Technology Review Submission

Amazon Comprehend

CS410 Text Information System

Authors:

Nishant Gulgulia (nishant9@illinois.edu)
Pradeep Khandelwal (pk5@illinois.edu)
Liu, Ling-Hsi (lhliu2@illinois.edu)

Nowadays, natural language processing(NLP) is very popular in various areas. NLP could help computer or devices to understand and analyze human languages. Then, computers and devices could generate response and communicate with human speech, such as echo dot and google home which are application by NLP. The goal of NLP is that when human communicate with device or computer, people could feel like interaction with human. This is very complicate steps for operating artificial language by computer. Firstly, computers need to recognize the speech or text and understand the meaning of text. Different NLP systems use different architectures or technologies, but the process is almost similar. Furthermore, computer needs to understand each word and know that it is noun, verb or preposition. NLP system also need to know grammar rules and do statistical machine learning for the documents and get meaning of natural language. The challenge sides are that there are several meanings (polysemy) or different words have similar meaning (synonymy.) NLP system is very complicated, but it is very useful and advantageous for our life.

Moreover, NLP could help us to know the sentiment and feeling of documents and diagnose their unhappy or happy like "conscious mind." NLP is not only making convenient in our current life but also teaching you how to manage our future life. There are some NLP libraries of open sources, such as Stanford's Core NLP Suite, Natural Language toolkit, NLTK, Apache OpenNLP and Apache UIMA. In addition, some companies create easy-to-use NLP API service which are created by Machine Learning models, and those NLP API could extract information, understand sentiment and recognize entities. I will introduce and compare four different Natural Language Processing Services, including Amazon Comprehend NLP, Google Could Natural Language, Microsoft Azure Text Analytics, IBM Watson NLU. All of them provides NLP API for people to use, we will use the same sentences to compare the output of features. Then, we will also compare the price of NLP tools for paid plan.

Using the same Sentences for testing is that "Amazon.com, Inc. is located in Seattle, WA and was founded July 5th, 1994 by Jeff Bezos, allowing customers to buy everything from books to blenders. Seattle is north of Portland and south of Vancouver, BC. Other notable Seattle - based companies are Starbucks and Boeing."

Compare Features of Different Could NLP Tools:

	Amazon Comprehend	Google Cloud Natural Language	Microsoft Azure Text analytics	IBM Watson NLU
Entities	V	V	V	V
Sentiment	V	V	V	V
Key Phrases	V		V	V
Syntax	V	V		
Language	V		V	
Categories	V	V		V
Concept				V

Table 1: Features of Different Cloud NLP Tools

- **Entities** – Retrieve noun and Recognize location, people, date, Organization for Entities
- **Sentiment** – Provide emotional sentiment by neutral, positive, negative, or mixed
- **Key phrases** – Retrieve the noun phrases and also calculate confidence
- **Language** – Detect the primary language for this text
- **Syntax** – Analyze each word belong in which speech
- **Categories** – Put almost similar words and noun in the same group
- **Concept** – Detect noun, location and organization similar to entity

1. **Amazon Comprehend NLP:**

   Amazon Comprehend NLP was announced at 2017. Amazon Comprehend NLP provides some features, such as entities recognition, key phrases extraction, language recognition, sentiment analysis and syntax analysis. It supports UTF-8 format of text files. Amazon Comprehend analyze and detect documents by English, German, Spanish, Italian, French and Portuguese. Amazon Comprehend also offers three document processing modes, such as single document, multiple document synchronous processing, and asynchronous batch processing. Moreover, we could build our own model by custom classification and training custom classifier. For instance, we could do classifier by our setting categories. And, we train different categories by label different group. Furthermore, Amazon Comprehend also supports Medical text analysis, such as clinical documents from doctor or clinical trial reports. We could see the output by Amazon Comprehend NLP as following.

Entity	Category	Confidence
Amazon.com, Inc	Organization	0.90
Seattle, WA	Location	0.89
July 5th, 1994	Date	0.99+
Jeff Bezos	Person	0.99+
Seattle	Location	0.97
Portland	Location	0.99+
Vancouver, BC	Location	0.94
Seattle	Location	0.99+
Starbucks	Organization	0.99+
Boeing	Organization	0.99+

Figure 1: Amazon Comprehend Entity

Key phrases	Confidence
Amazon.com	0.86
Seattle, WA	0.95
July 5th, 1994	0.91
Jeff Bezos	0.99+
customers	0.99+
books	0.99+
blenders	0.98
Seattle	0.99+
Portland and south	0.89
Vancouver, BC	0.95

Figure 2: Amazon Comprehend Key

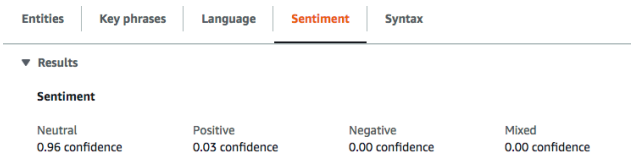

Figure 3: Amazon Comprehend Sentiment

Word	Part of speech	Confidence
Amazon.com	Proper noun	0.94
,	Punctuation	0.99+
Inc.	Proper noun	0.99+
is	Auxiliary verb	0.98
located	Verb	0.99+
in	Adposition	0.99+
Seattle	Proper noun	0.99+
,	Punctuation	0.99+
WA	Proper noun	0.99+
and	Conjunction	0.99+

Figure 4: Amazon Comprehend Syntax

2. Google Could Natural Language:

The first version of Google Could Natural Language was announced in 2016. Google Could Natural Language provides an easy to use REST API by machine learning models. Google Cloud Natural Language offers entities analysis, sentiment analysis, syntax analysis and categories classified. The syntax analysis is very special, it builds relation graph of entities extracted from documents. Google Cloud Natural Language could analyze and detect English, Korean, Portuguese, Japanese, German, Chinese, Italian, French and Spanish language. Google Cloud also supports different operation systems, such as Linux, Ubuntu, Red Hat, macOS and Windows. Client libraries offer different computer languages, including C#, Go, Java, Node.js, PHP, Python and Ruby. Then, we could use Google Could Natural Language to understand sentiments from social media or conversation of call center or message app. We got some output by testing the same sentences on Google cloud natural language as following.

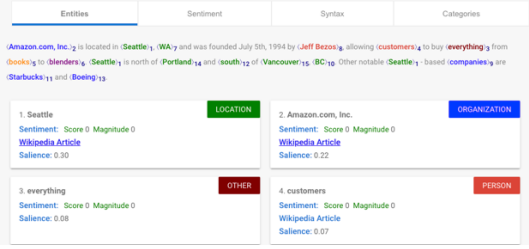

Figure 5: Google Natural Language Entities

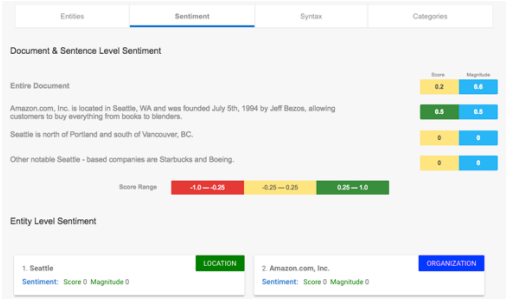

Figure 6: Google Natural Language Sentiment

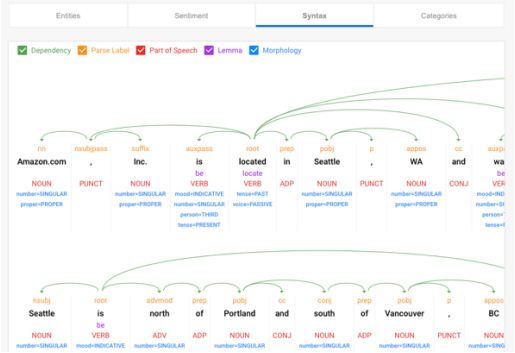

Figure 7: Google Natural Language Syntax

Figure 8: Google Natural Language Categories

3. Microsoft Azure Text Analytics:

Microsoft Azure Text Analytics was announced in 2017. Microsoft Azure Text Analytics provides language detection, key phases extraction, sentiment analysis and liked entities. In the sentiment part, the score of API is between 0 and 1 which is generated by classification techniques. The positive sentiment is close to score 1; the negative sentiment is close to score 0. Microsoft Azure Text Analytics could analyze and detect texts or documents up to 120 languages, such as Danish, Japanese, English, French, Italian, Korean, Finnish, German, Norwegian, Dutch, Spanish, Portuguese, Greek and Russian. The language score is between 0 and 1. If scores is 1, it indicates 100% which means language identifies true. Microsoft Azure Text Analytics also has some limits, such as 5,000 Maximum size of each single document, 1 MB maximum size of entire request and 1000 documents maximum number of documents in a request. Then, we could see the output of Microsoft Azure Text Analytics as following.

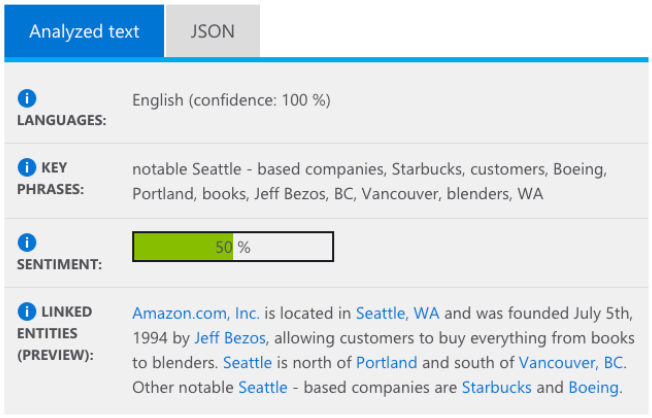

Figure 9: Microsoft Azure Text Analytics

4. IBM Watson NLU (Natural Language Understanding):

The first version of IBM Watson NLU announced in 2017. IBM Watson NLU provides various features, such as sentiment, emotion, keywords, entities, categories, concept and semantic roles. Watson NLU could also analyze text from webpages. People could use raw HTML or URL in NLU API request. Then, the NLU service will analyze the content of the page, such as blog context or news article. This is super great and convenient for people who want to analyze the web context and get sentiment of web comments. The emotion feature of IBM Watson NLU provides 5 levels, such as joy, anger, disgust, sadness and fear. The sentiment levels are very different with other cloud NLP tools.

Moreover, the category of IBM Watson NLU is a hierarchy category. It provides three level hierarchy, and we could see more detail of categories. Watson NLU supports over 80 different languages. Watson NLU have 50,000 single-byte, multibyte characters text limits, 30 concurrent request limits, and custom model size limit for Lite pricing plans. People could also create custom models to detect entities and relation by text on Watson Knowledge Studio. Watson NLU offer several SDK, including Android SDK, Java SDK, Node.js SDK, python SDK. We show the output of IBM Watson NLU as following.

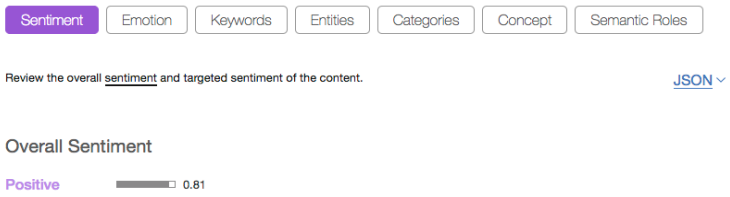

Figure 10: IBM Watson NLU Sentiment

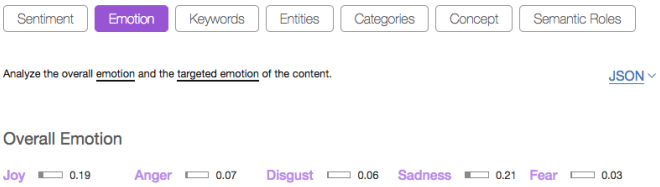

Figure 11: IBM Watson NLU Emotion

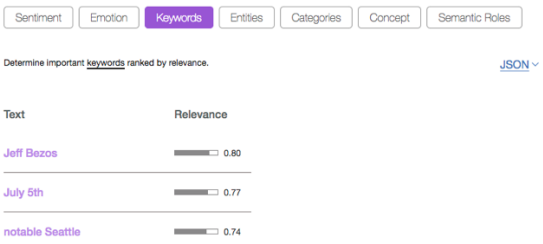

Figure 12: IBM Watson NLU Keywords

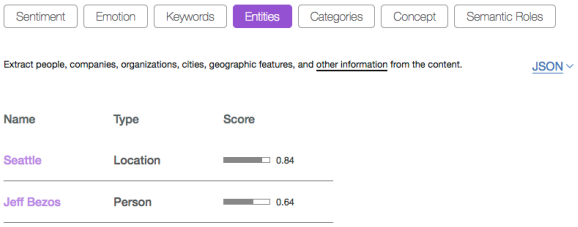

Figure 13: IBM Watson NLU Entities

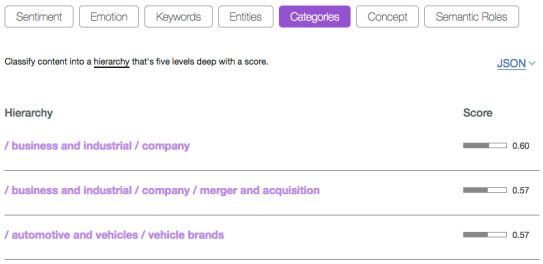

Figure 14: IBM Watson NLU Categories

After we tested and used the four different NLP API, we saw different sentiment analysis, Key Phrases extraction, syntax recognition, entities analysis and language detection from these different could NLP tools. We get a short analysis by entities, sentiment, key phrase, syntax and language as following.

- Entities: Amazon Comprehend not only provides entities but also offers category and score of entities. Google Natural Language separate county and state of location and separate year and moth of date. However, we think location and date shouldn't be separated. Therefore, we prefer the entities of Amazon Comprehend, and we could quickly know categories of entities.
- Sentiment: Although Cloud Natural Language from Google provides the sentiment analysis for each sentence, it doesn't show total sentiment score. So, we couldn't know how many percentage for positive, negative or natural. IBM Watson NLU provides five levels emotions. Two of them is positive; three of them is negative. However, it doesn't show natural sentiment. Therefore, the best sentiment of all could NLP tools is Amazon Comprehend which show well to understand sentiment situation.
- Key Phrases: After comparing key phrases, we think the best of key phrases of is Amazon Comprehend. It is more completed and correct.

- Syntax: Cloud Natural Language from Google creates relation diagram for documents. It is clear to understand the relation of syntax. The Syntax of Amazon Comprehend doesn't look well, it is similar to entities. And, Microsoft Text Analytics and IBM Watson NLU didn't provide syntax.
- Language: Only Amazon Comprehend and Microsoft Azure Text Analytics provides this function. Both of them show how many percentages and which language.

Compare Pricing of Different Could NLP Tools:

	Free Plan (Lite version)	Paid Plan (Standard version)
Amazon Comprehend	50k unites of text per month	Standard 1: 0-10M Key Phrase Extraction, Sentiment Analysis: $0.0001 Entity Recognition, Language Detection: $0.0001 Syntax Analysis: $0.00005 per unit Standard 2: 10M-50M Key Phrase Extraction, Sentiment Analysis: $0.00005 Entity Recognition, Language Detection: $0.00005 Syntax Analysis: $0.000025 per unit Standard 3: 50M+ Key Phrase Extraction, Sentiment Analysis: $0.000025 Entity Recognition, Language Detection: $0.000025 Syntax Analysis: $0.0000125 per unit
Google Cloud Natural Language	5k unites of text per month	Standard 1: 5K-1M Entity Analysis, Sentiment Analysis: $1.00 Syntax Analysis: 0.50 Entity Sentiment Analysis: $2.00 Standard 2: 1M-5M Entity Analysis, Sentiment Analysis: $0.50 Syntax Analysis: 0.25 Entity Sentiment Analysis: $1.00 Standard 3: 5M-20M Entity Analysis, Sentiment Analysis: $0.25 Syntax Analysis: 0.125 Entity Sentiment Analysis: $0.50
Microsoft Azure Text analytics	5k transactions per month	Standard 1: 1-0.5M text records - -$0.002 per text records Standard 2: 0.5M-2.5M text records - -$0.001 per text records Standard 3: 2.5M-10.0M text records - -$0.0005 per text records Standard 4: 10.0M+ text records - -$0.00025 per text records

IBM Watson NLU	30k NLU items per month I free custom model	Standard 1: 1-0.25M NLU item - -$0.003 per NLU Standard 2: 0.25-5M NLU item - -$0.001 per NLU Standard 3: 5M+ NLU item - -$0.0002 per NLU 1 custom model $800 per month

Table 2: pricing of different Cloud NLP Tools

Summary of comparing features and pricing on different Cloud NLP Tools

After comparing features of different cloud NLP Tools, we feel that Amazon Comprehend is the best tool service for analyzing the text sentiment, retrieving the entities, detecting language and extracting key phrase. It provides clearly and completed sentiment percentage, we could know that this documents or texts are positive, negative, natural or mixed. It also offers correct and obvious entities and key phrases extraction. And, we could know the entity categories in entity analysis. It also provides lots of different language for documents and texts.

Moreover, after comparing pricing of different Cloud NLP tools, Amazon comprehend API provides very good price for paid plan. Amazon Comprehend API not only provides paid plan and but also free plan. In the free plan, people could analyze 50k unites of text per month. This free plan is better than Google Cloud Natural Language and Microsoft Azure Text analytics. In paid plan, it only cost $0.0001 for first 10M units, cost $0.00005 for 10M to 50M units and cost 0.000025 for over 50M units. The price of paid plan from Amazon Comprehend NLP tools is the cheapest price of all cloud NLP tools. Therefore, after comparing the features and price, we think Amazon Comprehend is the best cloud NLP tools. Therefore, we use Amazon Comprehend to get sentiment of tweet from twitter. And, we recommend people to choose Amazon Comprehend to do more text analysis, such as twitter tweet, yelp food and movie review. In our final project, we analyze the tweets from twitter API by Amazon Comprehend. Then, we could see the sentiment results by bar chart, pie chart and map on Amazon QuickSight.