REPORT ON COVID-19 TWITTER DATA ANALYSIS

SUBMITTED BY
HRISHIKESH KALITA

Data download link :
https://spotleai.sgp1.digitaloceanspaces.com/course/zip/tweets_corona.txt.zip

BRIEF INFO

The data contain various tweets from Twitter regarding the present crisis of the Corona virus (COVID19). It showers light over the thoughts of the common mass or the public during this phase of imminent danger.

GOALS

- To bring out a WORDCLOUD in order to get an overview of the words used by the people.
- To compare the usage of various HASHTAGS.
- To spot the top 10 HANDLES used by people to address various organisations/people in power.

PYTHON LIBRARIES USED

- matplotlib
- wordcloud
- re
- pandas
- numpy
- nltk

PROCESS FLOW

- DATA CLEANING

  - **Motive**: The text data file is filled with various symbols, numbers, dates, URLs etc. These would definitely interfere in the upcoming process of analysing the data. Since these are practically not very useful in the analysis we are to perform, hence it's a good choice to remove them from the data to clean it and prepare it for further analysis. This was done in the following steps :

    - Clearing of newline and **white space** characters (using **re**).
    - Deleting **URLs** and other symbols from the data (using **re**). A copy of the data was kept before deleting the symbols from the data because we would need that to analyse the hashtags and the handles.

- DATA EXTRACTION

  - **Motive** : To proceed to the analysis part, the meaningful data need to be extracted from the original data. This was done by :

    - **NLTK** library was used in order to **omit** some **commonly used words** (such as 'the, he, she, an, a, etc') so that the **wordcloud** is not flooded with them and we get the proper view of the data.
    - **Hashtags** and **Handles** were extracted using Python's Regular Expression library called '**re'**.

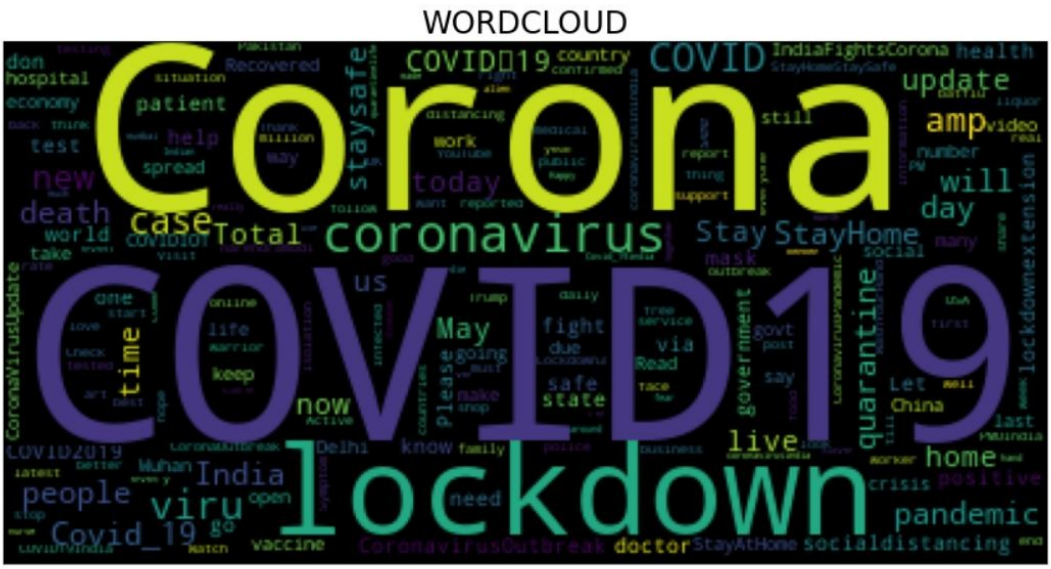
- DATA ANALYSIS

  - **Motive** : To bring out the facts hidden inside the data. This was done by :

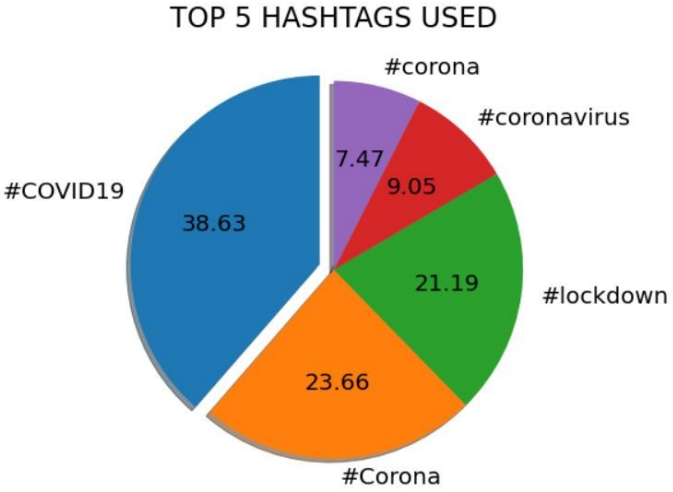
    - The Word Cloud was plotted.
    - Pie plots were plotted for Hashtags and Handles.

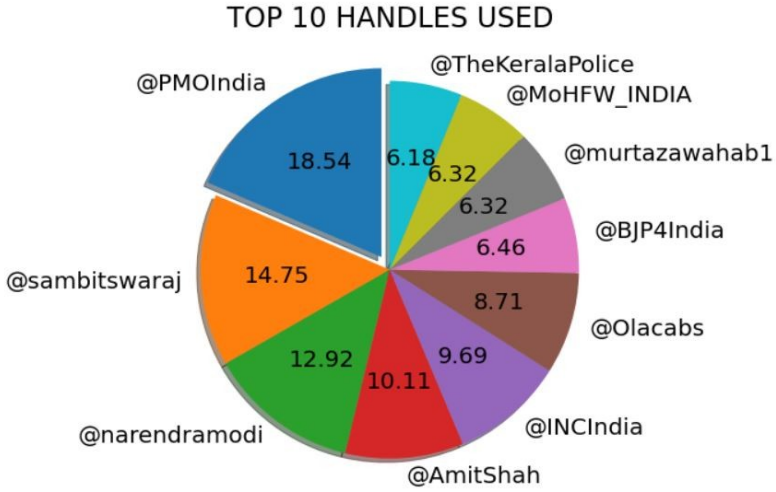
Submitted By
Hrishikesh Kalita

PLOTTINGS



USED HASHTAGS and HANDLES





Submitted By
Hrishikesh Kalita

CONCLUSIONS

1. From the wordcloud it's pretty clear that the tweets have been mostly about the **Corona Virus** and the ongoing **Lockdown.**
2. **#COVID19** seems to be the **most used hashtag** in the tweets.
3. **@PMOIndia** grabs the **most active handle**. The public would definitely look up to the leader of the nation in this time of moral crisis.
4. **Top 10 Handles** : '@PMOIndia', '@sambitswaraj', '@narendramodi', '@AmitShah', '@INCIndia', '@Olacabs', '@BJP4India', '@murtazawahab1', '@MoHFW_INDIA', '@TheKeralaPolice'

Submitted By
Hrishikesh Kalita