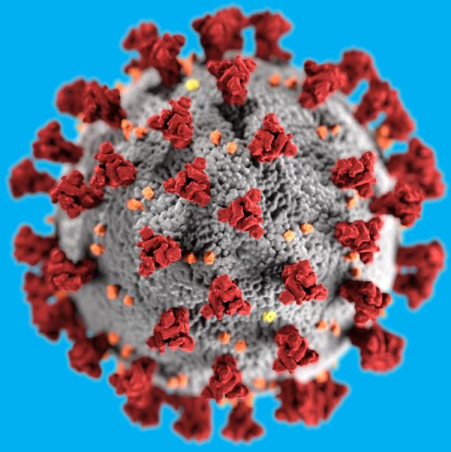


COVID19 Twitter Data Analysis

A detailed Report on Twitter Trend Analysis using python

By: SHIVAM KUMAR GIRI

COVID19 Twitter Data Analysis

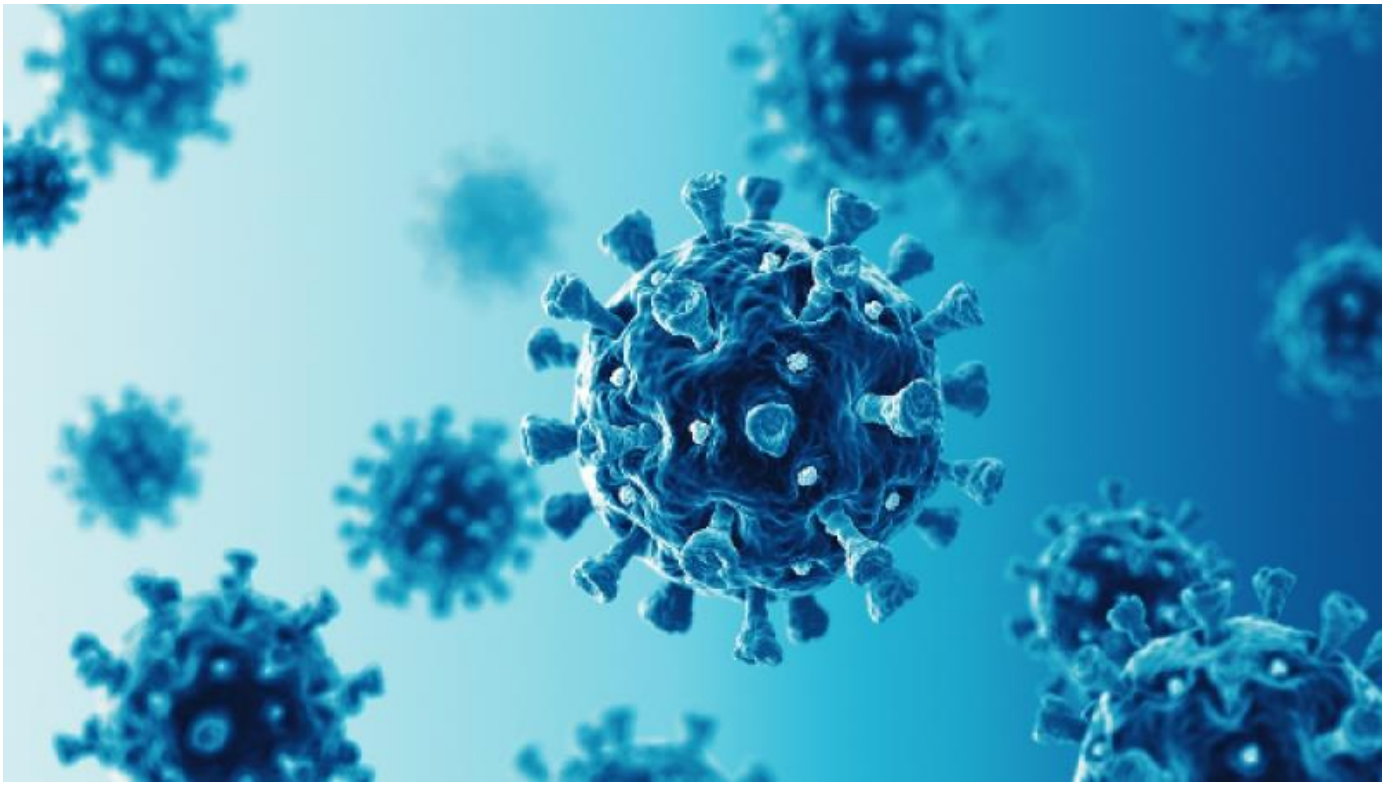
PROBLEM STATEMENTS

To **build a twitter trend analyzer** which 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:

- **A tag cloud** depicting what topics / Word were being talked about on Twitter
- **Trending hashtag**
- Twitter Handler **which dominated** conversation on Twitter

ABSTRACT

COVID-19 is a **humanitarian crisis on a global scale**. The virus continues to spread throughout the globe, placing health systems under unprecedented stress in the battle to save lives. The human scale of this tragedy is set to worsen as the virus spreads to lower income countries with weaker healthcare systems. Twitter is a huge social media platform. There are millions of tweets every day, ranging from politicians to celebrities on various social problems. Here we are going to build a twitter trend analyzer which can **analyze the tweets** seeded on COVID19 and show **the trending hashtag and dominated Twitter Handler**



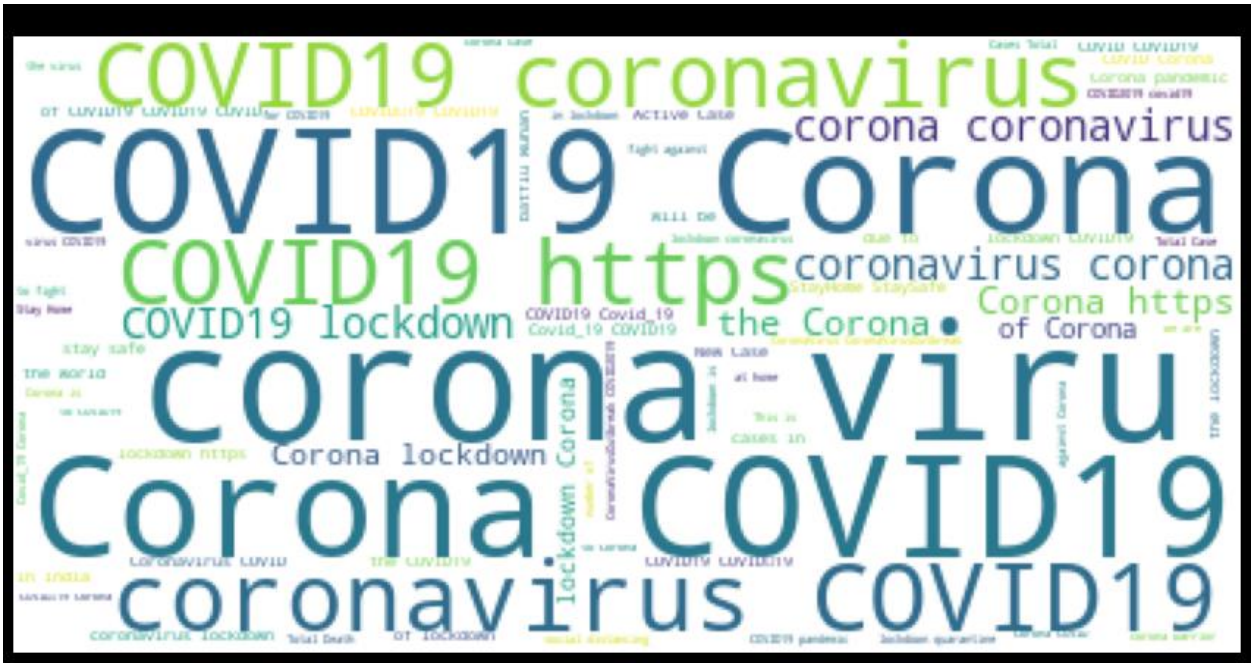
METHODOLOGY

1 a) Generate a word cloud based on the tweets

Step 1: Import required Libraries
Libraries such as PIL, nltk, pandas, matplotlib, numpy, wordclouds and re (regular Expressions are imported). Matplotlib for data visualization, nltk for NLP, padas and numpy for data exploration, PIL for image processing and wordcloud for word clouds.

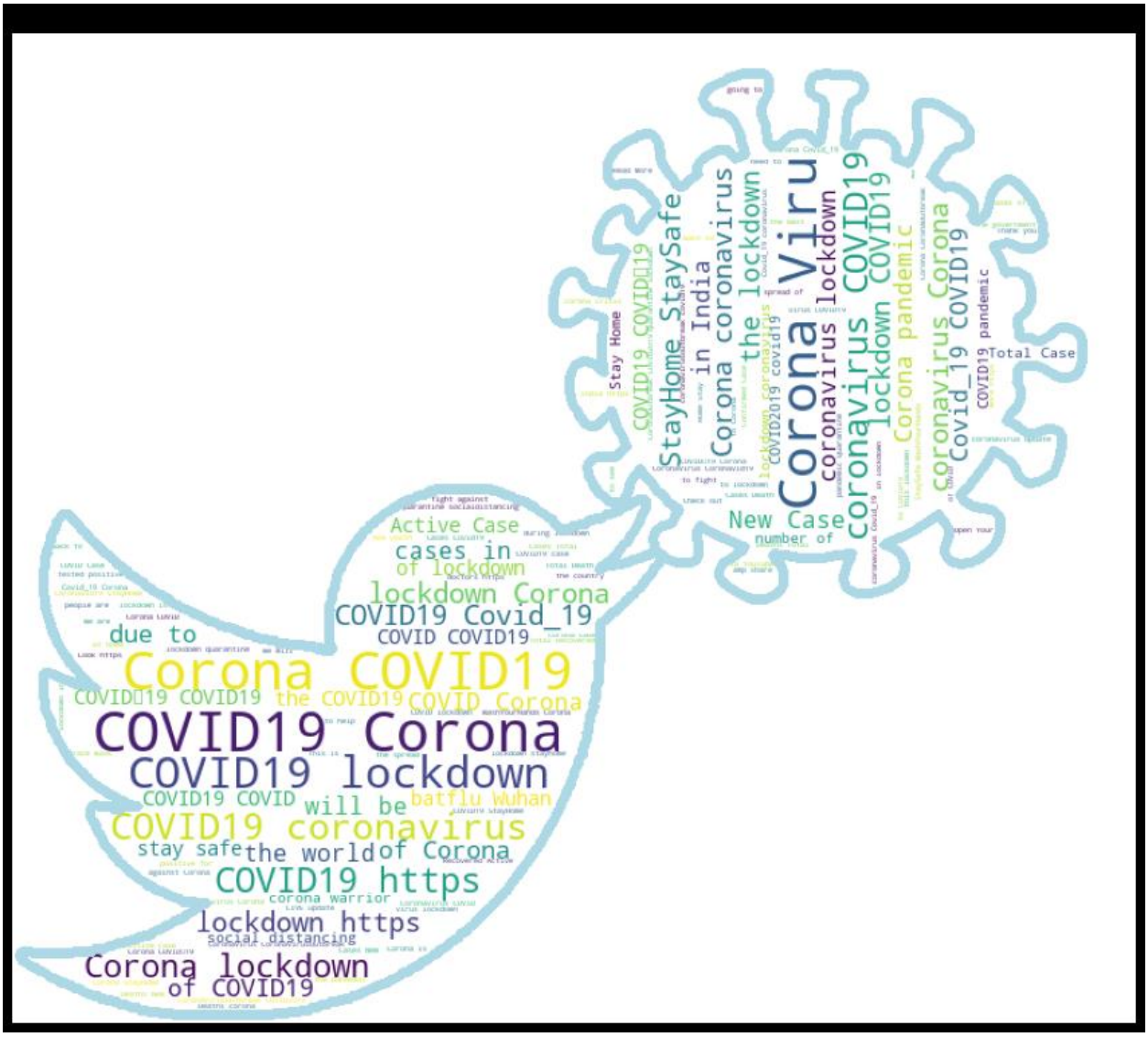
Step 2: Importing the tweets
Here we are importing files from tweets_corona.txt

Step 3: Constructing Word Cloud
Using Matplotlib and Word cloud Libraries we generate the word cloud. This can be modified according to different background styles, number of words, figure size etc.



We can see most common tweets like Covid19, coronavirus and Corona etc.

Step 4: ***Constructing Word Cloud with added stop word and twitter-Corona Mask***



1 b) Show the relative popularity of the hashtags

Step 5: Basic data Exploration and *Creating a hashtag list from tweets seeded Covid19*

Get the frequency of various hashtags. Output is as follows:

```
Total number of hashtags in the tweets: 584666
Total number of Unique tags: 18896

Total number of Twitter-handles in the tweets: 60126
Total number of Unique Twitter-handles: 6531
```

Now, we will fetch all the hashtags from the file and store in python dataframe.

Step 6: Display the count in descending (popular to unpopular)

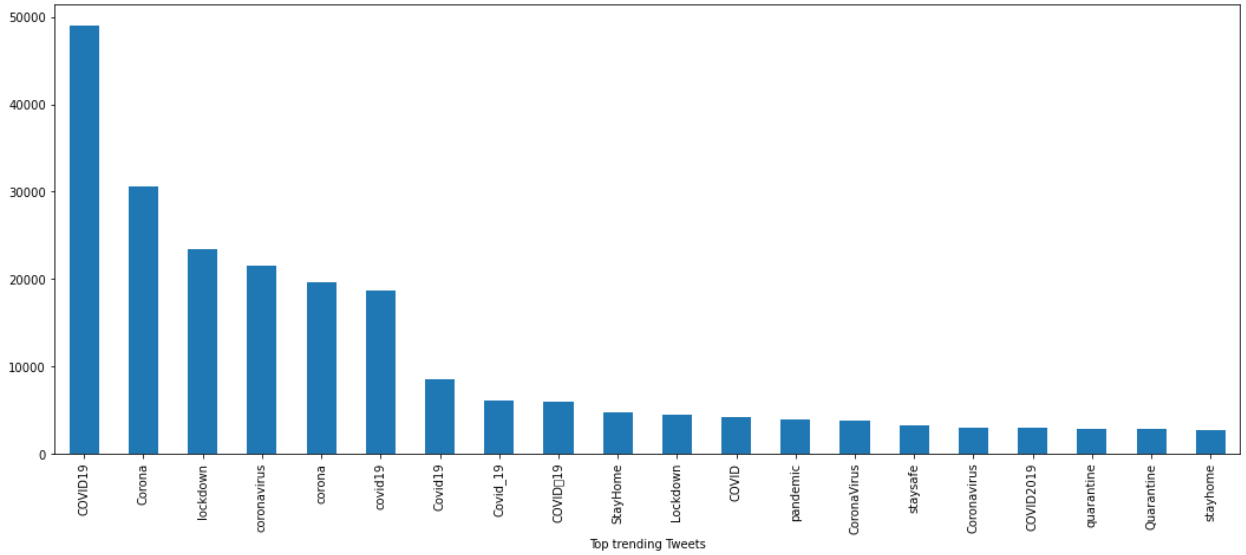
Here is the list of top 20 tweets:

COVID19	49016
Corona	30586
lockdown	23420
coronavirus	21504
corona	19609
covid19	18718
Covid19	8530
Covid_19	6145
COVID‿19	6056
StayHome	4796
Lockdown	4544
COVID	4225
pandemic	3919
CoronaVirus	3875
staysafe	3224
Coronavirus	3072
COVID2019	2953
quarantine	2879
Quarantine	2849
stayhome	2807

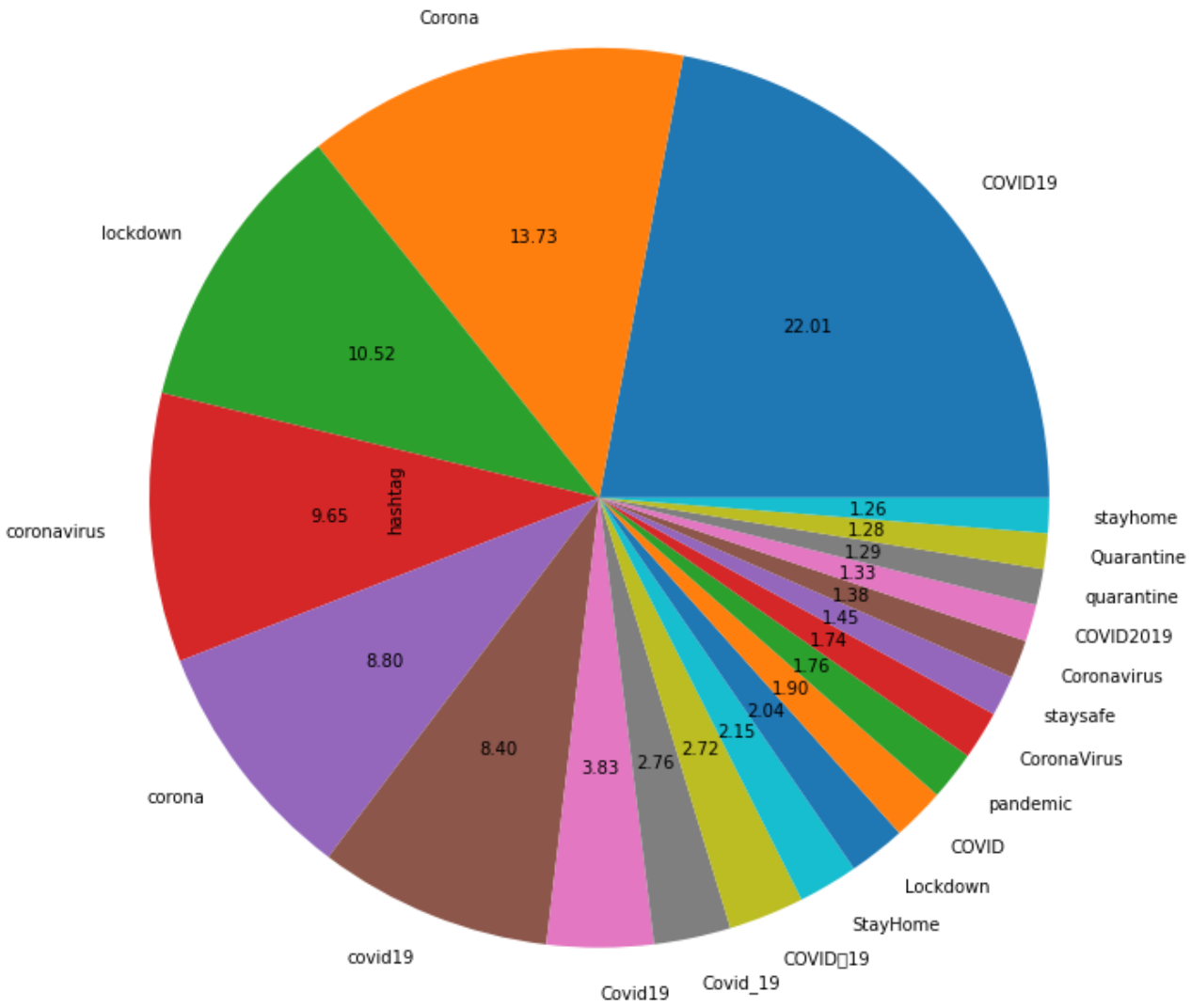
Here is graphical plotting of data

Here we can see the most trend topic is COVID19 and its related terminologies. The most widely used hashtag in Twitter is **#COVID19**

Step 7: ***Creating a bar Graph on relative popularity of Tags***



Step 8: ***Creating a pie Chart on relative popularity of Tags***



1 c) Show which Twitter handler had the maximum share of voice

Step 8: Optimize the Tweets as per NLP

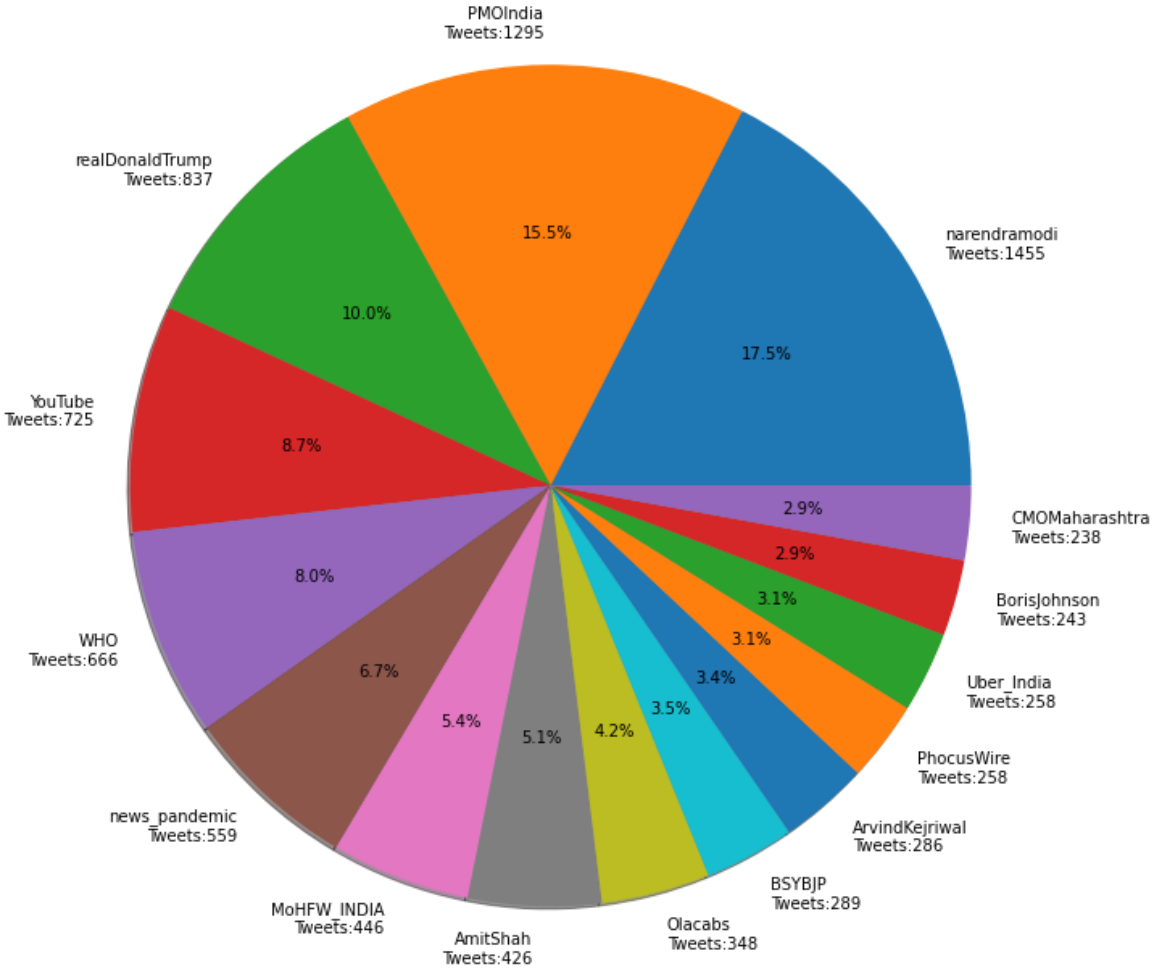
Here I used spacy to extract NLP out of tweets

Step 9: function to find Twitter Handlers which had the maximum share of voice

Output is as follows:

narendramodi	1455
PMOIndia	1295
realDonaldTrump	837
YouTube	725
WHO	666
news_pandemic	559
MoHFW_INDIA	446
AmitShah	426
Olacabs	348
BSYBJP	289
ArvindKejriwal	286
PhocusWire	258
Uber_India	258
BorisJohnson	243
CMOMaharashtra	238

Step 11: Plotting a Pie Chart of Twitter Handler had the maximum share of voice



COVID19 Twitter Data Analysis

OBSERVATION

Here are the following observations:

From Basic Data Exploration:

Total number of hashtags in the tweets: 584666
Total number of Unique tags: 18896
Total number of Twitter-handles in the tweets: 60126
Total number of Unique Twitter-handles: 6531

From Analysis of Data:

From the Graphs and Analysis of Data, I am with the conclusion that

Most Popular Hashtag: #COVID19 with 49016 hashtags
Most dominant Voice in Twitter: @narendramodi with 1455 tweets