Twitter Project Report

Wanting Tan

1. Topic Introduction

For this project, I picked Skin Care and Trade War as my two topics. The summary table is listed below. For the Skin Care topic, the reason why I choose it is because I'm thinking of changing my skin care product and I want to know what other people are using and their opinion. The key words I used are skin care, anti-aging and skincare routine. Since my key words are very specific, it took me 39 hours to collect a meaningful dataset. The final file size is 74.3 MB and number of tweets is 14044.

For the Trade War topic, the trade war between China and the US is a very popular topic these days. And especially in Dec.2, the US and China agreed to a temporary truce to deescalate trade tensions. Therefore, it's a great time to collect tweets to gather everybody's opinion. It took me around 30 hours and the number of tweets is 35758.

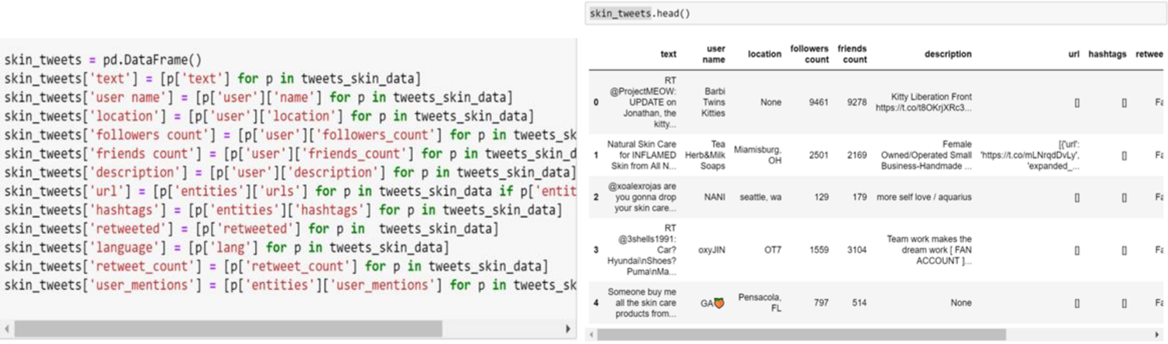
After getting all the data I need, I started my analysis.

Topic Name	Key Words	Time	File Size	Number of Tweets
Skin Care	Skin care Anti-aging Skincare routine	**Total: 39h** Dec.3 1h Dec.4 10h Dec.5 18.5h Dec.7 9.5h	74.3 MB	14044
Trade War	China trade China trump Trade war	**Total: 29.5h** Dec.3 1h Dec.4 10h Dec.5 12h Dec.7 6.5h	310 MB	35758

Table 1

2. Challenges

The challenges for these two topics are quite different. For the skin care topic, since the dataset is relatively small. I could just use data frame to deal with it. However, it's a little time-consuming when I transfer JSON file into data frame. I used many list generators to deal with it. Following are screenshot of code and the data frame I got.



```
skin_tweets = pd.DataFrame()
skin_tweets['text'] = [p['text'] for p in tweets_skin_data]
skin_tweets['user name'] = [p['user']['name'] for p in tweets_skin_data]
skin_tweets['location'] = [p['user']['location'] for p in tweets_skin_data]
skin_tweets['followers count'] = [p['user']['followers_count'] for p in tweets_sk
skin_tweets['friends count'] = [p['user']['friends_count'] for p in tweets_skin_d
skin_tweets['description'] = [p['user']['description'] for p in tweets_skin_data]
skin_tweets['url'] = [p['entities']['urls'] for p in tweets_skin_data if p['entit
skin_tweets['hashtags'] = [p['entities']['hashtags'] for p in tweets_skin_data]
skin_tweets['retweeted'] = [p['retweeted'] for p in  tweets_skin_data]
skin_tweets['language'] = [p['lang'] for p in tweets_skin_data]
skin_tweets['retweet_count'] = [p['retweet_count'] for p in tweets_skin_data]
skin_tweets['user_mentions'] = [p['entities']['user_mentions'] for p in tweets_sk
```

```
skin_tweets.head()
```

	text	user name	location	followers count	friends count	description	url	hashtags	retwee
0	RT @ProjectMEOW: UPDATE on Jonathan, the kitty...	Barbi Twins Kitties	None	9461	9278	Kitty Liberation Front https://t.co/t8OKrjXRc3...	[]	[]	Fa
1	Natural Skin Care for INFLAMED Skin from All N...	Tea Herb&Milk Soaps	Miamisburg, OH	2501	2169	Female Owned/Operated Small Business-Handmade ...	[{'url': 'https://t.co/mLNrqdDvLy', 'expanded_...	[]	Fa
2	@xoalexrojas are you gonna drop your skin care...	NANI	seattle, wa	129	179	more self love / aquarius	[]	[]	Fa
3	RT @3shells1991: Car? Hyundai\nShoes? Puma\nMa...	oxyJIN	OT7	1559	3104	Team work makes the dream work [FAN ACCOUNT]...	[]	[]	Fa
4	Someone buy me all the skin care products from...	GA🍑	Pensacola, FL	797	514	None	[]	[]	Fa

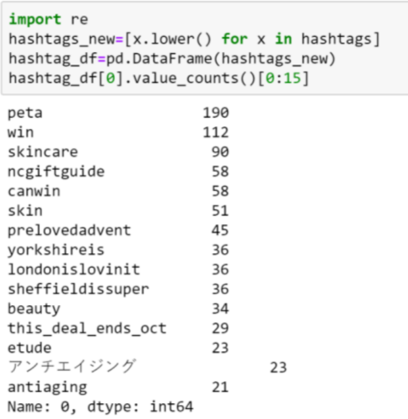
Figure 1

The other challenge is that the data is hard to clean because the data contains a lot of advertisements that many tweets have links, symbols and emojis.

For the trade war topic, the biggest problem is the dataset size. It's quite slow when I did in data frame, especially when I dealt with top important words. That's why I also tried Hadoop. However, it's a little bit hard to write map function for top significant words in Hadoop because I can't use packages like NLTK, stopwords, at least in my computer. It maybe because low version of Python in my virtual machine. Therefore, I need to find other method to clean data.

3. Topic 1 in Data Frame

- **Top hashtags**
  After getting all the hashtags from raw data by using list generator, I used value_count() in data frame to count and sort top hashtags. Following plot is the code and summary table of top 15 hashtags.



```
import re
hashtags_new=[x.lower() for x in hashtags]
hashtag_df=pd.DataFrame(hashtags_new)
hashtag_df[0].value_counts()[0:15]
```

```
peta                  190
win                   112
skincare               90
ncgiftguide            58
canwin                 58
skin                   51
prelovedadvent         45
yorkshireis            36
londonislovinit        36
sheffieldissuper       36
beauty                 34
this_deal_ends_oct     29
etude                  23
アンチエイジング                23
antiaging              21
Name: 0, dtype: int64
```

Figure 2

Peta is the top 1 hashtag. Peta is an organization where can protect and rescue animals. At first glance, it's hard for me connect it with skin care product. Then I looked back specific retweets and found that people were talking about "peta-approved" skin care product. The ingredients from this kind of products have no harm to animals. That's why many tweets contain "peta" when introduced products. "win" is also very popular in hashtags. Actually, it appears very frequently in advertisements to give a promotion to customers. Words like this also include "canwin" and "thes_deal_ends_oct". "skincare", "skin", "antiaging" and "beauty" also show up in the list, which is reasonable since they are my keywords (the Japanese word in the list means anti-aging).

- **Follower count vs friends count**
  I'm interested in the correlation between follower number and friends number, so I drew a correlation plot between the two (plot below) by using sns.jointplot(). We can see a positive correlation between them. It also gives pearson coefficient and corresponding p-value. They all show that follower number and friends number are highly positive correlated. Another fact is that the average value of follower number is higher than friends number.

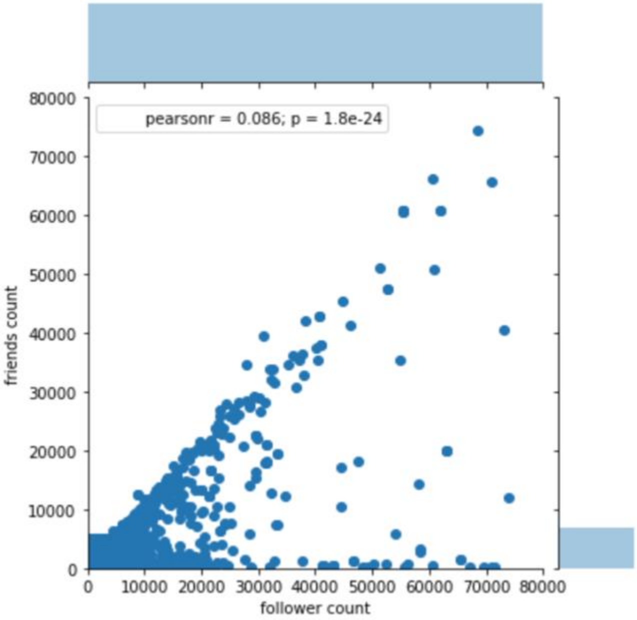


Figure 3

I also tried to plot density or histogram for follower number. However, it contains large number of large values. In order to squeeze the range, I took logarithm of it and then plot. It's clear to see there's a little hip at zero value and rest of the plot shows a normal distribution.

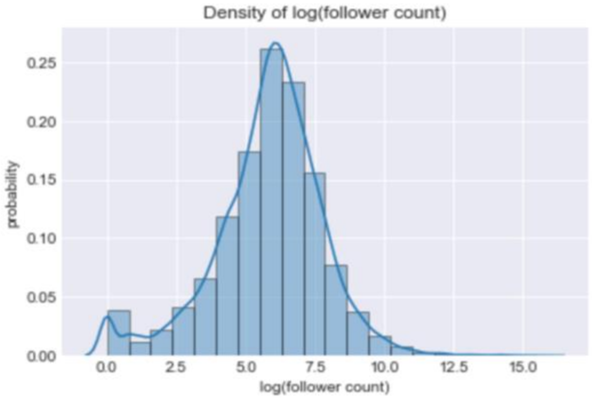
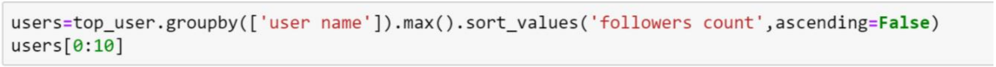
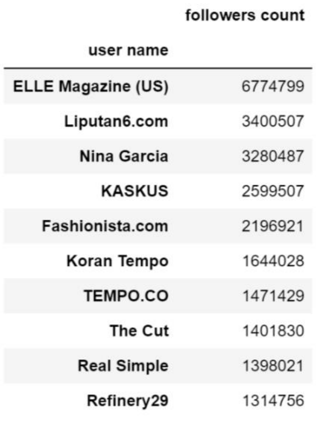


Figure 4

- **Top authoritative users**
  I choose users who have top number follower numbers as top authoritative users. However, during the time of collecting data, the follower number may change. So, I need to find out the maximum follower count for each user and then sort by it. Following is the code and summary table.



```
users=top_user.groupby(['user name']).max().sort_values('followers count',ascending=False)
users[0:10]
```

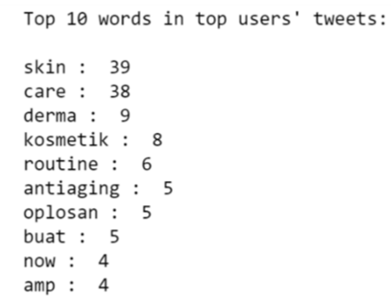


	followers count
user name	
ELLE Magazine (US)	6774799
Liputan6.com	3400507
Nina Garcia	3280487
KASKUS	2599507
Fashionista.com	2196921
Koran Tempo	1644028
TEMPO.CO	1471429
The Cut	1401830
Real Simple	1398021
Refinery29	1314756

Figure 5

Elle magazine (US) is a lifestyle magazine that focuses on fashion and beauty and Nina Garcia is the editor of this magazine. That's why they showed up together. Fashionista and The Cut are fashion industry who creates or promotes high fashion. Liputan 6 is an Indonesian flagship television news program. KASKUS is an Indonesian internet forum. Koran Tempo is an Indonesian newspaper and tmpo.co is their website. These Indonesian tweets show up frequently in my data maybe because I collected part of my data during the midnight and there were all kinds of promotions starting in these websites.

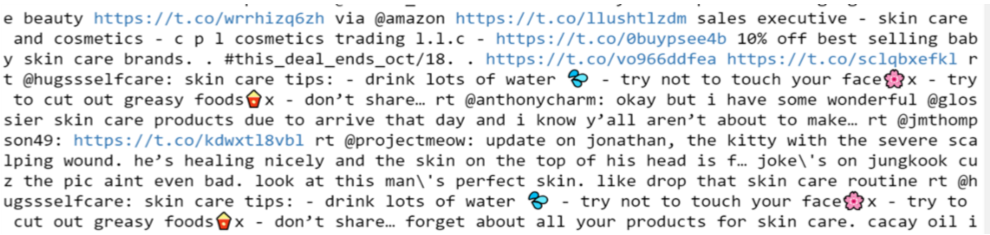
I also collected top 10 words in top 30 users' tweets (table below). "skin" and "care" are the most popular words, which is as expected. "derma", "cosmetic" and other words are all related to skin care. Since this is only small sample of data, we can see more detail in top significant words part.



```
Top 10 words in top users' tweets:

skin :  39
care :  38
derma :  9
kosmetik :  8
routine :  6
antiaging :  5
oplosan :  5
buat :  5
now :  4
amp :  4
```

Figure 6

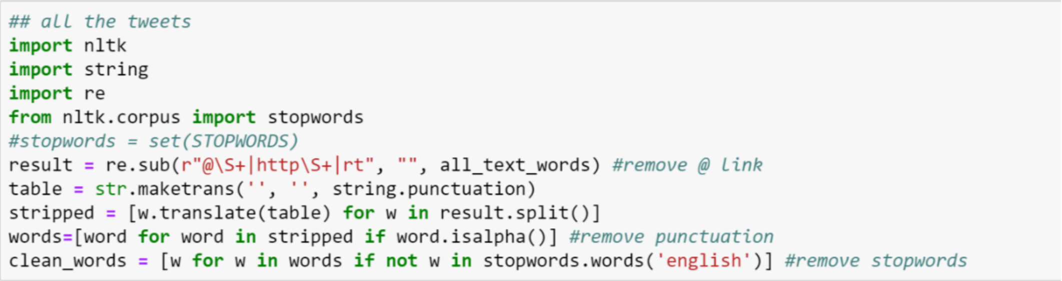
- **Top significant words**
  In this part, I put text of all tweets together. This is a screenshot of raw data.



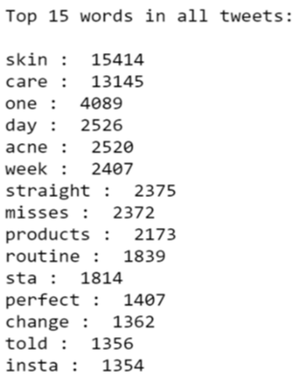
```
e beauty https://t.co/wrrhizq6zh via @amazon https://t.co/llushtlzdm sales executive - skin care
 and cosmetics - c p l cosmetics trading l.l.c - https://t.co/0buypsee4b 10% off best selling bab
y skin care brands. . #this_deal_ends_oct/18. . https://t.co/vo966ddfea https://t.co/sclqbxefkl r
t @hugssselfcare: skin care tips: - drink lots of water 💦 - try not to touch your face🌸x - try
 to cut out greasy foods🍿x - don’t share… rt @anthonycharm: okay but i have some wonderful @glos
sier skin care products due to arrive that day and i know y’all aren’t about to make… rt @jmthomp
son49: https://t.co/kdwxtl8vbl rt @projectmeow: update on jonathan, the kitty with the severe sca
lping wound. he’s healing nicely and the skin on the top of his head is f… joke\'s on jungkook cu
z the pic aint even bad. look at this man\'s perfect skin. like drop that skin care routine rt @h
ugssselfcare: skin care tips: - drink lots of water 💦 - try not to touch your face🌸x - try to
 cut out greasy foods🍿x - don’t share… forget about all your products for skin care. cacay oil i
```

Figure 7

Following is the code to clean data. First, I remove all the link (start with http), mentioned users (start with@) and retweet symbol ("rt"). Then, I removed all punctuation, lowered case and moved stopwords. Then I used a dictionary to record the number of appearance of each word. Finally, print out first top words in tweets text.



```
## all the tweets
import nltk
import string
import re
from nltk.corpus import stopwords
#stopwords = set(STOPWORDS)
result = re.sub(r"@\S+|http\S+|rt", "", all_text_words) #remove @ link
table = str.maketrans('', '', string.punctuation)
stripped = [w.translate(table) for w in result.split()]
words=[word for word in stripped if word.isalpha()] #remove punctuation
clean_words = [w for w in words if not w in stopwords.words('english')] #remove stopwords
```



```
Top 15 words in all tweets:

skin :  15414
care :  13145
one :  4089
day :  2526
acne :  2520
week :  2407
straight :  2375
misses :  2372
products :  2173
routine :  1839
sta :  1814
perfect :  1407
change :  1362
told :  1356
insta :  1354
```

Figure 8

The most popular words are skin and care. “one day” “week” is a part of skin care routine information. We can see all the words are related skin care topic, which means the data collection word is successful though costed plenty of time.

- **Word cloud**
  This is word cloud of skin care text data. We can see that most words have already showed in top significant word list. Other words like “buying, babe, insta” are used often in advertisement, therefore, they also showed in word cloud.

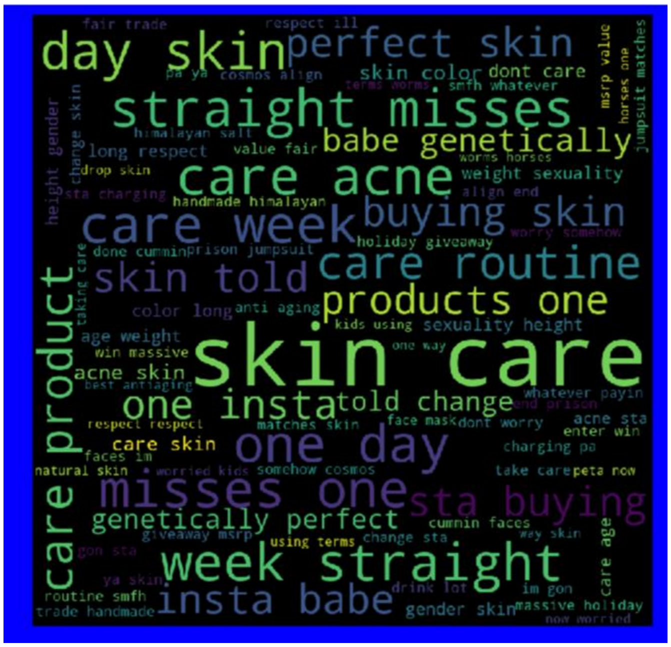
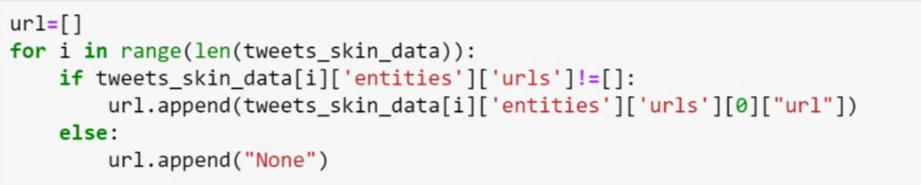


Figure 9

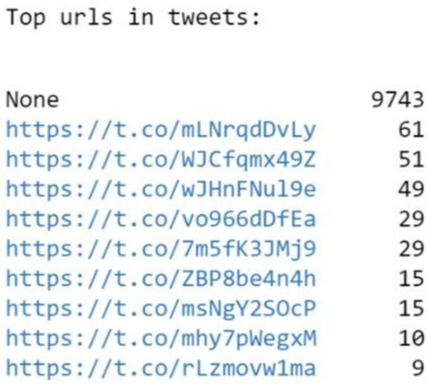
- **Top urls**
  I choose “url” in “entities” as my objective. If it exists, I used a list to record the text. Otherwise, it’s “none” for that tweet. Below is the code.



```
url=[]
for i in range(len(tweets_skin_data)):
    if tweets_skin_data[i]['entities']['urls']!=[]:
        url.append(tweets_skin_data[i]['entities']['urls'][0]["url"])
    else:
        url.append("None")
```

Figure 10

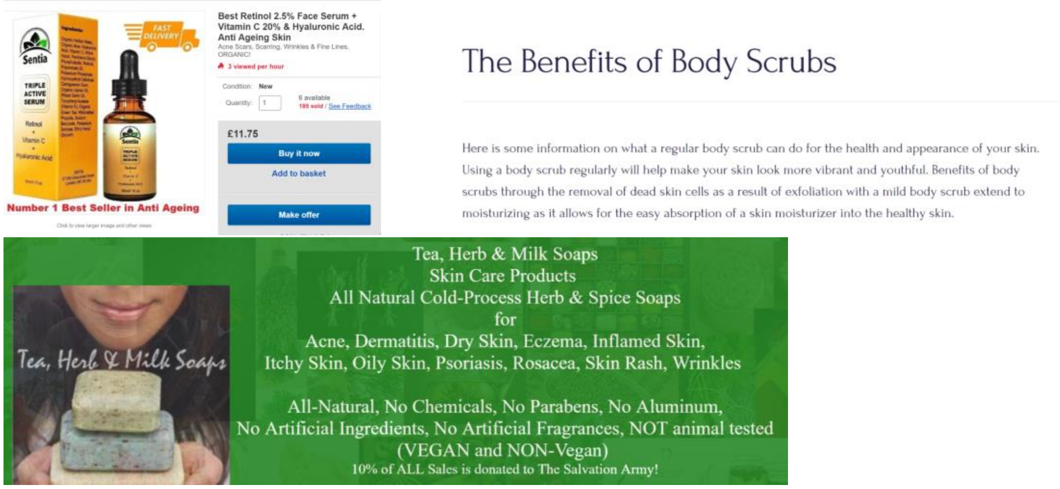
Here’s the top 10 urls in tweets.



```
Top urls in tweets:


None                       9743
https://t.co/mLNrqdDvLy      61
https://t.co/WJCfqmx49Z      51
https://t.co/wJHnFNul9e      49
https://t.co/vo966dDfEa      29
https://t.co/7m5fK3JMj9      29
https://t.co/ZBP8be4n4h      15
https://t.co/msNgY2SOcP      15
https://t.co/mhy7pWegxM      10
https://t.co/rLzmovw1ma       9
```

Figure 11

The 1st, 3rd and 4th are product promotions or introduction links.


Sentia
TRIPLE ACTIVE SERUM
Retinol
+
Vitamin C
+
Hyaluronic Acid
FAST DELIVERY
Number 1 Best Seller in Anti Ageing
Click to view larger image and other views

Best Retinol 2.5% Face Serum + Vitamin C 20% & Hyaluronic Acid. Anti Ageing Skin

Acne Scars, Scarring, Wrinkles & Fine Lines. ORGANIC!

3 viewed per hour

Condition: New

Quantity: 1

6 available

195 sold / See Feedback

£11.75

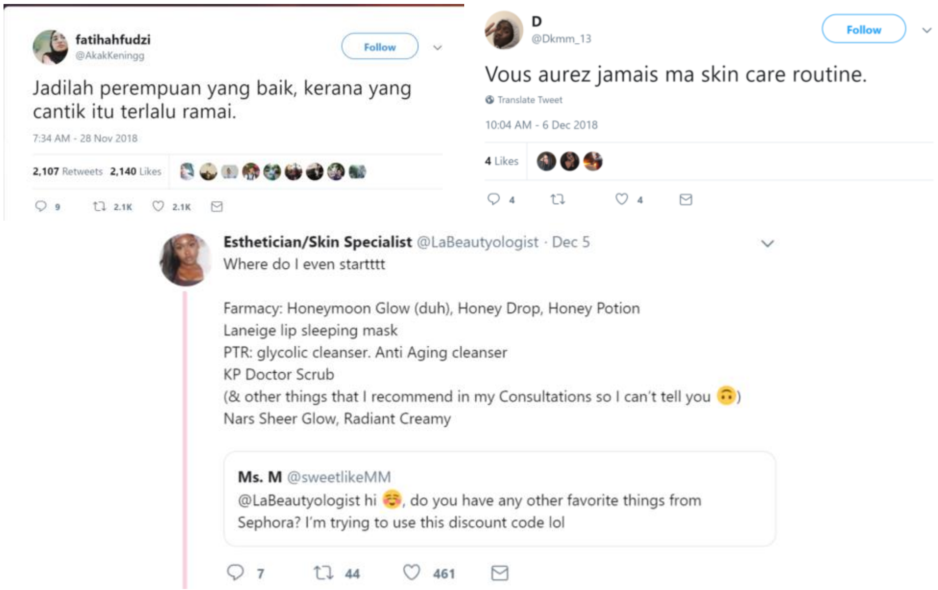
The Benefits of Body Scrubs

Here is some information on what a regular body scrub can do for the health and appearance of your skin. Using a body scrub regularly will help make your skin look more vibrant and youthful. Benefits of body scrubs through the removal of dead skin cells as a result of exfoliation with a mild body scrub extend to moisturizing as it allows for the easy absorption of a skin moisturizer into the healthy skin.

Tea, Herb & Milk Soaps
Tea, Herb & Milk Soaps
Skin Care Products
All Natural Cold-Process Herb & Spice Soaps
for
Acne, Dermatitis, Dry Skin, Eczema, Inflamed Skin,
Itchy Skin, Oily Skin, Psoriasis, Rosacea, Skin Rash, Wrinkles
All-Natural, No Chemicals, No Parabens, No Aluminum,
No Artificial Ingredients, No Artificial Fragrances, NOT animal tested
(VEGAN and NON-Vegan)
10% of ALL Sales is donated to The Salvation Army!

Figure 12

Rest are other people’s tweets. Some are talking about encouraging girls to be beautiful, some are introducing different ways and tips for skin improvement.



fatihahfudzi @AkakKeningg

Jadilah perempuan yang baik, kerana yang cantik itu terlalu ramai.

7:34 AM - 28 Nov 2018

2,107 Retweets 2,140 Likes

9 2.1K 2.1K

D @Dkmm_13

Vous aurez jamais ma skin care routine.

10:04 AM - 6 Dec 2018

4 Likes

4 4

Esthetician/Skin Specialist @LaBeautyologist · Dec 5
Where do I even starttttt

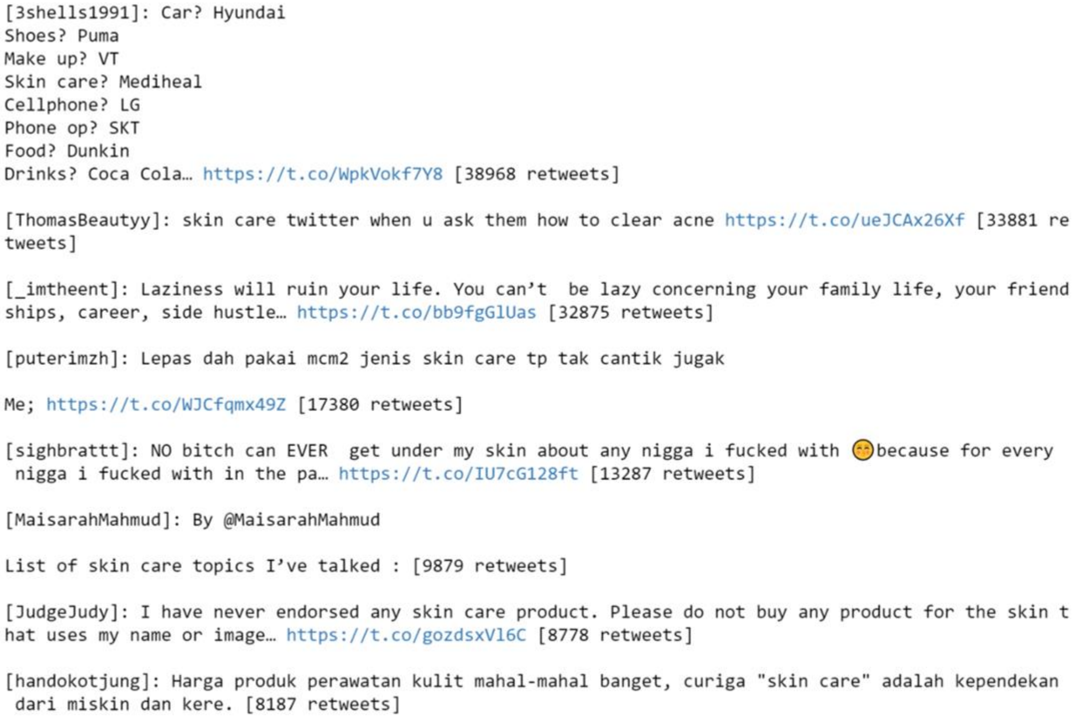
Farmacy: Honeymoon Glow (duh), Honey Drop, Honey Potion
Laneige lip sleeping mask
PTR: glycolic cleanser. Anti Aging cleanser
KP Doctor Scrub
(& other things that I recommend in my Consultations so I can’t tell you 🙃)
Nars Sheer Glow, Radiant Creamy

Ms. M @sweetlikeMM
@LaBeautyologist hi ☺️, do you have any other favorite things from Sephora? I’m trying to use this discount code lol

7 44 461

Figure 13

- **Top retweets**
  For retweets, I looked for retweeted_status and found id of this retweet and then recorded and sorted the retweet number. Following are top retweets.

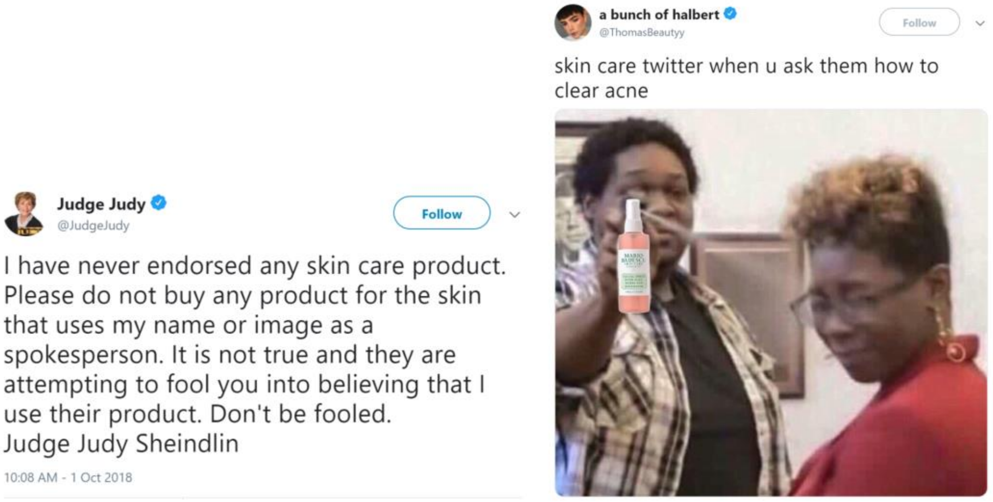


```
[3shells1991]: Car? Hyundai
Shoes? Puma
Make up? VT
Skin care? Mediheal
Cellphone? LG
Phone op? SKT
Food? Dunkin
Drinks? Coca Cola… https://t.co/WpkVokf7Y8 [38968 retweets]

[ThomasBeautyy]: skin care twitter when u ask them how to clear acne https://t.co/ueJCAx26Xf [33881 re
tweets]

[_imtheent]: Laziness will ruin your life. You can’t  be lazy concerning your family life, your friend
ships, career, side hustle… https://t.co/bb9fgGlUas [32875 retweets]

[puterimzh]: Lepas dah pakai mcm2 jenis skin care tp tak cantik jugak

Me; https://t.co/WJCfqmx49Z [17380 retweets]

[sighbrattt]: NO bitch can EVER  get under my skin about any nigga i fucked with 😁because for every
 nigga i fucked with in the pa… https://t.co/IU7cG128ft [13287 retweets]

[MaisarahMahmud]: By @MaisarahMahmud

List of skin care topics I’ve talked : [9879 retweets]

[JudgeJudy]: I have never endorsed any skin care product. Please do not buy any product for the skin t
hat uses my name or image… https://t.co/gozdsxVl6C [8778 retweets]

[handokotjung]: Harga produk perawatan kulit mahal-mahal banget, curiga "skin care" adalah kependekan
 dari miskin dan kere. [8187 retweets]
```

Figure 14

The first one is a fan account of BTS (BTS is a popular South Korean boy band named Bangtan Boys). They are popular in twitter during these days because one of them was celebrating birthday. And in this tweet, he mentioned skin care, so it ranked top one in retweets.

Some tweets are from famous people. Here's two examples below. The left one is from Judge Judy. She warned everyone don't be fooled that she never spoke for skin care product. The right retweet is from Thomas Halbert. He is an Instagram star as a professional Kris jenner cosplayer. He also mentioned skin care.



Judge Judy
@JudgeJudy

I have never endorsed any skin care product. Please do not buy any product for the skin that uses my name or image as a spokesperson. It is not true and they are attempting to fool you into believing that I use their product. Don't be fooled.
Judge Judy Sheindlin

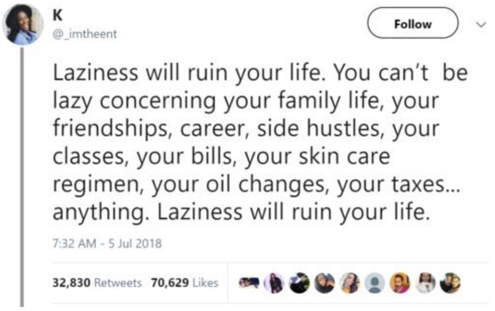
10:08 AM - 1 Oct 2018

a bunch of halbert
@ThomasBeautyy

skin care twitter when u ask them how to clear acne

Figure 15

Other tweets are also related to skin care. Take this one for example (figure 16) because I can't agree with her more. "Laziness will ruin your life". This satisfied not only in skin care but also in other fields.



K
@_imtheent

Laziness will ruin your life. You can't be lazy concerning your family life, your friendships, career, side hustles, your classes, your bills, your skin care regimen, your oil changes, your taxes... anything. Laziness will ruin your life.

7:32 AM - 5 Jul 2018

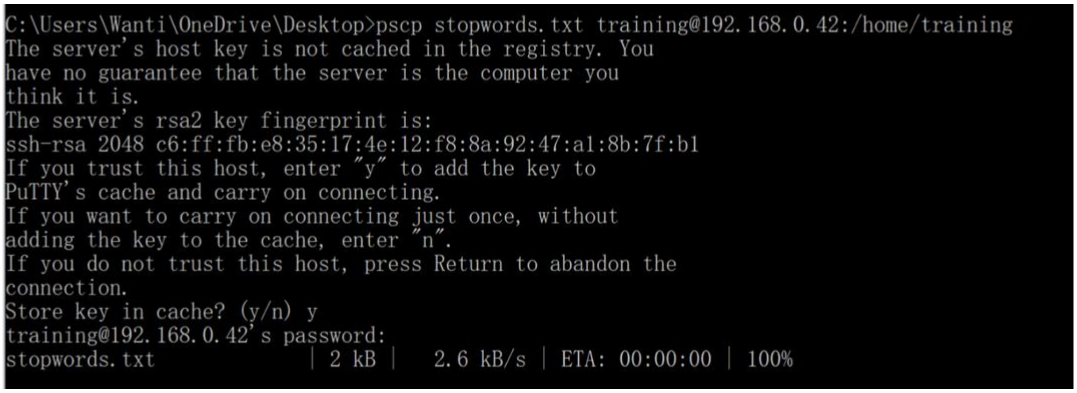
32,830 Retweets 70,629 Likes

Figure 16

Topic 2 in Hadoop and Data Frame

- For trade war topic, I have used data frame and Hadoop to implement. It took me a whole day to figure out how to transfer file to my virtual machine. The procedure given by this link: https://docs.google.com/document/d/1MZ_rNxJhR4HCU1qJ2-w7xlk2MTHVqa9lnl_uj-zRkzk/pub

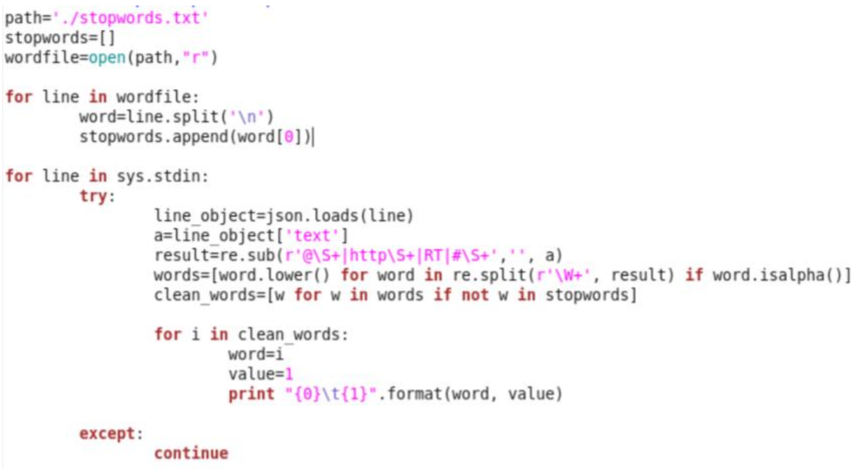
  It can help but it's a little different at the end for my computer. I had to use Putty to connect my computer with virtual machine IP address. And then type following code in my computer Command Prompt:



```
C:\Users\Wanti\OneDrive\Desktop>pscp stopwords.txt training@192.168.0.42:/home/training
The server's host key is not cached in the registry. You
have no guarantee that the server is the computer you
think it is.
The server's rsa2 key fingerprint is:
ssh-rsa 2048 c6:ff:fb:e8:35:17:4e:12:f8:8a:92:47:a1:8b:7f:b1
If you trust this host, enter "y" to add the key to
PuTTY's cache and carry on connecting.
If you want to carry on connecting just once, without
adding the key to the cache, enter "n".
If you do not trust this host, press Return to abandon the
connection.
Store key in cache? (y/n) y
training@192.168.0.42's password:
stopwords.txt             | 2 kB |   2.6 kB/s | ETA: 00:00:00 | 100%
```

Figure 17

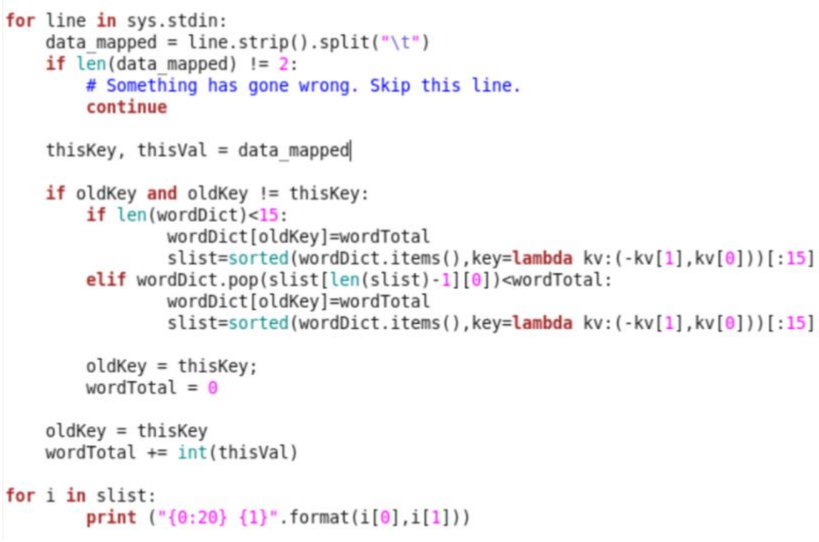
- **Top significant words**
  As I mentioned before, I can't use data clean packages in my computer. Therefore, I downloaded a stopwords txt file from the internet and wrote upload code in my map function (showed below). Then like in topic 1, I removed links, punctuation and stopwords. Finally, printed out each word with format (word, 1).



```
path='./stopwords.txt'
stopwords=[]
wordfile=open(path,"r")

for line in wordfile:
        word=line.split('\n')
        stopwords.append(word[0])

for line in sys.stdin:
        try:
                line_object=json.loads(line)
                a=line_object['text']
                result=re.sub(r'@\S+|http\S+|RT|#\S+','', a)
                words=[word.lower() for word in re.split(r'\W+', result) if word.isalpha()]
                clean_words=[w for w in words if not w in stopwords]

                for i in clean_words:
                        word=i
                        value=1
                        print "{0}\t{1}".format(word, value)

        except:
                continue
```

Figure 18

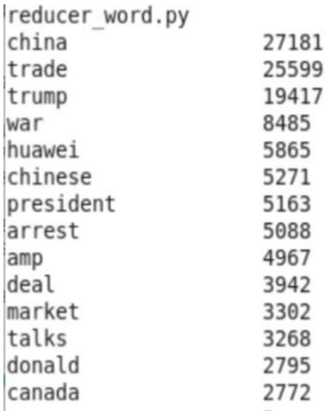
In the reduce function (showed below), I used dictionary to record top 15 words: key is each word, value is the number of appearance. I also used sorted function to make dictionary become a list and sort the value. Then printed out key and value in the sorted list.



```
for line in sys.stdin:
    data_mapped = line.strip().split("\t")
    if len(data_mapped) != 2:
        # Something has gone wrong. Skip this line.
        continue

    thisKey, thisVal = data_mapped

    if oldKey and oldKey != thisKey:
        if len(wordDict)<15:
                wordDict[oldKey]=wordTotal
                slist=sorted(wordDict.items(),key=lambda kv:(-kv[1],kv[0]))[:15]
        elif wordDict.pop(slist[len(slist)-1][0])<wordTotal:
                wordDict[oldKey]=wordTotal
                slist=sorted(wordDict.items(),key=lambda kv:(-kv[1],kv[0]))[:15]

        oldKey = thisKey;
        wordTotal = 0

    oldKey = thisKey
    wordTotal += int(thisVal)

for i in slist:
        print ("{0:20} {1}".format(i[0],i[1]))
```

Figure 19

Result is listed below. We can see “china”, “trade”, “trump” and “war” are raked the top since they are my key words. We can also see “Huawei”, “arrest” and “Canada”. These words remind me of a news: Meng Wanzhou, the CFO of Chinese tech giant Huawei, was arrested December 1 in Canada and faces extradition to the United States. This event shows in my trade war topic, which means people think this event is one part of trade war.



```
reducer_word.py
china                27181
trade                25599
trump                19417
war                  8485
huawei               5865
chinese              5271
president            5163
arrest               5088
amp                  4967
deal                 3942
market               3302
talks                3268
donald               2795
canada               2772
```

Figure 20

- **Word cloud**
  We can see most top significant words show in word cloud. Other words, such as deficit, bilateral, echoed, didn't count as top significant words but appear here.

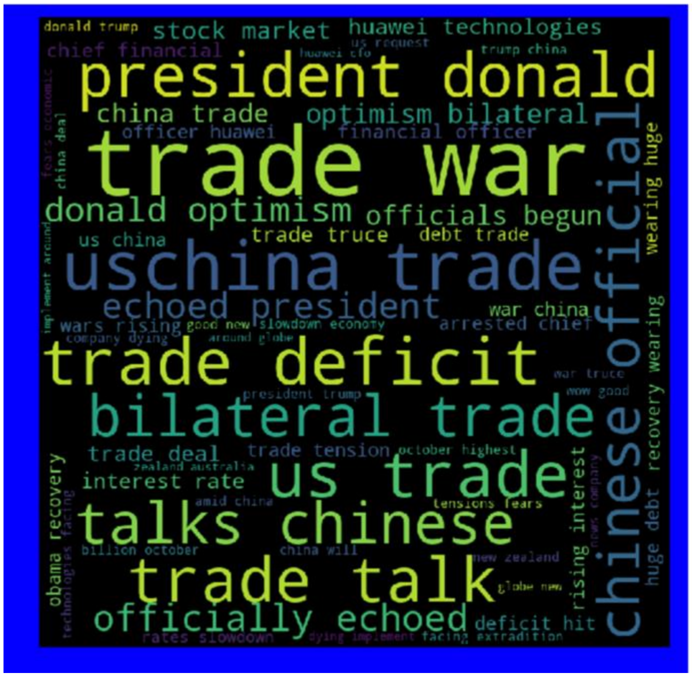
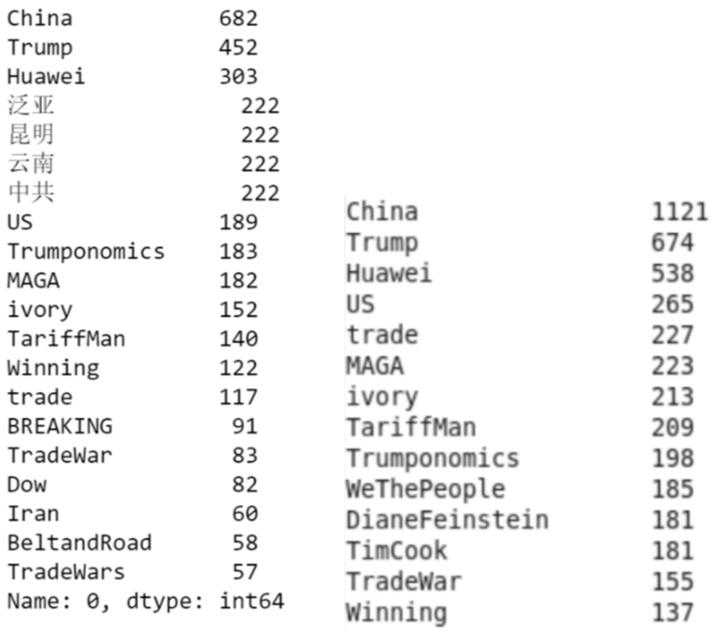


Figure 21

- **Top hashtags**
  I used data frame and Hadoop to count this. However, the result of the two is a little different. The left side is result from data frame and right side is from Hadoop. We can see except for Chinses words (Hadoop can't recognize multiple language), all the number in Hadoop are relatively large than number in data frame. So, I inspect that data frame missed some data and I'll show this again later.

  Top hashtags are almost the same. "China", "trump", "Huawei", "US", "trade" and "tariffman" are the top words. "MAGA" means "make America great again". Therefore, it's reasonable for those words to become top hashtags.



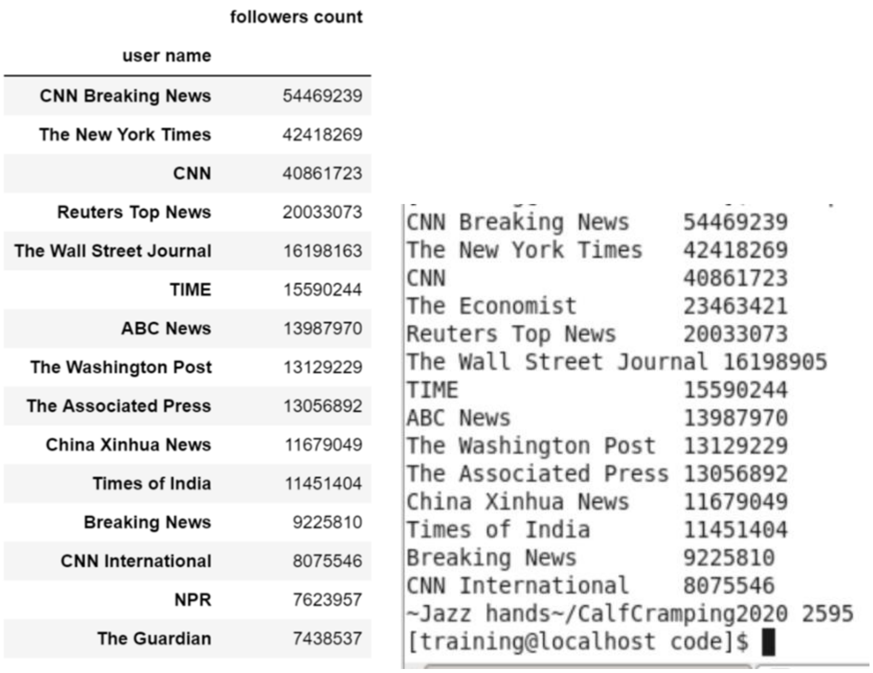
```
China           682
Trump           452
Huawei          303
泛亚              222
昆明              222
云南              222
中共              222
US              189
Trumponomics    183
MAGA            182
ivory           152
TariffMan       140
Winning         122
trade           117
BREAKING         91
TradeWar         83
Dow              82
Iran             60
BeltandRoad      58
TradeWars        57
Name: 0, dtype: int64
```

```
China               1121
Trump               674
Huawei              538
US                  265
trade               227
MAGA                223
ivory               213
TariffMan           209
Trumponomics        198
WeThePeople         185
DianeFeinstein      181
TimCook             181
TradeWar            155
Winning             137
```

Figure 22

- **Top authoritative users**
  The results from data frame and Hadoop are different again. The left side is the result from data frame and right side is from Hadoop. All the user names and followers number are totally the same except that data frame completely missed user “the Economist”. I even tried to search tweets from user whose name is The Economist but it showed nothing. That means data frame didn’t recognize The Economist at all. On the contrary, Hadoop can give more accurate result.

  From the result, all the users are media groups from US, China, UK and India. That means this event is important and people from all the world care about it.



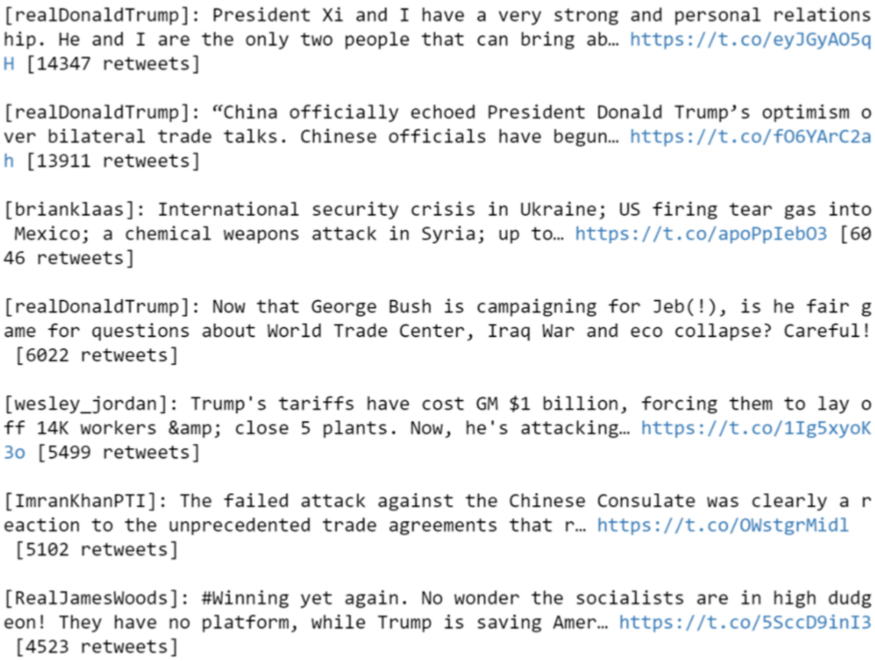
	followers count
user name	
CNN Breaking News	54469239
The New York Times	42418269
CNN	40861723
Reuters Top News	20033073
The Wall Street Journal	16198163
TIME	15590244
ABC News	13987970
The Washington Post	13129229
The Associated Press	13056892
China Xinhua News	11679049
Times of India	11451404
Breaking News	9225810
CNN International	8075546
NPR	7623957
The Guardian	7438537

```
CNN Breaking News    54469239
The New York Times   42418269
CNN                  40861723
The Economist        23463421
Reuters Top News     20033073
The Wall Street Journal 16198905
TIME                 15590244
ABC News             13987970
The Washington Post  13129229
The Associated Press 13056892
China Xinhua News    11679049
Times of India       11451404
Breaking News        9225810
CNN International    8075546
~Jazz hands~/CalfCramping2020 2595
[training@localhost code]$ █
```

Figure 23

- **Top retweets**

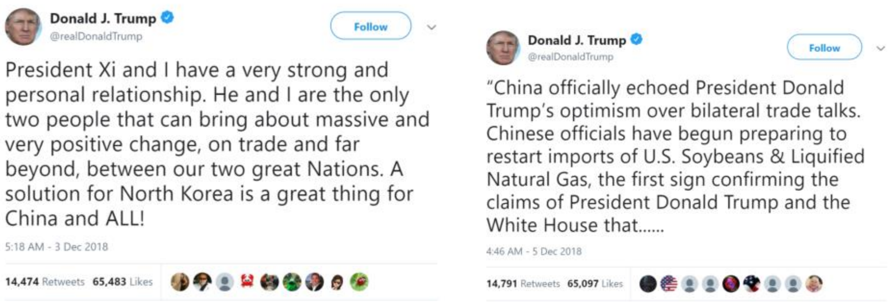
  These are top retweets. The code is almost the same as it in topic one.



```
[realDonaldTrump]: President Xi and I have a very strong and personal relations
hip. He and I are the only two people that can bring ab… https://t.co/eyJGyAO5q
H [14347 retweets]

[realDonaldTrump]: “China officially echoed President Donald Trump’s optimism o
ver bilateral trade talks. Chinese officials have begun… https://t.co/fO6YArC2a
h [13911 retweets]

[brianklaas]: International security crisis in Ukraine; US firing tear gas into
 Mexico; a chemical weapons attack in Syria; up to… https://t.co/apoPpIebO3 [60
46 retweets]

[realDonaldTrump]: Now that George Bush is campaigning for Jeb(!), is he fair g
ame for questions about World Trade Center, Iraq War and eco collapse? Careful!
 [6022 retweets]

[wesley_jordan]: Trump's tariffs have cost GM $1 billion, forcing them to lay o
ff 14K workers &amp; close 5 plants. Now, he's attacking… https://t.co/1Ig5xyoK
3o [5499 retweets]

[ImranKhanPTI]: The failed attack against the Chinese Consulate was clearly a r
eaction to the unprecedented trade agreements that r… https://t.co/OWstgrMidl
 [5102 retweets]

[RealJamesWoods]: #Winning yet again. No wonder the socialists are in high dudg
eon! They have no platform, while Trump is saving Amer… https://t.co/5SccD9inI3
 [4523 retweets]
```

Figure 24

Several tweets are from president Trump. It shows that he is working on this but doesn't show much about his opinion to this trade war. Here are two examples:



Donald J. Trump
@realDonaldTrump
Follow

President Xi and I have a very strong and personal relationship. He and I are the only two people that can bring about massive and very positive change, on trade and far beyond, between our two great Nations. A solution for North Korea is a great thing for China and ALL!

5:18 AM - 3 Dec 2018

14,474 Retweets 65,483 Likes

Donald J. Trump
@realDonaldTrump
Follow

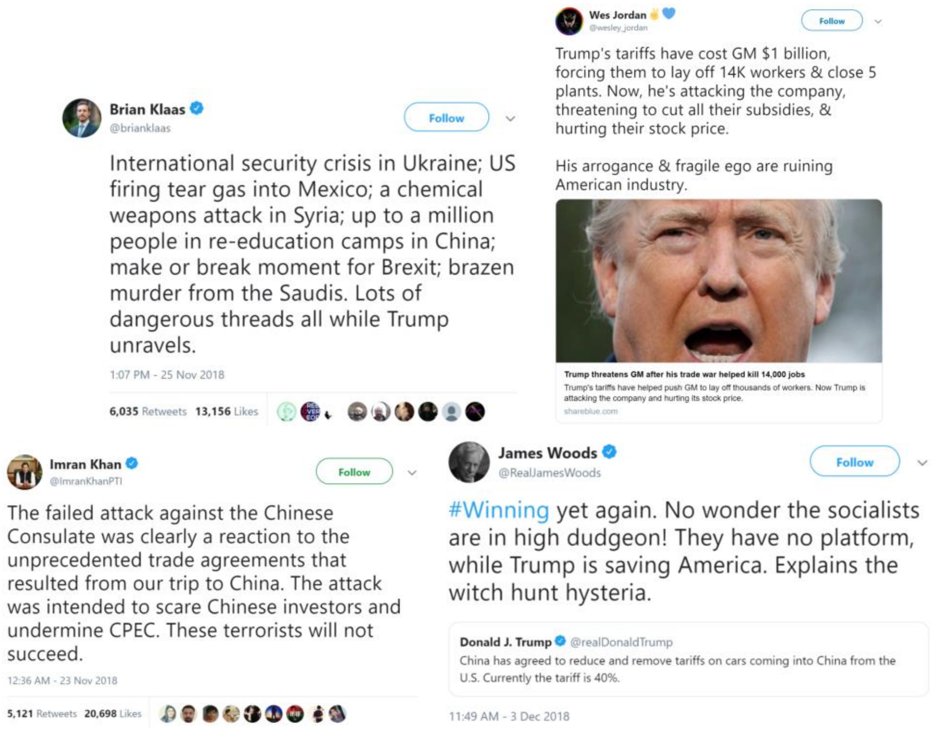
"China officially echoed President Donald Trump's optimism over bilateral trade talks. Chinese officials have begun preparing to restart imports of U.S. Soybeans & Liquified Natural Gas, the first sign confirming the claims of President Donald Trump and the White House that......

4:46 AM - 5 Dec 2018

14,791 Retweets 65,097 Likes

Figure 25

However, we can get more opinion from other top retweets. Here are several examples: One is from Prime Minister of Pakistan, Imran Khan, he showed an against point of view to this event. One is from Brian Klaas. He is a political scientist and he also showed an unsatisfaction to president Trump's reaction. Another one is Wes Jordan, who is just a normal people, also presents a dislike atitude to Trump. One is from James Woods, who is an american actor and producer, and he seems the only one who agreed with Trump. Therefore, we can see from retweets, more people think it's incorrect to start a trade war with China.



Wes Jordan
@wesley_jordan
Follow

Trump's tariffs have cost GM $1 billion, forcing them to lay off 14K workers & close 5 plants. Now, he's attacking the company, threatening to cut all their subsidies, & hurting their stock price.

His arrogance & fragile ego are ruining American industry.

Trump threatens GM after his trade war helped kill 14,000 jobs
Trump's tariffs have helped push GM to lay off thousands of workers. Now Trump is attacking the company and hurting its stock price.
shareblue.com

Brian Klaas
@brianklaas
Follow

International security crisis in Ukraine; US firing tear gas into Mexico; a chemical weapons attack in Syria; up to a million people in re-education camps in China; make or break moment for Brexit; brazen murder from the Saudis. Lots of dangerous threads all while Trump unravels.

1:07 PM - 25 Nov 2018

6,035 Retweets 13,156 Likes

Imran Khan
@ImranKhanPTI
Follow

The failed attack against the Chinese Consulate was clearly a reaction to the unprecedented trade agreements that resulted from our trip to China. The attack was intended to scare Chinese investors and undermine CPEC. These terrorists will not succeed.

12:36 AM - 23 Nov 2018

5,121 Retweets 20,698 Likes

James Woods
@RealJamesWoods
Follow

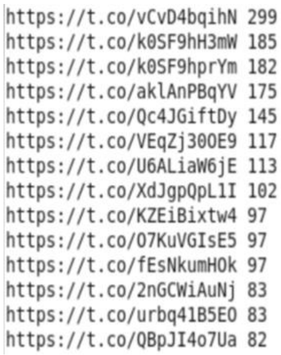
#Winning yet again. No wonder the socialists are in high dudgeon! They have no platform, while Trump is saving America. Explains the witch hunt hysteria.

Donald J. Trump @realDonaldTrump
China has agreed to reduce and remove tariffs on cars coming into China from the U.S. Currently the tariff is 40%.

11:49 AM - 3 Dec 2018

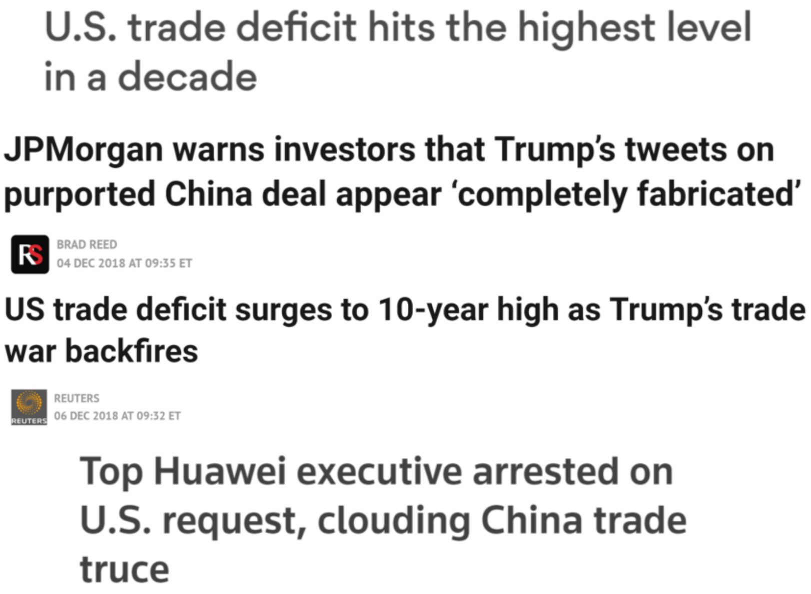
Figure 26

- **Top 10 urls**
  These are top 10 urls. Most of them are new report link. I showed several titles of these reports. These links mentioned deficit in the US because of this trade war, the financial market for investors and the correlation between Huawei event and trade war.



```
https://t.co/vCvD4bqihN 299
https://t.co/k0SF9hH3mW 185
https://t.co/k0SF9hprYm 182
https://t.co/aklAnPBqYV 175
https://t.co/Qc4JGiftDy 145
https://t.co/VEqZj30OE9 117
https://t.co/U6ALiaW6jE 113
https://t.co/XdJgpQpL1I 102
https://t.co/KZEiBixtw4 97
https://t.co/O7KuVGIsE5 97
https://t.co/fEsNkumHOk 97
https://t.co/2nGCWiAuNj 83
https://t.co/urbq41B5EO 83
https://t.co/QBpJI4o7Ua 82
```

Figure 27



U.S. trade deficit hits the highest level in a decade

JPMorgan warns investors that Trump's tweets on purported China deal appear 'completely fabricated'

BRAD REED
04 DEC 2018 AT 09:35 ET

US trade deficit surges to 10-year high as Trump's trade war backfires

REUTERS
06 DEC 2018 AT 09:32 ET

Top Huawei executive arrested on U.S. request, clouding China trade truce

Figure 28

4. Project Summary

- **Topic summary**

  This project requires to collect data by using key words. In order to get high quality tweets, key words are important. They can't be either too broad and general or too narrow and specific. Otherwise, you may get many irrelevant tweets or spend plenty of time to collect it.

  For the skin care topic, data frame is good way to deal with JSON file because it's easy to sort and plot. And NLTK and stopwords packages are very useful when cleaning data. From the content of tweets, people are interested in peta-approved skin care product and don't be lazy in skin care routine.

  For the trade war topic, Hadoop is quite useful for big data. Map function is used to clean data and print each word with format and reduce function is responsible to record and sort value. It turns out Hadoop can give more accurate result than data frame. As for the content, more people are unsatisfied with Trump's reaction to trade war. Deficit in US and Huawei event are two focus of the trade war.

- **Data frame vs Hadoop**

  The advantages of data frame are obvious. It's interactive and easy to implement because of plenty of packages to use. Data frame can also provide beautiful plot and tables and it can read multiple languages. Hadoop doesn't have these. However, it can deal with big data and the result is more accurate than data frame.