Using AlchemyAPI for Enterprise-Grade Text Analysis

by Joseph Turian, Ph.D.

8/2013

AlchemyAPI™
Transforming Text Into Knowledge

This paper was underwritten by AlchemyAPI.

Executive Summary

- Text analysis software unlocks the hidden value in text data, such as from blogs, social media, email, and mobile devices. This can help enterprise decision-makers to improve brand equity, increase revenue, and reduce operational costs.
- Decision-makers who understand the value of text analysis software will want to buy from a leading provider. They understand that building your own text analysis system is expensive, tough, and not the core competency of most companies.
- Given the potential untapped value in text analysis, decision-makers want to avoid risk and get a battle-tested, versatile solution that they can trust. AlchemyAPI offers a comprehensive suite of high-quality text analysis components. AlchemyAPI's deep expertise shows in areas like quality, reliability, and flexibility.
- AlchemyAPI components can be composed together into a straightforward pipeline, solving a variety of problems. We specifically illustrate a media analysis use-case, demonstrating how to ingest and analyze news articles using AlchemyAPI.

What Is Text Analysis?

Organizations have access to a lot of text data. This data can contain great value for enterprise. But it is not easy to leverage this resource.

Text data from blogs, social media, email, and mobile devices is “unstructured.” It is not quantitative, and it is hard to query and use in reports and charts. It needs to be prepared before it is ready for operational use. Text analysis is the process of adding structure to text, making it accessible to computers. Once structured, data can be distilled into business intelligence and used to drive business decisions.

> “AlchemyAPI is obsessed with speed. Our customers recognize the time value of information; therefore, we have a financial incentive to process data as quickly as possible.”—Elliot Turner, AlchemyAPI CEO and Founder

Given the scale of enterprise data sets, text analysis is done using software, not by hand. As explained in our companion whitepaper, “Text Analysis: A Crucial Part of Enterprise Data Initiatives,” building your own text analysis suite can be expensive and risky. There are many hidden costs and pitfalls, and building out text analysis does not match the core competency of most organizations. By comparison, buying a solution from a market leader can have inexpensive total cost of ownership, and allows deployment that is fast and simple.

In particular, AlchemyAPI is able to read and understand text at massive rates of speed, by using complex linguistic, statistical, and neural network algorithms. AlchemyAPI handles over three billion API requests per month, and serves requests with fast turnaround. Speed and volume are important because unstructured data is increasing in volume and velocity, so future needs cannot be easily predicted.

Speaking with TechCrunch, AlchemyAPI CEO and Founder Elliot Turner commented: “AlchemyAPI is obsessed with speed. Our customers recognize the time value of information; therefore, we have a financial incentive to process data as quickly as possible.”

In addition to speed, third-party research, such as Rizzo and Troncy (2011) and Saif et al (2012), demonstrate AlchemyAPI's state-of-the-art accuracy on tasks such as entity recognition and sentiment analysis.

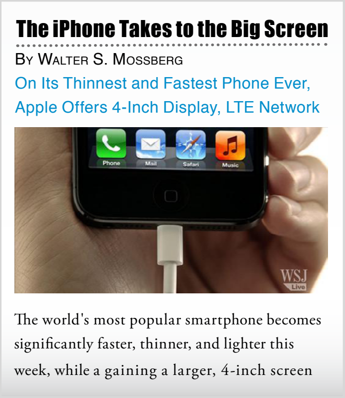
Text Analysis Example: Social Media Monitoring

In order to better understand the effectiveness with which AlchemyAPI handles text analysis in the real world, consider this example. In our companion whitepaper, we described a social media monitoring use case for text analysis. Subsequent sections will explore this use case as a running example.

Consider a media analysis firm who is responsible for analyzing the media response to iPhone 5 capabilities, in particular the pricing and screen. They crawl hundreds of blog posts every day that contain the keywords “iphone 5” and the company wants to use text analysis software to analyze these hundreds of blog posts.

They would first like to filter out the irrelevant blog posts, which are ones that contain the keywords but are merely mentioning it in passing, but don’t add anything relevant to the conversation. The company would like to prepare reports that answer questions such as:

- What is the volume of discussion on the iPhone 5?
- Which features of the iPhone 5 are most heavily discussed?
- Focusing specifically on iPhone 5 screen size and iPhone 5 pricing, what opinions are expressed in the media about these particular features?
- Who are the most vocal critics?
- Who are the most vocal supporters?
- What related topics are being discussed?



The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

As part of our running example, consider the article above from *The Wall Street Journal* entitled “The iPhone Takes to the Big Screen.” This is one of the online articles on which the media analysis firm will perform text analysis.

Focusing specifically on iPhone 5 screen size and iPhone 5 pricing; what opinions are expressed in the media about these particular features?

Components of Text Analysis

Text analysis can be performed on internal enterprise documents, as well as external web content such as social media. Web and social media content can be obtained through web scraping, as well as through services such as Google Alerts, Gnip, DataSift, Topsy, and Social Mention.

After collecting relevant text data, a text analysis pipeline is used. This text pipeline will include different text analysis components. Using the example above, we will show the structured output extracted by AlchemyAPI.

Language Detection

Generally, the first step of a text analysis pipeline is to categorize documents by language. Even a company doing multilingual analysis will want to have a separate report for each target language.

AlchemyAPI can identify almost 100 different languages. It can take as input any text, HTML, or web content.

The iPhone Takes to the Big Screen

By Walter S. Mossberg

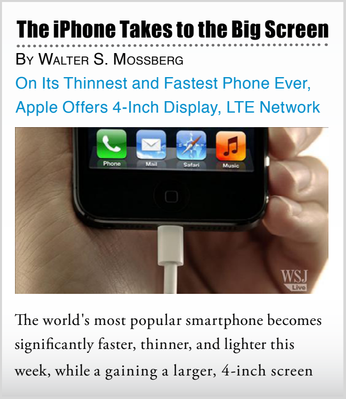
On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

→ Language: English

Text Extraction

When a URL comes in, it's a raw bundle of HTML code, containing boilerplates such as menus, links, and ads. Text extraction pulls out the important page text, and ignores ads, navigation bars, and other unimportant content.



The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

→

The world's most popular smartphone becomes significantly faster, thinner and lighter this week, while gaining a larger, 4-inch screen--all without giving up batter life, comfort in the hand and high-quality construction.

Keyword Extraction

Topic keywords are phrase variations that occur within the document. They are useful when it is important to capture broad linguistic variation, and to understand the diversity of terminology being used.

Concepts are the abstract ideas being discussed in the text, regardless of whether they are mentioned explicitly or not.

The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

→

Phrase: iPhone
Phrase: Apple
Phrase: price
Phrase: new maps
Phrase: screen

Concept Extraction

Concepts are the abstract ideas being discussed in the text, regardless of whether they are mentioned explicitly or not. For example, an article that contains "iPhone" and "Android" would have the concept "Mobile phone" extracted. Concepts have a single canonical name, and phrasal variations are combined into one canonical name.

The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

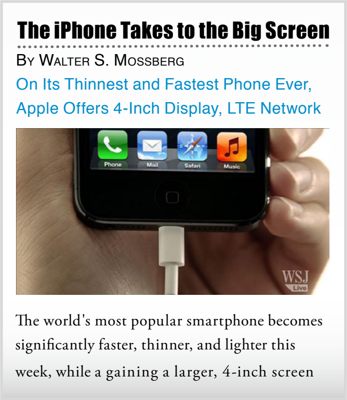
→

Concept: iPhone
Concept: App Store
Concept: Mobile phone
Concept: Google Maps
Concept: Google

Entity Extraction

This step identifies the people, companies, organizations, and other “typed” entities being discussed. Similar to concept extraction, phrasal variations are combined into a single canonical name, and entities are linked to DBpedia and Freebase URLs. The difference is that concepts are abstract and need not be explicit, whereas entities must be explicitly discussed in the article.

Additionally, entities are tagged with their type, including subtypes. For example, company “Apple” includes many subtypes, such as Operating System Developer, Processor Manufacturer, etc. These subtypes are useful when searching for very specific areas of interest.



The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

→

Technology: iPhone
Company: Apple
Company: AT&T
Person: Walt Mossberg

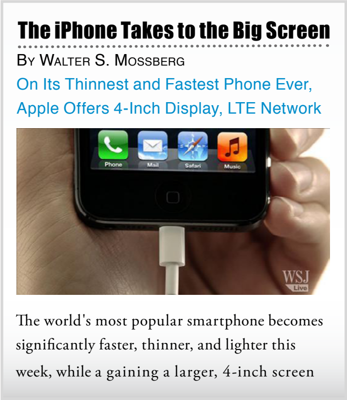
Sentiment Analysis

Often, one wants to determine whether the sentiment is positive or negative. There is document-level sentiment, which is computed over the entirety of the article for all topics discussed.

However, knowing the coarse-grained sentiment of an entire article is misleading. For example, the article could have positive sentiment overall as it discusses your market, but the last sentence contains a sharp attack on your product. For this reason, AlchemyAPI automatically includes entity-level sentiment when doing entity extraction and keyword-level sentiment when extracting keywords.

> Knowing the coarse-grained sentiment of an entire article is misleading. For example, the article could have positive sentiment overall as it discusses your market, but the last sentence contains a sharp attack on your product.

Consider the example of determining the sentiment in the article towards the iPhone 5 screen. The article mentions that “the iPhone 5's screen likely won't suffice” and then goes on to say that “I found the new iPhone screen much easier to hold and manipulate than its larger rivals and preferred it.” Ultimately, this is slightly positive .



The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

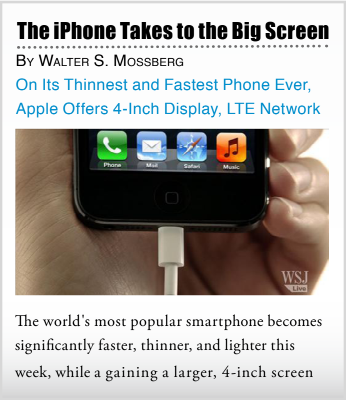
→

Document Sentiment: Positive

Keyword-Level Sentiment
  new maps: Positive
  Price: Negative
  Screen: Positive
Entity-Level Sentiment
  iPhone: Positive
  Apple: Positive
  Google Maps: Positive

Relation Extraction

For some applications, it is important to identify Subject-Action-Object relations. For example, one might be interested in automatically detecting company acquisitions, in which case one would search for relations in which the Action is "acquire."



The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

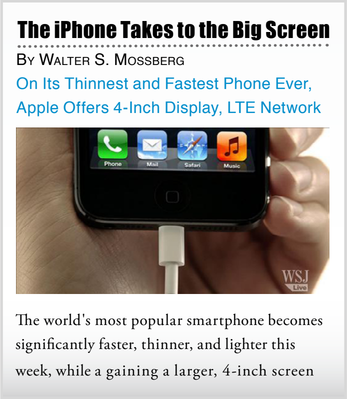
→

Subject: The iPhone
    Technology: iPhone
Action: Takes
    Verb Lemma: take
    Verb Tense: present
Object: to the big screen

Text Categorization

Text can be assigned to one of twelve high level categories, including "Arts & Entertainment," "Business," "Computers & Internet," etc. These categories are similar to the major sections of a newspaper.

For some applications, it is important to identify Subject-Action-Object relations.



The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

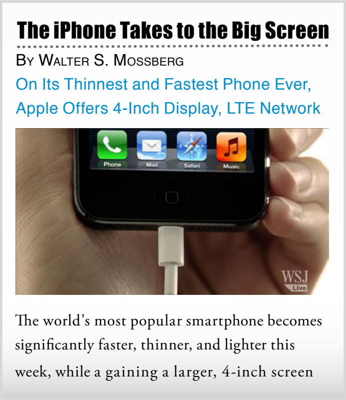
The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

→

Category: computer_internet

Author Extraction

By determining the author of the article, it can be associated with other articles by the same author.



The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

→

Author: Walter S. Mossberg

Closing the Loop with Text Analysis

Having run the different text analysis components, a variety of structured information has been extracted and can be stored in a database for further analysis. For example, entities and keywords can be associated with sentiment, and cross-referenced against author (Table 1).

Author	TextAttribute	TextAttributeValue	Relevance	Sentiment
Walter S. Mossberg	Entity	iPhone	0.84	Positive
Walter S. Mossberg	Entity	Apple	0.76	Positive
Walter S. Mossberg	Keyword	price	0.51	Negative
Walter S. Mossberg	Keyword	new maps	0.54	Positive
Walter S. Mossberg	Concept	App Store	0.57	
Walter S. Mossberg	Concept	Mobile Phone	0.56	

Table 1

When delivering a report, the media analysis firm can now easily answer questions such as:

- How much negative content was there about the iPhone 5?
- Is negative sentiment trending up or down?
- Who are the most vocal critics of Apple? Who are the biggest proponents of Apple?
- What is overall sentiment towards iPhone 5 pricing? Towards iPhone 5 screen size?
- What concepts and keywords are trending, which we didn't know in advance would be hot topics of discussion?

Linked Data and the Semantic Web

AlchemyAPI was one of the first API platforms to embrace RDF and linked data standards. For users interested in building upon semantic web functionality, AlchemyAPI concepts and entities are linked to DBpedia, Freebase, OpenCyc, GeoNames, and other linked data URLs. These URLs contain additional structured information about people, companies, and concepts.

For users interested in building upon semantic web functionality, AlchemyAPI concepts and entities are linked to DBpedia, Freebase, OpenCyc, GeoNames, and other linked data URLs.

An example usage of linked data using AlchemyAPI was an academic research group's Disease Tracker. This application tracked the outbreak of diseases across different cities, states, and countries. They leveraged the linked data to augment the data within the text documents they analyzed; the linked data URLs provided regional demographics, life expectancy statistics, and other information.

Outlook

Data drives business decisions. It enables organizations to increase revenue and decrease costs. If you ignore the trove of information hidden in text data, you risk leaving value on the table. Our example showed a media analysis use-case, but AlchemyAPI is equally applicable to other valuable business applications, such as:

- Analyzing the sentiment of tweets
- Driving automated stock trading by processing SEC filings
- Competitive intelligence
- Contextual advertising
- Semantically-powered SEO

To learn more about different applications of text analysis, we encourage you to read our companion whitepaper, "Text Analysis: A Crucial Part of Enterprise Data Initiatives."

About Joseph Turian

Joseph Turian, Ph.D., heads MetaOptimize LLC, which consults on data science, NLP, and machine learning. He also runs the MetaOptimize Q&A site, where Machine Learning and Natural Language Processing experts share their knowledge. He specializes in large data sets.

Joseph Turian holds a Ph.D. in computer science (with a focus on Machine Learning and Natural Language Processing) from New York University since 2007. During his graduate studies, he developed a fast, large-scale machine learning method for parsing natural language. He received his AB from Harvard University in 2001.

As a scientist, Joseph Turian has over 14 refereed publications in top NLP + ML conferences. His team submitted the best parser in EVALITA 2009 Main+Pilot tasks. He is an advocate for open-notebook science, releasing his research code on his github, and for broader scientific collaboration through the internet.

If you ignore the trove of information hidden in text data, you risk leaving value on the table.

About AlchemyAPI

The product of over 75 person years of engineering effort, AlchemyAPI is a text mining platform providing the most comprehensive set of semantic analysis capabilities in the natural language processing field. Used over 3 billion times every month, AlchemyAPI enables customers to perform large-scale social media monitoring, target advertisements more effectively, track influencers and sentiment within the media, automate content aggregation and recommendation, make more accurate stock trading decisions, enhance business and government intelligence systems, and create smarter applications and services. If you would like to learn more about our company and services, please call us at 1-877-253-0308 or email info@alchemyapi.com.