

INTERNSHIP REPORT

COVID19 TWITTER DATA ANALYSIS

By

TEAM ALPHA TITANS

Manjeet Singh Lakhanpal

Sahibjot Singh

Harkirat Singh



At

SPOTLE.AI

INTRODUCTION

OBJECTIVE:

To build a Twitter Trend analyzer that will analyze a set of tweets using NLP and text-processing techniques. The trend analyzer will work on a given set of tweets, seeded on COVID19 / CORONA.

1. A tag cloud depicting what topics / word was being talked about on twitter.
2. Which hashtag trended most on twitter.
3. Which twitter handler dominated the conversation on twitter.

BACKGROUND:

Every time you log in to twitter, it is likely that you are checking the trends list to see what is trending right then. From Salman Khan to Beyonce, from #COVID to #LOCKDOWN everything and everyone worth talking about have trended on Twitter. Twitter follows a sophisticated mechanism to uncover trends. This tutorial demonstrates the analysis of tweets on the Assembly elections 2019 in India:

<https://spotle.ai/feeddetails/twitter-trend-elections/4130>

DATA:

Download the set of tweets for analysis here:

https://cdn.spotle.ai/zip/Twitter_Data_IN.csv.zip

The dataset will be updated periodically. You have to work on the latest dataset.



Tweeter_Data_IN.csv

COVID19 TWITTER DATA ANALYSIS



TEAM ALPHA TITANS

- Manjeet Singh Lakhanpal
- Sahibjot Singh
- Harkirat Singh

To analys the COVID19 twitter data given by the SOPTLE.AI

Importing the Dependencies/Libraries

In [1]:

```
import numpy as np
import pandas as pd
import re
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS
```

In [2]:

```
tweets = pd.read_csv("Tweeter_Data_IN.csv")
```

In [3]:

```
tweets.head()
```

Out[3]:

	created_at	hashtags	favorite_count	id	lang	place	retweet_count	
0	Wed Mar 25 06:20:02 +0000 2020	NaN	0	1242697733175220000	en	Pune, India	0	Meditation
1	Wed Mar 25 06:36:25 +0000 2020	NaN	0	1242701857253980000	en	Gandhinagar, India	0	Con C
2	Wed Mar 25 06:18:32 +0000 2020	NaN	0	1242697356249880000	en	Bidhan Nagar, India	0	@Dell @
3	Wed Mar 25 06:05:46 +0000 2020	SSC_UFM_MAT_KARONA Corona UFM	9	1242694142242650000	en	Maharashtra, India	19	#SSC_UFM_M
4	Wed Mar 25 06:31:10 +0000 2020	Corona pritamkumarmurari Voice	0	1242700536752700000	en	Bokaro Steel City, India	0	https://i



```
len(tweets)
```

Out[4]:

44179

In [5]:

```
data = [] # Data contains tweet without tweet link in the end
for i in range(len(tweets)): # Splitting each tweet from https to remove link from tweet

    text = tweets.text[i]
    og_tweet = text.split("https")[0]
    data.append(og_tweet)
```

In [6]:

```
len(data)
```

Out[6]:

44179

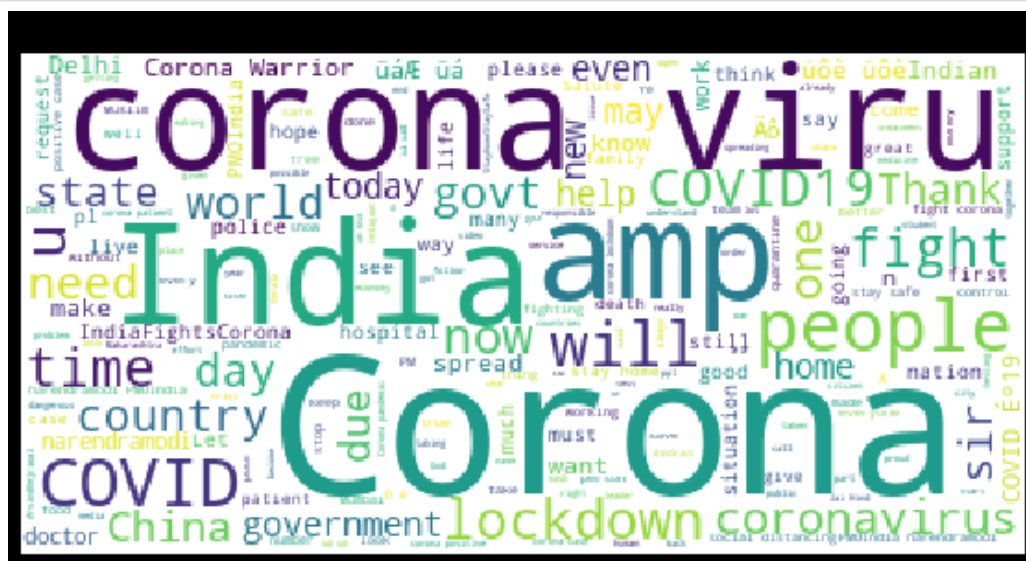
Task 1 - A tag cloud depicting what topics were being talked about on Twitter

In [7]:

```
def word_cloud(tweets):
    stopwords = set(STOPWORDS)
    wordcloud = WordCloud(background_color="white", stopwords=stopwords, random_state = 42)
    .generate(" ".join([tw for tw in tweets]))
    plt.figure( figsize=(10,5), facecolor='k')
    plt.imshow(wordcloud)
    plt.axis("off")
    plt.title("Twitter WordCloud")
```

In [8]:

```
word_cloud(data)
```



Task 2 - Which hashtag trended more on twitter

In [9]:

```
data[:5]
```

Out[9]:

In [10]:

[illegible]

In [11]:

```
#Corona                2428
#COVID19                1684
#corona                 1605
#coronavirus            1316
#IndiaFightsCorona      1075
#COVID,„É°19           955
#lockdown               885
#COVID                 697
#Covid_19              488
#StayHomeStaySafe      457
Name: hashtag, dtype: int64
```

In [12]:

In [13]:

Out[13]:

	hashtag	val
0	#Corona	2428
1	#COVID19	1684
2	#corona	1605

3	#coronavirus	1318
4	#IndiaFightsCorona	1075
5	#COVID_19	955
6	#lockdown	885
7	#COVID	697
8	#Covid_19	488
9	#StayHomeStaySafe	457

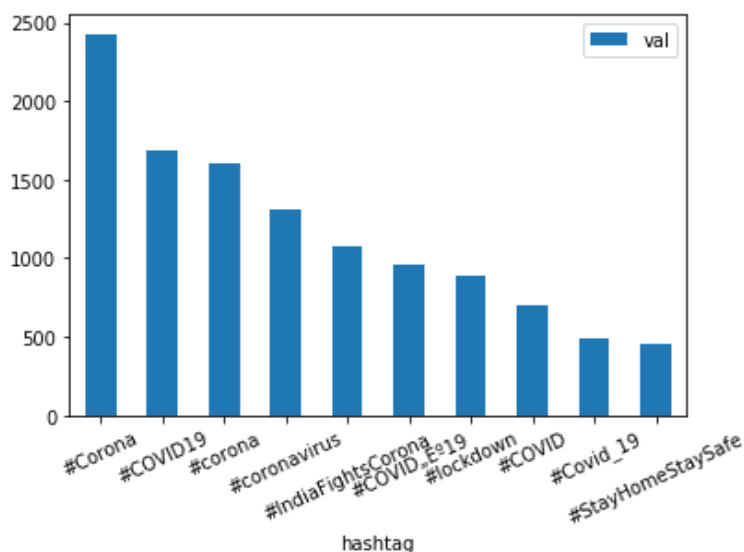
Bar Graph for Top Hastags

In [14]:

```
hashtag1.plot.bar(x="hashtag" , y= "val", rot=25)
```

Out[14]:

<matplotlib.axes._subplots.AxesSubplot at 0x1504de5e668>



Task 3 - Which Twitter Handler dominated conversation on Twitter

In [15]:

```
# Extracting handlers from raw data
raw = ' '.join(data)
tags = [re.sub(r"(\W+)$", "", j[1:]) for j in [i for i in raw.split() if i.startswith("@")
) and len(i) != 1 ]]
d = pd.DataFrame({"handler": tags})
print(d)
```

```

      handler
0    DelhiPolice
1    DCPEastDelhi
2      msisodia
3    AamAadmiParty
4    ArvindKejriwal
...
57007    NITIAayog
57008    FinMinIndia
57009    PMOIndia
57010    ZeeNews
57011    sudhirchaudhary
```

[57012 rows x 1 columns]

In [16]:

```
# Counting handlers by using .value_count() function from pandas
top10_handlers = d['handler'].value_counts().head(10)
top10_handlers
```

Out[16]:

```
narendramodi      3622
PMOIndia          2655
AmitShah           722
ArvindKejriwal     708
MoHFW_INDIA        511
myogiadityanath     504
CMOMaharashtra     470
RahulGandhi        454
aajtak             416
BJP4India          337
Name: handler, dtype: int64
```

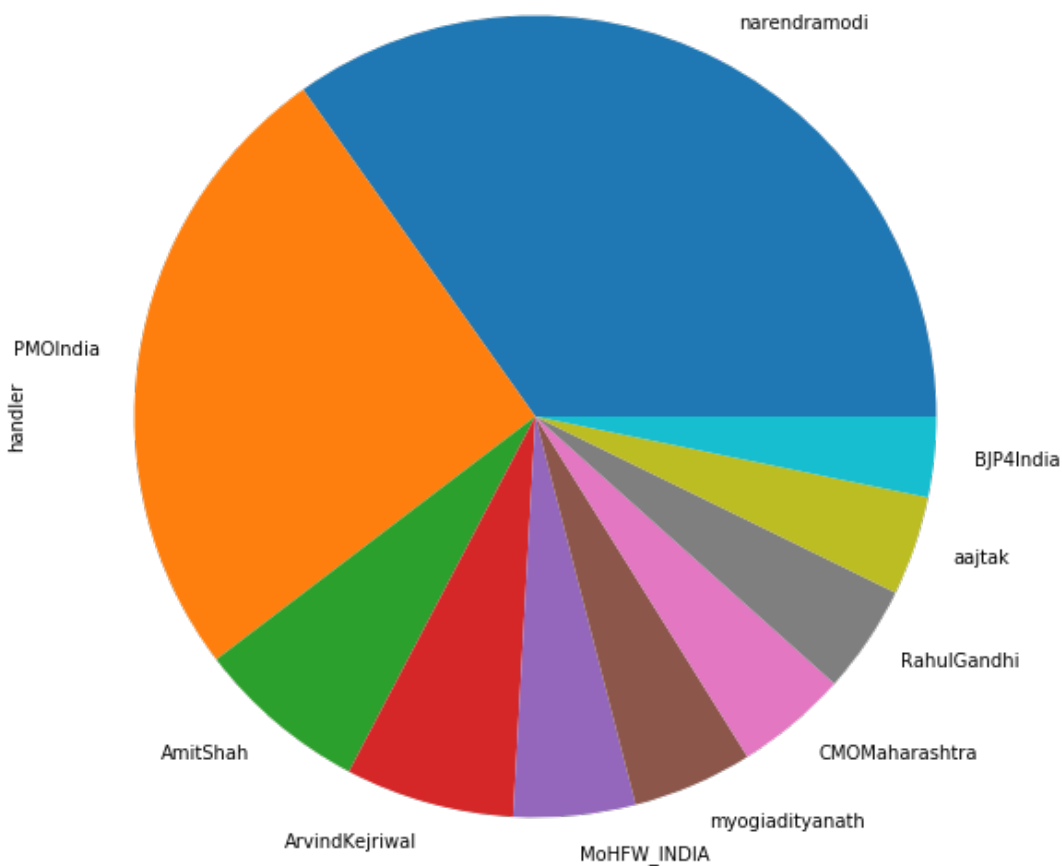
Plotting top 10 handlers using pie function of pandas

In [18]:

```
top10_handlers.plot.pie(y="handler",figsize=(10, 10))
```

Out[18]:

<matplotlib.axes._subplots.AxesSubplot at 0x150453adcf8>



STORING THE PREPROCESSED DATA FOR FUTURE USE

In [19]:

```
import pickle
```

In [20]:

```
with open("data.pkl", "wb") as f:  
    pickle.dump(data, f)
```

In [21]:

```
with open("data.pkl", "rb") as f:  
    data = pickle.load(f)
```

In [22]:

```
len(data)
```

Out[22]:

44179

□

THE END