

# **Natural Language Processing Lab**

## **Mini Project Document**

### **COVID-19 Tweet Analysis**

Group No: 32

1604031	Juhi Gianani
1604032	Priya Gianchandani
1604051	Ankita Kar

# **1. Problem Statement**

COVID-19 analysis by analyzing tweets related to COVID-19 using Named Entity Recognition (NER).

# **2. Problem Definition**

COVID-19 disease is an infectious disease caused by a newly discovered coronavirus. The new Coronavirus is related with coronaviruses causing SARS and MERS. Real time information about infected people is important in order to overcome this disease. Micro-bloggers like Twitter can be a great source to get real time information about the number of infected people in different locations. This project takes the streaming COVID-19 tweets on Twitter as an input and gives the approximate count where COVID-19 might have affected the most. Our project therefore helps in analyzing how COVID-19 is rapidly increasing and affecting various locations all over the world.

# **3. Input**

Tweets from Tweeter related to COVID-19 is the input to this project. We are taking tweets on real-time basis. At a given time, top 1000 tweets containing word “coronavirus” is being retrieved and given as input. Some example tweets are given below.

Timestamp	Tweet
Sun Mar 29 12:26:34 +0000 2020	Coronavirus: Italy becoming impatient with lockdown - and social unrest is brewing <a href="https://t.co/LleYNRDF2d">https://t.co/LleYNRDF2d</a>
Sun Mar 29 12:26:34 +0000 2020	Coronavirus, Chloroquine & Zinc @NHSuk @CDCgov @realDonaldTrump #coronavirus #stayathomeandstaysafe <a href="https://t.co/M5G7eZ4iL7">https://t.co/M5G7eZ4iL7</a> via @YouTube
Sun Mar 29 12:26:34 +0000 2020	RT @cinefilo_K: “morre em SP jovem de 26 anos vítima do coronavírus” “bebê de um ano morre nos EUA vítima de coronavírus”
Sun Mar 29 12:26:35 +0000 2020	RT @MarcACaputo: Before Trump even took office in Jan 2017, the nation’s former Ebola czar was warning he would mismanage a pandemic crisis of coronavirus
Sun Mar 29 12:26:35 +0000 2020	RT @realDonaldTrump: On the recommendation of the White House CoronaVirus Task Force, and upon consultation with the Governor’s of New York
Sun Mar 29 12:26:36 +0000 2020	RT @LeaveEUOfficial: Shocking news that China is celebrating the end of their coronavirus lockdown by reopening the disgusting meat markets
Sun Mar 29 12:26:37 +0000 2020	RT @mrsrats: YOU CANNOT, I repeat YOU CANNOT get coronavirus from your pets. People are dumping their animals at shelters out of ignorance
Sun Mar 29 12:26:37 +0000 2020	RT @Nedunaija: China is an irresponsible country. The world should deal with them after this coronavirus.
Sun Mar 29 12:26:38 +0000 2020	RT @AP_Europe: Germany's Angela Merkel is rising to the occasion, facing the coronavirus pandemic with scientific facts and calm determinat
Sun Mar 29 12:26:38 +0000 2020	RT @NorbertElekes: NEW: Princess Maria Teresa of Spain, 86, has died from coronavirus. World's first royal to die from coronavirus.

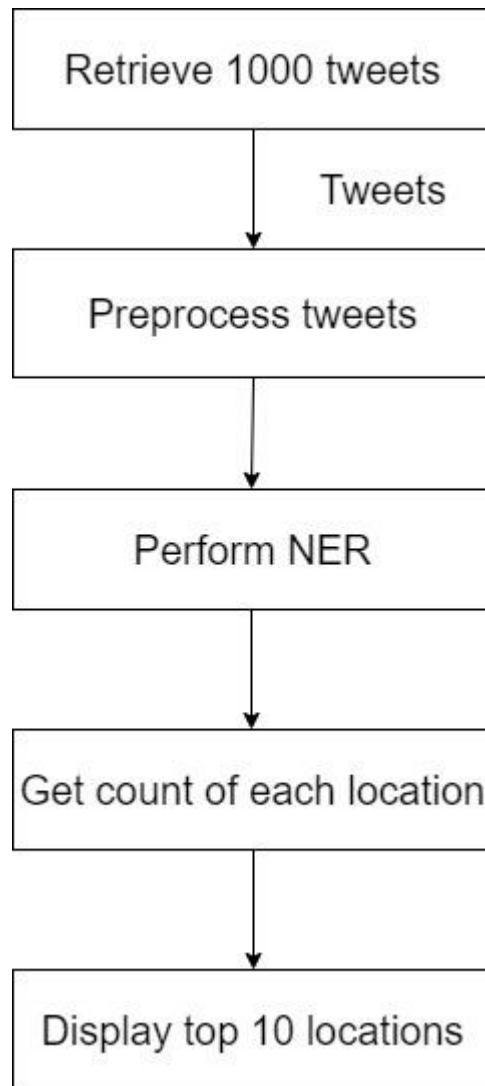
## 4. Output

The output of the project will be a list of top 10 places with highest number of tweets. The base of our analysis is that the we will get more number of tweets from those places which are more effected by COVID-19. Following is the output when we executed our model on March 29 2020 at 12:30:10.

Top 10 places with highest number of tweets:-  
China: 36  
US: 35  
Washington: 21  
Wuhan: 10  
Germany: 9  
Spain: 9  
Italy: 8  
France: 6  
Florida: 6  
India: 5

## **5. Algorithm Used**

One of the main requirements of this project is a Twitter Developer Account. Since we are retrieving real-time tweets from twitter, we have to use Twitter API for extracting the tweets which are available only for Twitter Developer Account holders. After creating Twitter Developer Account, we retrieved 1000 real-time tweets which had the word “coronavirus”. The following image shows the flow diagram of our project.



## 6. Libraries Used

The following libraries are used in this project.

### 1. Tweepy:

This module is used for connecting the python program with Twitter Developer Account.

### 2. Algorithmia:

This module is used for retrieving tweets from Twitter and performing NER on the retrieved tweets.

## 7. Code

```
import re
from collections import defaultdict, Counter
import Algorithmia
import tweepy
import openpyxl

client = Algorithmia.client("api_key")          ## API key ##

## function for retrieving 1000 tweets ##
def pull_tweets():
    input = {
        "query": "coronavirus",                ## Keyword for searching ##
        "numTweets": "1000",                  ## Number of tweets ##
        "auth": {
            "app_key": 'consumer_key',         ## Twitter keys ##
            "app_secret": 'consumer_secret_key',
            "oauth_token": 'access_token',
            "oauth_token_secret": 'access_token_secret'
        }
    }

    twitter_algo = client.algo("twitter/RetrieveTweetsWithKeyword/0.1.3")
    result = twitter_algo.pipe(input).result
    tweet_list = [tweets['text'] for tweets in result]

    workbook = openpyxl.Workbook()            ## Excel sheet for entering tweets ##
    sheet = workbook.active
    sheet.title = "Tweets"
    cell1 = sheet.cell(row=1, column=1)
    cell1.value = "Timestamp"
    cell1 = sheet.cell(row=1, column=2)
    cell1.value = "Tweet"
```

```

for i in range(len(result)):
    c1 = sheet.cell(row=i+2, column=1)
    c1.value = result[i]['created_at']
    c2 = sheet.cell(row=i+2, column=2)
    c2.value = result[i]['text']
workbook.save(r"your_file_path")          ## Add the project folder path here Eg.
C:\Users\Juhi\Desktop\Coronavirus\InputTweets.xlsx (InputTweets.xlsx is the file name) ##
return tweet_list

```

## function for preprocessing tweets ##

```

def process_text():
    """Remove emoticons, numbers etc. and returns list of cleaned tweets."""
    data = pull_tweets()
    regex_remove = "(@[A-Za-z0-9]+)|(^0-9A-Za-z \t)|(\w+:\V\S+)|^RT|http.+?"
    stripped_text = [
        re.sub(regex_remove, "",
            tweets).strip() for tweets in data
    ]
    return ' '.join(stripped_text)

```

## function for performing Named Entity Recognition (NER) ##

```

def get_ner():
    """Get named entities from the NER algorithm using clean tweet data."""
    data = process_text()
    ner_algo = client.algo(
        'StanfordNLP/NamedEntityRecognition/0.1.1').set_options(timeout=600)
    ner_result = ner_algo.pipe(data).result
    return ner_result

```

```
## function for getting counts of each location and organization ##
```

```
def group_data():  
    data = get_ner()  
    default_dict = defaultdict(list)  
    for items in data:  
        for k, v in items:  
            if 'LOCATION' in v or 'ORGANIZATION' in v or 'NAME' in v:  
                default_dict[v].append(k)  
    ner_list = [{keys: Counter(values)}  
                for (keys, values) in default_dict.items()]  
    return ner_list
```

```
## MAIN FUNCTION ##
```

```
if(__name__ == '__main__'):  
    result = group_data()  
    data = result[0]  
    for i in data.keys():  
        data[i] = dict(data[i])  
    if('LOCATION' in data.keys()):  
        locations = data['LOCATION']  
        f = open('Output.txt','w')  
        ## Text file for saving output  
        ##  
        if(len(locations)>=10):  
            print("Top 10 places with highest number of tweets:-\n")  
            f.write("Top 10 places with highest number of tweets:-\n")  
            for _ in range(10):  
                place = max(locations.items(), key = lambda x: x[1])  
                print("{0}: {1}".format(place[0], place[1]))  
                f.write(place[0]+": "+str(place[1])+"\n")  
                del locations[place[0]]  
            else:  
                print("Places with highest number of tweets:-\n")  
                f.write("Places with highest number of tweets:-\n")
```

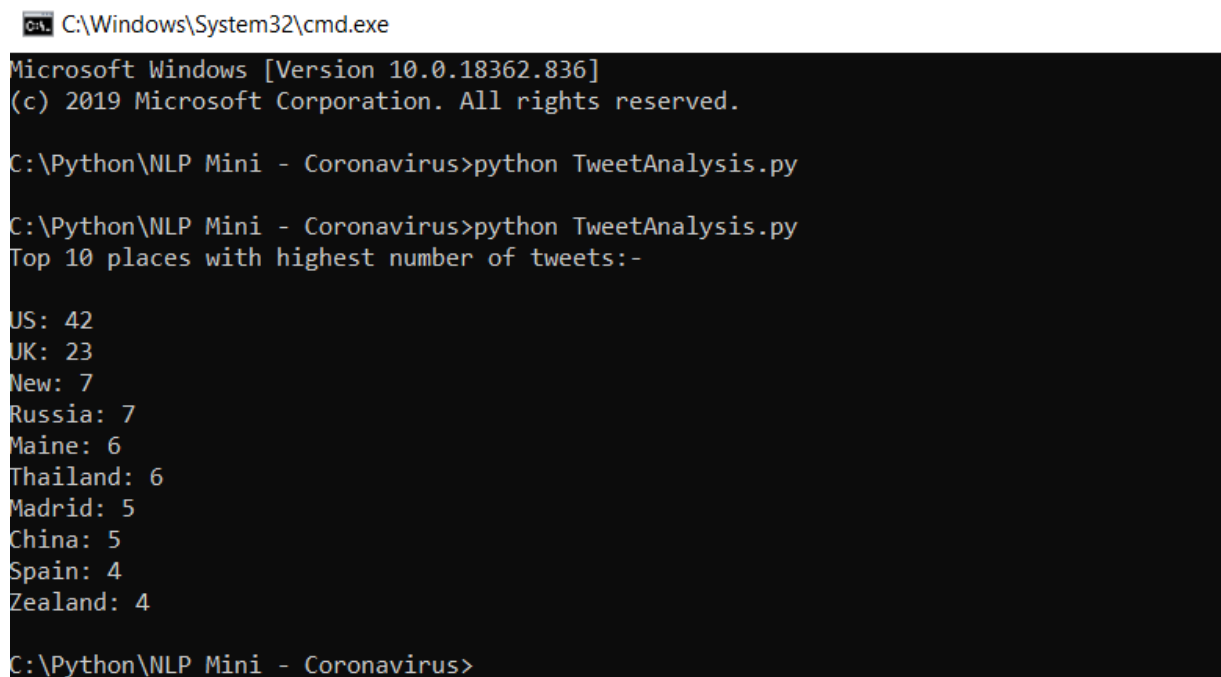


```
for place in locations:
    print("{0}: {1}".format(place, locations[place]))
    f.write(place+": "+str(locations[place])+"\n")
f.close()
```

## 8. Output

Once the TweetAnalysis.py file is run, an output file will be created at the same place whose path we had specified in the code. Suppose **Output.txt** is the name of the file that we had specified in the code. It will be created in the same folder as of **Tweet.xlsx** where all the tweets were retrieved. The following images show output in command prompt and in Output.txt.

### Command Prompt



The screenshot shows a Windows Command Prompt window titled "C:\Windows\System32\cmd.exe". The prompt displays the following text:

```
Microsoft Windows [Version 10.0.18362.836]
(c) 2019 Microsoft Corporation. All rights reserved.


C:\Python\NLP Mini - Coronavirus>python TweetAnalysis.py

C:\Python\NLP Mini - Coronavirus>python TweetAnalysis.py
Top 10 places with highest number of tweets:-

US: 42
UK: 23
New: 7
Russia: 7
Maine: 6
Thailand: 6
Madrid: 5
China: 5
Spain: 4
Zealand: 4

C:\Python\NLP Mini - Coronavirus>
```

## Output.txt

 Output - Notepad

File Edit Format View Help

Top 10 places with highest number of tweets:-

US: 42

UK: 23

New: 7

Russia: 7

Maine: 6

Thailand: 6

Madrid: 5

China: 5

Spain: 4

Zealand: 4