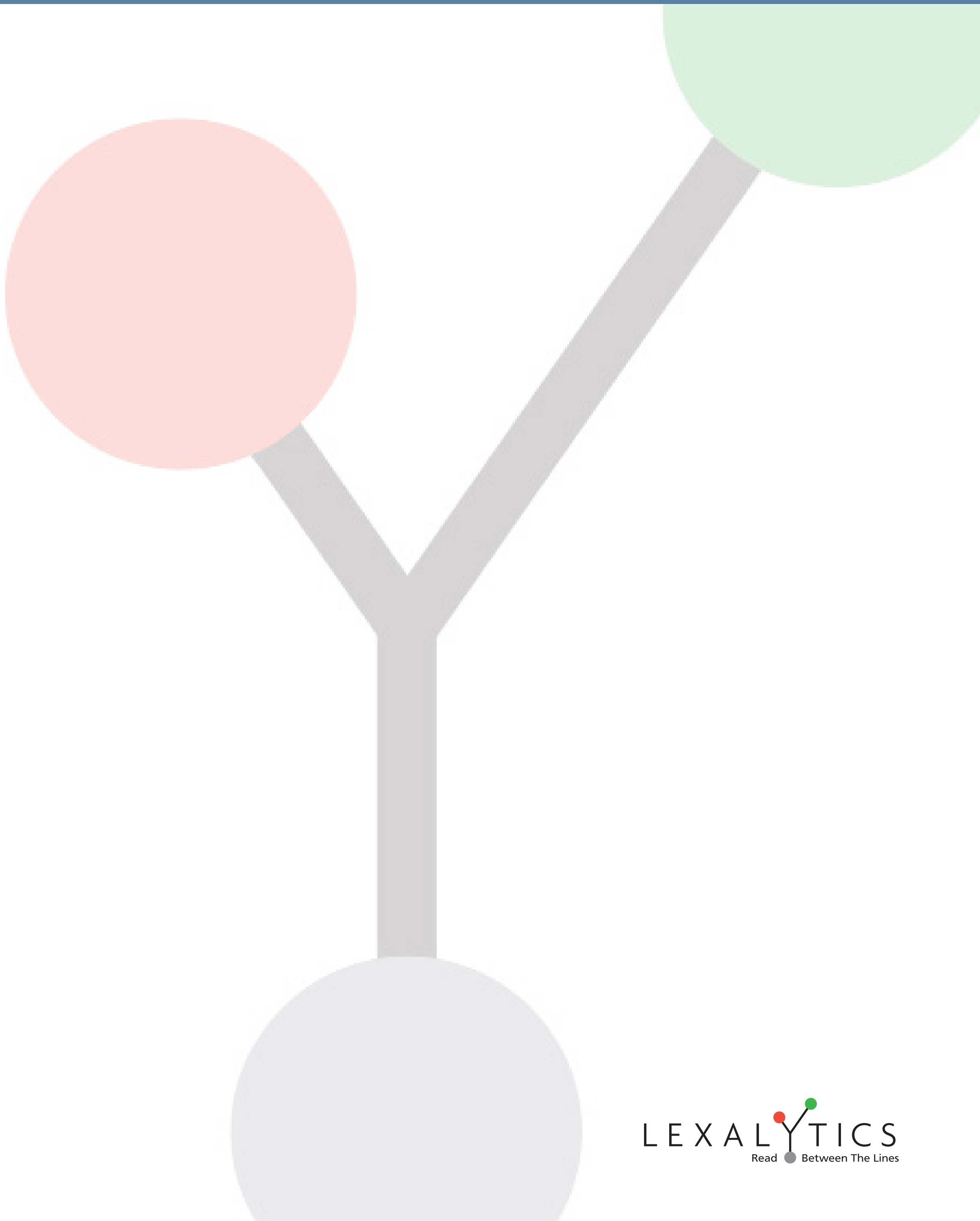


Categorization of Text

Whitepaper

LEXALYTICS
Read Between The Lines

Categorization is a core function of text mining software, and a key component in sentiment analysis of text

It's hard to measure the content and tone of consumer conversations without knowing the topics they're talking about: this is where categorization comes in. Lexalytics categorizes your input content to deliver actionable insights into the entities, themes, and ideas that fill your customers' conversations, and the frequency with which these mentions occur. For example, the following text:

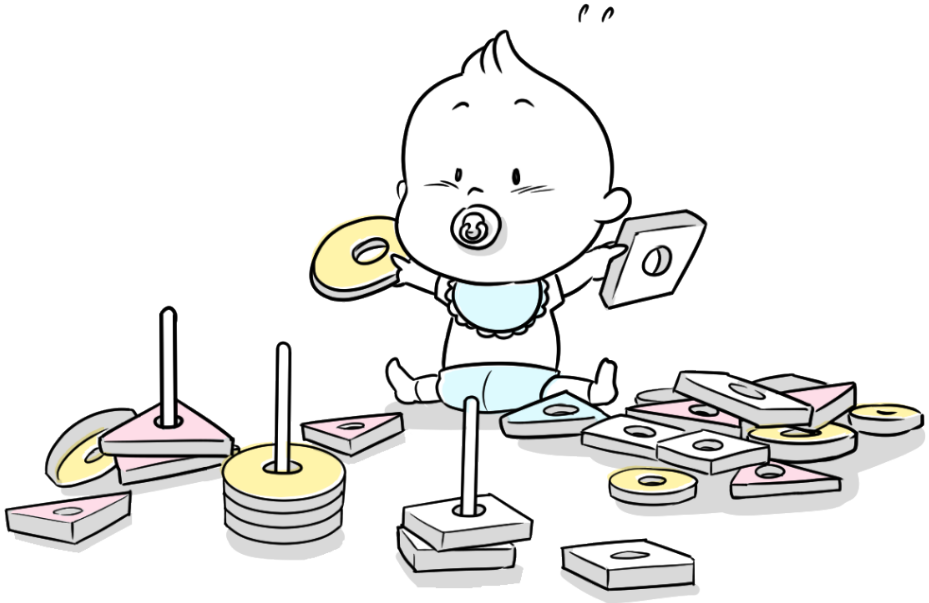
"I wonder who will win in the California mid-term congressional elections?"

will be classified or associated with a topic called "Politics". If your input contains many documents similar to the above, our text mining tools will show you that your customers are very interested in politics, without you ever having to read a single document yourself. In fact, if you tune the classification models further, they'll show just which area of politics people are discussing: in this case, the California congressional mid-terms.

Everyone categorizes content, but few text analytics tools do it as well as Lexalytics. Our categorization processes are effective, reliable, and fully customizable, capable of showing you everything from the broader picture down to the minutiae that drive informed business decisions. We understand the importance of content categorization, so much so that we've developed three powerful methods for classifying.
They are as follows:

- Query Topics
- Model-Based Classifiers
- Concept Topics

We'll discuss them in that order.



Query topics are simple: they're search categories delineated by Boolean classifiers (AND, OR, NOT, WITH and NEAR).

When an item is found that matches a query topic's content, the item is sorted into that query topic's "bucket".

To create a politics category, you might use a query topic that looked like this:

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

In this case, the query topic is "Politics" and the tags that clarify its contents are "elect", "congress", "president", "senate", and "representative". This topic searches for phrases, entities, and themes that are associated with those specified words.

Query topics are clear and simple to use. They are surgical and completely transparent, but work on basic principles and contain little depth unless painstakingly refined. Query topics work well if you have an exact set of words to look for. For example, if you were looking for all occurrences of the word "iPhone", a query topic would be the method of choice.

Lexalytics provides the following Boolean operators for query topics:

- **AND**
- **OR**
- **NOT**
- **WITH**: where WITH must be qualified with another word
    - In the above, "Aruba WITH wireless" would only count mentions if Aruba was found in the same sentence as "wireless".
- **NEAR**: where NEAR must be qualified with a distance
    - In the above, "Aruba NEAR25 wireless" would only count mentions if Aruba was within 25 words of the word "wireless", regardless of how many sentences that encompassed.

However, query topics search for exactly what you request, not what you mean by your request. In the above example, "president NOT ceo" is included to prevent any mentions of business news from contaminating political topics. Without the clarifier, the computer wouldn't know if "president" referred to the governmental position or an executive of a company. In other cases, the words you look for may have different meanings in different contexts, or you can't define a set of words that embody everything you're looking for.
Query topics also take time to set up and refine: time that could be spent acting on insights that our text analytics have already delivered.

MODEL-BASED CLASSIFIERS

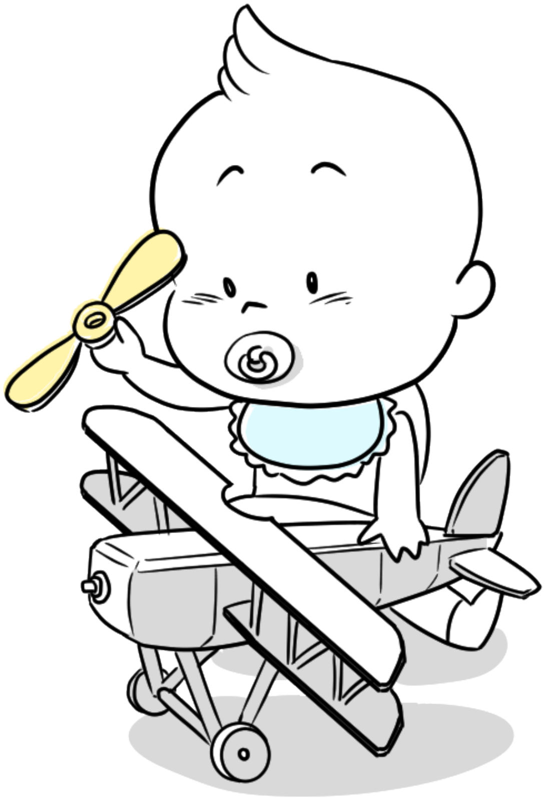
Our model-based classification system takes time to set up but requires less maintenance than query topics once configured.

Once you input a set of "training" content (Twitter posts, blog articles, etc.) for each category you create, this classification model identifies the words that are most statistically significant for that particular category. **This is a fairly simple process for the user, but allows for as much (or as little) customization as you need.**

Model-based classification works best on medium-length content (around a page in length), when you are trying to classify the content into relatively wide, well-separated categories.

Say you have gathered a bunch of content about diseases and wish to classify your documents. Mentions of individual diseases make up only a few documents per disease, so the model relies on other words: as it turns out, the word "with" has a high probability of containing content about diseases (as in you're diagnosed "with" something). A human putting together a query topic for diseases would not have made this connection; this is a completely counter-intuitive result, and shows the strength of models developed by a computer.

Model-based classifiers work well with enough tweaking and in the right situations, but are not always appropriate for a given situation. For more information, please contact us at Lexalytics directly.



Query-based categorizers are simple and functional, but a system with 100 “buckets” can take as many as 50 hours just to get up and running

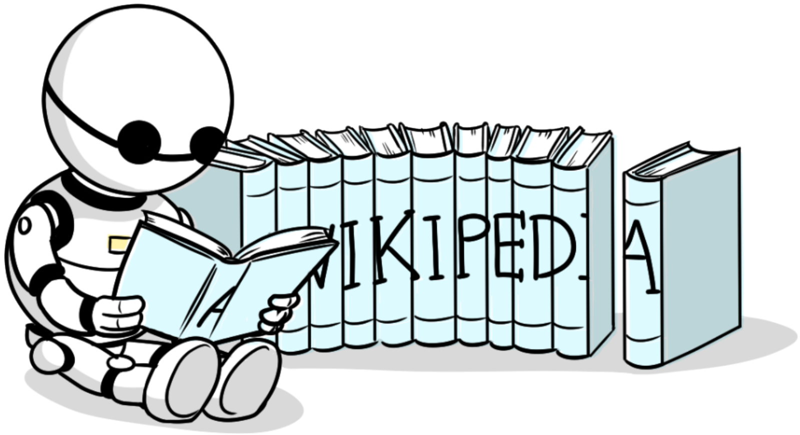
-- and once established, you’ll need to continually tweak your category definitions as pieces of rogue content get misclassified. Model-based classifiers are powerful and customizable, but work best for broader categorization of medium-to-long content. Both systems take time to configure, and are not always the best option for your needs.

Lexalytics recognized this gap in modern text analytics offerings, so we developed the most powerful text mining tool on the market: the Lexalytics Concept Matrix, generated from the top 640,000 articles on Wikipedia™.

That’s right, more than 640,000 Wikipedia™ pages.

The Concept Matrix utilizes the links between Wikipedia™ articles to develop a matrix of associations. Think of it this way: the “Politics” article on Wikipedia™ contains dozens of links to other articles, all related to the concept of politics. Each of those articles, in turn, contains links to other pages, and so on. The closer an article is in this chain to the original topic, the more closely related it is to that topic, and the stronger the association is.

We took those 640,000-plus articles and found 1,100,000 words and bi-grams (two-word combinations) that have 56,000,000 links between them. Based on these links, we developed a matrix of associations and relations that form the basis for a number of interesting technologies. Most of these technologies are still under lock-and-key, so for now we’ll discuss one major application of this technology: **Lexalytics Concept Topics.**



Concept Topics represent, well, concepts. They’re larger, bigger-picture categories that we developed from our Concept Matrix, and contain a number of keywords relating to the larger concept. Each of those keywords is associated with more keywords, and so on. The strength of a document’s association with a particular Concept Topic depends on how close its keywords are to the original topic.
Here’s an example of how this works:
The Salience 6 release package ships with a number of example Concept Topics, including the two below. The words next to “Food and Agriculture” represent the first level of the Concept Topic:

- **Food**: food, meals, vegetable, meat, fruit
- **Agriculture**: farming, agriculture, farmer

The number next to each of the following sentences represents the strength of that sentence’s match with the given concept topic (on a scale of 0 to 1).

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

Here are some other Concept Topics from the Salience Six 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

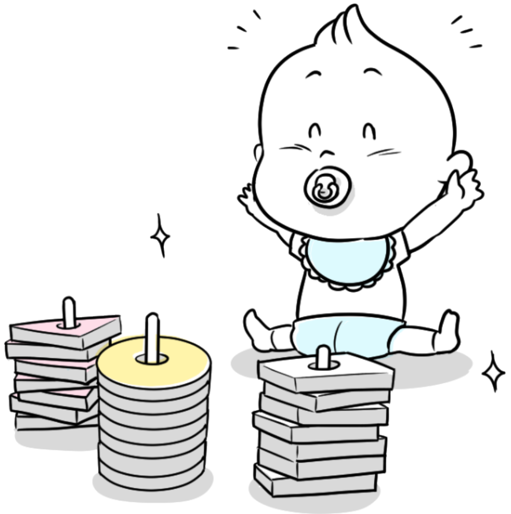
So with our system, the sentence “American Airlines had to announce a gate change” correctly categorizes to Aviation, even though none of those words directly occur in the Aviation category.

SUMMARY

Lexalytics provides the most powerful, reliable content categorization systems available. Our classification techniques deliver meaningful information on the themes and topics that your consumers are focusing on — so that you can act immediately, safe in the knowledge that you are making an informed decision to further your business.

Lexalytics® is the industry leader in translating text into profitable decisions. Lexalytics deploys state-of-the-art on-premise and in-the-cloud text and sentiment analysis technologies that process billions of unstructured documents every day globally, transforming customers' thoughts and conversations into actionable insights. The on-premise Salience® and SaaS Semantria® platforms are implemented in a variety of industries for social media monitoring, reputation management and voice of the customer programs.

Lexalytics is based in Boston, MA, and has offices in the U.S. and Canada. For more information, please visit www.lexalytics.com, email sales@lexalytics.com or call 1-617-249-1049. Follow Lexalytics on Twitter, Facebook, and LinkedIn for updates and insights on the world of text mining.



LEXALYTICS
Read Between The Lines

320 Congress St
Boston, MA 02210

General Inquiries
1-800-377-8036

Sales
sales @lexalytics.com
1-800-377-8036 x1

International
1-617-249-1049