Statistics Software Final Report
--The mystery behind ‘Discount’ and ‘Le Accor’
Kailong Wang

Introduction

Since Twitter launched in July of 2006, the service rapidly gained worldwide popularity. In 2012, more than 340 million tweets were posted by over 100 million active users per day, and the service handled an average of 1.6 billion search queries at the same time. In 2013, it was one of the ten most-visited websites and has been described as "the SMS of the Internet”. Since 2015, and continuing into 2016 and future years, Twitter has also been the home of debates, and news covering Politics of the United States, especially during the 2016 U.S. presidential election. People describe President Trump run American by Twitter, or Trump make Twitter great again. Since that, Twitter proved to be the largest source of breaking news on the day of the 2016 election.

Behind great successful, Twitter also faces big challenge. The free open source platform not only gives people access to express themselves, but also attracts the dark side of world. During the election, rumors indicate that there were thousands spies and hackers trying to influence result by tweets. Common restricted messages are also easy to spread through the website. And there are thousands of hundreds zombie accounts tweeting trash messages every day. It is impossible for twitter to filter such large amount web traffic manually. Design a powerful automatic tool is urgent. Text message analysis with open API and large data operation technique is the start of it.

This project using Twitter developer API collected two datasets from Dec 6th to Dec 11th. The topics are ‘discount’ and ‘Le Accor’. Techniques mainly implemented are python 3.6 with jupyter notebook. Although Hadoop and Spark are better to handle large dataset, my datasets are not necessary to use such techniques. With Occam's Razor principle, Hadoop and Spark are discarded. By implementing such methods, it is clear to see how different a topic could be through different view and how many advertisements is spreading on the Internet.

Experiment

Topic	Discount	Le Accor
Keywords	discount, deal, promotion, freebie, doctorofcredit, slickdeals, amexoffer, chaseoffer, coupon, cashback	leaccor, accorhotel, leclub, Fairmont, ibis, Sofitel
Keywords Discard	black friday, offer, gift card, gc, ms, mo, bonus	presidentclub
Collecting Time	8hrs (Dec 6th)	72+hrs (Dec 7th~Dec 11th)
Size	2.26G	43M
Number of Tweets	300,000+	8808

Table 1

The first topic was chosen by personal interests. There are many discounts and promotions going on in America. Since last year I was able to buy 10,000 printing paper for $60 with $80 rebates, I begin to enjoy digging on-going deals and take advantage of it. Keywords I selected are directly related to topic, by meaning or contents. I discard Black Friday because it is an annually event which is not common. GC, MO, MS means gift card, money order, manufacture spend respectively. However, shortcuts may lead to irrelevant information significantly so they must be discarded to keep dataset clean. Since manufacture spending is generally illegal, people are not likely to talk about how to implement it but against it, which is not the topic. And thus, it needs to be discarded. Although I carefully selected my keywords, 'deal' still makes a lot trouble in analysis because it is a verb, but also display various results which is identical. The challenge I face in this topic is that I reach hourly data traffic limits multiple times. This makes my collecting Jason files contains special entities (Figure 1). To eliminate these entities, I use filter, lambda and list function filter out unusual result and get a clean json file.

```
tweets_data[38176]
```

```
{'limit': {'track': 5, 'timestamp_ms': '1544121146334'}}
```

Figure 1

The second topic I choose is a hotel brand. 'Presidentclub' is discarded because it refers to Mr. Trump more than topic with a sample test. There is not much challenge in data collecting. But due to low volume of data traffic, I have to collect it long time to get enough data point. The real challenge is in analysis.

Analysis

Topic 1

Top 10 Common Words	
Deal	2986
Will	574
Brexit	392
Amp	295
New	283
Now	268
People	251
Bonus	239
EU	228
One	226

Table 2

To do analysis, we need to first import collected dataset into python environment. With for loop and list construction, data is successfully imported and has 307589 data points. After removing 5 'limit' datapoint and stop-words, we first take a look at word frequency. From top 10 most common words, there is not much information related to

our topics. Although ‘bonus’ still appears 239 times, it is far less than the number of whole datasets which is more than 300,000. ‘Deal’ is the most common word, but given it is also a verb, the usage here should relate to short phrase like ‘deal with’ more than my topic. Besides, ‘brexit’ and ‘eu’ indicate that the hottest topic on Dec 6th is British Brexit event. This is further demonstrated by word cloud. (Figure 2) On the left image, we still see many keywords, so I update stop-word set and get the image on the right. Theresa May is British Prime Minster, Jim Jordan is American congress officer. People mention these two politicians mainly talking about ‘Brexit’. There are also many tweets regarding food and supplement which showed British people worrying about economic drops. The Canadian related words here is about Meng Wanzhou, Huawei CFO and Founder's Daughter was arrested at Canada airport. People believe this is a ‘deal’ between American government and Canadian Government to slow down China’s development on 5G technology. Over all, it may because the time had many political deals on-going, but there is hint that common reflecting my topic.

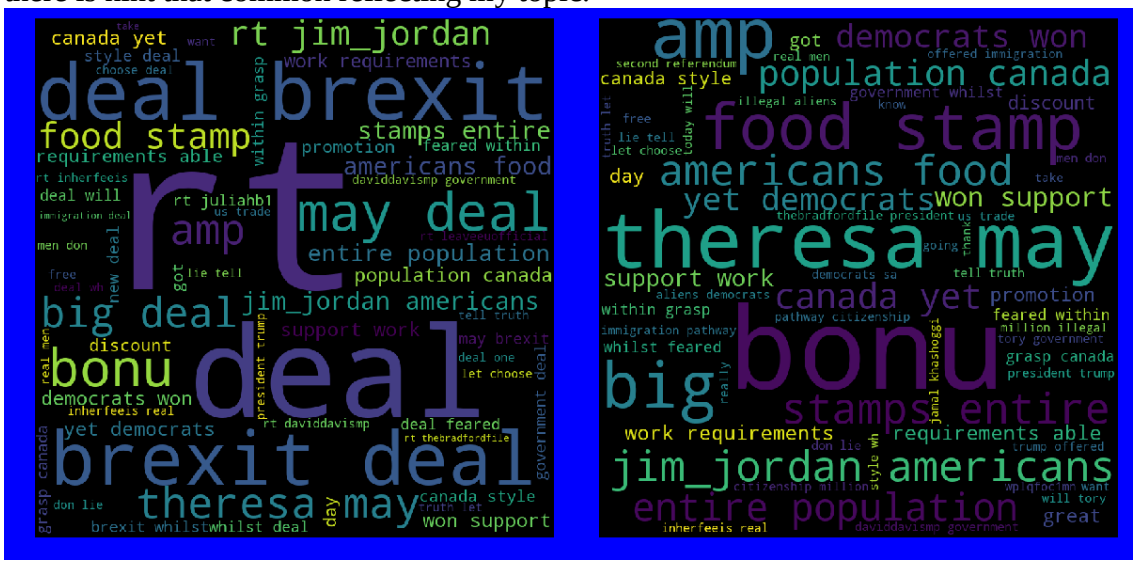


Figure 2

Moving to top retweets (Table 3), we see another trend. Not showing many political perspectives, instead people who shared their thought and emotions got most retweets. The reason behind this has three explanations. First, regarding hot topics, people prefer expressing their own ideas rather than retweeting others. In this case, although ‘Brexit’ and ‘Huawei’ is hot at the time, they didn’t gain enough retweets on individual tweets. Second, people are easy to be attracted by beauties, famous and animals. These are top 3 retweets’ original twitters. If you check the rest 7 tweets, this rule is still applied. The third reason may because different people have different twitter preference. Those who retweet others’ tweets doesn’t share similar preference on tweets with those who write their own tweets. More specific, people who retweet most prefer slogans, animals, shows while people who write tweets most prefer political, hot topic tweets. No matter which the right explanation is, top retweets shows totally different trend with frequents word and both are not related to the topic, ‘discount’. From top retweet, they are caught because there is deal as verb in content. This is what I didn’t expected and should be considered for next twitter scraping.

Top Retweets	
You can go to the gym, drink your water and take your vitamins, but if you don't deal with the shit going on in you…	87042
Want more?! @bts_twt gives an incredible bonus performance of "I'm Fine"	79195
When your partner falls asleep in an awkward position but you love them so much you just deal with it.	72911
How I deal with my depression.	65548
The ocean is the scariest place on earth, don't @ me.	65097
if you remember this you're entitled to a veteran's discount	55273
Some heros wear capes. Some wear Saint Laurent.	53277
Alec Smith died 3 days before his next paycheck, waiting to buy insulin. • He aged off his mom's insurance • His j…	50484
Yeah sex is cool...But can you deal with my depression? Will you calm me down when i have anxiety attacks? Will You…	48892
Can you guys make sure your dads are home so I don't have to deal with this	48408

Table 3

User profile analysis is also important in twitter study. By doing this, it is possible to locate user with unusual activity pattern. These users are generally talk bots or zombie account. For example, if a user never posts any tweets but always like a few twitters' tweet, they are possible to be sold as commercial account. Or if an account post twitter every second, it is definitely a robot rather than a person. From twitter json file, we can get all user information from user entities. Figure 3 shows the most common language used in this collection. Since English twitter is way more than the third language, French, the third to twelfth common language is displayed on the right. It is clear that English has 100 times more usage than French. The reason behind it can be tracked through location.

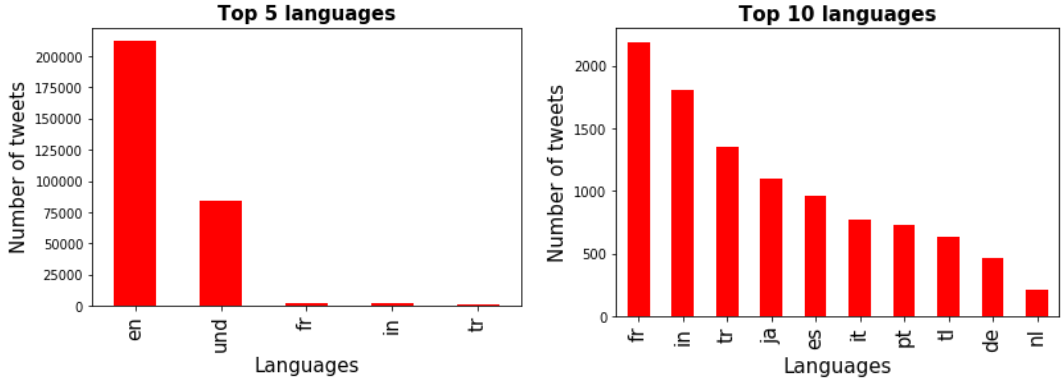


Figure 3

From user locations (Figure 4), US and UK are most common countries and London is the most common city. Although UK and US have even performance on city base locations, US still surpass UK a lot on country base. This is because US has more population and area than UK. Although 'Brexit' definitely is a topic from UK, it cannot omit the impact from US person. Since both US and UK speak English, there is no wonder why English dominate twitter language rank.

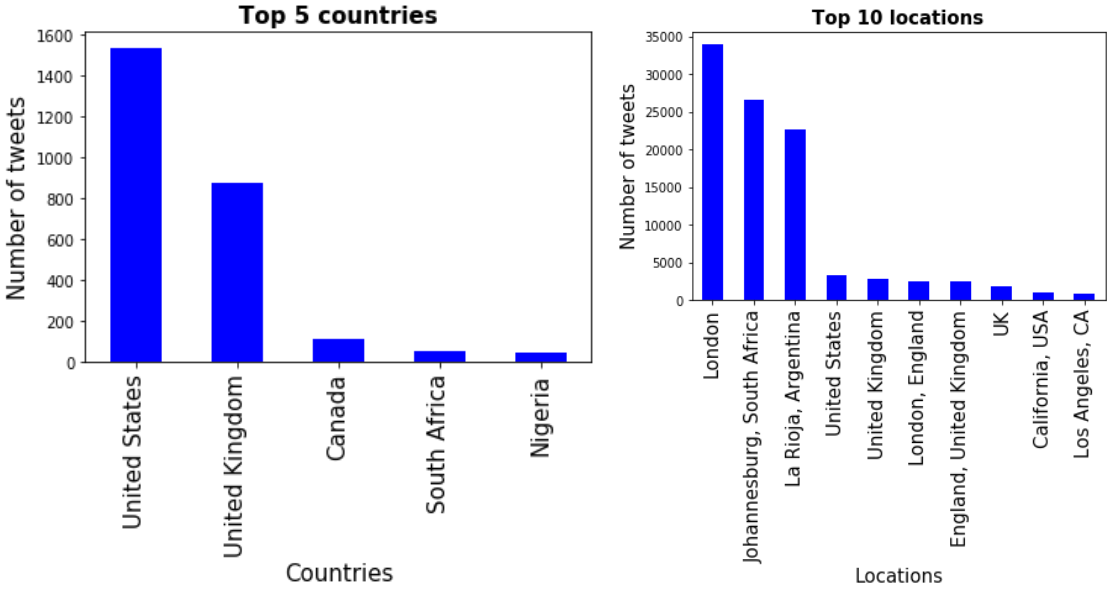
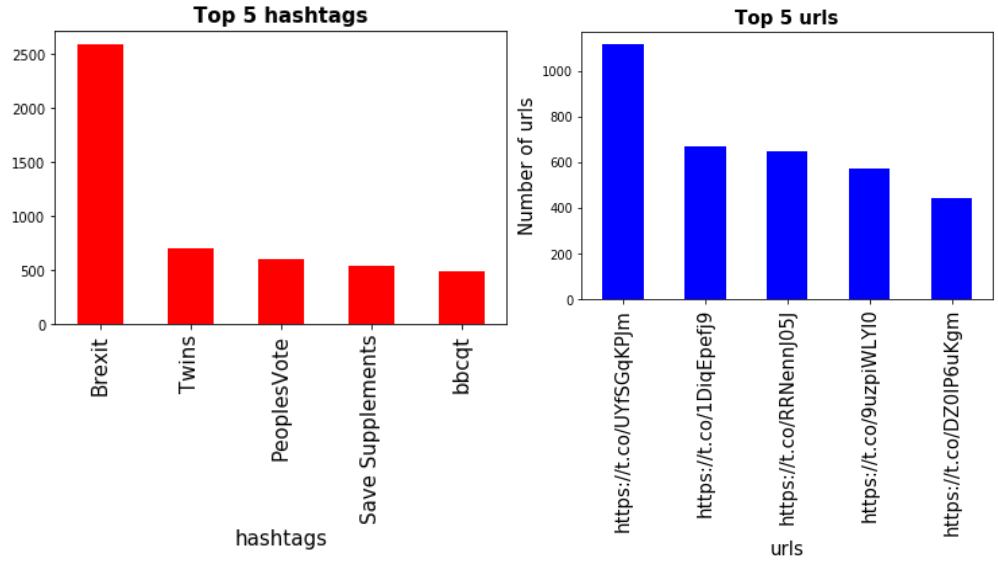


Figure 4

Then which user post most twitter? The number one user has significant unusual pattern. Given my collection time is less than 8 hours, he/she post more than 4000 tweets per hour. That is more than one tweet per second. That is why his/her twitter was deleted by twitter and has no activity since Dec 7th. Similar result also applies to the second and third user though their account is still activated. And finally, the fourth user is related to the topic.

Tweet Most Often	
BareNoize	31383
sarah_koopman	26424
SergioGCasas	22605
DealDonkeyUS	267
amruthasuri	212
thewoogyu	138
rezaisl74859661	110
womanorium	101
MdAlvi05023981	101
michela_tan	90

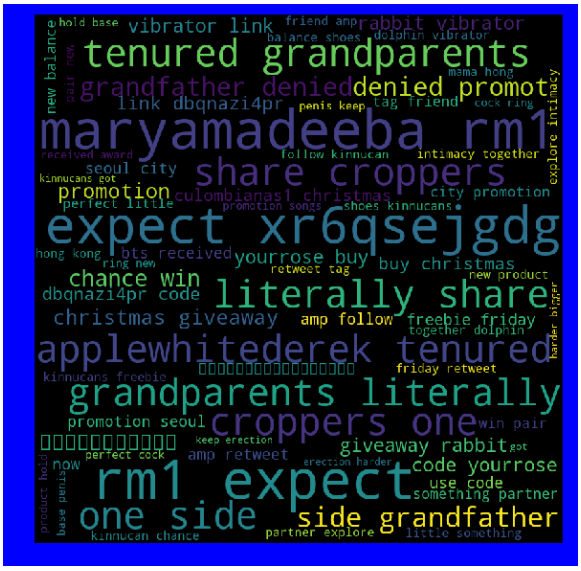


Top hashtag shows strong correlation with most common words. ‘Brexit’, ‘Vote’, ‘Supplements’ are also viewed in word-cloud. On the other hand, top URLs are very mystery with no clue. The top three urls are no access right now. A hypothesis states that they are shared commonly by top three users, which has been denied and blocked by twitter. The fourth url is about Julian Assange from British. This further increase tweets number published from English region and UK.

From above analysis, there are several conclusions can be established.

1. Don’t use verb in keywords and choose keywords carefully
2. After removing stop-words, word frequency does show the hottest topic
3. Top retweets may not relate to hot topics because of various reasons
4. Beauties, Famous, Cute animals are easy to get retweets
5. The more you tweet doesn’t means the hotter you are. Don’t waste your time
6. People tweet in English 100 times more than the second languages
7. US is too large, UK is too small. UK have tried hard, but…
8. People really like to talk about politics

These conclusions can be further proved in the second topic. Before that, a new data collection is generated on Dec 14th without keyword ‘deal’. There are small hint showing ‘discount’ topic like ‘giveaway’ and ‘expect’. But in general, there are more words with no sense. Finding great deals through twitter text analysis seems unreliable.

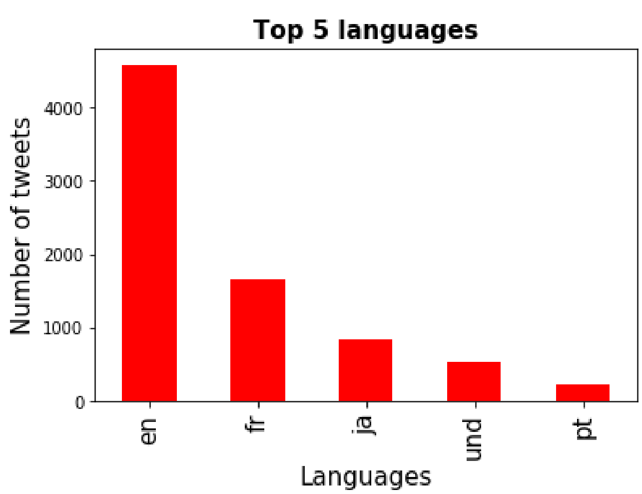


Topic 2

Comparing with the first topic, the second one is really low heat. Since data collected is separated in different files, a couple for-loop is designed to append all files in one. Word-cloud has showed difference. There is no trend of hot topic displayed from high frequent word. Instead, they are nonsense words like the new discount word-cloud. It is possible that, the fewer sense word-cloud is, the better keywords picked.

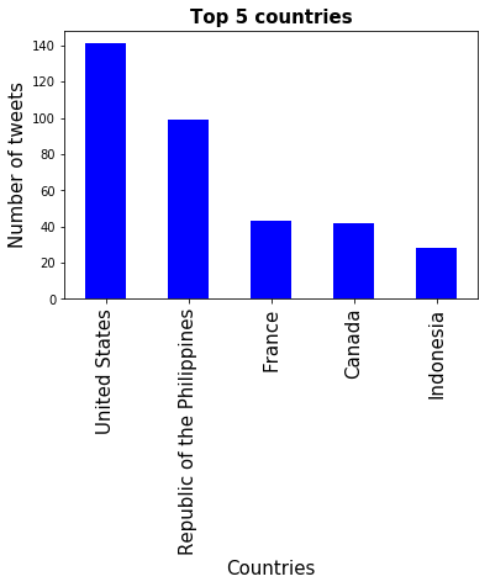
There is not much retweets in this dataset. I want to show top 10 retweets, but there are only four retweets in total. And none of them is related to the topic.

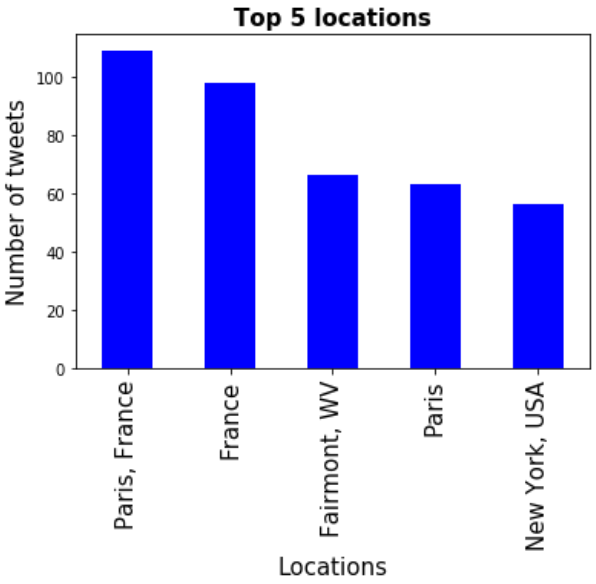
Top Retweets	
OP Neyshen after SPS / Sofitel Date HAHAHAHA #ALDUBNewbornBabies	3115
มายืนยัน นอนยันค่ะว่าแอพ ibis paint ดีจริงงง เป็นแอพวาดรูปในโทรศัพท์ที่ใช้มานานที่สุดเลย มีแปรงเยอะมาก ไม่ต้องเสียเ	2614
เมื่อกูโหลด ibis ตามคนในทวิตมาลองเล่นๆ... เล่นๆไรมึ้งง	2333
そんなわけでご無沙汰してます。『Ibis Pink』トキさんのお話その１です。前回よりちょいと長めです。	1431



The user file from this dataset is very interesting. English still dominate language ranking but the difference between English and French is much less. This is because Accorhotels

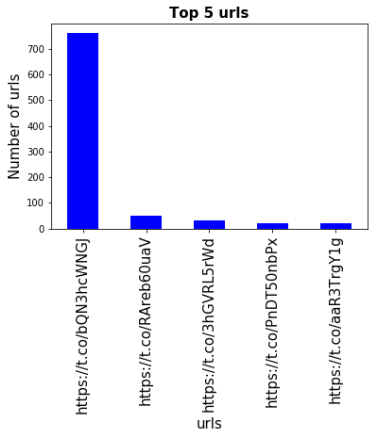
is a French company. There is a large French user base. However, the second popular country to post tweet is from Philippine not from France or Paris. But Paris ranked as number one on location. So, what happened? Basically, Paris people still post most tweets. But there is a city in West Virginia called Fairmont. And considering population and area of America, it is not strange that US is still the most active country in this dataset. Regarding post from Philippine, this can be explained by most tweet user.





Tweet Most Often	
travelbagtours	47
gccjobs4u	45
LilianP24	43
trendingwwwandw	27
OCFA_Bot	23
ashabakah_com	19
magnifyk	16
gunz11895894	16
TheMountainEast	15
SugeChaos_ACbot	13

The first three user post tweets with regular pattern every day. And they are all from Manila, Philippines doing travel agent and recruiter job. It is no wonder why they are so active and make Philippines take the second place. On the other hand, their contents have high similarity, which increase word frequency significantly but doesn't offer much information.



Look into the top urls, all of them are from accorhotel website with advertisement and special offers. The discount I'm looking for in discount topic showed here. Similarly, advertisements have most usage, which rank number one in hashtags, and it is even more than the sum of the rest.

Top Hashtags	
Offer_Hotels Makkah movenpick Swiss Conrad Hayat_regency Hilton_Suites Hilton_convention Marriott Fairmont	368
spiralbuffet sofitelmanila	49
一年で絵が上達してるか見てみましょう 糸会 オリジナル 絵描きさんと繋がりたい	43
FairmontScottsdalePrincess Arizona	40
ぽこあーと オシャレになりたいピーナッツくん	23
ReformAndOpeningUp	20
Vijayawada	20
wvprepbb	15
NLA2018	14

Overall, this experience is pretty valuable. To get most from twitter text contents, choose a right topic is more important than design well-constructed keywords.