Project 1

Social Media Analytics Report

Team 30

Jessica Tsai (30547137)
jessica.yc.tsai@gmail.com
Peter Hung (18704156)
ubc.peterkhung@gmail.com

COMM 337-202

1.0 Data Collection

The chosen keywords are: **USA vs. China**

2.0 Preliminary Analysis

2.1 What were the ten most popular words with/without stop words?

Strategy

While collecting tweets, we realized that many words included user mentions (@username) and URLs (http://). While iterating through each tweet, we ran another "for" loop that split the tweets into individual words and removed words that did not including user mentions and URLs. We also realized that punctuation affected the results of the tweets and removed the punctuation.

We tested various stemmers to determine what worked best for our research based on the keyword, "China":

China									
None		**Lancaster**		**Porter**		**Snowball**		**WordNet**	
china	6691	chin	6697	china	7168	china	7669	china	6698
kim	621	thi	743	thi	743	kim	641	kim	621
amp	577	wil	704	kim	630	trump	619	amp	577
korea	559	hav	632	amp	577	korea	604	korea	559
trade	532	kim	630	korea	576	amp	577	trade	533
trump	508	new	626	trump	560	trade	541	trump	508
china's	492	amp	577	new	560	new	489	china's	492
new	489	trump	560	trade	541	india	475	new	489
chinas	477	kore	559	china's	497	countri	435	india	470
india	470	trad	542	india	473	one	433	chinas	470

Based on the results of the different stemmers, we decided to use Snowball as it created the least amount of errors. The only notable error is "amp" and "countri".

Without stemmers, there were too many repeat words. Out of all the stemmers, Lancaster performed the worse as for majority of the words, it cut off the last letter. Porter and WordNet both contained a significant number of repeat words. Surprisingly, none of the stemmers work perfectly.

China					
No Stopwords Filter		**Using Stopwords Filter**		**Removing “amp”**	
china	7669	china	7669	china	7669
the	5344	kim	641	kim	641
and	2384	trump	619	trump	619
for	1451	korea	604	korea	604
with	1357	amp	577	trade	541
that	1314	trade	541	new	489
you	923	new	489	india	475
are	875	india	475	countri	435
from	755	countri	435	one	433
this	744	one	433	chines	413

Choosing between using a stopwords filter, it is evident to us that without the filter, the top ten words are meaningless. However, we believe that “amp” is a frequent Twitter convention and it does not hold any meaning. Therefore, we removed it (by adding it to the stopwords list) and the next word that appears is “chines”. Upon further investigation, it is supposed to say “chinese”; however, the stemmer incorrectly stemmed it.

Below are the results for USA, based on Snowball with stopwords and excluding “amp”:

USA	
Word	Occurrences
usa	4622
trump	816
new	705
come	531
releas	515
final	495
congratul	494
download	480
android	470
superstarbt	464

Over the collection period, we recognized that many words did not make sense to us and in order to figure out the true meaning, we looked into each tweet containing the unique word and ensured it was relevant (such as

amp). Instead of manually viewing the tweets in a JSON file, we wrote a piece of code that collected all tweets containing our search term.

2.2 What are the ten most popular hashtags?

Top 10 Hashtags			
China		**USA**	
China	885	USA	1146
TRON	141	SuperStarBTS	464
TRX	141	Trump	170
NorthKorea	126	usa	128
Qanon	114	ReleaseTheVideo	116
G20	103	WrestleMania	114
blockchain	91	PJNET	89
crypto	91	Congress	88
cryptos	86	jobs	86
DICE	86	Cruz	85

Note: For the hashtags, we did not run a stemmer to collect as hashtags are often a collection of words. It would have also removed the true meaning behind each hashtag.

2.3 What are the ten most frequently tweeting person about the keyword?

Top 10 User Mentions			
China		**USA**	
realDonaldTrump	402	SuperStarBTStwt	465
spectatorindex	198	realDonaldTrump	260
AllenWest	196	TeaPainUSA	221
prayingmedic	147	kylegriffin1	121
justinsuntron	141	SethAbramson	119
mitchellvii	121	inittowinit007	110
DICEsup	86	GOP	106
XHNews	80	smartvalueblog	100
CGTNOfficial	71	JoyceWhiteVance	98
Ghanasfinestx	71	WWE	91

2.4 Whos is the most frequently tweeting person about the keyword?

Uses Twitter Too Much			
China		**USA**	
blacksmoke26	23	usajoblist	146
TheTomar	23	smartvalueblog	78
SihabBD	21	SalesCandidate	65
TeachDotJobs	20	WazeTrafficATL	28
compre_en_china	18	AlHamad02703934	23
yuuji_K1	15	GetaHotJob	23
jailingtrump	13	GetAJobUSA	22
MSpan10	12	discoversiberia	21
ClaudiaGhimel	12	KarenChestney	20
NapaDoji	11	USA_Biz	18

2.5 Which is the most influential tweet

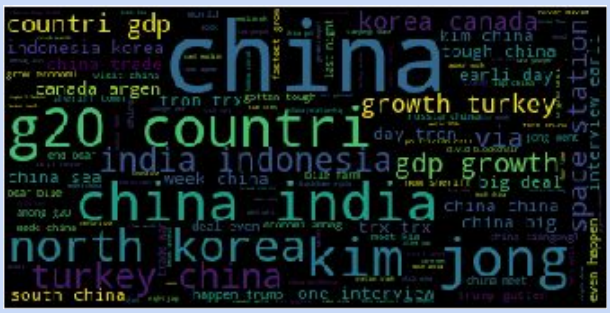
Strategy

We created a influence score variable that would be updated each time a tweet with a higher sum of retweet count, reply count and quote count appeared. Then, the text of the top tweet would be collected.
However, as all the tweets were streamed, there is no chance for each tweet to accumulate a retweet, reply, nor quote count. All these values are 0, and therefore, there is no influential tweet.

3.0 Word Cloud

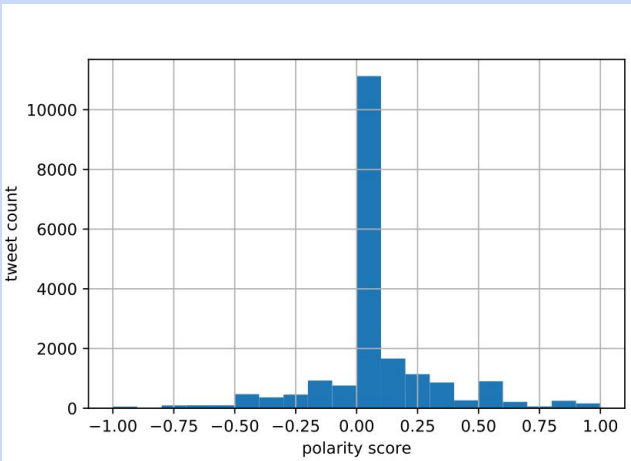
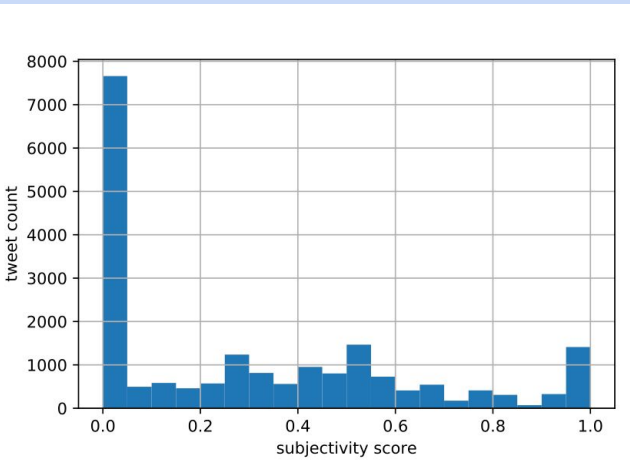
Strategy

In order to effectively collect words in the word cloud, the words have to be placed into a string. We ran a “for” loop that transferred all the words in the original list into one string; however, it resulted in sentences in the word cloud. Therefore, in the middle of “for” loop, we created a temporary variable that added a space after each word which fixed the sentence issue.

China	USA
	

4.0 Sentiment Analysis

Strategy
Instead of running two separate polarity/subjectivity tests on both JSON files, we opened both files at the same time and added each subjectivity/polarity value into one list.

Polarity	Subjectivity
	


```
Polarity Average: 0.06254648495572096
Subjectivity Average: 0.30659864243666857
```

5.0 Insights

Based on the top ten words under `China`, it is obvious that most of the tweets are controversial – as `kim` is the second most popular word, we can assume that most tweeters are referencing the relationship between China and North Korea. Unsurprisingly, `trump` also appears as another top word – this is likely a reference to the recent trade tariffs put in place by the POTUS. With `trade` as another keyword, it reinforces the idea that this is the topic tweeters are talking about.

Although the previous words are expected, the word that we did not expect was `india`. There has been no recent story between China and India, so it was surprising to see this word appear. When we ran a code to look at the tweets individually including the word `india`, it appears that most are referencing to the G20 summit. The tweets do not show any unique news story involving both the countries.

For *USA*, unsurprisingly, `trump` appears as the top word, but the other words do not appear to have any significance. The word `new` also appeared in the China analysis and when we looked into all the tweets individually, `new` was involved in various topics, without a significant pattern worth noting.

It was surprising to see `android` in the list, and after investigating this word, all the other words in the list started to make sense. A sample of a tweet involving `android:`

```
Download
☞ ios : https://t.co/cFiKBqmgUA
☞ Android : https:…
RT @SuperStarBTStwt: Congratulations!
#SuperStarBTS Finally Released In #USA!
```

Despite the massive number of controversial topics involving Trump and the USA, the largest recurring tweets were retweets of a Korean band “SuperStarBTS” and their release in the US. Therefore, the words `releas`, `final`, `congratul`, `download`, `android`, `superstarbt`, are all interconnected.

Under Hashtags, there are several interesting hashtags that are worth noting:

`#TRON, #TRX, #blockchain, #crypto,` are referencing the recent cryptocurrency market correction which affected all cryptocurrencies. TRX is the trade name for the TRON cryptocurrency. Although, Bitcoin and Ethereum are the most arguably the most popular cryptocurrencies, they did not appear in the hashtags. This is likely due to the suspected collapse of TRON. Cryptocurrencies are not a surprising topic in relation to China as they are commonly known to be one of the largest cryptocurrency mining countries in the world.

For USA, the most interesting hashtag is `#ReleaseTheVideo` as it references the hot topic with the relationship between Trump and a pornographic actress. Supposedly there is a video that was recorded of their affair and it appears that Twitter users are advocating for its release.

The top user mentions were unsurprising, as `@realDonaldTrump` is the most popular tag, whereas `@SuperStarBTStwt` represented USA’s most popular tag for this dataset. Meanwhile, there is no significant insight from the most active Twitter user as the numbers are relatively low without any significant names.

It was interesting to see polarity leaned on the positive side, and subjectivity was rather low, despite the controversy connected to China and USA. This may be due to the limitation of TextBlob to fully recognize the actual polarity/subjectivity. This was observed from the significant number of 0s in both sentiment analyses.

In conclusion, as the TweetStreamer API runs over a very short time frame, the tweets collected are not representative of what users are truly tweeting about. This is why SuperStarBTS appeared as one of the most popular topics despite all the other news stories around USA. If the TweetStreamer can be run over a period of one week to evenly spread out daily trends, a more accurate dataset can be collected.