CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Prof. Biplav Srivastava, Spring 2021

Finals / For Undergraduates Only/ Apr 29, 2021/ Instructions

• Return answer to quiz as .pdf and GitHub link by 1:00 pm on Tuesday, May 4, 2021 by posting to your shared folder (e.g., Google folder mentioned in spreadsheet) and email to biplav.s@sc.edu.

• Ask question by email. Or, office hour of Monday, May 3, 2021 can be used to clarify questions. Timing: 11:30am-12:30pm.

• Total points = 100, Obtained =

Student Name: Terric Taylor

GitHub link with final solution: https://github.com/terricT/TerricTaylorCSCE590Final

Things to prepare:

a) A .pdf with answers to the questions

b) A GitHub folder with structure:

./doc: pdf file with answers

./data/input: news of cities you select

./data/output: plots you generate

./code: jupyter notebook if you use python

c) Put information about Github in Google sheet:

https://docs.google.com/spreadsheets/d/1vNQo_uG0t7lUxVGPSr1yzxt0MWxGmHjeQOE9Txmr98/edit#gid=0 against your name and send email to me when done.

Situation:

Suppose that you are a journalist for a major newspaper and have been asked to cover Covid situation in a foreign country – India – on the ground in the first week of May 2021. You can choose which subset of cities you visit as long as you visit at least three cities that week. Using techniques you have learnt in this class, you will try to decide how to safely complete your foreign assignment.

The cities you are allowed to choose from are:

• National Capital Region of Delhi (also called Delhi, New Delhi), state: Delhi

• Bengaluru (also called Bangalore), state: Karnataka

• Mumbai (also called Bombay), state: Maharashtra

• Chennai (also called Madras), state: Tamil Nadu

• Hyderabad, state: Telangana and Andhra Pradesh.

Comment: the city was part of unified Andhra Pradesh but the state was split into Telangana and Andhra Pradesh with joint membership until a few more years. For analysis, you can choose either of the two states. Just state which state you are using for Hyderabad

• Lucknow, state: Uttar Pradesh

About terminology: states in India are made up of districts. A city can have one or more districts. For the purpose of our analysis, cities and districts will be interchangeable.

The COVID statistics for the states and raw data are at:

Data sources: https://www.covid19india.org/

Github: https://github.com/covid19india/api

The news for the cities are at:

News

https://news.google.com/search?q=mumbai

https://news.google.com/search?q=delhi

https://news.google.com/search?q=lucknow

https://news.google.com/search?q=hyderabad

https://news.google.com/search?q=bengaluru

https://news.google.com/search?q=chennai

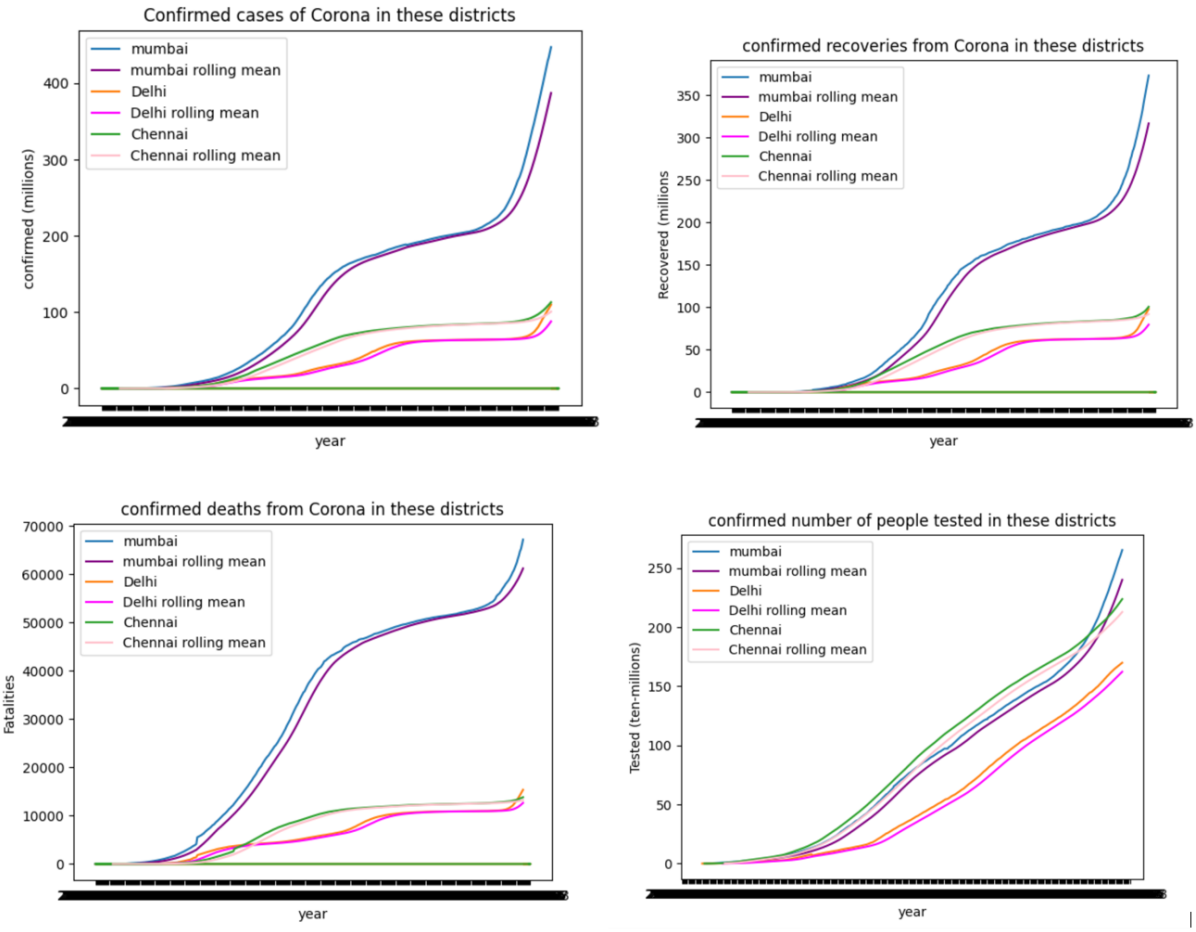
Question 1: Analyze quantitative data from states

[20 + 20 = 40 points]

Consider daily statistics for states in https://api.covid19india.org/csv/latest/states.csv

a) Plot the important COVID statistics for all the states in scope of your decision (i.e., Confirmed, Recovered, Deceased, Tested) and their moving averages

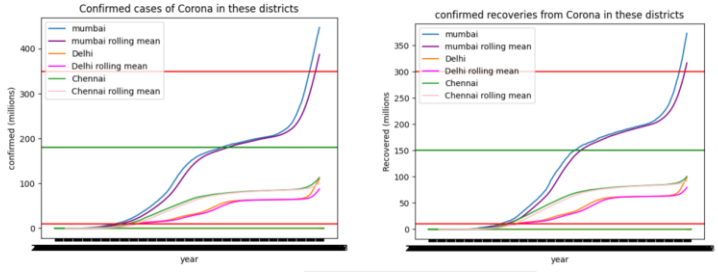
Note: date in data follows the dd-mm-yyyy format.

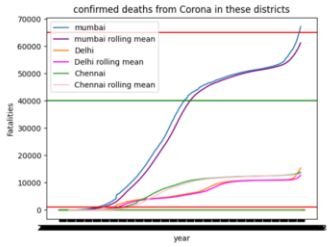


Couldn’t quite edit the year format

b) Define a metric called safety metric to convey how safe that city might be to visit. Now rank the states, and by proxy the cities, on a low to high scale using the metric.

Hint: for safety metric, you can define a weighted sum of columns in data.





Using the safety metrics provided that I've added into the previously created graphs I can say that Delhi and Chennai rank higher within my safety metric. This is due to both states staying below an "acceptable risk" line without crossing it nor the upper limit.

Meanwhile Mumbai would be ranked lower, seeing as how 'currently' it's not only past the "acceptable risk" line it's normally going above the upper limit in all three scenario's only the rolling mean of confirmed deaths stays below this limit.

Question 2: Analyze textual data – news – about the cities

[4 x 10 = 40 points]

Now look at the news for the cities aggregated by Google. Pick a news article related to COVID

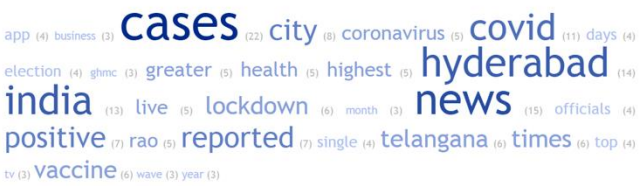
of your choice. (In your result GitHub, include the document in data sub-folder.)

a) Draw a word-tag cloud for each city's news.

Mumbai

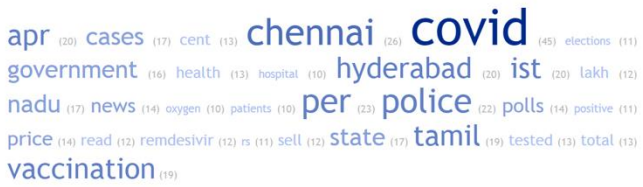
Delhi

Lucknow



Hyderabad

Bengaluru



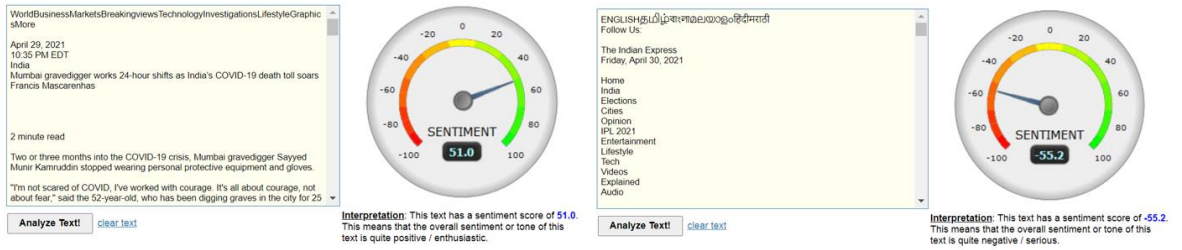
Chennai

b) Perform sentiment analysis for each news

Using this (https://www.danielsoper.com/sentimentanalysis/default.aspx) analyzer I come up with a certain score out of 100 for each states newspage; this helps me determine how serious/negative headlines are.

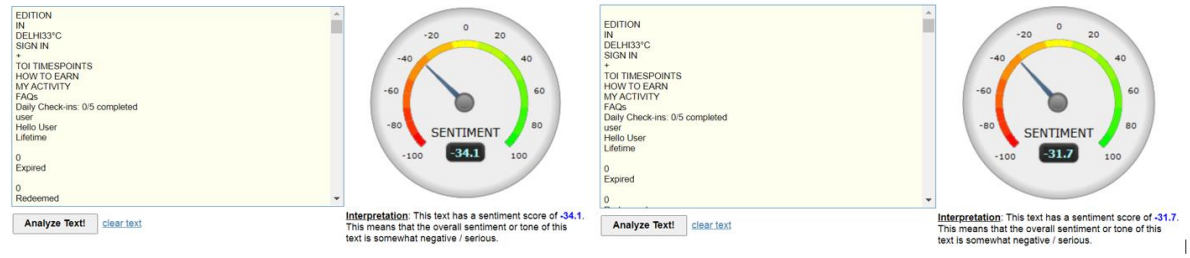
c) Extend your safety metric to include sentiment score.

Using the above I'm looking for sentiment scores close to if not **above 0 points** for each states news article since this analyzer features a negative and positive range.



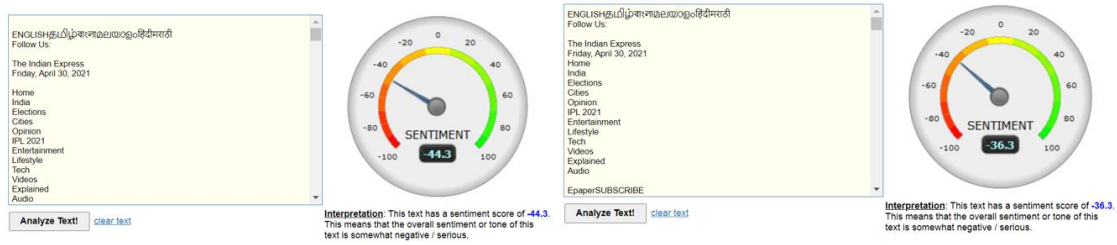
Mumbai

Delhi



Lucknow

Hyderabad



Bengaluru

Chennai

d) Re-rank your 6 cities in the order of safety

For the most part all news headlines were negative with Delhi being an acception, that said in order of safest to least safe I'd rank the cities as follows

Hyderabad

Lucknow

Mumbai - (oddity since this was the only one positive of the entire group; as such I'll place this one in the middle.)

Chennai

Bengaluru

Delhi

Question 3: Answer the business question

[2 x 10 = 20 points]

a) Which three cities will you visit and why?

Of the six options given I feel as though I'd visit Hyderabad, Chennai and Mumbai.

Hyderabad would be my first choice seeing as how the chosen news article was closest to being positive given most of the other states had a lower safety rating.

I'd Choose Chennai seeing as how through the data provided from my graphs it seems to consistently be in a safe spot in regards to my safety metric.

Finally I'd go to Mumbai, sure it may be an outlier in my graphical data, not to mention it's an outlier when it comes to the sentiment analysis; though I feel as though the first issue is due to high population whilst the second problem comes from the use of overly positive wording.

b) What are the assumptions you are making in your selection that may invalidate your

decision?

I'm assuming that people in these states/districts wear masks a majority of the time whilst in public. Another assumption I'm making is that of the data I've analyzed there are no 'odd' data entries that can skew the data. Lastly I'm assuming that