CS 410 Progress Report

1) Which tasks have been completed?

As of now, I have completed the code that will be used to generate the dataset. The code basically scrapes the Google news page. I basically input a Covid related query plus the country and then extract the headline. The final dataset consists of a series of countries and the corresponding headline.

Since I haven't really used the choropleth map feature available in Python, I wanted to implement that with respect to a mock dataset consisting of countries and their values. After generating the mock dataset, I set up code that could be reused to generate a map when I finally finish the sentiment analysis on the actual dataset.

2) Which tasks are pending?

The following tasks are pending:

1) Implement sentiment analysis on the dataset (using the NLTK library)
2) After getting the values from the sentiment analysis for each country, I will display the results on the choropleth map and also showcase the vaccination rates for those countries
3) Lastly, I have to learn how to use dash, so that I can display my choropleth on a web hosted dashboard.

3) Are you facing any challenges?

Yes, the biggest challenge was that I was not able to get my Twitter API use case approved by Twitter. Because of this, I had to shift my focus to implementing web scraping in Google News. I ended up scraping data from the Google News page to generate a list of headlines for each country. This is different from my original plan because initially I wanted to generate a list of tweets for each country. Other than the change in the data source and the type of data, my plan is exactly the same. I will still be applying sentiment analysis to each of the headlines.

Questions from the Project Proposal:

1) Please be more specific about the output. It is not very clear now

After I apply sentiment analysis to the dataset, I expect the output to be "positive" or "negative". For each country, I will average the number of positive and negative reviews to generate one value. This will then lead to a value for each country and that will be mapped to a choropleth map that reflects the values for each country. I will also add the vaccination rates for each country (from a dataset available online), so that we can see the relationship between the two.

2) The method used to generate the data set is not very clear. Are you going to do web crawling? Do you plan to save the data in what format, database or local files?

I was initially going to implement the Twitter API which gives us tweets and then from the output of that API, I was going to generate a list of tweets and countries.

Because my use case for the API was not approved by Twitter, I am now using web crawling techniques on the Google news page to generate a list of countries and their Covid-19 related headlines. The dataset will consist of a headline followed by the country and I will apply sentiment analysis on this. This requires no use of APIs, I am using the web crawling techniques taught to us in class.

3) What is the user input?

After working on the project, I realized that no user input is needed because I am trying to do a sentiment analysis on headlines. In this use case, I already know what I have to input, I want to generate values and create visualizations to display the analysis.