Endsem Take home (Marks-50)

Deadline: Sunday May 10, 2020 23:59:00

Instructions :

- Languages allowed: Python/Java
- You are free to use API documentation but if referring to any other sources, please cite them
- Please write your own code. All codes will be tested for Plagiarism and if found, institute policy for plagiarism will be followed.
- If you are using Python, it will be good to use Jupyter notebook, to show analysis, graphs, and code.
- Document your code properly.
- You can use any database for storing the data but it will be tested at the time of demo.
- Write all the analysis along with graphs, charts, etc in **analysis.pdf**
- Make a readme.txt file with instructions on how to run the code. All libraries, sources, etc used should be properly mentioned in it.
- Do the exam in groups of at most 3.
- Zip all your code files along with analysis and readme file in RollNo_Name_TakeHome..zip format. Example 201402230_Swati_TakeHome.zip
- P.S: **Saturday May 09, 2020 1600hrs - 1700hrs are TA hours for query clarification in Link embedded in Cal. Event. In case of mail, related queries to be asked in the same thread as floated by TAs for the purpose with Subject: "[ENDSEM-TAKEHOME] Queries Thread".**

Question:

Given below the dataset of COVID-19:
https://api.covid19india.org/csv/latest/raw_data1.csv
The data constitutes of confirmed cases till April 19, 2020

Dataset description:

1. *Patient Number*: the unique identification number of the patient.
2. *State Patient Number:* the unique identification number of the patient for a particular state.
3. *Date announced:* the date on which COVID is detected.
4. *Notes*: contains the textual reason for the infection spread.
5. Source_1, Source_2, Source_3: Sources from where the information was gathered
   The rest of the fields are self-explanatory.

a) Draw the bar chart of all different sources (including Source_1, Source_2, Source_3) where x-axis: source(Facebook, Twitter, etc), y-axis: number of samples in data collected from that source.
Example: Twitter: 10000, Facebook: 400, Hindustan times: 300, etc

[5 marks]

b) Which is the major information source in the dataset in terms of OSM and other sources? Please put the quantitative numbers for both sources.

[2 marks]

Consider all the Sources from the dataset where the Social media platform used is twitter. Collect the Tweets from the Tweet id present in status URL and answer the below questions:

c) List down the 10 most recent Tweets from the data. **[3 marks]**

d) Plot the word Cloud of all the Tweets. What are the most prominent words? **[5 marks]**

e) How many tweets are from verified sources and Non-Verified Sources? How credible are the non-verified sources? Given quantitative numbers. Pick 5 non-verified Twitter handles who post about confirmed cases. Identify and analyze user profiles. Comment on why we should believe or not believe these user's claims of COVID19. Write at least 5 observations.

[5+5 marks]

f) Identify the PII information from the Tweets. What is the major PII information revealed from the Tweets? State at least 5 PII and the relevant tweets. **[15 marks]**

g) What percentage of Tweets are revealing about the location (as mentioned in the Tweet content itself) of the Patient which matches the Location Column (Detected City/ Detected State) vs (Notes column) in the dataset? **[5 marks]**

Hint: You can make use of NER (Named-Entity Recognizer) and regex matching techniques for identifying location from the text.

h) Which Twitter user profile has the most number of COVID19 posts? Is it a verified handle?

[5 marks]