I never actually took a speed reading class, never experienced Evelyn Wood's techniques first-hand.

Looking back, I assumed speed reading was a fad, interesting in its day, but not applicable now. However, a quick search revealed numerous recent articles and YouTube videos touting the benefits of speed reading and promoting various techniques to help speed up your reading. These techniques can be contradictory. Run your finger under sentences or never run your finger under sentences. Use the techniques on fiction. No, only use them on non-fiction. I'm guessing that Kennedy and Carter concentrated on non-fiction. I'm also thinking that speed reading is not applicable to poetry. Of course, if you're thinking of speed reading in a business context, poetry is unlikely to be part of your agenda.

Text Analytics and NLP Boost Comprehension

At its heart, text analytics is a technology-boosted form of speed reading. Speed reading on steroids, if you will. The advantages of having a computer read through a very large number of documents, identify the main points, see patterns that a human reader could miss, and spit out a comprehensive analysis of those documents is immense. As SearchBlox's CEO, Robert Selvaraj, points out, the benefit of text analytics has increased during our pandemic times because the volume and velocity of data produced has skyrocketed. To understand all this data, we need text analytics and natural language processing (NLP).

Selvaraj lists five concrete examples of how text analytics and NLP can lead to actionable insights. He starts with simplification. Not only is today's data complex, it occurs in both structured and unstructured formats. It's not simply text documents, it's email, websites, surveys, social media, and many other forms of unstructured data. With all these varied sources, text analytics and NLP can simplify the process.

Think about how much has changed in the past few months. Data from pre-pandemic times may have reduced relevance now. Language has changed. A year ago, Zoom was not a verb and positive medical test results were not viewed as having negative outcomes. Thus, machine learning models that have not evolved to accommodate changed circumstances and language have limited applicability for decision making. It takes NLP to understand new patterns.

Bots have gained traction now that so many people are working from home. They don't sleep, they're always available, and they have infinite patience with answering repetitive questions. Intelligent search has also come into its own, in an environment where people have more questions than answers. Whether it's information about health, government policies, or something else, people are floundering, given the barrage of data. NLP recognizes implicit intent, corrects for misspellings, and returns relevant results despite users' not really knowing what they were asking for.

Text analytics and NLP boost productivity. Lacking physical proximity in the workplace, it's imperative that employees have the tools they need to do their jobs—and those tools should be intuitive and helpful. The overnight transformation of work environments from office to home encourages using text analytics and NLP to discover new and important insights.

"At its heart, text analytics is a technology-boosted form of speed reading. Speed reading on steroids, if you will. The advantages of having a computer read through a very large number of documents, identify the main points, see patterns that a human reader could miss, and spit out a comprehensive analysis of those documents is immense."

Language Analysis for Speed Learning

Access Innovations' approach of using text analytics and NLP centers on language attributes and analysis. It's speed learning, not simply speed reading. The underlying concept here is to bring targeted thesauri and dictionaries to bear in document text scanning and using a set of tools based on artificial neural networks to understand the nature of the document and distinguish between the worthwhile information and nonsensical gibberish.

Daniel Vasicek, Data Scientist at Access Innovations, cautions that documents that "have been passed through enough layers of misinformed character encoding transformations" present problems, since they can no longer be decoded due to a high error rate. He has more positive comments on the ability of tools to automatically detect language and sentiment. Thesaurus tools identify standardized concepts and named entities in documents.

And let's not forget grammar. Vasicek explains how grammars can be recognized by their Bayesian fingerprints of the words in a document. One of those fingerprints involves vocabulary; another is frequency counts for phrases. Text analytics and NLP tools speed the learning and understanding of documents.

Text analytics and NLP hold so much promise for human understanding of vast quantities of information. It's a new generation of speed reading and speed learning. Evelyn Wood, eat your heart out. ▮

Marydee Ojala is conference program director for Information Today, Inc. She works on conferences such as Enterprise Search & Discovery, which is co-located with KMWorld, and WebSearch University, among others. She is a frequent speaker at U.S. and international information professional events. In addition, she moderates the popular KMWorld webinar series.

Ojala is based in Indianapolis, Indiana and can be reached at marydee@infotoday.com.

5 Ways Text Analytics and NLP Provide Insight in a Pandemic

By Robert Selvaraj, CEO, SearchBlox

Before the pandemic, about 15% of U.S. employees worked from home and only some of the time. By the middle of April, half of U.S. employees were doing all of their work remotely. Today each employee's home is a little data silo, complicating information sharing and data discovery.

Humans can't keep pace with the volume and velocity of data this pandemic is producing. But text analytics using natural language processing (NLP) can help. Let's take a closer look at how NLP helps us keep up—and even get ahead—as distributed teams working through the coronavirus crisis.

This image illustrates how various technologies can help organizations process high volumes of data, turning text into actionable insights.

1. Text analytics using NLP simplifies the process of analyzing complex structured and unstructured data from many different sources, including email, websites, customer surveys and social media. It's important to note that as much as 80% of your data is unstructured. Imagine the insight you're ignoring if you don't tap into that!
2. Machine learning models built with pre-COVID data are of limited use to decision makers today. Text analytics powered by NLP understands linguistic ambiguities and context well enough to identify new patterns and recognize vocabulary that pre-pandemic models couldn't or might incorrectly identify as outliers. Think about how a "positive" test result is now a negative, Zoom is a verb, and "plandemic" is more than a typo.
3. Text analytics and NLP enable bots to converse with users. According to the marketing experts at Hubspot, chat volume for its 70,000+ customers has steadily risen week-over-week since the pandemic began. The week of July 27, for instance, Hubspot customers engaged in 64% more chatbot conversations than pre-COVID averages.

Why? Chatbots offer a better employee and customer experience in a crisis because:

- They don't sleep.
- They're on duty 24/7.
- They're patient: They don't mind answering the same questions all day long.

They allow you to scale customer and employee support.

Robert Selvaraj

Prior to starting SearchBlox in 2003, CEO Robert Selvaraj was a leader at Valtech, Mimecast and Grassroots Group (now part of Blackhawk). "From day one our mantra has been 'Search is simple,'" he explains. "We want to make search less complex to deploy and manage, while simultaneously making it more powerful and capable."

4. We have way more questions than answers right now. One of our customers, the World Health Organization, for instance, saw a 10X increase in search traffic on their site in the first month of the pandemic. NLP helps people search for information they don't fully understand, whether it's about the virus, the Paycheck Protection Program (PPP) or private learning pods. Intelligent search using NLP "understands" implicit intent, forgives misspellings and fetches connected content, providing more relevant results even if the user doesn't know exactly what they're looking for.
5. Text analytics and NLP boost productivity. Before the pandemic, enterprise employees spent an average of 1.8 hours every day searching for the information and data they need to do their jobs. Now that they're physically separated from teammates and collaborators, it's probably more. That's not only terribly inefficient, it's a horrible employee experience. And now that lines between home and work are blurrier than ever, your teams want the tools they use to do their jobs to be as intuitive and helpful as the ones they use in their personal lives.

They expect answers that are:

- Relevant like Google's search engine results pages (SERP)
- Insight-driven like Amazon's recommendation engine
- On-demand like Spotify's personalized playlists
- Convenient like Apple's Siri

During Microsoft's quarterly earnings call in April, CEO Satya Nadella said: "We've seen two years' worth of digital transformation in two months." As that rapid transformation forces you to consider new data sources to discover important insight, start your search with text analytics and NLP. ▮

SearchBlox builds intuitive and intelligent insight engines based on open source technologies.

www.searchblox.com
1-866-933-2626
info@searchblox.com

Text Analytics and Natural Language Processing

By Daniel Vasicek, Data Scientist, Access Innovations Inc.

Daniel Vasicek

Daniel Vasicek joined the Access Innovations team after spending more than 25 years as a geophysicist for British Petroleum. He received bachelor's and master's degrees in engineering from Purdue University before earning his doctorate in aerospace engineering sciences from the University of Colorado. His professional affiliations include the Society for Industrial and Applied Mathematics, the Professional Engineers of Colorado, and the Oklahoma Society of Professional Engineers.

If we are given a document, what information would we like to have about that document? We would like to have a bit more than the title, author name, and date of publication. For example, we might like to know the important concepts and entities discussed in the document? A thesaurus can help us condense all of the different ways (synonyms) of expressing concepts into a set of standardized concepts. How coherent is the document? If there are a great many wildly different concepts in the document, then it might be less coherent. A great many nonsense generators are available on the internet that produce incoherent nonsense with the casual appearance of erudite text. How can we detect this kind of nonsense? Can we measure the coherency of a document? Can we detect and measure sentiment? Is the author optimistic or pessimistic? What language and encoding is used in the document?

Build Your Toolbox

Like any expert, we need appropriate tools to help us answer these questions. One set of tools is an efficient document text scanner (Access Innovations MAI) coupled with targeted thesauri and dictionaries. Another set of tools is based on artificial neural networks trained using sorted sets of examples. These two sets of tools compliment one another. Both require maintenance and human support as society and language evolve.

Often the character encoding is specified as part of the document header information. If it is not explicitly available, we must detect it using appropriate dictionaries. Unfortunately, some documents have been passed through enough layers of misinformed character encoding transformations to produce documents that cannot be decoded without error using any encoding. The encoding that produces the least number of errors might be preferred.

Watch Your Language

Automatic detection of language can be done by counting words from the document that are in language dictionaries. The language that recognizes the most words can be assigned to the document.

Concept and entity identification is the function of a thesaurus. A hierarchical list of concepts can be created using frequency sorted lists of phrases from appropriate documents. The more interesting concepts are likely to be captured by more frequently used phrases. These can be sorted into synonymous concepts and encoded into regular expressions and nested within IF … ELSEIF … ENDIF blocks. Proximity relationships have a powerful influence on the meaning of words. The logical tests must incorporate the capability of expressing proximity relationships such as "adjacent to," "in the same sentence," "in the same paragraph," as well as upper and lower case letters. These can be an important component of the meaning of words. For example, "AIDS" is probably a disease, and "aids" is probably a verb. Sorting the resulting concepts into a hierarchy of related concepts may require some human intervention.

"Often the character encoding is specified as part of the document header information. If it is not explicitly available, we must detect it using appropriate dictionaries. Unfortunately, some documents have been passed through enough layers of misinformed character encoding transformations to produce documents that cannot be decoded without error using any encoding."

Caring and Plagiarism

Sentiment can be detected using sentiment dictionaries of optimistic and pessimistic words and scored according to the number of hits in each dictionary.

Some instances of plagiarism can be detected by comparing index terms. Clusters of documents in index term space (or references space) can be explored for plagiarism. Once you have candidates for plagiarized articles (perhaps documents with the same [or almost same] index terms?) two candidate documents can be compared by entering the sentences of the document into a hash table and counting the number of sentences that are replicated between two documents.

Documents generated by random production grammars can be recognized by Bayesian fingerprints of the words and phrases used by the grammar. The vocabulary used by the production grammar is a fingerprint that can be used to identify the grammar. Often it will be much smaller than the vocabulary of a real human being. And the frequency of the phrases used by the grammar will be characteristically different from the frequency of those same phrases in "normal" text. If you can gain access to the production rules used to generate the documents, then you can count the instances of those phrases and words in documents and estimate the Bayesian probability that the document was produced by the grammar or by observed occurrence of the words and phrases in some acceptable corpus. If you do not have access to the production grammar, then you can use examples of the documents produced by the grammar to develop frequency counts for the words and phrases in the documents for use in estimating Bayesian probabilities.

Bring it Together

In summary, character encoding and language can be detected by comparison with dictionaries having known encoding and known language. Lists of standardized concepts and entity names can be obtained by scanning documents with thesauri designed to recognize interesting concepts and entities. Sentiment can be evaluated by counting instances of positive and negative words in text. Plagiarism can be detected by counting identical sentences in different documents leveraged by selecting documents that cluster in concept or references space. And documents generated by production grammar-based nonsense generators can be recognized by the relative frequency of phrases from the production grammar. ▮