

Blog (<https://www.lexalytics.com/lexablog/>) › Technology (<https://www.lexalytics.com/lexablog/category/technology/>) ›

Context Analysis in NLP: Why It's Valuable and How It's Done



(<https://twitter.com/intent/tweet?url=https%3A%2F%2Fwww.lexalytics.com%2Fflexablog%2Fcontext-analysis-nlp&text=Context+Analysis+in+NLP%3A+Why+It%26%238217%3Bs+Valuable+and+How+It%26%238217%3Bs+Done>)

## Context Analysis in NLP: Why It's Valuable and How It's Done



(<http://www.facebook.com/sharer.php?u=https%3A%2F%2Fwww.lexalytics.com%2Fflexablog%2Fcontext-analysis-nlp>)



February 18, 2019 (<https://www.linkedin.com/shareArticle?mini=true&url=https%3A%2F%2Fwww.lexalytics.com%2Fflexablog%2Fcontext-analysis-nlp&title=Context+Analysis+in+NLP%3A+Why+It%26%238217%3Bs+Valuable+and+How+It%26%238217%3Bs+Done>)

🕒 11 m, 50 s

Context analysis in NLP involves breaking down sentences to extract the n-grams, noun phrases, themes, and facets present within. In this article, I'll explain the value of context in NLP and explore how we break down unstructured text documents to help you understand context.

1. The Value of Context in NLP
2. The Foundations of Context Analysis
3. Using N-grams for Basic Context Analysis
4. Using Stop Words to Clean up N-gram Analysis
5. Noun Phrase Extraction
6. Themes and Theme Extraction with Relevancy Scoring
7. Facets: Context Analysis Without Noun Phrases
8. Summary & Further Reading

### The Value of Context in NLP

The goal of natural language processing (NLP (<https://www.lexalytics.com/lexablog/what-is-natural-language-processing/>)) is to find answers to four questions: *Who is talking? What are they talking about? How do they feel? Why do they feel that way?* This last question is a question of **context**.

Imagine a pundit's tweet that reads:

*"Governor Smith's hard-line stance on transportation cost him votes in the election."*

Entity-based [sentiment analysis](https://www.lexalytics.com/technology/sentiment-analysis) (<https://www.lexalytics.com/technology/sentiment-analysis>) of this sentence will indicate that "Governor Smith" is associated with negative sentiment.

But that's all you have: the knowledge that the Governor is viewed negatively. Wouldn't it be even more helpful to know *why*?

Performing theme extraction on this sentence might give us two results:

- “hard-line stance”
- “budget cuts”

Suddenly, the picture is much clearer: Governor Smith is being mentioned negatively in the context of a hard-line stance and budget cuts.

Notice that this second theme, “budget cuts”, doesn’t actually appear in the sentence we analyzed. Some of the more powerful NLP context analysis tools out there can identify larger themes and ideas that link many different text documents together, even when none of those documents use those exact words.

Remember: it’s not uncommon to find data analysts processing tens of thousands of tweets like this every day to understand how people feel. But without context, this information is only so useful. Contextual analysis helps you tell a clear, nuanced story of **why** people feel the way they do.

## The Foundations of Context Analysis

The foundation of context determination is the **noun**. Of course, this is true of [named entity extraction](https://www.lexalytics.com/lexablog/public-portal/entity-extraction) (<https://www.lexalytics.com/lexablog/public-portal/entity-extraction>) as well. But while entity extraction deals with proper nouns, context analysis is based around more general nouns.

For example, where “Cessna” and “airplane” will be classified as entities, “transportation” will be considered a theme (more on themes later).

Lexalytics supports four methods of context analysis, each with its merits and disadvantages:

- N-grams
- Noun phrases
- Themes
- Facets

Let’s start with the first and work our way down.

## Using N-grams for Basic Context Analysis

N-grams are combinations of one or more words that represent entities, phrases, concepts, and themes that appear in text. Generally speaking, the lower the value of “n”, the more general the phrase or entity.

N-grams form the basis of many text analytics functions, including other context analysis methods such as Theme Extraction. We’ll discuss themes later, but first it’s important to understand what an n-gram is and what it represents.

There are three common levels of n-gram:

- 1 word = mono-gram
- 2 words = bi-gram
- 3 words = tri-gram

To get an idea of the relative strengths and weaknesses of mono-grams, bi-grams, and tri-grams, let’s analyze two phrases and a sentence:

- “President Barack Obama did a great job with that awful oil spill.”

Here’s what we extract:

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

What conclusions can we draw?

First, the **mono-grams** (single words) aren’t specific enough to offer any value. In fact, monograms are rarely used for phrase extraction and context. Instead, they offer other value as entities and themes.

**Tri-grams** aren’t much use either, for the opposite reason: they’re too specific. If you happen to be searching for very particular phrases with a document, specificity can be useful. But tri-grams usually offer too narrow a lens to look through. In the end, tri-grams do get used for phrase extraction, but not as frequently as bi-grams.

That leaves us with **bi-grams**. If n-grams are bowls of porridge, then bi-grams are the “just right” option.

As demonstrated above, two words is the perfect number for capturing the key phrases and themes that provide context for entity sentiment.

But be warned: N-grams can come a lot “noise”. Phrases such as “with that”, which technically are bi-grams, offer no value in context determination, and do little more than clutter your view.

## Using Stop Words to Clean Up N-Gram Analysis

**Stop words** are a list of terms you want to exclude from analysis. Classic stop words are “a”, “an”, “the”, “of”, and “for”.

In addition to these very common examples, every [industry \(https://www.lexalytics.com/lexablog/5-industries-taking-advantage-text-analytics-2\)](https://www.lexalytics.com/lexablog/5-industries-taking-advantage-text-analytics-2) or vertical has a set of words that are statistically too common to be interesting.

Take the phrase “cold stone creamery”, relevant for analysts working in the food industry. Most stop lists would let each of these words through unless directed otherwise. But your results may look very different depending on how you configure your stop list.

If you stop “cold stone creamery”, the phrase “cold as a fish” will make it through and be decomposed into n-grams as appropriate.

If you stop “cold” AND “stone” AND “creamery”, the phrase “cold as a fish” will be chopped down to just “fish” (as most stop lists will include the words “as” and “a” in them).

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

Offers thematic insight at different levels (mono, bi-, tri-grams)

Simple, easy to conceptually understand

## Drawbacks to N-grams

Indiscriminate: requires a long list of stop words to avoid useless results

Simple count does not necessarily give an indication of “importance” to text or of its importance to an entity

## Noun Phrase Extraction

The form of n-gram that takes center stage in NLP context analysis is the **noun phrase**.

Noun phrases are [part of speech patterns \(https://www.lexalytics.com/technology/text-analytics#pos-tagging\)](https://www.lexalytics.com/technology/text-analytics#pos-tagging) that include a noun. They can also include whatever other parts of speech make grammatical sense, and can include multiple nouns. Some common noun phrase patterns are:

- Noun
- Noun-Noun..... -Noun
- Adjective(s)-Noun
- Verb-(Adjectives-)Noun

So, “red dog” is an adjective-noun part of speech pattern. “Cavorting green elk” is a verb-adjective-noun pattern.

Other part-of-speech patterns include verb phrases (“Run down to the store for some milk”) and adjective phrases (“brilliant emerald”).

Verb and adjective phrases serve well in [analyzing sentiment \(https://www.lexalytics.com/technology/sentiment-analysis#pos\)](https://www.lexalytics.com/technology/sentiment-analysis#pos). But nouns are the most useful in understanding the context of a conversation. **If you want to know “what” is being discussed, nouns are your go-to.** Verbs help with understanding what those nouns are doing to each other, but in most cases it is just as effective to only consider noun phrases.

Noun phrase extraction takes part of speech type into account when determining relevance. Many stop words are removed simply because they are a part of speech that is uninteresting for understanding context. Stop lists can also be used with noun phrases, but it's not quite as critical to use them with noun phrases as it is with n-grams.

### Advantages of Noun Phrase Extraction

Restricts to phrases matching certain part of speech patterns

Fewer stop words needed, less effort involved

### Drawbacks to 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 and Theme Extraction with Relevancy Scoring

Noun phrases are one step in context analysis. But as we've just shown, the contextual relevance of each noun phrase itself isn't immediately clear just by extracting them. This is where **themes** come into play.

In a nutshell: Themes are noun phrases with contextual relevance scores.

Themes are themselves noun phrases, which we identify and extract based on part of speech patterns. Then we score the relevance of these potential themes through a process called **lexical chaining**. Lexical chaining is a [low-level text analytics process \(https://www.lexalytics.com/lexablog/text-analytics-functions-explained\)](https://www.lexalytics.com/lexablog/text-analytics-functions-explained) that connects sentences via related nouns. (To learn more about lexical chaining, read this piece on [The 7 Basic Functions of Text Analytics \(https://www.lexalytics.com/lexablog/text-analytics-functions-explained\)](https://www.lexalytics.com/lexablog/text-analytics-functions-explained).)

Once we've scored the lexical chains, themes that belong to the highest-scoring chains are assigned the highest relevancy scores.

To demonstrate theme extraction for context analysis, let's use this old CNN article (originally at <https://edition.cnn.com/2010/TECH/web/10/27/yahoo.faster.email.cnet/index.html>):

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.

Via <https://edition.cnn.com/2010/TECH/web/10/27/yahoo.faster.email.cnet/index.html>

This article's top 5 themes are:

Theme	Score
Cloud computing technology	4.11
Including window docking	2.976
Mail service	2.672
Top users requests	2.66
Rich media previews	2.635

You can see that those themes do a good job of conveying the context of the article. And scoring these Themes based on their contextual relevance helps us see what's really important. **Theme scores** are particularly handy in comparing many articles across time to identify trends and patterns.

Lexalytics goes one step further by including **sentiment scores** for every theme we extract. This is key in differentiating between the opinions expressed towards different entities within the same body of text. Remember this sentence: "President Barack Obama did a great job with that awful oil spill."

Lexalytics' scoring of individual themes will differentiate between the positive perception of the President and the negative perception of the theme "oil spill".

### Advantages of Theme Extraction and Scoring

Restricts to phrases that match certain part-of-speech patterns

Scored based on contextual relevance and importance

Includes sentiment scores for themes

### Drawbacks to Theme Extraction and Scoring

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

Sometimes your text doesn't include a good noun phrase to work with, even when there's valuable meaning and intent to be extracted from the document. Facets are built to handle these tricky cases where even theme processing isn't suited for the job.

Think about this sentence: "The bed was hard."

There is no qualifying theme there, but the sentence contains important sentiment for a hospitality provider to know.

Now imagine a big collection of reviews. They may be full of critical information and context that can't be extracted through themes alone. This is where facets come into play.

We should note that facet processing must be run against a static set of content, and the results are not applicable to any other set of content. In other words, facets only work when processing collections of documents. **This means that facets are primarily useful for review and survey processing**, such as in Voice of Customer and Voice of Employee analytics.

Noun phrase extraction relies on part-of-speech phrases in general, but facets are based around "Subject Verb Object" (SVO) parsing. In the above case, "bed" is the subject, "was" is the verb, and "hard" is the object. When processed, this returns "bed" as the facet and "hard" as the attribute.

The nature of SVO parsing requires a collection of content to function properly. Any single document will contain many SVO sentences, but collections are scanned for facets or attributes that occur at least twice.

Here's an example of two specific facets, pulled from an analysis of a collection of 165 cruise liner reviews:

Facet	Positive Documents	Neutral Documents	Negative Documents
Ship	45	127	14
Food	36	44	0

Each of these facets, in turn, are associated with a number of **attributes** that add more contextual understanding:

Top 5 Ship Attributes For "Ship":		Top 5 Attributes for "Food":	
Attribute	Count	Attribute	Count
Beautiful	22	Excellent	14
Clean	8	Good	12
New	8	Great	10
Huge	6	Best	4
Nice	6	Fabulous	4

Judging by these reviews, this is a new ship with great food — the kitchen seems to be doing an excellent job.

Our facet processing also includes the ability to combine facets based on semantic similarity via our Wikipedia™-based Concept Matrix. We combine attributes based on word stem, and facets based on semantic distance.

Here's an example:

Enterprise			Enterprise/ Company	
Profitable	2		Growing	3
Growing	3		Acquired	2
Company			Bought	2
Fun	3		Enjoyable	2
Enjoyable	2		Fun	3
Acquired	2		Makes	2
Bought	2		Produces	2
Makes	2		Profitable	2
Produces	2		Scaling	2
Scaling	2		Employee	
Employee			Motivated	4
Motivated	4		Intelligent	3
Smart	3		Smart	3
Intelligent	3		Attrition	2
Turnover	2		Turnover	2
Attrition	2			

Facet  
Rollup →

Enterprise and Company are combined into a single facet, providing you with even richer information by combining the attributes from both.

## Summary & Further Reading

Business intelligence tools use natural language processing to show you who's talking, what they're talking about, and how they feel. But without understanding *why* people feel the way they do, it's hard to know what actions you should take.

Context analysis in NLP involves breaking down sentences into n-grams and noun phrases to extract the themes and facets within a collection of unstructured text documents.

Through this context, data analysts and others can make better-informed decisions and recommendations, whatever their goals.

[Download this article as a PDF white paper \(https://www.lexalytics.com/content/whitepapers/Lexalytics-WP-Context.pdf\)](https://www.lexalytics.com/content/whitepapers/Lexalytics-WP-Context.pdf)

[See how Gunnvant Saini used NLP to perform context analysis on news article headlines \(https://towardsdatascience.com/how-i-used-natural-language-processing-to-extract-context-from-news-headlines-df2cf5181ca6\)](https://towardsdatascience.com/how-i-used-natural-language-processing-to-extract-context-from-news-headlines-df2cf5181ca6)

[Glance through this 2016 MIT Technology Review perspective on context, language and reasoning in AI \(https://www.technologyreview.com/s/602658/context-language-and-reasoning-in-ai-three-key-challenges/\)](https://www.technologyreview.com/s/602658/context-language-and-reasoning-in-ai-three-key-challenges/)

[Read this Quora response for a slightly more technical perspective and more links \(https://www.quora.com/How-do-we-train-AI-to-understand-context-in-text\)](https://www.quora.com/How-do-we-train-AI-to-understand-context-in-text)

Categories: [Technology \(https://www.lexalytics.com/lexablog/category/technology\)](https://www.lexalytics.com/lexablog/category/technology)

Tags: [context \(https://www.lexalytics.com/lexablog/tag/context\)](https://www.lexalytics.com/lexablog/tag/context), [context analysis \(https://www.lexalytics.com/lexablog/tag/context-analysis\)](https://www.lexalytics.com/lexablog/tag/context-analysis), [context extraction \(https://www.lexalytics.com/lexablog/tag/context-extraction\)](https://www.lexalytics.com/lexablog/tag/context-extraction), [contextual \(https://www.lexalytics.com/lexablog/tag/contextual\)](https://www.lexalytics.com/lexablog/tag/contextual), [contextual analysis \(https://www.lexalytics.com/lexablog/tag/contextual-analysis\)](https://www.lexalytics.com/lexablog/tag/contextual-analysis), [determining context \(https://www.lexalytics.com/lexablog/tag/determining-context\)](https://www.lexalytics.com/lexablog/tag/determining-context), [facets \(https://www.lexalytics.com/lexablog/tag/facets\)](https://www.lexalytics.com/lexablog/tag/facets), [n-gram \(https://www.lexalytics.com/lexablog/tag/n-gram\)](https://www.lexalytics.com/lexablog/tag/n-gram), [n-grams \(https://www.lexalytics.com/lexablog/tag/n-grams\)](https://www.lexalytics.com/lexablog/tag/n-grams), [noun phrase extraction \(https://www.lexalytics.com/lexablog/tag/noun-phrase-extraction\)](https://www.lexalytics.com/lexablog/tag/noun-phrase-extraction), [noun phrases \(https://www.lexalytics.com/lexablog/tag/noun-phrases\)](https://www.lexalytics.com/lexablog/tag/noun-phrases), [technology \(https://www.lexalytics.com/lexablog/tag/technology\)](https://www.lexalytics.com/lexablog/tag/technology), [themes \(https://www.lexalytics.com/lexablog/tag/themes\)](https://www.lexalytics.com/lexablog/tag/themes)

## One Response to “Context Analysis in NLP: Why It’s Valuable and How It’s Done”



Wonderful article the best I have seen till date. Would have loved to see a Python sample project highlighting the concepts but great article none the less...

[Reply](#).

## Leave a Reply

Name (required)

Email (will not be published) (required)

Website

Comment

**XHTML:** You can use these tags: `<a href="" title="">` `<abbr title="">` `<acronym title="">` `<b>` `<blockquote cite="">` `<cite>` `<code>` `<del datetime="">` `<em>` `<i>` `<q cite="">` `<s>` `<strike>` `<strong>`

Submit Comment



Tim Mohler

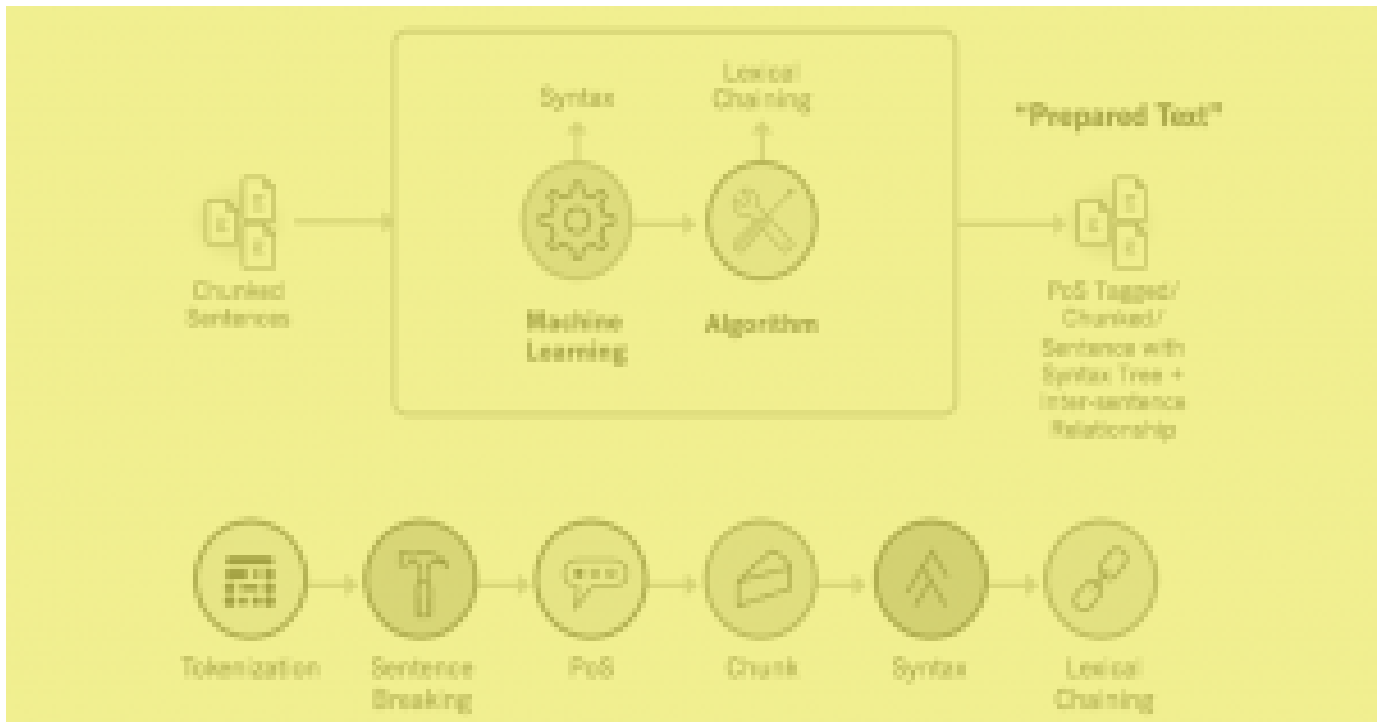
[in \(http://www.linkedin.com/in/tim-mohler-3090512/\)](http://www.linkedin.com/in/tim-mohler-3090512/)

Tim has been at Lexalytics since the early Bronze Age. During this time, he's played an integral part as COO, developing solutions that continue to change the way natural language processing happens. He can also hold his breath for 48 minutes.

## Related articles



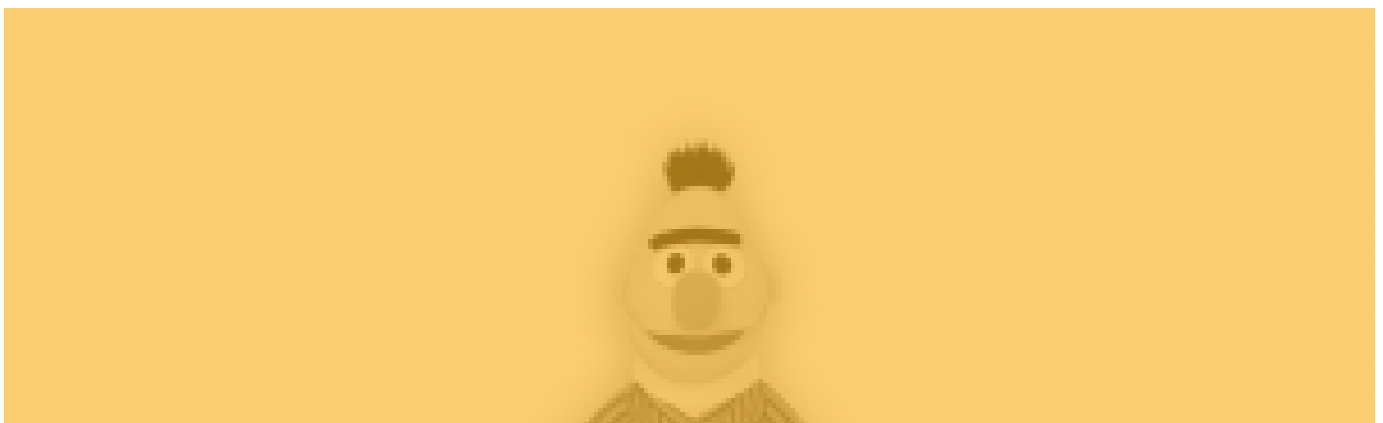
What is Tokenization in NLP? (<https://www.lexalytics.com/lexablog/tokenization>)



The 7 Basic Functions of Text Analytics & Text Mining (<https://www.lexalytics.com/lexablog/text-analytics-functions-explained>)



Build vs Buy: Text Analytics & NLP (<https://www.lexalytics.com/lexablog/build-or-buy-natural-language-processing>)



[BERT Explained: Next-Level Natural Language Processing \(https://www.lexalytics.com/lexablog/bert-explained-natural-language-processing-nlp\)](https://www.lexalytics.com/lexablog/bert-explained-natural-language-processing-nlp)

[Talk about solution fit \(/contact?action=demo\)](/contact?action=demo)

Or call us at **1-800-377-8036**

#### Platform & Services

[Lexalytics Intelligence Platform \(/platform\)](/platform)  
[NLP On-Premise \(/salience\)](/salience)  
[NLP Cloud API \(/semantria\)](/semantria)  
[Semi-Custom Applications \(/semi-custom-apps\)](/semi-custom-apps)  
[Professional Services \(/services\)](/services)

#### Resources



[Demo \(/nlpdemo/index.html\)](/nlpdemo/index.html)  
[Blog \(/lexablog\)](/lexablog)  
[Technology \(/technology/text-analytics\)](/technology/text-analytics)  
[Case Studies & Papers \(/resources\)](/resources)  
[Industries \(/industries\)](/industries)  
[Applications \(/applications\)](/applications)

#### Support

[Support home \(/support\)](/support)  
[Tuning guides \(/support/tuning\)](/support/tuning)  
[API integration \(https://semantria-docs.lexalytics.com/docs\)](https://semantria-docs.lexalytics.com/docs)  
[Salience Integration \(https://salience-docs.lexalytics.com/docs\)](https://salience-docs.lexalytics.com/docs)

#### Company

[About \(/about\)](/about)  
[Company News \(/news/company-news\)](/news/company-news)  
[Press \(/news/press-releases\)](/news/press-releases)  
[Privacy Policy \(/privacy-policy\)](/privacy-policy)  
[Security Overview \(/security-overview\)](/security-overview)  
[Contact us \(/contact\)](/contact)

 <https://www.facebook.com/pages/Lexalytics/124936814233570>  
 <http://www.twitter.com/lexalytics>

(<https://www.youtube.com/user/Lexalytics>),

(<https://www.quora.com/topic/Lexalytics>),

 (<https://www.linkedin.com/company/lexalytics-inc->),



(<https://www.facebook.com/pages/Lexalytics/124936814233570>),

 (<http://www.twitter.com/lexalytics>),

(<https://www.youtube.com/user/Lexalytics>),

(<https://www.quora.com/topic/Lexalytics>),

 (<https://www.linkedin.com/company/lexalytics-inc->),