Topic Modeling on Twitter Using K-Mean And LDA

Springboard Capstone Project 2

Abstract

The purpose of this study is to extract mainstream topics from twitter using NLP tools. The first tool we utilize is the pyLDAvis which allows us to visualize the sophisticated dimensions of text mining. In order to perform pyLDAvis, we have to go through many processes of cleaning the text like converting the text into lowercase, collecting only English letters, tokenizing and removing stop words. The pyLDAvis helps us understand that there are six topics that are reasonable and cohesive during the duration of the tweets. To improve the model, LDA and TF-IDF are used to perform the topic mining with weight factor incorporated into the model. In addition, we use PCA to reduce the dimensions of the corpus dictionary and we have the result of six topics. Among these topics, four topics stand out the most: president Donald Trump, celebrities, human emotions and sexual contents.

Keang Cheang Ung

keangcheang@gmail.com

979-240-2429

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]

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.

2. 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 data was collected during August, 2017. 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

[1] Isaac, Mike; Ember, Sydney (November 8, 2016). "For Election Day Influence, Twitter Ruled Social Media". *The New York Times*. Retrieved November 20, 2016.

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.

3. Modeling Design and 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.

Table 2: List of Words in the 5 Topics

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Word	**Weight**	**Word**	**Weight**	**Word**	**Weight**	**Word**	**Weight**	**Word**	**Weight**
love	0.008	ladi	0.006	video	0.007	trump	0.010	shit	0.005
vote	0.005	mtv hottest	0.005	girl	0.006	follow	0.009	look	0.005
people	0.003	gaga	0.004	youtube	0.005	people	0.006	think	0.005
time	0.003	win	0.003	sex	0.005	retweet	0.004	time	0.005
right	0.003	love	0.003	ever	0.004	white	0.004	love	0.004
feel	0.002	fuck	0.003	porn	0.003	nazi	0.004	fuck	0.004
taylor	0.002	zara	0.002	year	0.003	check	0.003	never	0.004
omg	0.002	larsson	0.002	cri	0.003	year	0.002	people	0.004

Figure 1: pyLDAvis Visualization for 5 Topics

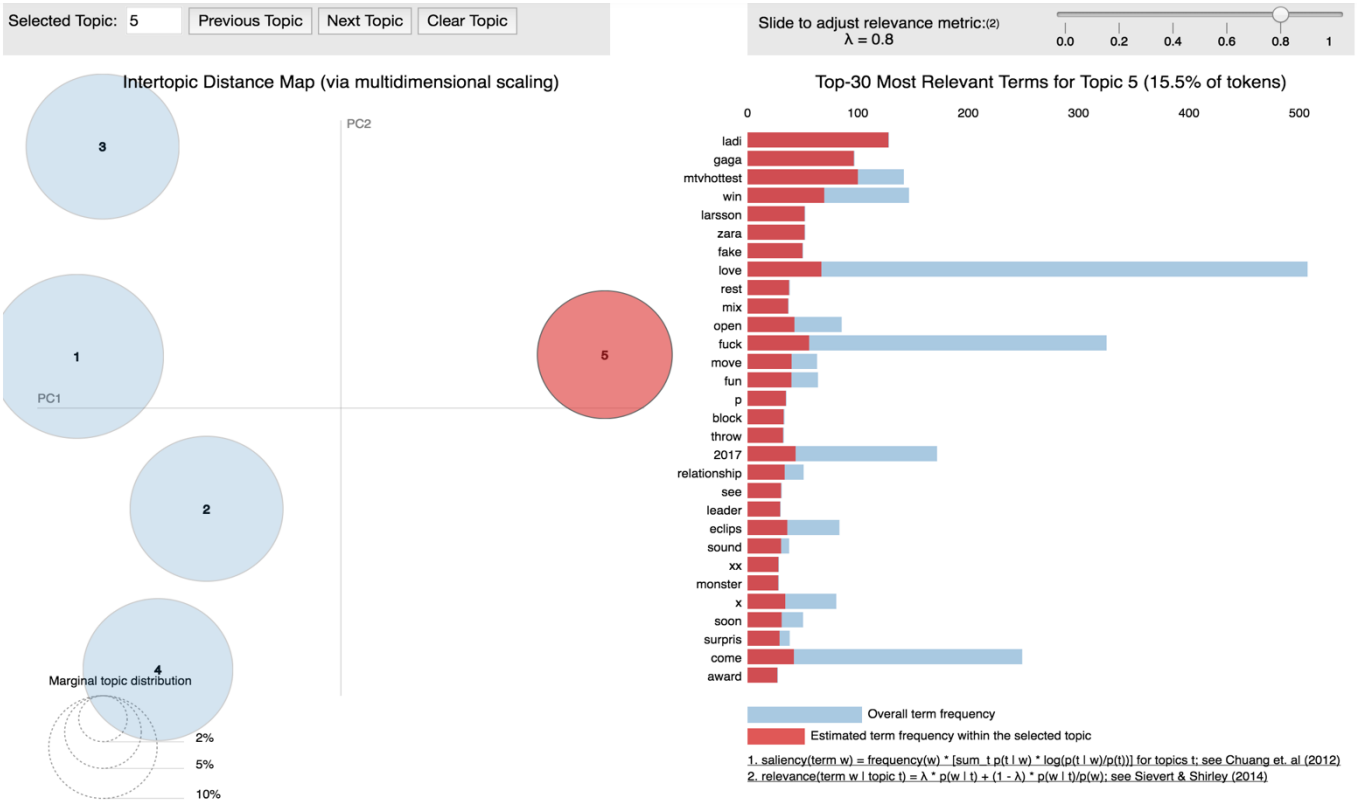
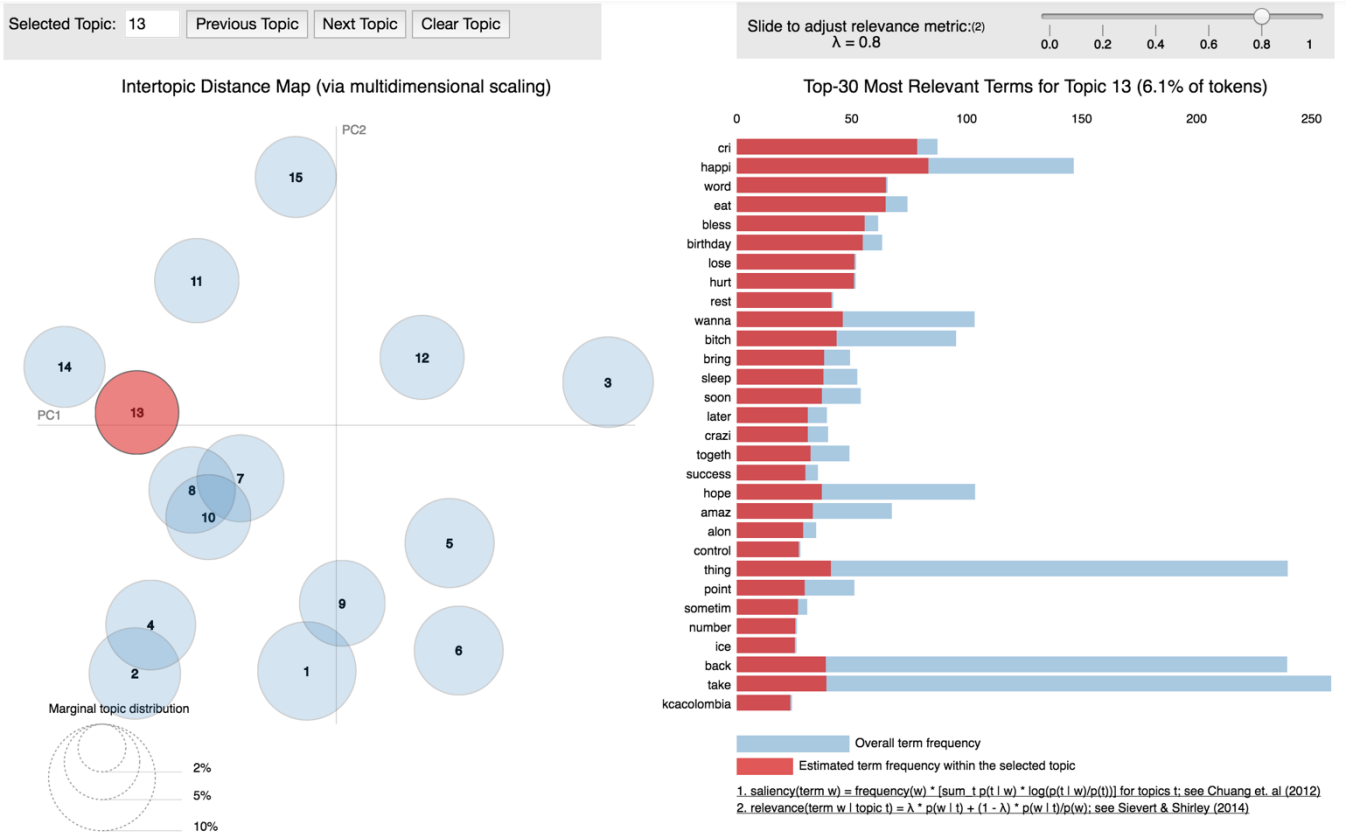


Figure 2: pyLDAvis Visualization for 15 Topics



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]:

- 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.

From Table 1, Figure 1 and 2, it had shown that 15 topics gave a more dispersed point on the pyLDAvis and it included all the topics covered in 5 topics. With these 15 topics, the most frequent topics appeared many time were listed below:

- 2nd topic talked more about President Trump and around racial issues: trump, white, nazi, bad, eclips, presid, racist, sumpremacist

[2] https://medium.com/@alexisperrier/topic-modeling-of-twitter-timelines-in-python-bb91fa90d98d

- 3rd topic mentioned sexual contents on the internet: girl, sex, fuck, video, porn, beauty, ad, teen, hot, pic
- 4th topic talked more about being back to school: never, first, time, high, people, school, learn, public, start
- 12th topic talked about youtube media: youtube, talk, like, video, Justin, friend, Bieber, exo
- 13th topic showed some emotions on internet: cri, happi, bless, birthday, lose, hurt, rest
- 15th topic talked about celebrities: ladi, gaga, mtvhottest, Larsson, zara, tri

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. TF-IDF (term frequency–inverse document frequency) was used in this project as a weighted factor which reflected how vital the words in the corpus were. There were parameters which needs optimization for the best performance like cut off values of the weights in each word. These cut off values was set by max_df and min_df. When we set the min_df by 0.0001, the number of words dropped from 196,941 to 17,818 which has simplified the model tremendously and helped improve the accuracy score. Finding the best parameters was very difficult, thus we wrote a loop function to repeat our testing with different parameters. The results were shown below in Table 4 with Silhouette score explanation in Table 3:

Table 3: Silhouette Score Explanation

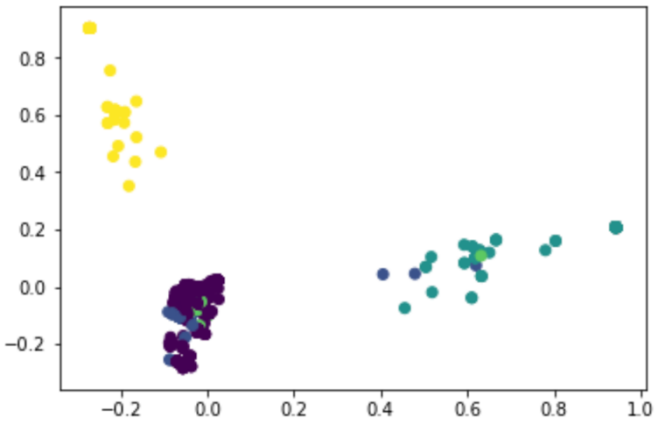
Range	Interpretation
0.71 - 1.0	A strong structure has been found.
0.51 - 0.7	A reasonable structure has been found.
0.26 - 0.5	The structure is weak and could be artificial.
< 0.25	No substantial structure has been found.

Table 4: Result of K-mean with k=10, max_df=1.00 and min_df: 0.003

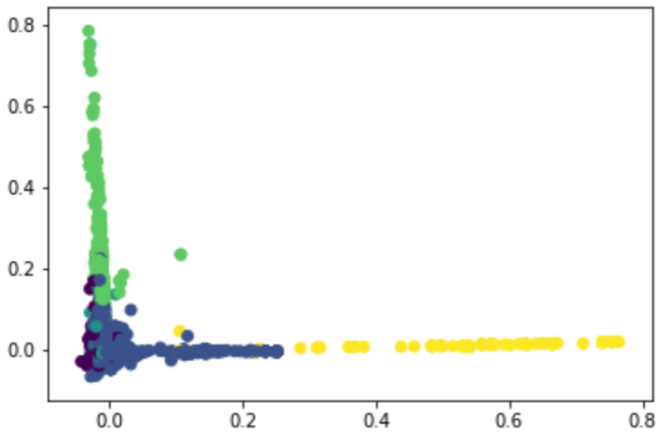
Cluster	Score	Silhouette Score
2	-13180.47	0.02745
3	-13103.95	0.02789
4	-13088.60	0.02682
5	-13022.45	0.02707
6	-12961.73	0.03006
7	-12967.42	0.02835
8	-12912.95	0.03059
9	-12855.11	0.03131

Table 4 showed that cluster 6 gave the highest silhouette score among the 2nd cluster and 5th cluster, thus we considered this the best option for our model and shall use 6 clusters as the parameter for visualization PCA. With the visualization of our clustering, we would try 3 options with the following parameters and results:

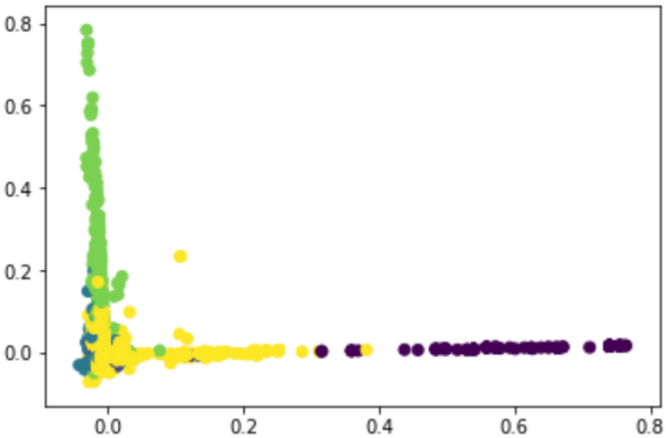
1. k=5, max = 1.00, min_df = 0.015 and number of PCA components = 3



2. k=5, max = 1.00, min_df = 0.003 and number of PCA components = 5



3. k=6, max = 1.00, min_df = 0.001 and number of PCA components = 5



The first seemed very reasonable since the clusters were distinguishable between each other yet if we looked at the number of words that were used in this model; it was too small. Thus, we would reject this option. The second option and last option were very similar in the graph and the colors were understandable to where the clusters were. However, in the 3rd option the dark green color seemed to mixed up with the yellow color. The 3rd option was considered the best one so far from our testing due to its highest Silhouette score and better picture of PCA visualization. We shall examine thoroughly the topic in the 3rd option as below:

Table 5: Top Terms Per Cluster

Cluster 0: ladi gaga gaga ladi mtvhottest ladi mtvhottest ladi gaga mtvhottest gaga zara gaga zara larsson

Cluster 1: thing thing ever ever love happen seen thing happen tell bad life cutest thing peopl

Cluster 2: love hate peopl beauti love love birthday heart happi life us world happi birthday fall

Cluster 3: come come back come soon soon back special video come special video video come drop come drop

Cluster 4: follow trump pleas retweet donald donald trump everyon automat check automat check presid

Cluster 5: look time fuck peopl think video girl shit take neve

With 6 clusters and 5 components of PCA, the model showed clustering words very similar to the graph produced by pyLDAvis. These clusters portrayed many popular terms people use in their tweets but there were only four logical topics that showed the mainstream news during August 2017 like singers (Lady Gaga…etc.) Donald Trump as well as other miscellaneous topics like pornography and human emotions.

4. Conclusion

Clearly processing a text from twitter presented a very challenging problem for topic modeling due to its time-consuming and various procedures. The tweets from August, 2017 that we collected from CrateDB were focusing on popular topics such as Donald Trump, sexual contents, singer names, school start and random emotions. In the future, one can improve upon this project by improving the K-mean to be smarter by feeding labels for text among

these topics and then rerun K-mean again. This might improve the accuracy score of the model and train a better classifier.

Reference:

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