Keang Cheang Ung

keangcheang@gmail.com

979-240-2429

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

Topic Modeling on Twitter Using K-Mean And LDA

Springboard Capstone Project 2

1. **Problem and Objective**

Nowadays social media has been very influential to all sorts of aspects in human lives that people use this mass media to express their sentiments and views to the publics. These sharing thoughts and emotions has been very useful for the researcher or businesses to analyze and harvest some insights to further understand the needs and demands of the writers. This is very true to social media like Twitter and Facebook. For instance, in 2016 twitter alone had 319 million monthly active users and there were 40 million election-related tweets on the U.S. presidential election day.[[1]](#footnote-1)

This spectrum of influence from social media was the inspiration of this project which aimed to explore what topics that people are talking about on this platform. This project will use only twitter to achieve this goal.

1. **Methodologies and Data Wrangling**

The methodologies for this project was very challenging because it involved many steps just to get the texts ready for cluster modeling. Firstly, we had to collect our tweets from Crate.io because only with this method that we could get the most tweets out of the twitter. This method would include the data from different regions different language. There were more than 50,000 tweets collected from CrateDB with 5 columns. We were only interested in the text column.

Table 1: A Portion of Data Set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **created\_at** | **Id** | **retweeted** | **source** | **text** |
| 2017-08-15 17:32:31 | 89751133  4123892736 | False | <a href="http://twitter.com/  download/android" ... | RT @qikipedia: What on earth could be more lux... |
| 2017-08-15 17:32:31 | 89751133  4085926912 | False | <a href="http://twitter.com/#!/  download/ipad" ... | RT @Medicis1917: 撃ち落としたら日本がどうなるかを語れよ。日米同盟の為に戦争... |
| 2017-08-15 17:32:31 | 89751133  4107009024 | False | <a href="http://twitter.com/  download/android" ... | @SLandinSoCal @foxandfriends @realDonaldTrump ... |

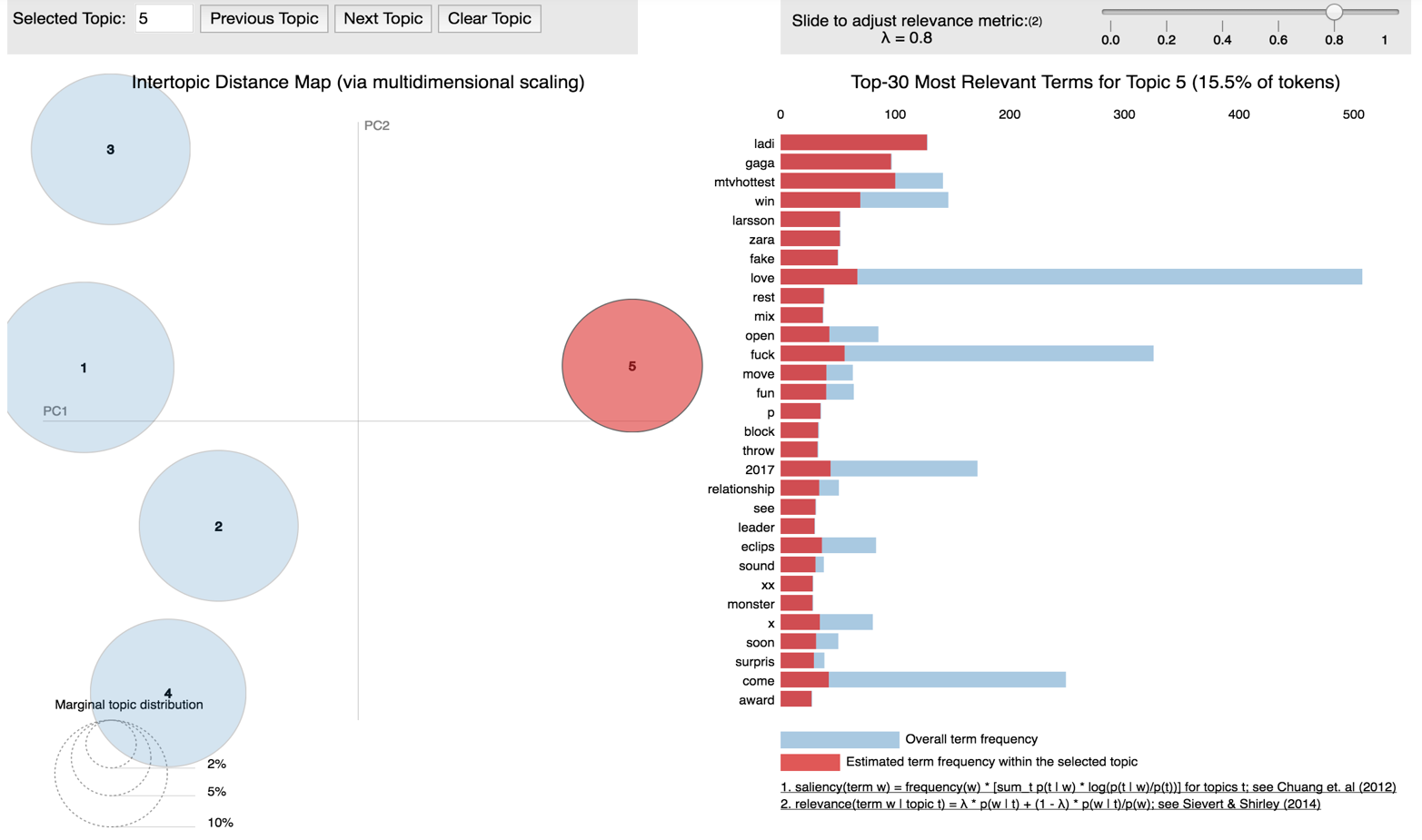
As shown in Table 1, there were tweets with different languages, thus we had to filter only English tweets using Textblob language detection library. Next, we cleaned the texts by turning every character into lowercase, tokenizing the words and then removing the stop words. We also used Porter stemmer to stem the texts so that these text would follow their base or root forms.

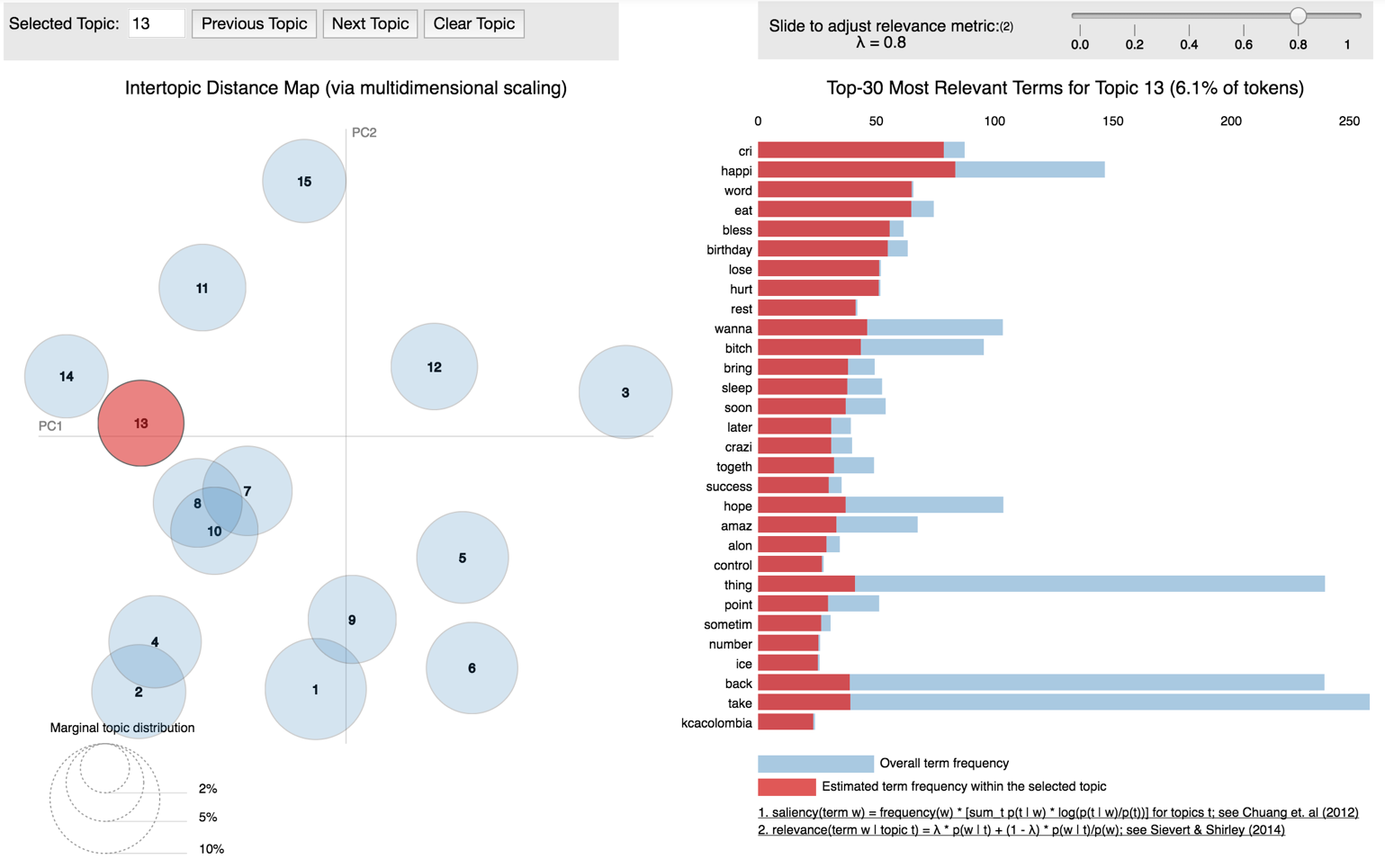
1. **Result**

In order to run LDA model on the cleaned texts, we had to create a python dictionary for the text corpus as a structured set of text. With this corpus, we could feed the text into LDA model to check for their topics and visualize it with a library called pyLDAvis.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic 1** | | **Topic 2** | | **Topic 3** | | **Topic 4** | | **Topic 5** | |
| **Word** | **Probability** | **Word** | **Probability** | **Word** | **Probability** | **Word** | **Probability** | **Word** | **Probability** |
| love | 0.08% | ladi | 0.06% | video | 0.07% | trump | 0.10% | shit | 0.05% |
| vote | 0.05% | mtv  hottest | 0.05% | girl | 0.06% | follow | 0.09% | look | 0.05% |
| people | 0.03% | gaga | 0.04% | youtube | 0.05% | people | 0.06% | think | 0.05% |
| time | 0.03% | win | 0.03% | sex | 0.05% | retweet | 0.04% | time | 0.05% |
| right | 0.03% | love | 0.03% | ever | 0.04% | white | 0.04% | love | 0.04% |
| feel | 0.02% | fuck | 0.03% | porn | 0.03% | nazi | 0.04% | fuck | 0.04% |
| taylor | 0.02% | zara | 0.02% | year | 0.03% | check | 0.03% | never | 0.04% |
| omg | 0.02% | larsson | 0.02% | cri | 0.03% | year | 0.02% | people | 0.04% |

Table 2: List of Words in the 5 Topics

Figure 1: pyLDAvis Visualization for 5 Topics

Figure 2: pyLDAvis Visualization for 15 Topics

With LDA modeling, there was no ground rule to what indicated how many topics were ideal, but based on experiences and personal judgment on the topic itself, and by looking how cohesive the topics were followed by their probabilities. In order to do this, we had to test the LDA model with different parameters like k (topics), beta and alpha. Below is the descriptive explanation quoted from medium website[[2]](#footnote-2):

* K: the number of topics
* Alpha which dictates how many topics a document potentially has. The lower alpha, the lower the number of topics per documents
* Beta which dictates the number of word per document. Similarly, to Alpha, the lower Beta is, the lower the number for words per topic.

After that, we also need split the data into train and test data set and then filter out the important words by using TF-IDF model. Lastly, we run the K-mean model with different k topics to find the best score and visual the feature using PCA to reduce the dimensions of the data.

1. **Modeling Design**
2. **Conclusion**

**Reference:**

1. Isaac, Mike; Ember, Sydney (November 8, 2016). "For Election Day Influence, Twitter Ruled Social Media". *The New York Times*. Retrieved November 20, 2016. [↑](#footnote-ref-1)
2. https://medium.com/@alexisperrier/topic-modeling-of-twitter-timelines-in-python-bb91fa90d98d [↑](#footnote-ref-2)