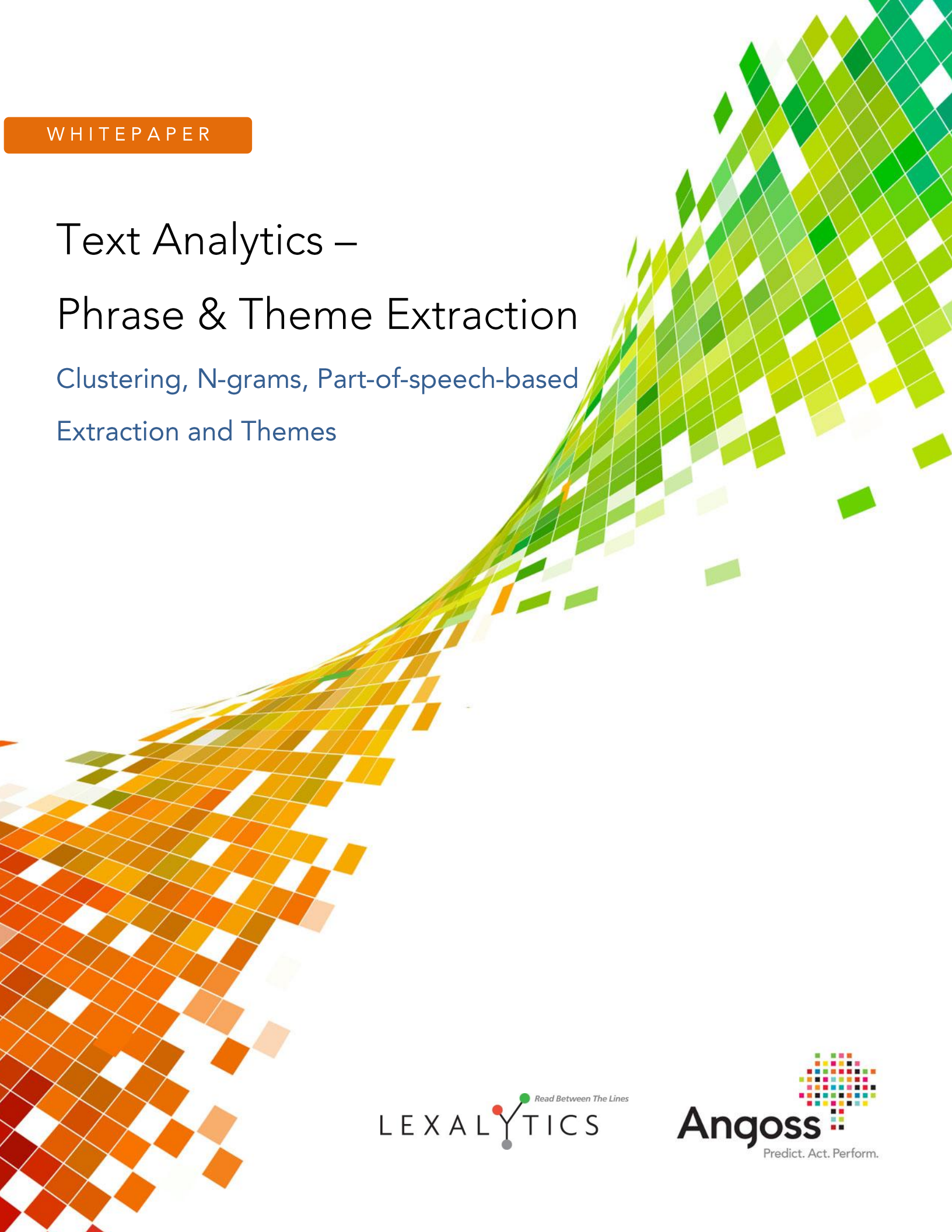
WHITEPAPER

Text Analytics – Phrase & Theme Extraction

Clustering, N-grams, Part-of-speech-based Extraction and Themes



LEXALYTICS
Read Between The Lines

Angoss
Predict. Act. Perform.

WHITEPAPER

Table of Contents

Angoss
Predict. Act. Perform.

Introduction

Once you know who (entities) is being discussed, the natural next step is to understand the context and content of the conversation.

This paper focuses on nouns and noun phrases. To be more specific, it's those nouns that you're not getting to through entity extraction. Entity extraction deals, roughly speaking, with proper nouns. Here we are considering those entities that, for the most part, are not proper nouns. There may be proper nouns that are picked up as part of the "noun analysis" that you're doing, but that is because they were not identified as an entity.

Consider the following sentence. It's politically controversial, but gives a good example of how important it is to separately recognize the entities and the context.

"*President Obama did a great job with that awful oil spill.*"

Entity extraction will give you "President Obama" as a person. Sentiment analysis will note a positive sentiment pointed back towards the person "President Obama". However, without understanding the additional nouns, you'll have no idea of the context in which President Barack Obama is receiving praise.

And so, other than a vague positive sentiment, you don't really know anything; as opposed to knowing that some author (or someone being quoted by some author) is giving thumbs up to President Barack Obama's oil spill handling skills.

Extracting non-entity phrases is an excellent next step to greater understanding of the content.

WHITEPAPER

Mechanical Techniques for Phrase Extraction

We're going to talk about 4 computational techniques for extracting phrases: **clustering**, **N-grams**, **noun phrase extraction** and **themes**.

Clustering

Clustering dynamically creates topic categories across a set of content. Clustering works by simultaneously examining many documents and automatically extracting a set of phrases that best represent the relationships between the documents. The phrases are typically some sort of n-gram (see the next section for N-grams).

Clustering is less useful than other techniques for ongoing analysis of content in that it is designed to show a snapshot when presented with some large data set. The results of the clustering will change as documents are added and/or deleted from the dataset.

Clustering is most useful as a navigational technique to augment search. Clustering can also be used as an analytical technique for looking at a large set of data and "getting the lay of the land". As such, clustering algorithms are typically optimized around how many documents can be clustered in a given period of time (say, the amount of time you're willing to wait for search results).

The main problem with clustering is that its very nature requires more than one piece of text, and the addition of more text changes the clusters. In order to do things like "emerging topic detection", you need to analyze a single piece of text in isolation from other pieces of text and then store the results for that text. Then when you get another piece of text, analyze it and store the results from that. This way you have results that you can trend over time.

You can still "cluster" when using results from text processed a piece at a time, but you can also do many other interesting operations that aren't easy with pure clustering.

Advantages of Clustering

- Very low latency processing of many thousands of documents for navigational purposes

Disadvantages of Clustering

- Not useful for trending
- Any change in documents changes the clusters
- Limited to words that appear in the text

N-Grams

N-grams are combinations of 1 or more words. For example, 1: monogram, 2: bi-gram, 3: tri-gram, etc. Rarely is it more than 3, unless looking for a specific slogan or turn of phrase. Words are not taken from any part of speech class, so, you're going to get any and all strings.

Monograms vs. bi-grams vs. tri-grams - Consider the phrases: "*crazy good*" and "*stone cold crazy*"

	Mono-grams	Bi-grams	Tri-grams
Phrases Extracted (crazy good, stone cold crazy)	crazy (2) cold good stone	crazy good cold crazy stone cold	stone cold crazy
Phrases Extracted (President Obama)	a awful did great job obama oil president spill that with	a great awful oil obama did a great job job with obama did oil spill president that awful with that	a great job awful oil spill obama did did a great great job with job with that obama did a president obama that awful oil with that awful

Results:	Not specific enough	Just right	Very specific, misses important phrase
	Generally not used for "phrase extraction" good for other things	Most often used	Used, gives very specific phrases

N-grams and stop words

The biggest problem with n-grams as phrase extraction is that it is a promiscuous algorithm.

Stop words let you make a list of terms to exclude from analysis. Classic stop words are things like: a, an, the, of, for, and… In addition to these very common examples, each domain has a set of words that are statistically too common to be interesting.

With most stop lists, all of the words "crazy, good, stone, cold" would probably make it through. Unless, perhaps, you were working on data for the "Cold Stone Creamery" (for those not in the USA, that's an ice cream parlo(u)r.), and you'd stopped the words in your name.

Now, it's important to note that if you "stopped" the phrase "cold stone creamery" that's very different than stopping "cold", "stone", and "creamery", as follows:

In the "cold stone creamery" case, if you got the phrase "cold as a fish", that phrase would make it through and be decomposed into n-grams as appropriate.

In the "cold", "stone", and "creamery" case, if you got the phrase "cold as a fish", that phrase would be chopped down to just "fish" (as most stop lists will also have the words "as" and "a" in them along with "cold", "stone", and "creamery".

N-gram stop words generally stop entire phrases in which they appear. For example, the phrase "for example" would be stopped if the word "for" was in the stop list (which it generally would be).

Advantages of N-grams

- You'll catch everything that you don't stop out, without any regard to parts of speech or anything else
- Computationally simple, easy to conceptually understand

Disadvantages of N-grams

- Promiscuous: requires long list of stop words to be interesting
- Simple count does not necessarily give an indication of "importance" to text or of its importance to an entity.
- Limited to words that appear in the text

Noun Phrase Extraction

Noun phrases are part of speech patterns that include a noun. They can include whatever other parts of speech make sense, and can include multiple nouns.

As a consequence of English language ordering, a noun generally ends the phrase.

Some common noun phrase patterns are:

- Noun
- Nouns
- Adjectives Noun
- Verb (Adjectives) Noun

Note that there is absolutely no reason why you can't have verb phrases or whatever other part of speech patterns you care to. However, nouns are most generally useful to understand the context of a conversation – if you want to know "what" is being discussed. Verbs help with understanding what those nouns are doing to each other, but it simplifies things a lot to just consider and work with noun phrases.

This does take into account parts of speech. Many stop words are stopped simply because they are a part of speech that is uninteresting from a statistical standpoint of understanding meaning. Because you're being very specific about classes of words that are interesting, most common stop words are instantly eliminated automatically. Stop lists can also be used with noun phrases, but it's not quite as critical to use them as it is with n-grams.

Noun phrase extraction would provide both phrases (assuming appropriate patterns).

Input Phrases	Extracted Phrases
crazy good, stone cold crazy	crazy good stone cold crazy
President Obama	great job awful oil spill

Consider this article:

Yahoo wants to make its Web e-mail service a place you never want to -- or more importantly -- have to leave to get your social fix.

The company on Wednesday is releasing an overhauled version of its Yahoo Mail Beta client that it says is twice as fast as the previous version, while managing to tack on new features like an integrated Twitter client, rich media previews and a more full-featured instant messaging client.

Yahoo says this speed boost should be especially noticeable to users outside the U.S. with latency issues, due mostly to the new version making use of the company's cloud computing technology. This means that if you're on a spotty connection, the app can adjust its behavior to keep pages from timing out, or becoming unresponsive.

Besides the speed and performance increase, which Yahoo says were the top users requests, the company has added a very robust Twitter client, which joins the existing social-sharing tools for Facebook and Yahoo. You can post to just Twitter, or any combination of the other two services, as well as see Twitter status updates in the update stream below. Yahoo has long had a way to slurp in Twitter feeds, but now you can do things like reply and retweet without leaving the page.

If asynchronous updates are not your thing, Yahoo has also tuned its integrated IM service to include some desktop software-like features, including window docking and tabbed conversations. This lets you keep a chat with several people running in one window while you go about with other e-mail tasks.

http://edition.cnn.com/2010/TECH/web/10/27/yahoo.faster.email.cnet/index.html

There are scads of noun phrases in this article. Which ones are the important ones? You can simply frequency count them; but that doesn't give any indication of whether these were lexically important or not.

Advantages of Noun Phrase Extraction

- Restricts to phrases matching certain part of speech patterns, fewer stop words needed

Disadvantages of Noun Phrase Extraction

- No way to tell if one noun phrase is more contextually relevant than another noun phrase
- Limited to words that occur in the text

Themes

Themes are noun phrases with contextual relevance scores. Themes extract exactly as described above in noun phrase extraction. Once extracted, themes are then scored for contextual relevance using lexical chaining.

Lexical Chaining

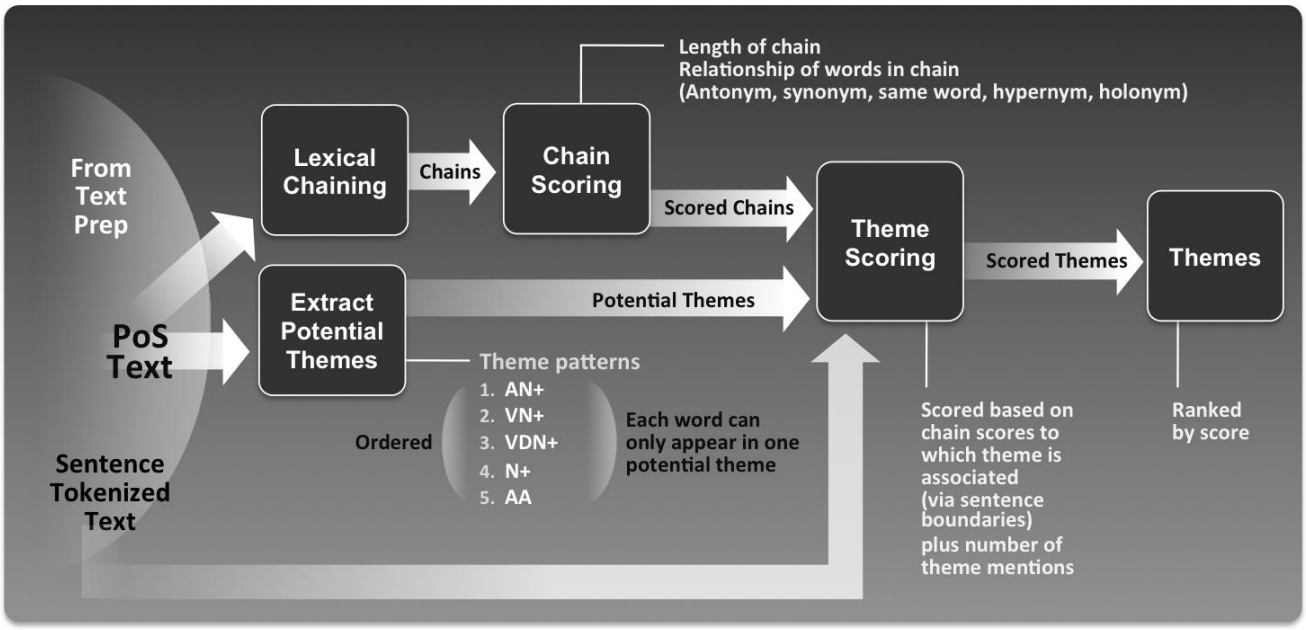
Lexical chaining relates sentences via thesaurally-related nouns. Consider the following:

I like beer. Miller just launched a new pilsner ale. But, because I'm a beer snob, I'm only going to drink pretentious Belgian beers.

Those 3 sentences are related through beer->ale->beers. Even if those sentences are not adjacent to each other in the text, they are lexically related to each other and can thus be associated with each other. The "score" of a lexical chain is directly related to the length of the chain and the relationships between the chaining nouns (same word, antonym, synonym, meronym, hyper/holonyms, etc.)

Theme Extraction and Scoring

First, potential themes are extracted based on the part-of-speech patterns. Then, the chains are scored, and themes that belong to the highest scoring chain (sentences chained together), get the highest scores. If there are fewer than 4 chains, the algorithm gracefully degrades to scoring purely by count.



"*crazy good*" and "*stone cold crazy*"

Noun phrase extraction would provide both phrases (assuming appropriate patterns). With theme extraction, their scores would be different depending on where they occurred in the text (e.g. if they were associated with a central theme or if they were associated with a tangential thread.).

"*President Obama did a great job with that awful oil spill.*"

Yields the same noun phrases as with straight noun phrase extraction (great job, awful oil spill). However, the score for each would be highly dependent on where this sentence fit in the grand scheme of things. Meaning, if there were further sentences that referenced concepts relating to oil, that will boost the score of the "oil spill" theme.

Consider the same article as above:

Yahoo wants to make its Web e-mail service a place you never want to -- or more importantly -- have to leave to get your social fix.

The company on Wednesday is releasing an overhauled version of its Yahoo Mail Beta *client that it says is twice as fast as the previous version, while managing to tack on new features like an integrated Twitter client, rich media previews and a more full-featured instant messaging client.*

Yahoo says this speed boost should be especially noticeable to users outside the U.S. with latency issues, due mostly to the new version making use of the company's cloud computing technology. *This means that if you're on a spotty connection, the app can adjust its behavior to keep pages from timing out, or becoming unresponsive.*

Besides the speed and performance increase, which Yahoo says were the top users requests, the company has added a very robust Twitter client, which joins the existing social-sharing tools for Facebook and Yahoo. You can post to just Twitter, or any combination of the other two services, as well as see Twitter status updates in the update stream below. Yahoo has long had a way to slurp in Twitter feeds, but now you can do things like reply and retweet without leaving the page.

If asynchronous updates are not your thing, Yahoo has also tuned its integrated IM service to include some desktop software-like features, including window docking and tabbed conversations. This lets you keep a chat with several people running in one window while you go about with other e-mail tasks.

http://edition.cnn.com/2010/TECH/web/10/27/yahoo.faster.email.cnet/index.html

In this case, the top 5 themes are:

Theme	Score
Cloud computing technology	4.11
Including window docking	2.976
Mail service	2.672
Top users requests	2.669
Rich media previews	2.635

You can see that those themes do a reasonable job of conveying the actual context of the article. The addition of contextual scoring information is hugely useful in determining what's really important in the text, and is useful to compare across many articles across periods of time (to see what's emerging, etc).

Specific to Lexalytics, themes also carry the advantage of being scored for sentiment. This is particularly important when considering a case like the President Obama sentence where it's important to be able to distinguish between the positive perception of the President and the negative perception of the theme "oil spill".

Advantages to Theme Extraction and Scoring

- Restricts to phrases matching certain part of speech patterns, more wheat from the chaff
- Scored based on contextual importance
- Sentiment scores/theme

Disadvantages

- Limited to words in the text (true for all algorithms)

Summary

Theme extraction and scoring provides a highly valuable combination of context scored noun phrases. The theme extraction algorithm degrades nicely with shorter content. There is nothing to prevent you from running multiple algorithms on your text, Lexalytics supports both n-gram and theme extraction.

The primary direction for future development is to reach beyond the boundaries of a single piece of text – right now, all themes are extracted exactly from words in this text. The next big challenge is to map these themes to higher-level concepts that can then be easily "rolled up" and compare across different texts that use different words to denote the same concepts.

In the meantime, themes still provide an excellent view of the context of conversations, and are useful on all lengths of content – from tweets up to hundred-page secondary research reports.

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