CIS 668 NLP Data Analysis Project Report - on COVID-19

Instructor: Professor Lu Xiao

Students: Leah Luo (lluojr), Bo Li (bli158)

Content

Abstract

Public opinion analysis is applied in various fields in society as an increasing trend. With such analysis, decisions shall be made with a wiser and more comprehensive consideration by the government, companies, and research scholars.
The goal of our project is to apply the algorithms and methodologies learned from this course to analyze the public opinions towards COVID-19 under different trending topics from Twitter and Sina Weibo, as well as conduct a comparative study between them. Our study will focus mainly on sentiment analysis. As the popular social media platforms in American and China, we believe that by analyzing the tweets and weibos, we shall conclude about how different are the trending topics in China from the other country and how the sentiment on COVID-19 different towards the topics daily.

Introduction

1. Public Opinion Analysis

While public opinion analysis has become significant in decision making among various fields, the potential of using social media as a data resource seems reasonable and necessary. Twitter, as the large social media platform in American and other countries, people post and interact with messages whenever they want. All the tweets are visible to public unless the sender set a restriction to them. Twitter users can forward the tweets from others, and "favorite" the tweets, the number of how many users "favorite" and forward are shown under each tweet. Tweets are labeled with various kinds of topics. Those features allow us to extract only the popular tweets under the popular trending topics, in other words, the tweets that most people are interested.
Given the user base of social media, the analysis of them could reasonably imply public preference and inclination.
Such findings could help policymakers, business and practitioners gain large insights and therefore make intelligent decisions.

2. Coronavirus Disease 2019

The severe special infectious pneumonia (Coronavirus disease 2019, abbreviation: COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2.
Figure 1 shows the total number of active cases and closed cases.

Figure 1

At the end of 2019, the first patient was diagnosed in Wuhan, China and the disease has spread rapidly in various countries over the world since then. Figure 2 below shows the countries that are impacted by COVID-19 as of March 2020. And as of May 6, 2020, it has spread to 226 countries. It is widely believed that the severe special infectious pneumonia epidemic is the most severe crisis since the Second World War.

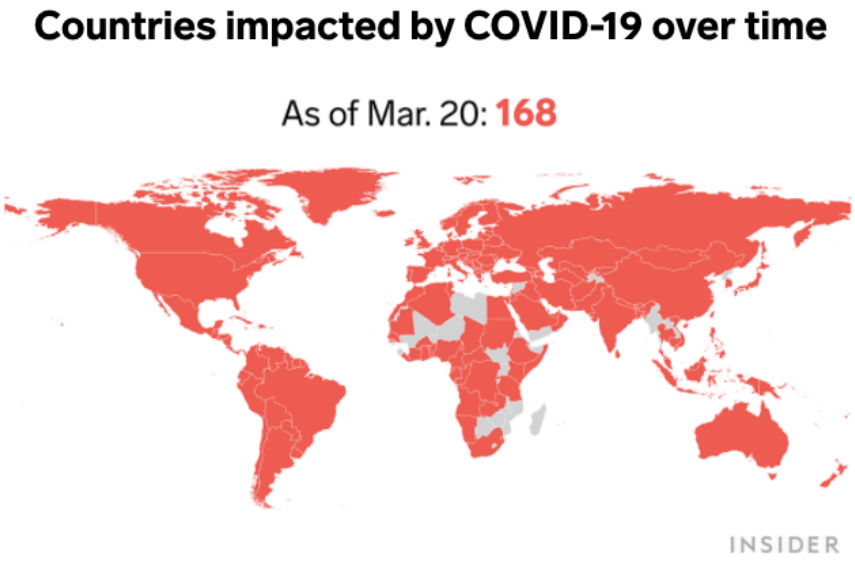

Figure 2

Despite the serious situation in China since January 2020, CDC did not recommend the public to wear a face mask (cloth face covering, surgical mask, N95, etc.) until April 2020. This study aims to analyze public's attentions and opinions towards COVID-19 on Twitter. To be specific, the public's opinion from March to April.
We crawl the most popular tweets (people "favorite" most) under the top 30 trending topics, extract the tweets related to COVID-19 from them. Sentiment analysis will be conducted on the tweets related to COVID-19 then.

The analysis result could imply how the virus affects people's daily life and how people react to it.
Besides, the government and CDC have the responsibility to raise the public awareness of the virus, so that the precautions could be more efficient. This study could also be used to analyze how CDC and the government performed their jobs.

Dataset Description

1. Collection

a. Twitter

The tweets dataset was collected and generated by Twitter with its own Search API. A developer account is required to access Twitter APIs. The developer could simply use the twitter account and declare the purpose while applying and should get the result within a few hours. The API key information, which is consumer_key and consumer_secret, will be given to the new developer. Copy the values of them in the python program and access the Search API.
Unnecessary information should be excluded before extracting since the original Twitter dataset that includes user information, text contents, date time, hashtags, and emoji might be too large for data analysis.
We used Tweepy, a Python library, to extract only the trending topics, date time, and text contents.
We want to collect the tweets that had great attention, like the most popular tweets under the top trending topics. If a tweet that talks about COVID-19 but no one likes it or forwards it, it means that it does not attract other's attention.
The top 30 trending topics and the most popular 30 tweets under each of them were extracted every day and stored in a csv file with data frame format. Popular tweets from March 11 to April 10, 2020 were collected and stored in twitter.csv file, which is about 8.5 MB. Note that Tweepy has its own algorithm to extract the tweets tag as popular.
Figure 3 shows below is the first few tweets of the first day we collected.

	Date	Topic	Comment
0	03/11/2020, 01:56:25	['TheBachelor']	Retweet if you want #thebachelor to film the W...
1	03/09/2020, 15:22:31	['TheBachelor']	Tonight...you WILL find out who Barb is crying...
2	03/10/2020, 00:18:25	['TheBachelor']	Wait are these Pete's parents? #TheBachelor ht...
3	03/11/2020, 00:56:57	['TheBachelor']	Hannah you can get engaged again.. Angelina fr...
4	03/10/2020, 00:56:34	['TheBachelor']	Madi is saying she's saving herself for marria...
...	...	...	...
44438	['04/10/2020, 22:37:03']	['JusticeForJan']	RT @skinnylegendari: #JusticeForJan https://...
44439	['04/10/2020, 22:37:03']	['JusticeForJan']	RT @AndrewBarretCox: Anyone else? @jansportnyc...
44440	['04/10/2020, 22:37:03']	['JusticeForJan']	@jansportnyc WAS ROBBED, PERIODT.\n\n#JusticeF...
44441	['04/10/2020, 22:37:03']	['JusticeForJan']	The deserved winner of tonight's episode #Just...
44442	['04/10/2020, 22:37:03']	['JusticeForJan']	I love and respect all the queens but my two f...

44443 rows × 3 columns

Figure 3

b. Sina Weibo

The Weibo Dataset is used to collect data such as news which users posted to Weibo every moment. To collect such data and generate it into readable form, the program mainly implemented BeautifulSoup and pandas as a technique to call every single section via attributes in HTML of the Weibo webpage.

```
link = 'https://s.weibo.com'+links[i]
url = urllib.request.urlopen(link).read()
soup = bs4.BeautifulSoup(url,'lxml')

all_contents = soup.findAll("div", {"class": "content"})

for content in all_contents:
    if content.find("p", attrs={'node-type': 'feed_list_content_full', 'class': 'txt'}):
        expended_content = content.find("p", attrs={'node-type': 'feed_list_content_full', 'class': 'txt'}).get
        data.append(expended_content)
    else:
        data.append(content.find("p", {"class": "txt"}).get_text())
```

Figure 4

In figure 4, it indicates that the soup (Invoked by BeautifulSoup) has functions called find/findAll. The program above could expose the functionality that it could jump to the specific webpages or areas of them where we expected to enter. Using BeautifulSoup, we could simply retrieve all data in each topic starting from the main webpage of Weibo.resou (Top Trends of Weibo).

We used the program to retrieve the top 30-40 trends every day from March 11 – April 10. Then we implemented DataFrame to store such data/trends into a csv files inserting columns of "Data", "Topic" and "Comment". The final size of weibo.csv is 12.6MB and it has 34,150 comments as data. (See figure 5).

Date	Topic	Comment
03/10/2020-21:01:35	['意大利紧急求助中国']	【#加拿大航空公司暂停往返意大利航班#】加拿大航空公司10日宣布，将从11日起停飞往返意大利的航班直至5月1日。加拿大新冠肺炎疫情最近一周出现暴延趋势。#意大利累计确诊升至10149例#。截至9日，加拿大累计新冠患
03/10/2020-21:01:35	['意大利紧急求助中国']	【意大利紧急求助，中方将提供口罩等医疗物资】3月10日，国务委员兼外长王毅应约同意大利外长迪马的通电话。王毅说，我们不会忘记，在中国抗击疫情最困难的时候，意大利给予了中方宝贵支持。现在，我们也愿和意大利
03/10/2020-21:01:35	['意大利紧急求助中国']	【#意大利紧急求助中国#，王毅：将向意方提供口罩等医疗物资】 据外交部网站消息，2020年3月10日，国务委员兼外长王毅应约同意大利外长迪马的通电话。迪马的表示，当前意大利疫情形势十分严峻，意政府正密切关注和
03/10/2020-21:01:35	['意大利紧急求助中国']	【#意大利紧急求助中国# 王毅：将向意方提供口罩等医疗物资】 据外交部网站消息，2020年3月10日，国务委员兼外长王毅应约同意大利外长迪马的通电话。迪马的表示，当前意大利疫情形势十分严峻，意政府正密切关注和学
03/10/2020-21:01:35	['意大利紧急求助中国']	#印度当街烧新冠病毒怪物塑像# 是迷信吗 尊重国家的文化 但还是要做好防护才行 #意大利紧急求助中国# 意大利真的太严重了 中国毕竟大国
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#阿呵呵，意大利这么多尊华言论和行为，还好意思求助，
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#emmmmm捐赠医疗物资就可以了吧，不允许出动我国的白衣天使，本来这段时间都很累很累了，超负荷工作，如果再去支援他国倒也不必
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国# 为中国的这个举动点赞，病毒不分国界，各国应团结一致，共同对抗疫。 #印度当街烧新冠病毒怪物塑像# 精神上的胜利是短暂的，最终要还是需要在现实中去战胜。
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#这些评论啥意思啊？？？
03/10/2020-21:01:35	['意大利紧急求助中国']	产能蹭上了，最近口罩容易抢了！ 还得说一句社会主义中国牛逼！ #意大利紧急求助中国#
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#有多余能力就帮 换回思考下
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#提供医疗物资和食物就行了！让医护人员们休息休息吧
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#不见棺材不落泪 真是又当又立
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#咱们最近损失非常大，意大利也是第一个对中国闭关锁国的国家，现在既然求到中国，咱们是口罩生产国，建议国家对外统一口罩出口价格，医用N95卖50欧一个，一次性医用外科卖10欧一个，一次性医
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#？？？？干嘛干嘛不是不在意吗
03/10/2020-21:01:35	['意大利紧急求助中国']	【#意大利紧急求助中国#，中方将提供口罩等医疗物资】#周刊君与你共同战疫# 3月10日，国务委员兼外长王毅应约同意大利外长迪马的通电话。王毅说，我们不会忘记，在中国抗击疫情最困难的时候，意大利给予了中方宝贵
03/10/2020-21:01:35	['意大利紧急求助中国']	#意大利紧急求助中国#网友: 如果美国求助，请启用扁鹊三连：治不了，告辞了，等死吧

Figure 5

2. Extraction

a. Twitter

Though the official name is COVID-19, people might call it differently. We noticed that some people called it "COIVD" or "coronavirus" on Twitter. To extract all the tweets related to it despite of how people called it with different names, we created a list contains related keywords.

```
keywords = ['coronavirus', 'COVID-19', 'Pandemic', 'epidemic', 'outbreak',
            'COVID19', 'COVID 19', 'corona virus', 'COVID']
```

Figure 6

Extract the tweets contain the keyword from the dataset, and store it in a new file. There are 2589 tweets related to the virus out of 47072 tweets in total. Figure 8 shows the ratio of tweets about the virus to tweets about others, which is extremely low.

	Date	Topic	Comment
0	03/11/2020, 03:22:34	['LoseWithBiden']	RT @7458Boyz: Like And Subcribe \nWatch Now!!\...
1	03/10/2020, 21:45:20	['Sears Tower']	I saw "Sears Tower" trending and was afraid it...
2	03/11/2020, 01:01:08	['Catholics for Trump']	As Bernie Sanders and Joe Biden scale back the...
3	03/11/2020, 03:22:44	['Rask']	Defiance amid coronavirus and other #Flyers o...
4	03/11/2020, 03:22:44	['Rask']	RT @PatMcLoone: Defiance amid coronavirus and ...
...	...	...	...
2584	['04/10/2020, 22:37:03']	['NationalSiblingDay']	Missing everyone on this #nationalsiblingday i...
2585	['04/10/2020, 22:37:03']	['FridayNight']	#SiblingsDay dance party Pt.1 because it's my ...
2586	['04/10/2020, 22:37:03']	['FridayNight']	Can anyone relate???? #COVID−19 #FursuitFrida...
2587	['04/10/2020, 22:37:03']	['FridayNight']	GETAnalysis: In the midst of a #GlobalCrisis, ...
2588	['04/10/2020, 22:37:03']	['FridayNight']	If @realDonaldTrump isn't charged with crimes ...

2589 rows × 3 columns

Figure 7

```
# Propotional Sector Diagram
import matplotlib.pyplot as plt

# Data to plot
labels = 'COVID-19', 'Others'
sizes = [len(virus_df.index), len(df.index)]
colors = ['yellow', 'springgreen']
explode = (0.1, 0)  # explode 1st slice

# Plot
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
autopct='%1.1f%%', shadow=True, startangle=140)

plt.axis('equal')
plt.show()
```

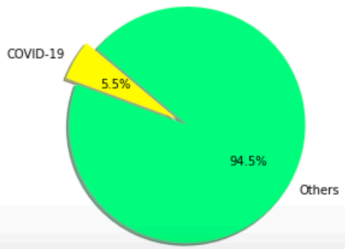

Figure 8

The number of tweets related to coronavirus and the growth rate is calculated. We used Python to draw the line charts for virtualization, which figure 9 shown below.
To our surprise, the number seems to be decreasing as the situation getting more severe.

```
plt.figure(figsize=(15, 5))
plt.title('Number of Tweets related to COVID-19')
plt.plot(dates_list, steps);
```

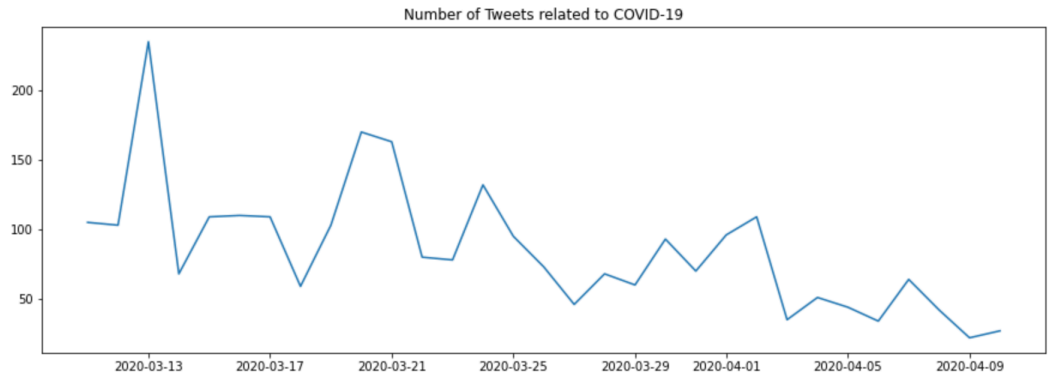

Figure 9

```
plt.figure(figsize=(15, 5))
plt.title('Growth Rate of Tweets related to COVID-19')
plt.plot(dates_list, trend_lst, color='green');
```

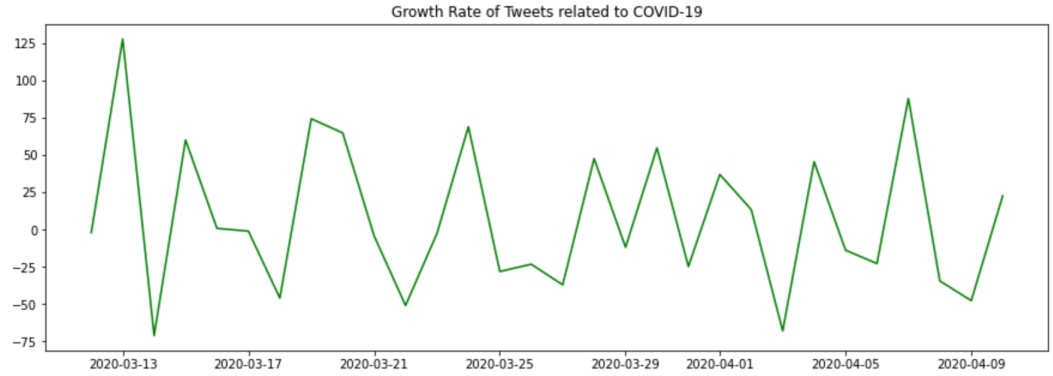

Figure 10

b. Sina Weibo

Similar to what we did for the Twitter dataset, we created a character list of all possible keywords.

```
keywords = ['肺炎', '新冠', '新型冠状', '病毒', '疫情', 'COVID', '感染']
```

Figure 11

Extract weibos contain the keyword from dataset and store in a new file. There are 9114 weibos related to the virus out of 34150 tweets in total. Figure 13 shows the ratio of weibo about the virus to weibo about others. The ratio is much higher than Twitter's.

	Date	Topic	Comment
0	03/10/2020-21:01:35	['意大利紧急求助中国']	\n 【#加拿大航空公司暂停往返意大利航班#】加拿大航...
1	03/10/2020-21:01:35	['意大利紧急求助中国']	\n 【意大利紧急求助，中方将提供口罩等医疗物资】3月...
2	03/10/2020-21:01:35	['意大利紧急求助中国']	\n 【#意大利紧急求助中国#，王毅：将向意方提供口罩...
3	03/10/2020-21:01:35	['意大利紧急求助中国']	\n 【#意大利紧急求助中国# 王毅：将向意方提供口罩...
4	03/10/2020-21:01:35	['意大利紧急求助中国']	\n#印度当街烧新冠病毒怪物塑像# 是迷信吗 尊重国家的文化 但还是要做好防护才行 #意大利...
...	...	...	...
9109	03/31/2020, 00:00:00	['美国监狱疫情肆虐多地在押人员抗议']	\n#美国监狱疫情肆虐多地在押人员抗议# 按照现在的数学模型来统计，在今年的夏天，我们把这个...
9110	03/31/2020, 00:00:00	['美国监狱疫情肆虐多地在押人员抗议']	\n 【疫情播报	4月13日早间重要疫情消息】09...
9111	03/31/2020, 00:00:00	['美国监狱疫情肆虐多地在押人员抗议']	\n#美国监狱疫情肆虐多地在押人员抗议#里外都不安全，啥措施也没有啊 ...
9112	03/31/2020, 00:00:00	['美国监狱疫情肆虐多地在押人员抗议']	\n#美国监狱疫情肆虐多地在押人员抗议# 当初没人预料到会严重，但是严重以后的做法才是最主要...
9113	03/31/2020, 00:00:00	['美国监狱疫情肆虐多地在押人员抗议']	\n#美国监狱疫情肆虐多地在押人员抗议#做得不错？因为成功送某些犯人去见上帝？ ...

9114 rows × 3 columns

Figure 12

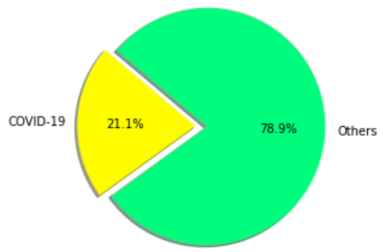

Figure 13

Figures 14 and 15 show the daily number of weibo related to COVID-19 and the growth rate.

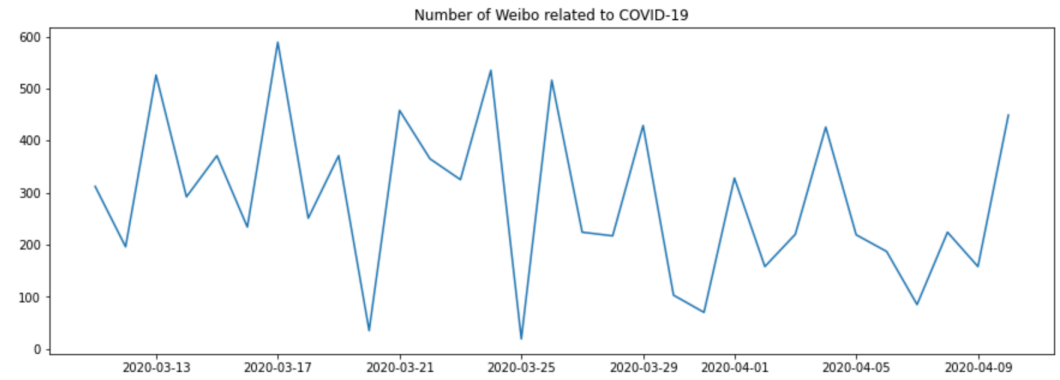

Figure 14

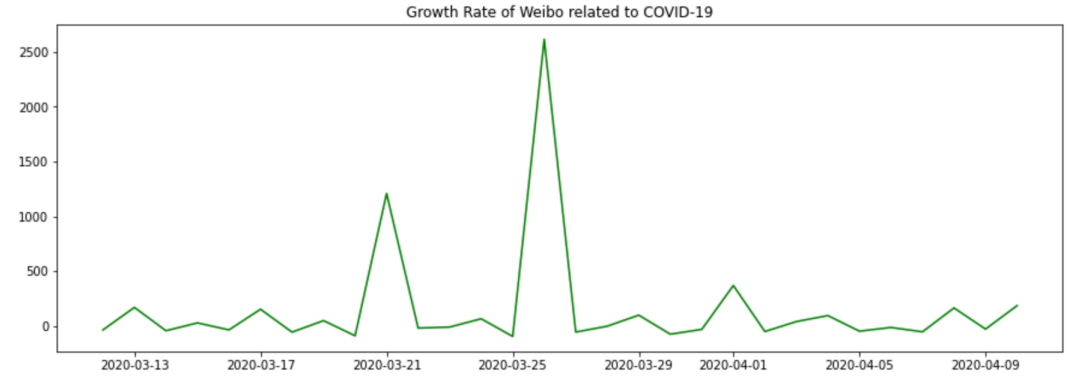

Figure 15

Implementation

1. Model Training for Twitter

We used the training set named "twitter samples" from NLTK package, which is a dataset of sample tweets. There are 5000 tweets with positive sentiments, 5000 tweets with negative sentiments, and 20000 tweets with no sentiments. We downloaded it and applied methods on it for cleaning. A model will be created and trained on the dataset. Evaluation of this new model should be analyzed before we apply it on the Twitter dataset.

negative_tweets.json: 5000 tweets with negative sentiments
positive_tweets.json: 5000 tweets with positive sentiments
tweets.20150430-223406.json: 20000 tweets with no sentiments

a. Data Cleaning

Tokenization

```
# Normalizing the Data
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')

from nltk.tag import pos_tag
from nltk.corpus import twitter_samples
tweet_tokens = twitter_samples.tokenized('positive_tweets.json')
print(pos_tag(tweet_tokens[0]))
```

```
[nltk_data] Downloading package wordnet to /Users/LXIN/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /Users/LXIN/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

```
[('#FollowFriday', 'JJ'), ('@France_Inte', 'NNP'), ('@PKuchly57', 'NNP'), ('@Milipol_Paris', 'NNP'), ('for', 'IN'),
('being', 'VBG'), ('top', 'JJ'), ('engaged', 'VBN'), ('members', 'NNS'), ('in', 'IN'), ('my', 'PRP$'), ('community',
'NN'), ('this', 'DT'), ('week', 'NN'), (':)', 'NN')]
```

Removing the noise. We used the regular expressions in Python to remove the noises, which are the hyperlinks, twitter handles in replies, punctuation, and special characters.

```
    for token, tag in pos_tag(tweet_tokens):
        token = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+#]|[!*\(\),]|'\
                       '(?:%[0-9a-fA-F][0-9a-fA-F]))+','', token)
        token = re.sub("(@[A-Za-z0-9_]+)","", token)
```

Removing the stopwords

```
# StopWords
stopwords = nltk.corpus.stopwords.words('english')
stop_words_list = [word for word in alpha_words_list if word not in stopwords]
```

Create a Subjectivity feature and function applies the feature

```
SL_featuresets = [(SL_features(d, word_features, SL), c) for (d, c) in documents]
```

b. Creating the training and test set for model
Label each tweet with either “positive” or “negative”. Create a dataset that has all the positive and negative tweets.

```
train_data = dataset[:5000]
test_data = dataset[5000:]
```

c. Build and train the model. We used the NaiveBayesClassifier in this project. The accuracy is 78.2, which shall ensure the accuracy of sentiment analysis.

```
classifier = nltk.NaiveBayesClassifier.train(train_data)
nltk.classify.accuracy(classifier, test_data)
```

```
0.782
```

```
classifier.show_most_informative_features(10)
```

```
Most Informative Features
          contains(flat) = True              neg : pos    =     23.1 : 1.0
    contains(engrossing) = True              pos : neg    =     18.3 : 1.0
       contains(routine) = True              neg : pos    =     15.7 : 1.0
      contains(mediocre) = True              neg : pos    =     15.1 : 1.0
       contains(generic) = True              neg : pos    =     14.4 : 1.0
     contains(inventive) = True              pos : neg    =     14.3 : 1.0
        contains(boring) = True              neg : pos    =     13.3 : 1.0
     contains(intimate) = True              pos : neg    =     12.9 : 1.0
    contains(refreshing) = True              pos : neg    =     12.3 : 1.0
  contains(refreshingly) = True              pos : neg    =     12.3 : 1.0
```

2. Model Evaluation
The accuracy we got for the model in the previous section [1.c] is calculated with a simple algorithm, which is not convincing enough to choose it for our dataset. Some more evaluations were conducted as follows.
The matrix in figure 16 shows below is the result of a test for the number of actual class labels (“Yes” and “No”) that match with the predicted class.
TP (true positive) means that the predicted class matches the actual one as “No”, FN (false negative) means that the predicted class is “No” which should be “Yes”, FP (false positive) means that predicted class is “Yes” which should be “No”, and TN (true negative) means that the predicted class matches the actual one as “Yes”. In the matrix, the number of these four types are listed.

		Predicted Class	
		Class=Yes	Class=No
Actual Class	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

Figure 16

This measure idea is proposed in the Information Retrieval field and there are two commonly used measures from it. Note that these two measures can be combined into another measure named “F-measure”.

Precision = TP / (TP+FP); Recall = TP/(TP+FN)

F-measure = 2 * (Recall * Precision) / (Recall + Precision)

Apply this method to the newly trained model. Figure 17 shows the computing result.

```
printmeasures('pos', ref1, test1)
```

```
pos precision: 0.7730923694779116
pos recall: 0.7857142857142857
pos F-measure: 0.7793522267206477
```

Figure 17

Sentiment Analysis on Twitter and Weibo

1. Apply the trained model

a. On Tweet dataset

The analysis result indicates there are 923 tweets tagged as "positive" and 1666 tweets as "negative".

```
pos_tokens = []
neg_tokens = []

for index, row in virus_df.iterrows():
    texttokens = nltk.word_tokenize(row['Comment'])
    inputfeatureset = SL_features(texttokens, word_features, SL)

    if classifier.classify(inputfeatureset) == 'pos':
        pos_tokens.append((row['Date'], texttokens))
    elif classifier.classify(inputfeatureset) == 'neg':
        neg_tokens.append((row['Date'], texttokens))


print(len(pos_tokens))
print(len(neg_tokens))
```

```
923
1666
```

b. On Weibo dataset

We chose to use an open-source tool named "bixin" for Chinese sentiment analysis. The analysis result indicates there are 3426 weibo tagged as "positive", 4934 weibo as "negative", and 754 weibo as "neutral".

What is bixin: It is a classifier for Chinese sentiment analysis based on dictionary and rules. It was tested with 6226 tagged corpus mixed with shopping reviews, Sina Weibo and news.

Accuracy of bixin: The latest update was on May 27, 2018, and the current accuracy is 0.827771.

Reference: https://github.com/bung87/bixin

Accuracy

Test with 6226 taged corpus mixed up with shopping reviews 、Sina Weibo tweets 、hotel reviews 、news and financial news

accuracy: **0.827771**

Notice:neutral texts are all ignored.

details about test dataset see wiki 关于测试数据集

```
for txt in cleaned_data:
    try:
        if predict(txt[1]) > 0:
            pos.append(txt)

        elif predict(txt[1]) == 0:
            neu.append(txt)

        elif predict(txt[1]) < 0:
            neg.append(txt)
```

```
Positive:  3426
Neutral:   754
Nagetive:  4934
```

2. Overall Proportion Analysis

a. Twitter

Figure 18 shows the overall proportion of the sentiments (positive and negative) from March 11 to April 10, 2020.
The negative sentiment accounts for the larger part which is 64.3%, while positive sentiment accounts for 35.7%. We could reasonably conclude that the number of people in the United States who has a negative attitude towards the coronavirus is larger than who has a positive attitude.

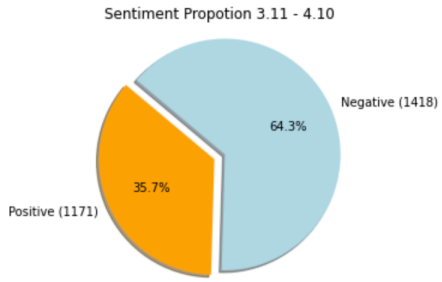

Figure 18

b. Weibo

Figure 19 shows the overall proportion of the sentiments from March 11 to April 10, 2020. From the figure, we could see that over half of the people who posted weibo had negative attitudes towards the COVID-19.

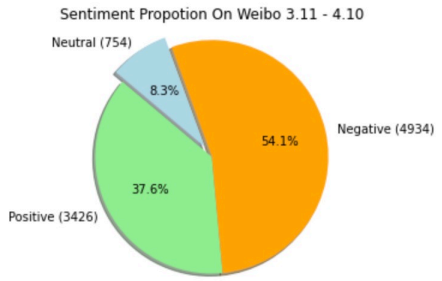

Figure 19

3. WordCloud

We used word cloud to conduct data virtualization for representing the text data. Word Cloud is a technique that represents the text with different size based on the frequency or importance. We downloaded and installed the three necessary modules, which are matplotlib, pandas, and wordcloud.
Noises like stopwords needed to be removed from the data, as they have a high frequency and should not be treated as useful information. We then tokenize the tweets and weibos, and analyzed them using the Python API WordCloud.

a. Twitter

```
    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()

        tokens[i] = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+#]|[!*\(\),]|'\
                        '(?:%[0-9a-fA-F][0-9a-fA-F]))+','', tokens[i])
        tokens[i] = re.sub("(@[A-Za-z0-9_]+)","", tokens[i])


    comment_words += " ".join(tokens)+" "


wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                stopwords = stopwords,
                min_font_size = 10).generate(comment_words)
```

Draw the figure for the wordcloud.

```
# plot the WordCloud image
plt.figure(figsize = (6, 6), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```

From the figure, we could see that many tweets are highly related to Alex Jones and Donald Trump. It seems that people talked about the COVID-19 on Twitter because of the speech from Alex Jones and Donald Trump most likely.

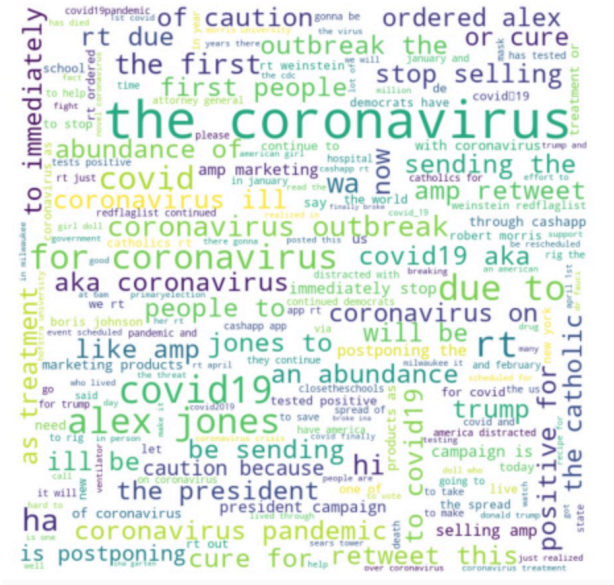

Figure 20

b. Weibo

A Chinese font should be downloaded first since wordcloud does not support the Chinese words.

Similar to how we drew the picture for Twitter, the wordcloud for Chinese weibo is shown below.

Figure 21

4. Daily Sentiment Proportion
We would like to see the detailed changes in people's opinions towards COVID-19 from March to April. An analysis of the daily sentiment proportion is conducted as follows. The time period is one month, therefore a classification by day should be fine and reasonable. We summarized the daily number of positive sentiment and negative sentiment of tweets (and weibo) and drew a line chart to represent the result.

a. Twitter
The label on the x-axis is the data time, and the label on the y-axis is the proportion of positive sentiment to the negative sentiment. In figure 22, the blue line represents the positive sentiment proportion and the red line represents the negative sentiment proportion. The trend of them seems not that obvious. However, we could notice that the trend of positive one seems to decrease and the negative one seems to increase after March 29. Besides, the ratio of them is getting larger, which means that more and more people had a negative attitude towards the COVID-19.

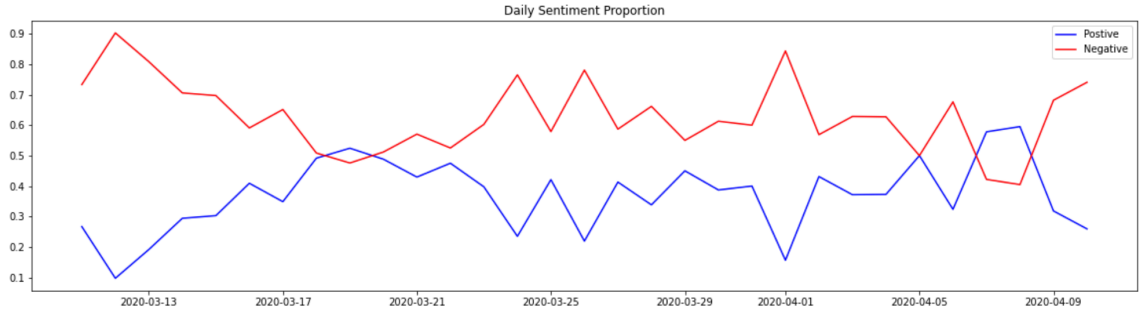

Figure 22

b. Weibo
The green line represents the positive sentiment proportion, the negative line represents the negative sentiment proportion, and the pink line represents the neutral proportion. The proportion of negative weibo has the highest proportion with a quite great difference. An interesting finding is that the negative sentiment proportion has an extremely high value at the end of March, and the negative sentiment proportion of tweets has a sharp increase at the same time. There could be some important news on that day, which is worth studying why most people in China and U.S. both had negative attitudes out of sudden.

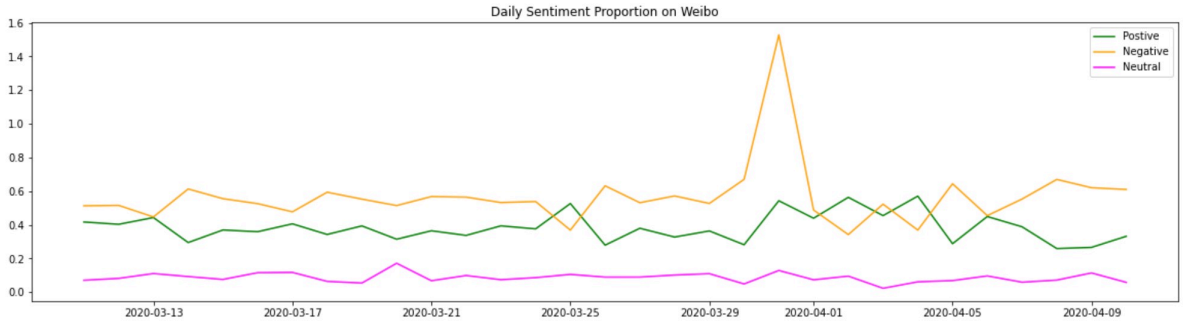

Figure 23

5. Daily Growth Rate on Sentiment Proportion
To have a better grasp of how public opinions changed during that period, we would like to conduct an analysis of the daily growth rate of the sentiment proportion.
In order to observe the change of sentiment more apparently, a changing rate is needed. So, we made another graph which named the weekly sentiment changing rate.

a. Twitter

In this graph, positive values represent the increase of proportion compared to the previous day. The range of changing rate is in [-15%,20%], which is not very large. However, the increment of negative sentiment and decline of neutral sentiment during 2.24-3.1 are very easy to find.

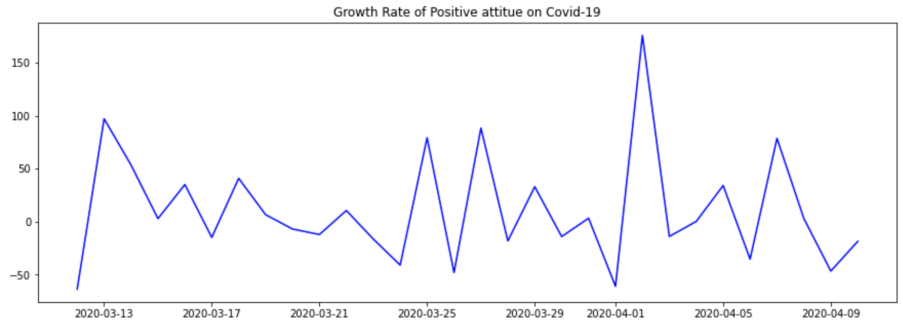

Figure 24

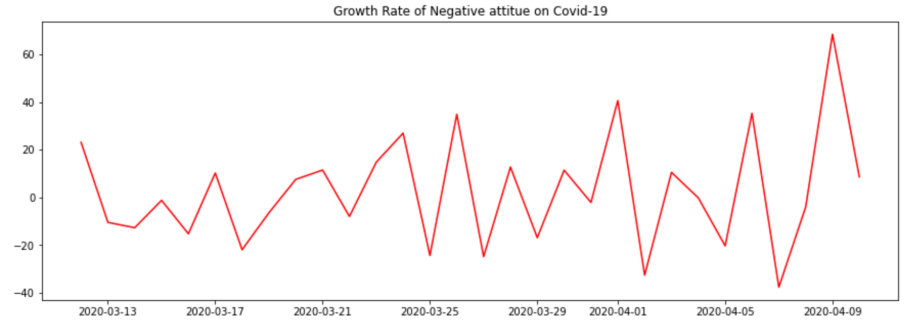

Figure 25

b. Weibo

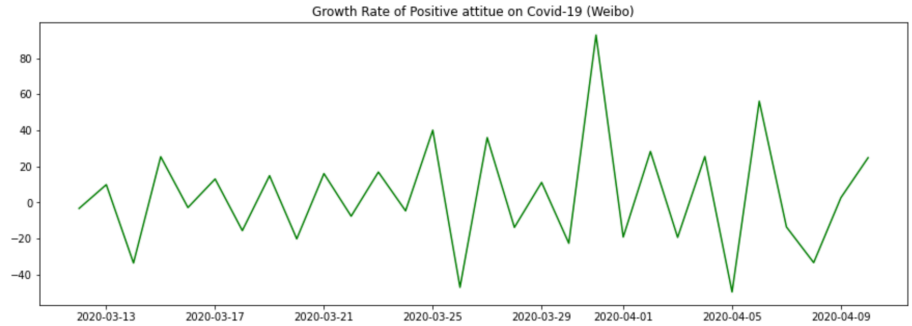

Figure 26

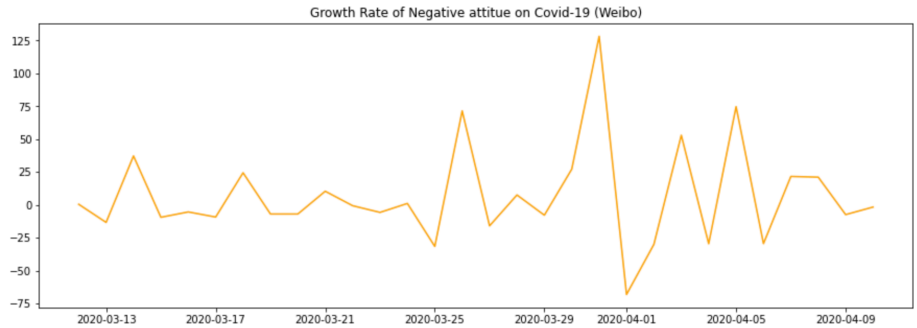

Figure 27

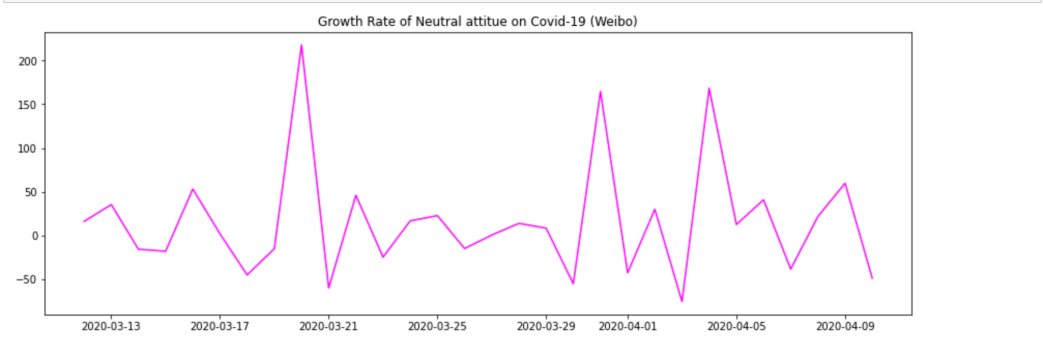

Figure 28

Conclusion

In this study, we collected the most popular tweets and weibos, extracted those containing keywords related to COVID-19, and implemented sentiment analysis on them. We want to see how people in China and the United States thought of the virus, for example, whether they believed it is something serious and urgent, whether they continued paying great attention to it, and whether they had positive or negative attitudes towards it.

There is a great difference in the attitudes of the Chinese and Americans towards the virus. The Chinese people were paying more attention to this matter obviously, and keywords such as pneumonia appeared on Weibo's hot searches much higher than Twitter. Americans did not seem to pay much attention to it though. The proportion of virus-related tweets on Twitter was always low in March. From the analysis of the high-frequency words, when people were talking about pneumonia, most of the talks were not that related to pneumonia.

People generally held a pessimistic attitude towards pneumonia. From the results of the sentiment analysis, the proportion of pessimism is higher than that of optimism. But the growth trends are different in the two countries. In China, the pessimistic growth rate gradually declined, and the optimistic growth rate gradually increased. But in the United States, the proportion of pessimistic growth has been rising. This might be because pneumonia had been effectively controlled in China after March, and the situation has stabilized, while pneumonia has just begun to erupt in the United States.

We believe that this study and other studies similar to this, is of great significance for studying the spread of the COVID-19 in the world and the prevention and control capabilities of various countries.

Reference

1. "Coronavirus Cases:" 2020. *Worldometer*. Accessed May 6. https://www.worldometers.info/coronavirus/.
2. Secon, Holly. 2020. "An Animated Map Tracks the Spread of the Coronavirus as Cases Were Reported in More than 180 Countries." *Business Insider*. Business Insider. https://www.businessinsider.com/map-tracks-novel-coronavirus-spread-in-countries-around-the-world-2020-3.