DS577 Final Project – Twitter Data Analysis

Ruixuan Song, Tai Yang

December 12, 2018

Abstract

Our initial goal of this project is to analyze twitter data in a form of JSON file. We define two of our key words for data filtering in order to perform further analysis including top authoritative accounts, top URLs, top numbers of retweets, top hashtags, top significant words and word clouds. In this case, our two unique key words are Bitcoin and Thanksgiving. The scripting language we use is python and data processing tool we use is Hadoop and PySpark.

1 Introduction

In order to analyze twitter data on a narrower scope, our query words are "Bitcoin" and "Thanksgiving". The initial interest of having "Bitcoin" as one of the key words is that cryptocurrency has been a hot topic in recent years, after the price of the Bitcoin reached a peak at the end of 2017 for $19,783.06, the price of Bitcoin then decreased all the way down to $3,562.01 as for December 9th, 2018. This could leads to a lot of discussions among societies, and we could even based on the analysis in the twitter dataset, such as retweet number, top significant words, etc. to make some reasonable explanations.

Since twitter is all about things happened around people, we think that Festival such as Thanksgiving might be a good key word to explore in the twitter data. The word "Thanksgiving" could related to anything topics in twitter, such as "Food makes for thanksgiving", "what do people do during Thanksgiving", these wild topic give us a variety choices of what angle do we want to see from the data.

There are generally six main analytic metrics we want to address in the datasets:

- Top hashtags
- Top significant words (exclude words such as "and", "a", "or", etc.)
- Top URLs
- Top authoritative users
- Top retweets

- Word Cloud of significant words

The method we use for scripting the data is python and since the size of the data is relatively large, we use hadoop and spark to achieve above goals.

2 Experiment

2.1 Days of collection period

The total days of collecting the data is around 11 hours for both datasets, including creating a twitter developer account and use sample scripting code for collecting data for the two selected topics. Once the tasks have been created, we put them onto the iLab server under the computer science department at Rutgers University, so that we do not have to sit and wait for the data to be finished collecting.

2.2 General information about datasets

Name	Data size	No. of tweets	Avg. of producing tweet
Bitcoin	302 MB	~112.9K	2360 tweets/hour
Thanksgiving	2.1 GB	~652.4K	13.6k tweets/hour

Table 1: General information of datasets

Some challenges we met during data collecting stage include questions related to multiple languages and special characters. Since twitter data may not limit to only one languages, in fact, the table listed in the appendix shows that it include 64 different languages. Due to the nature of the language, some of them do not belong to the same language family, for instance, Chinese belongs to Sino-Tibetan languages while English belongs to the family of Westgermanisch. When dealing with "top significant words", the separators of words are not always the same for all languages.

3 Analysis

3.1 Design of code

After creating twitter developer account, we use a sample code called "twitter_streaming.py" listed on the course website to collect for Bitcoin and Thanksgiving datasets. Despite the length of the data collection periods, after replacing the token we get from our own developer account with the token given in the sample code, we gathered the code together and put them onto the iLab (a server under CS department at Rutgers University).

Since there's no stopping criteria on when to stop collecting the data, we decide to collect them overnight. After finished the data collection step, we move forward to analysis part. In order to analyze those 6 topics (stated before under introduction), we decided to use Hadoop to analysis topics "Top Hashtags"," top URLs" and "Top significant words", use Spark to analysis topics "Top authoritative account" and " Top number of retweets".

3.2 Snippet of code

```
#!/usr/bin/python3

import sys
import json

# Hashtags
for line in sys.stdin:
    try:
        tweet = json.loads(line)
        for item in tweet['entities']['hashtags']:
            print(item['text'], 1, sep=' ')
    except:
        continue
```

Figure 1: Mapper for top hashtags

```
#! /usr/bin/python3

import sys
import json
import string


for line in sys.stdin:
    try:
        tweet = json.loads(line)
        if 'lang' in tweet and tweet['lang'] in ('zh', 'ko', 'ja'):
            for word in tweet['text']:
                print(word, 1, sep=' ')
        else:
            for word in tweet['text'].split():
                print(word.strip(string.punctuation), 1, sep=' ')
    except:
        continue
```

Figure 2: Mapper for top significant words

Figure 1 is a mapper code screenshot for analyzing top hashtags, figure 2 is a mapper code screenshot for top significant words.

4 Results

In this section, we'll discuss and show our results obtained using the 5 analyzing metrics stated previously, and the word cloud for both bitcoin data and thanksgiving data.

4.1 Hashtags

The top hashtags we obtained are generally as expected. We can find that the Top 3 hashtags in both topics are highly related to what we intend to analyze. First outlier, however, comes from Thanksgiving dataset at rank 4. When looking into the table in detail. We find that Thanksgiving dataset, though with huge volume comparing to the Bitcoin one, has very few hashtags. This helps the irrelevant hashtags bursting out. Regarding to these irrelevant hashtags, we find that ThankUNext, Arianators and ThankUNextChallenge are related to an American singer-songwriter Ariana Grande, who released a new song called Thank U Next on Nov. 3, 2018. This date is very close to the date we extract the data. Therefore, the trend had not yet vanished then. Camilizers, on the other hand, is related to an American-Cuban singer-songwriter Camila Cabello, who also released a new album this year. Moreover, as Google indicates, Camila and Ariana had a talk in November. This may again push them into the top trend. Rank 10 in Thanksgiving dataset is related to a current TV series, S03 E08 of which titled "Six Thanksgivings" was aired on Nov. 20, 2018. That pattern was definitely captured by our scraper.

The 7-10 hashtags in Bitcoin is Agile, DevOps and UX. They are indeed related to Bitcoin. Bitcoin Agile is a real-time data platform, DevOps is a type of job in Bitcoin industry and User Experience (UX) is related to bitcoin software interface.

Ranking	Bitcoin		Thanksgiving	
	Hashtags	numbers	Hashtags	numbers
1	IoT	8064	Thanksgiving	17122
2	bitcoin	7379	thanksgiving	2588
3	Crypto	7188	ThanksgivingWeek	869
4	ICO	6827	ThankUNext	834
5	ThingCoin	6511	Arianators	832
6	Bitcoin	7510	SweetenerWorldTour	832
7	BlockChain	4144	Camilizers	830
8	Agile	3771	ThankUNextChallenge	830
9	DevOps	2887	HappyThanksgiving	597
10	UX	2776	ThisIsUs	493

Table 2: Top Hashtags

4.2 Top URLs

In this section, we are going to discuss the top URLs. There are some quite interesting results. The results are listed in Figure 3 and Figure 4. First we'll cover the Bitcoin dataset.

In Bitcoin dataset, some of the URLs are just other tweets. We only focus on the other hyperlinks. First one is a online shop, selling goods related to pilots. Quite interestingly, the online shop even sells helicopters and aeroplanes among other things. The second one, "discord.gg", is a privacy-preserved chat platform widely used in bitcoin trade. Third one, the medium article and the tenth one, the YouTube video, are both some introduction of Bitcoin. The sixth one, "catex.io", is a cyptocurrency trading platform, including Bitcoin. They are more or less related to Bitcoin.

https://aerotrips.com/shop	253
http://discord.gg/SBM97UK	144
https://link.medium.com/CHR09k8AXR	98
http://t.me/queen_signalsss	80
https://twitter.com/fredCwam/status/1065171503...	79
http://catex.io	78
https://twitter.com/f4izalhassan/status/106450...	72
https://www.cryptohedgef.com/	71
https://twitter.com/newsycombinator/status/105...	68
http://youtu.be/e_9QouZWUvc	67

Figure 3: Bitcoin Top URLs

In Thanksgiving dataset, though majority of them are tweets link, there is one irrelevant link pointing to a new book called Dragon Racer 2, however. Published on October 29th, 2018, the book was on sale during Thanksgiving period.

4.3 Top Significant Words

The top significant words are quite normal, as listed in Figure 5. We can compare them with the hashtags (Table 2). In Bitcoin dataset, they are quite similar. They all point to the cryptocurrency we extract.

However, in Thanksgiving dataset, some outliers, like singers, are gone, replaced by some emotional words in such festival. Trump, though, is always there, since he is a very big fan of Twitter. Japanese word "の", meaning "of",

https://twitter.com/iilcollegegirl/status/1064...	5287
https://twitter.com/joshscampbell/status/10649...	2360
http://ABCNews.com/live	1362
https://twitter.com/vidcon/status/106454558479...	747
https://twitter.com/kylegriffin1/status/106498...	688
https://twitter.com/mikeseidel/status/10648878...	688
https://twitter.com/darrenrovell/status/106506...	473
https://twitter.com/mommmasaid/status/10628705...	465
https://twitter.com/VidCon/status/106454558479...	369
https://www.amazon.com/dp/B07GNPN629	241

Figure 4: Thanksgiving Top URLs

is the only non-Latin character in the top 10 lists. We decided to keep it though it is actually a stopping word.

RT	28204
Bitcoin	22796
bitcoin	10906
Crypto	8712
IoT	8114
ICO	7044
ThingCoin	6511
UserExperienceU	6204
BlockChain	3803
BTC	3751

(a) Bitcoin

RT	202294
Thanksgiving	158449
thanksgiving	60299
Happy	19043
Trump	18157
days	18062
の	17580
year	17298
will	16145
family	15372

(b) Thanksgiving

Figure 5: Top significant words

This also leads to the discussion of the technique issues here. A brief discussion is on page 2. Since different languages have different separators, we deal them with different methods, as shown in Listing 1. We treat East Asian languages as a group and assume other languages the similar way as English (space-as-separator).

```
#! /usr/bin/python3

import sys
import json
import string


for line in sys.stdin:
    try:
        tweet = json.loads(line)
        if 'lang' in tweet and tweet['lang'] in ('zh', 'ko', 'ja'):
            for word in tweet['text']:
                print(word, 1, sep=' ')
        else:
            for word in tweet['text'].split():
                print(word.strip(string.punctuation), 1, sep=' ')
    except:
        continue
```

Listing 1: Top significant word mapper

4.4 Top Authoritative Account

We now discuss the authoritative accounts. From Figure 6 on page 7, we can see that all but one are News media. "ye" in Thanksgiving dataset, is the Twitter account of American rapper Kanye West. He himself participated the Thanksgiving parade. That's why he was in the list of our data. Since he is a very popular star, the huge number of his followers is not surprising.

Reuters Top News	19991612
Forbes	14996057
NDTV	10995618
La Patilla	6828592
NBC News	6260067
ABS-CBN News	5918143
The Hindu	5005645
Bloomberg	5002254
MarketWatch	3599891
CNBC	3008531

(a) Bitcoin

The New York Times	42341389
CNN	40775502
ye	28642036
BBC News (World)	24087709
National Geographic	22352000
Reuters Top News	19992916
The Wall Street Journal	16149026
ABC News	13964238
Vogue Magazine	13549559
The Washington Post	13070373

(b) Thanksgiving

Figure 6: Top Authoritative Account

4.5 Top Number of Retweets

Since the length of the retweet varies, we insert stars "*" to separate different tweets in Figure 7 and Figure 8. Table 3 on page 8 indicates the specific number of retweets for each of these retweets' context in two data sets.

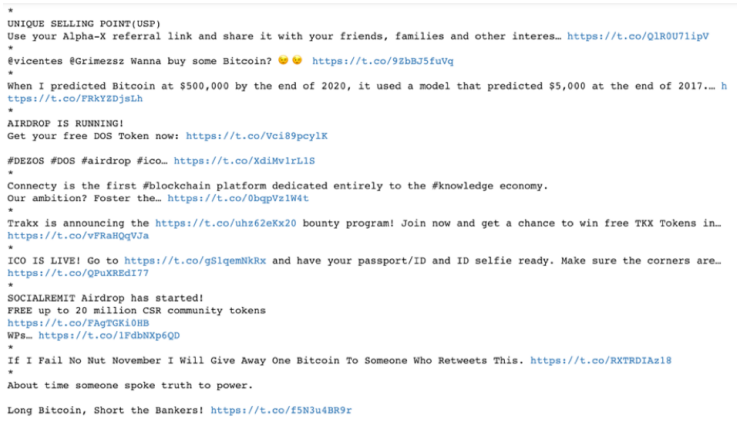

Figure 7: Bitcoin Top number of retweet

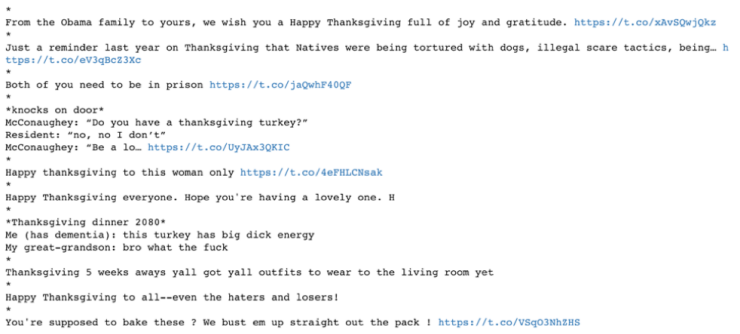

Figure 8: Thanksgiving Top number of retweet

Ranking	Bitcoin	Thanksgiving
1	13363	208932
2	8951	143189
3	7042	106868
4	5812	98105
5	5031	90144
6	3813	80388
7	3431	77569
8	3269	69519
9	3196	68009
10	2863	64668

Table 3: Number of Retweets

We can see that in Bitcoin dataset, people talked about the Bitcoin. Either how they were good at investment or how to trade Bitcoin.

While in Thanksgiving dataset, some very emotional words appeared. They were all expressing best wishes to beloved or cared people.

4.6 Word Cloud

In this section, we will discuss the word cloud generated by the top words in each topic. The words make sense because they all related to the topic. (See Figure 9 on page 9 and Figure 10 on page 9.

Figure 9: Bitcoin Word Cloud

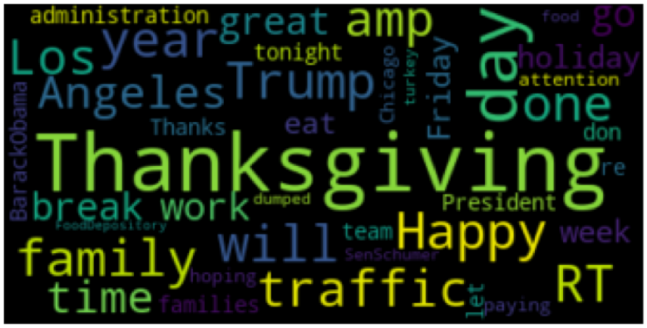

Figure 10: Thanksgiving Word Cloud

4.7 Top Language

We add one analyzing metric in our project, which is most used languages in each of the two datasets, see Figure 11. We can see that in both datasets, English is the most used language. It make sense since English is indeed the most

commonly used language across the world. One interesting details we found is that the second commonly used language for both datasets is "undefined". We figured that it might because the system could not recognize user's language. The reason is that it either contains too much emojis, less full sentences with only a few words contains in the tweet, or it begins with numbers or other special characters.

en	44801	English
und	3018	undefined
ja	2270	Japanese
es	1609	Spanish
fr	1171	French
tr	640	Turkish
in	621	Indonesian
de	432	German
pt	324	Portuguese
it	316	Italian

(a) Bitcoin

en	313184	English
und	4987	undefined
ja	3152	Japanese
es	1833	Spanish
fr	672	French
tl	377	Tagalog
pt	317	Portuguese
de	245	German
in	199	Indonesian
ht	160	Haitian

(b) Thanksgiving

Figure 11: Top Language

5 Conclusion

During this project, the biggest challenges we met is when analyzing the top significant words metric for each datasets. Due to the nature of languages, we have to take different separators under account in order to successfully partitioning. Moreover, we have to eliminate frequent used words such as "a", "the", "and" and punctuation in each tweet. We also need to consider cases when trying to filter out those preposition words.

Data size is another challenge we met, since our Thanksgiving data is relatively large, when we generate the word cloud, it cost some times to converge. Although we've met some unexpected challenges during this project, the outcome is quite satisfying and some interesting facts were being discovered.

In this project, we are able to use what we learned from class such as Hadoop, Spark and some python packages such as numpy, pandas, etc. to apply on a real world data analyzing task which gave us more hands-on experience on dealing with various data problems, we believe it's going to help us improve our data analyzing and wrangling skills in our future careers.