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

Everything Old Is New Again

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

I'm entranced by old technologies being rediscovered, repurposed, and reinvented. Just think, the term artificial intelligence (AI) entered the language in 1956 and you can trace natural language processing (NLP) back to Alan Turing's work starting in 1950. Text analytics has its antecedents in data mining. Data mining itself has a long history, all the way back to Thomas Bayes, who died in 1761, and his eponymous theorem that still informs algorithms regarding inference, probability, and predictions.

Even when the concept and the phrases are the same, the increased power, bandwidth, and sheer computing power available today changes the applications significantly. Whenever I read about the pattern-matching feats of NLP, machine learning (ML), and AI, I'm reminded of a student I knew at university. He manually counted how many times certain words appeared in a Shakespeare play. As you can imagine, this consumed an enormous amount of his time—and I've never been convinced that his professor thought it was worth the effort. With today's technology, this task would take maybe a minute, maybe even less time.

Data Quality Matters

Jen Snell, VP of Product Marketing at Verint, understands that the fundamental purpose of text analytics and NLP remains constant: "They help get the right information to employees at the right time." Sounds simple, right? It's certainly an admirable goal, but one too often thwarted by the amount of information available. Sifting through zettabytes of data is not a task for humans; it has to be done by computers.

For computers do this well, regardless of how robust their text analytics and NLP capabilities, the quality of the underlying data needs to be assessed and optimized. It doesn't matter whether that data is structured or unstructured, its quality determines the success or failure of a KM project.

Snell also cautions that not all NLPs are the same. Two approaches to using an NLP engine are statistical and symbolic. In the former, you train the system to identify patterns, generate a model, and predict word meanings based on a large data corpus. The latter used hard-coded linguistic rules, which originate with people and are then taught to machines. Neither, according to Snell, are sufficient by themselves. Regardless of how you deploy NLP, data quality makes all the difference.

Ask NLP

One major change from the early days of NLP is the increasing reliance on more informal language rather than the written word of structured letters or memos. Customers expect to interact with companies in a conversational way, either by actual phone calls or by text messages. As Susan Kahler, SAS's Global AI Product Marketing Manager, points out, you can use NLP engines to teach and guide machines to examine these types of audio and text data.

Identifying relationships and patterns in the data puts you in a better position to meet customer needs and delivery better experiences, personalized to the individual customer. Additionally, pulling together data from all your customer communications channels and feeding it into an NLP engine gives you insights that will streamline future customer communications.

"Identifying relationships and patterns in the data puts you in a better position to meet customer needs and delivery better experiences, personalized to the individual customer. Additionally, pulling together data from all your customer communications channels and feeding it into an NLP engine gives you insights that will streamline future customer communications."

Oh See Our Data

Optical character recognition (OCR) is another of those old-time technologies that has morphed into a new world of AI. Daniel Vasicek, Senior Data Scientist, Access Innovations, Inc., sees huge improvements in OCR performance and in automatic translation from one language to another. However, real data has a degree of uncertainty that causes it to be non-conforming.

Because data isn't always perfect (a person's job affiliation from several years ago, for example), models built to reflect and draw inferences from that data aren't always perfect, either. Thus, fitting data to the model requires some balancing of measurement errors with model errors. This enhances model predictions.

This may sound somewhat heretical, but Vasicek thinks that exact fits are not only no longer possible, but also not desirable. As he writes, "An exact fit to noisy data means that the model is fitting the noise." Although you'd like to reduce the uncertainty in predications, models should still be able to predict from noisy data even without a perfect fit. He warns, too, about overfitting, where the ML algorithms get overly specific, weeding out information that is actually germane and thus giving erroneous or misleading results.

Improving Internal Search

Internal search is one area deeply affected by text analytics, NLP, and ML. Sean Coleman, CTO and Chief Customer Officer at BA Insight, suggests adding semantic search to the list of "new" technologies. In his view, semantic search lets people "search as they speak," creates a single unified index, and applies conceptual knowledge to search queries. This leads to higher relevance of search results.

Associated technologies for better search include content intelligence, which gives employees detailed information about files, multimedia content, and taxonomies, along with the textual content. A single index incorporating ML features autosuggestion and autocorrection. Personalization is gained by showing what other people viewed, based on location, department, or interests. Sentiment analysis, to be meaningful internally, can focus on social media inside the enterprise and on emails and other electronic communications. Coleman also mentions search bots as coming of age for internal search.

Not everything that was old is new again, with good reason. Few people miss Clippy or rotary dial phones. AI in 1956 played checkers and that was the extent of its "intelligence." We've moved on, and I think we can all agree that the evolution of AI, NLP, ML, and text analytics benefits us both personally and professionally. ▮

Marydee Ojala

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.

Text Analytics and Natural Language Processing: Knowledge Management's Next Frontier

Jen Snell

Jen Snell is VP of product marketing at Verint, where she leads a product strategy team focused on intelligent self-service, conversational AI, automation, and analytics. She is a frequent speaker and a leading contributor on topics shaping the development and design of interactive technologies.

By Jen Snell, Vice President, Product Marketing, Verint

Text analytics and natural language processing are not new concepts. Most knowledge management professionals have been grappling with these technologies for years. From the KM perspective, these technologies share the same fundamental purpose: They help get the right information to employees at the right time.

Yet many organizations still struggle to deploy these tools effectively. The primary challenge 10 years ago was that the tools themselves were in their infancy. Most businesses didn't see the value, particularly relative to the risk of adopting a new technology. Today, the challenge is scale. We are creating data at astonishing rates. In fact, some researchers predict that we will be producing 44 zettabytes of data *per day* by 2025.

This means the typical KM professional faces an unprecedented amount of data to make sense of, and there isn't a human being on the planet with the capacity to do it. To find signals in the loud noise of big data, you need intelligent machines and intelligent interfaces.

That's where text analytics and natural language processing (NLP) come in. If your organization is interested in deploying these tools, here are a few best practices to help you get started.

Data Quality Is Key

Before you can begin any other work, you need to assess and optimize the quality of your data.

While data quality is not a new concept, it has quickly become a big problem. Some researchers estimate that up to 96 percent of enterprises report running into trouble with data quality in the process of incorporating AI in their projects. And this isn't just an IT problem. There are estimates suggesting that bad data costs the US over $3 trillion per year.

If you want to succeed with text analytics and NLP, data quality has to be the starting point. Further, it has to be a business priority. The most successful enterprises have an interdisciplinary team dedicated to this work.

Of course, with both text analytics and NLP, we're often working with unstructured data. Unfortunately, organizations often believe that because they are working with unstructured data, the quality doesn't matter. Nothing could be further from the truth. It's precisely because you're working with unstructured data that great care needs to be placed into the data you put into and derive from these technologies.

> ***"Some researchers estimate that up to 96 percent of enterprises report running into trouble with data quality in the process of incorporating AI in their projects. And this isn't just an IT problem. There are estimates suggesting that bad data costs the US over $3 trillion per year."***

The good news is that we have international standards for data quality assessment that many companies have adopted. And the business case for this work has never been stronger.

NLP Is Not a Commodity

Another misconception is the notion that all NLP solutions are the same. While it's true that most natural language recognition systems have comparable levels of accuracy, NLP is an entirely different technology. Recognizing words is a basic function compared to assembling meaning and intent. And there's a lot of debate about the various approaches to NLP.

The two main approaches are statistical and symbolic. The former is exactly what it sounds like: You train the system on a huge corpus of data; it identifies patterns, generates a model, and then predicts the meaning of some piece of language based on probabilities. Symbolic is based on hard-coded linguistic rules that are developed by people and taught to machines. At a high level, symbolic NLP seeks to teach the meaning of words to the machine, while statistical seeks to predict appropriate responses to inputs based on what worked for similar inputs in the past.

While there's much debate about the merits of both, neither approach is sufficient on its own. A solution that blends the two approaches offers more flexibility and long-term utility than a solution that sticks with one approach. However, the simple best practice is to determine your NLP engine based on your needs and your data.

After all, it comes down to data. These systems train on data, they are fueled by data, and they generate enormous amounts of data. It is therefore essential that you are data-driven in your assessment and selection of any solution.

Ultimately, data is the resource that will drive business and economic development for the next century. It's fitting then that knowledge management is at the center of this new frontier in business. KM professionals led digital transformation efforts as the world embraced the "information economy." Now, armed with technologies such as text analytics and NLP, they are leading the charge into the intelligence economy.

Contact Us:
1-800-4VERINT
www.verint.com/engagement
www.verintblog.com

Keeping It Personal With Natural Language Processing

By Susan Kahler, Global AI Product Marketing Manager, SAS

Susan Kahler

Susan Kahler is currently responsible for AI product marketing at SAS, with an emphasis on deep learning, computer vision and natural language processing. In her 15 years at SAS, Kahler has held roles in persona development, quantitative and qualitative research, competitive intelligence, risk product management and user interface design and evaluation. Prior to joining SAS, she was an advisory human factors engineer at IBM for 14 years and worked on healthcare projects. Kahler received her PhD from North Carolina State University.

Consumers are increasingly using conversational AI devices (e.g., Amazon Echo and Google Home) and text-based communication apps (e.g., Facebook Messenger and Slack) to engage with brands and each other. Yet many companies have not figured out how to analyze data from these communication tools to make it easier to interact with customers and create a more positive customer experience. In these days of customer choice, it is the customer experience that can make or break a company.

Today, customers expect their interactions with companies to be simple, intuitive, efficient and actually solve the issue at hand. In some cases, they even want to have fun when engaging with their favorite brands. The customer experience is a prominent factor in customer loyalty and longevity with a brand and a direct influence on a company's net promoter score. People want an immediate response from you whether they're on your website or phoning your call center. Social media messaging and web chats are becoming more prevalent forms of customer contact, and the nature of those communication streams demands faster, more accurate responses, all while still maintaining that personal touch.

"In an era with technology just a voice away, customers are unwilling to tolerate long wait times, engage in complex searches to find basic information or tolerate a bad experience."

NLP Improves the Customer Experience

But it's difficult for contact centers to understand and respond to customer requests using manual approaches, which leave staff overwhelmed and customers frustrated with long wait times and inaccurate resolutions. Add a variety of channels for customer feedback and the challenge grows exponentially for businesses. In an era with technology just a voice away, customers are unwilling to tolerate long wait times, engage in complex searches to find basic information or tolerate a bad experience. Another layer of complexity is added for global companies that must also identify regional trends and accurately respond to feedback in a variety of languages and dialects.

Natural language processing (NLP) helps companies make sense of all that noise and really understand what their customers are saying and take appropriate action to give customers the best experience possible. NLP is a branch of artificial intelligence that draws from many disciplines, including computer science and computational linguistics. Its focus is understanding, interpreting and emulating human language.

By using NLP to analyze large volumes of unstructured text data from all your customer channels, you can rapidly examine information to gain meaningful insights and use those insights to provide a better customer experience. Rapid categorization of unstructured text in call center transcripts, social media conversations and web chat information can identify the topics that are meaningful to customers and provide real-time recommendations that provide a quick and accurate response to their needs. Analyzing social media sites can provide insights to public sentiment with topic discovery that can reveal dimensions of interest and how those are changing over time. These insights then allow a company to act objectively and quickly based on the voice of the customer.

Tailor the Customer Experience With NLP

Why use natural language processing? Natural language processing can be used to teach and guide machines to analyze unstructured text at a scale and speed that humans simply cannot match. In turn, this allows humans to focus on more strategic, higher-level tasks that machines can't do, like applying reason or employing creativity, including understanding the impact and implementing a viable action plan.

Customer interactions can take many forms–a phone call, an email, a social media message, a web chat or interaction with a chatbot–and these interactions help companies understand customer demand and external brand perspective. Modern NLP capabilities allow you to scale personalized customer experiences by determining their context and recommending an effective action quickly regardless of communication channel or native language.

Intelligent algorithms and NLP techniques can help quickly determine which actions to take by automatically identifying relationships and patterns in data that may have previously gone unnoticed. You can meet multiple customer needs at once by deploying NLP capabilities to analyze unstructured text, recommending next best offers and providing assistance in the form of a customer service chatbot. All of which can help you improve your customer retention and build brand loyalty.

NLP Facilitates the Decision-Making Process

Considering the status of unstructured text as the largest data source produced by people that vastly increases every day, text analytics should figure prominently in any organization's strategy to harness the power of AI. Natural language processing transforms unstructured data to help find the information you need to make informed decisions. To get started with this process, make sure to determine and clearly articulate the objectives of your NLP application. To do this effectively, gather the technical and business users to identify the key terms and reporting requirements that are needed for productive decision making. From there, you can identify the data sources, the type of information you are gathering and the rate at which the data needs to be analyzed. This provides a strong foundation for you to get started with natural language processing.

Look at your organization and consider the unstructured text or audio data you gather and the possible revelations it may hold. That data reflects the voices of those you serve and holds the potential to help you deliver better experiences, improve quality of care and enrich human engagement. There are powerful stories to be told from your unstructured text data. And the best way for you to find them is with natural language processing. ▮

Data Uncertainty, Model Uncertainty, and the Perils of Overfitting

By Daniel Vasicek, Senior Data Scientist, Access Innovations, Inc.

Why should you be interested in artificial intelligence (AI) and machine learning? Any classification problem where you have a good source of classified examples is a candidate for AI. Historically, optical character recognition (OCR) was a difficult problem. We have recently experienced enormous improvement in the performance of OCR because, at least in part, we have a very large collection of already classified examples. Similarly, automatic translation between languages has made tremendous advances because we have access to enormous collections of translated documents that can be used to train the classifier. Other contexts that seem to recommend themselves to machine intelligence and AI learning are concept identification in texts, entity extraction, assigning peer reviewers to submitted documents, sentiment analysis, quality evaluation, and priority assignment.

Data Uncertainty

Real data has measurement errors or has noise that makes it non-conforming to the correct, intended or original values. Data veracity has been acknowledged since at least 2012 as an issue in using AI to support business decisions. Some examples of uncertain data include:

- Rooms are often not square even though they were designed to be
- A person's address in my contact management system from 5 years ago
- The official temperature reading in my city and my backyard thermometer reading

In these examples, the uncertainty can be caused by any number of factors: the carpenters measured wrong or misread a specification, or the ground beneath the building has shifted, or there was an earthquake that broke a supporting structure, or any number of possibilities. There are just as many possibilities for the other examples.

Model Uncertainty

Our models are never perfect, rather they are useful approximations. Consider geocentrism, the model of the universe where the Earth is the center around which other celestial bodies orbit. This model dates from the ancient Greeks, was further developed by Ptolemy in Egypt around the 2nd Century AD. This was the accepted model until 1543 AD, when Copernicus advocated Aristarchus' concept of heliocentrism—the model where the sun is the center of our planetary system. Debates raged for centuries as more and more information was collected, and finally around the late 18th and early 19th centuries, a confluence of empiric evidence overwhelmed the scientific community. What is important to note here is that the geocentric model was used for somewhere between 22 and 24 centuries until a heliocentric model was shown to be "better." And now we have better models where the sun is traveling in an orbit around the center of our galaxy and the universe is expanding.

Fitting the Data to the Model

Balancing measurement errors with model errors enhances model predictions.

Carl Friedrich Gauss, a German mathematician and physicist, made two major changes to a model that his predecessors had tried to use to rediscover the dwarf planet Ceres. One is that he adjusted the orbit parameters to minimize the sum of the squared error between the observed measurements and the model's elliptical orbit predictions, allowing him to improve his estimate of the orbit parameters. The second is that he forced the model to be an elliptical orbit. Using only elliptical orbits, he eliminated some wilder variations available in straight line and circular motion. Ceres would move in an elliptical orbit if it was moving under the gravitational influence of only the sun. However, Ceres is moving under the gravitational influence of the planets as well as the sun which perturbs it to wobble about an approximately elliptical solar orbit. An elliptical orbit model is better than either a straight line or a circle model—but it remains imperfect.

Once we allow for noise in measurements, we must allow for the fact that exact fits are no longer possible **or even desirable**. An exact fit to noisy data means that the model is fitting the noise. Each new data point will require an increase in complexity of the model. Noise means that repeated measurements of the same thing will not produce the same values, implying uncertainty in predictions and **violent** numerical behavior. We want to reduce the effect of noise on the model because we want to reduce the uncertainty in our predictions. Models can still predict from noisy data, but they **must not** fit perfectly.

Overfitting

When the uncertainty in predictions becomes less than the uncertainty in data, overfitting should be considered as a possibility. Overfitting is adjusting the model to fit the data exactly, even though we know that data to be uncertain. In overfitted models noise has large effects on model predictions.

"When the uncertainty in predictions becomes less than the uncertainty in data, overfitting should be considered as a possibility. Overfitting is adjusting the model to fit the data exactly, even though we know that data to be uncertain."

For example, machine learning is applied to a physics journal corpus to develop author profiles.

Overfitting leads the AI to assign extremely granular concepts to authors (e.g., "Quantum entanglement" rather than just "Physics"). The highly precise sub-topic leads the machine to NOT suggest this author's submission for review because they are not an expert in "Physics," rather, they are an expert in a highly specialized area of physics.

Overfitting is a constant challenge with any machine learning task. Because of the neural network basis of machine learning, and the fact that an overly complex model will often fit the same data "better," we must constantly be on guard against overfitting, balancing the errors associated with measurements with errors associated with overfitting. ∎

5 Ways Text Analytics and NLP Make Internal Search Better

By Sean Coleman, CTO and Chief Customer Officer, BA Insight

Implementing AI-driven internal search can significantly impact employee productivity by improving the overall enterprise search experience. It can make internal search as easy and user-friendly as internet search, ensuring personalized and relevant results.

Following are some solutions that can accomplish this:

Semantic Search

Semantic search has meaningfully impacted the usability of the web by understanding the intent and contextual meaning of the words a person is using, resulting in the delivery of much more relevant results. Until now, semantic search has not been available to internal search users for multiple reasons:

- Numerous disparate systems in an organization make implementing semantic search across them a daunting task.
- The majority of critical information is buried within documents.
- Applications are not integrated and there is no single search interface.

Semantic search requires the following to work with internal search:

- A user interface that takes advantage of natural language query to enable users to "search as they speak."
- Indexing/ingesting connectors from third parties such as BA Insight for critical systems to create a single unified index.
- NLP platforms such as Microsoft Cognitive Services, Google Cloud, Amazon Comprehend or Open Source to provide a deep understanding of the concepts within documents to add intelligence to the index.

The following are some opportunities that the above strategy enables:

- Use of natural language to search across multiple systems rather than using different user interfaces and search functionalities for each system.
- Finding the right documents based on concepts within them as opposed to their filenames/titles.
- Finding images or documents that contain images using the additional metadata.
- Discovery of videos through automatic creation of transcripts.

When done correctly, semantic search delivers highly relevant search results by providing them based on user intent as opposed to keyword search, where there is no understanding of user intent.

Machine Learning

Machine learning works well on the web as internet traffic generates large amounts of data that can be used to understand user behavior patterns and make relevant recommendations. This is not the case with internal search because the amount of consolidated data is limited and often stored in multiple applications.

Connecting the most important internal systems to a single index with a single UI solves this challenge. It makes it possible to capture search data and take advantage of machine learning to improve the user's search experience and make it more internet-like. Examples include:

- "Search as you type," which presents information based on previous searches and content suggestions based on location, project, department, etc.
- Automatic correction of search queries based on previous productive searches.
- "Viewed by others" personalizes users' search results based on information about them such as their location, department or interests.

A single index that unifies vast amounts of organizational data enables machine learning to take advantage of user search patterns and proactively provide recommendations and results that improve relevancy, similar to that on the web.

Content Intelligence

Content intelligence provides employees with critical knowledge locked within documents, including the reason they exist and their most important attributes. Imagine a dashboard that provides:

- Information about the file itself (who created it and when, who last edited it and when, number of versions, etc.)
- The textual content of the document (key concepts, document summary, named entities, etc.)
- Multimedia information (images and videos, including identified text and transcripts, objects, etc.)
- Taxonomy information (company categories, specific terms, data classification, etc.)

Imagine a search experience that provides users with access to important information about the documents that show up in their search results. Taking advantage of AI technologies such as natural language processing, image/video processing, extracted keywords and phrase search can deliver important intelligence about documents, without the need to open and read them. Although this can be done individually with each system, a simpler approach is to implement it as part of enterprise search with a single index and single UI.

Intelligent Search Bot

Search bots are all over the internet, but not many are used for internal search. A search bot combines a digital assistant-type user interface with natural language processing, a range of text analytics capabilities, and machine learning to act as a search assistant or even replace the search bar so users can find relevant information faster than using keyword search.

Below are scenarios in which a search bot can help users quickly find relevant information:

- Too many results:
  - The bot recommends alternate queries or refiners to narrow down the result set, even showing the number of results per option.
- Too few results:
  - The bot recommends alternate queries or the removal of refiners to deliver better results.
- Conversations:
  - The bot recognizes the question the user is asking in natural language and converses to fully understand what is desired to provide a precise result. For example, a user may be looking for support cases and the bot can ask the desired priority level to narrow results.

Bots automate multiple search processes, providing assistance for even the most novice users so they can find the right information quickly.

Sentiment Analysis

Sentiment analysis is used heavily in customer-facing applications. However, the same doesn't apply to employee-facing applications because the majority of information inside organizations is captured in documents, for which sentiment is rarely relevant. Content types for which sentiment analysis can be meaningful include:

- Social technologies like Yammer, Teams and Workplace by Facebook.
- Emails and other electronic communications.

Your employees are talking, using collaboration platforms such as Yammer and Workplace by Facebook, but how productive are those conversations? Sentiment analysis helps you improve employee engagement and morale by identifying, and taking action to correct, areas of employee dissatisfaction. ∎