WHITEPAPER

Text Analytics – Categorization and Concept Topics

Read Between The Lines

LEXALYTICS

Angoss

Predict. Act. Perform.

Categories/Tags/Topics/Classification

There are many different ways to say the same thing. For example, if you see the following text, 'I wonder who will win in the California mid-term congressional elections?', chances are you are going to want to classify or associate it with a topic named "Politics". How do you sort content into different buckets?

Categorizing or "bucketing" content is common, as it's a feature of many of the popular and highly trafficked websites. Sites like Google News (http://news.google.com), The Wall Street Journal (http://www.wallstreetjournal.com) and TripAdvisor (http://www.tripadvisor.com) all use various categorization techniques to segment content on their sites.

Lexalytics Salience Engine has 3 methods that you can employ to accomplish bucketing:

- Query Topics
- Model-Based Classifiers
- Concept Topics

Query Topics

Query Topics are easy to understand; they are simply Boolean queries that, if matched, will associate a piece of content with a particular topic. To correctly classify the above example, you would probably have a query topic that looked like this:

Politics	elect* OR congress* OR (president NOT ceo) OR senate* OR representative*

Query topics are clear and simple to use. They are completely transparent but only work on strings and do not interpret what you meant - in the above example, we use "president NOT ceo" to try to eliminate any mentions of business news from contaminating political topics.

If you know the exact set of words that you want to look for, then query topics are great (for example, if you were looking for all occurrences of the word "iPhone"). However, sometimes it's not clear what words you want, and words may have different meanings in different contexts.

Lexalytics provides the following operators: AND, OR, NOT, NEAR (where NEAR has to be qualified with a distance – so you can say something like "Aruba NEAR25 wireless" – which would only count as a hit if Aruba was within 25 words of the word "wireless".

Model-Based Classifiers

Model-based classification works best on longer form content (at least a page in length) and when you're trying to classify the content into relatively wide, well-separated buckets.

This works by collecting a significant set of training content for each of your buckets, and allowing the classification model to select words that are statistically significant with respect to that particular bucket. This is a fairly simple process for the user, but there can be varying amounts of manual work needed to refine the model.

Let’s say you wanted to classify content about diseases and you gathered a bunch of content about on the subject. Each disease may make up only a few documents, so the model will probably rely on other words within the content. It turns out that the word "with" actually occurs to a high level of frequency in content about diseases (as in:you're diagnosed "with" something). This method of categorizing content is not inuitive, and shows the strength of models.

Concept Topics

Concept Topics is a new method that offers the best characteristics of both Query Topics and Model-Based Classifiers, as other types of categorizers require lots of effort to set up and/or are inherently brittle.

A conservative estimate for deployment of a query based categorizer with 100 buckets is probably on the order of 50 hours to get it up and running. This is followed by ongoing efforts to ‘tweak’ your category definitions as that one piece of ‘rogue’ content gets misclassified.

Lexalytics’ Concept Topics are designed to reduce the burden of this configuration through the use of the new Concept Matrix that has been generated from all of the content in Wikipedia™.

Salience 5.0 release ships with a number of sample Concept Topics. Here are 2 of them. The words next to "Food and Agriculture" are literally all there is to the definition of the Concept Topic:

Agriculture	farming, agriculture, farmer
Food	food, meals, vegetable, meat, fruit

Consider the following sentences and you can see how they match to each of the concept topics:

	Food	Agriculture
I like chicken.	0.58	No match
I like chickens.	No match.	0.71
I like to eat chickens.	0.59	0.51

Here are a few other examples from the Salience 5.0 release:

Aviation	aviation, airplane, flying
Banking	banking, bank, mortgage, checking, savings
Beverages	beverage, alcohol, soda
Biotechnology	biotech, biotechnology, applied_biology, gene_therapy, genetic_engineering
Business	business, management, executive, company, shareholder, mba
Crime	crime, murder, arrested, theft, burglary, criminal, arraignment
Disasters	disaster, tornado, earthquake, volcano, meteor, apocalypse, explosion, devastation
Economics	economics, economist, GDP, game_theory, demand_curve

The sentence, "American Airlines had to announce a gate change." correctly categorizes to Aviation, even though the word "Airline" doesn't occur anywhere in the aviation category.

Concept Topics will revolutionize categorization. Read on if you'd like to understand how they are easier to use, than alternative techniques.

Concept Topics vs. Query-Based Categorizers

To see how Concept Topics work—and why they are so important—let's look at a very simple example. We will build a very simple categorizer for a single bucket (Travel) and try tocorrectly classify the following review from TripAdvisor:

> *We were on the ship by 11:45. About 10 minutes after my VIP parents. Went to Lido deck for lunch. It was hard to find a table for ten or two table near each other so we went to the back of the ship near the pizza. The ship was full and you could tell. Very crowded. We got to our room at 1:30 met Edguardo our room steward. Loved him got everyone's name but had a hard time with mine and called me misses all week. The room was the same as on the Splendor which we did two years ago. 4 of us had plenty of room and loved the balcony. We had early seating in Washington Dining room 3rd floor in the middle so no view for us. I wanted to do anytime dining but with ten it wouldn't of worked. The boys and us went to club sign up. My 14 year old was going to be 15 in September and I wanted him moved to club O2 they said to write it on the form. He was able to switch no problem. Club started at 9:00pm. They had a great time all week and came home 12:30-1:00 every night.*

	Query based Method	Concept Topics
Definition	Motel OR hotel OR show OR resort OR pool OR travel OR vacation	Travel, Tourism
Time to define	~ 5 minutes	~ 10 seconds
Succeed / Fail	Fail	Succeed

It would be easy to modify the basic categorization query to make it hit for the sample review, but the problem is that you'd have to test and refine this definition for quite a long time before you got a sufficiently broad query for Travel where you would not miss almost as many as you hit.

The Concept Topic on the other hand required minimal time and effort to create, and it worked the first time. It worked because the Concept Matrix understands that words like cruise and food and entertainment are related to Travel, so it is reasonable to categorize the review into Travel.

Concept Topics vs. Model-Based Classifiers

Considering the same review text, it is perfectly reasonable to believe that one could build a model-based classifier that would do a great job of finding tourism-related content. But it would take significantly less time to simply configure a Concept Topic. And what if you want to change a definition by a bit? You don't have to 'retrain' a whole Concept Topic, you just simply add and subtract from the definition.

Where Concept Topics Do Not Work Well

Given the ease of defining Concept Topics, and the high degree of accuracy in broad topic areas like Food and Travel and Business, it would be easy to assume that Concept Topics are a magic bullet for any and all categorization problems.

While they represent a huge advance in categorization technology, they do have some limitations. If for example you were conducting a very detailed, low-level categorization of drug interaction reviews and wanted to bucket content by drugs for disease classes (for example Liver Cancer Drugs), then Concept Topics wouldn't get you there.

In these detailed cases a query based or training based model will be more effective and will require less effort than trying to define a broad/general category.

Given the recent development of this new technology, that has profound implications for how content is categorized, we are still working to understand all of its limitations and continue to aim to improve its usefullness.

While not perfect for every situation, Concept Topics are a significant enhancement for general-purpose categorization as theysolve the very difficult problem of bucketing into generic high-level categories that are difficult to handle because of the breadth of things that fall into them.

In other words, if you're doing any categorization, you're going to want to try Concept Topics.

About Angoss

Angoss is a global leader in delivering predictive analytics to businesses looking to improve performance across risk, marketing and sales. With a suite of big data analytics software solutions and consulting services, Angoss delivers powerful approaches that provide you with a competitive advantage by turning your information into actionable business decisions.

Many of the world's leading organizations in financial services, insurance, retail and high tech rely on Angoss to grow revenue, increase sales productivity and improve marketing effectiveness while reducing risk and cost. Headquartered in Toronto, Canada, with offices in the United States, United Kingdom and Singapore, Angoss serves customers in over 30 countries worldwide. For more information, visit www.angoss.com.

For more information, visit www.angoss.com

About Lexalytics, Inc.

Lexalytics, Inc. is a software and services company specializing in text and sentiment analysis for social media monitoring, reputation management and entity-level text and sentiment analysis. By enabling organizations to make sense of the vast content repositories on sources like Twitter, blogs, forums, web sites and in-house documents, Lexalytics provides the context necessary for informed critical business decisions. Serving a range of Fortune 500 companies across a wide spectrum, Lexalytics partners with industry leaders such as Endeca, ThomsonReuters, Radian 6 and TripAdvisor to deliver the most effective sentiment and text analysis solutions in the industry.

Angoss Corporate Headquarters
111 George Street, Suite 200
Toronto, Ontario M5A 2N4 Canada
Tel: 416-593-1122

Angoss European Headquarters Enigma House
30b Alan Turing Road The Surrey Research Park
Guildford, Surrey GU2 7AA Tel: +44 (0) 1483-661-661
www.angoss.com

Angoss
Predict. Act. Perform.