Project Proposal CS 410

1. ***What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.***
   I am choosing to do this project alone. My name is Sukhween Bhullar and my NetID is bhullar5.
2. ***What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?***
   For my free topic I plan on doing sentiment analysis. Specifically, I plan on implementing the Twitter API (https://developer.twitter.com/en/docs/twitter-api) to make specific Covi-19 related tweets, then apply sentiment analysis to those results based on that input.
   I find this topic to be interesting because it applies theory and text analysis to a very serious public health crisis. I want to see how Twitter highlights the sentiment around Covid-19 and vaccinations. Is the coverage positive or negative in different countries? Do Twitter users in different countries have different sentiments around the vaccine? How do those sentiments reflect the vaccinations in those regions? I will either concentrate on the global scale (by country) or stay within the United States (by state)
   My planned approach is as follows:
      Part 1) This will involve exploring the API and the structure of its results, then processing those results to create some kind of a usable dataset that consists of either a list of tweets or a list of articles and their corresponding geographic regions. In the case of a tweet, we will use the geo-tagging feature, but for the google query we will use the location used in the query itself (5 hours). This step alone is pretty time consuming and tedious because it requires a thorough understanding of the different fields that are available in the API and choosing which ones to fully leverage.
      Part 2) This will involve doing the actual sentiment analysis on the datasets aggregated by geographical regions. I plan on doing a geographical analysis of positive and negative sentiments on the vaccine.
      Part 3) This will involve displaying the results, for this I will use the Dash framework.
      Part 4) I will test various queries and make sure the results are displayed properly.
   I plan on using the NLTK library and using many of its functions, such as nltk.corpus.stopwords.words, nltk.sentiment, nltk.FreqDist and many more as I begin working on the project. I will also be using the Python Dash framework to display the results of sentiment analysis, preferable through a map that displays the results of the analysis alongside actual vaccination data for that region.
   The expected outcome would be to see a trend of negative sentiments when discussing vaccinations and a lower vaccination rate in that region. I would also evaluate the actual results of the sentiment analysis on a tweet by tweet basis. In the context of tweets, I will go through the specific tweets and tag them myself at first to create a test set. I will then

split the data into a subset of training data. I plan on doing this with a mixture of regions and then applying that algorithm to the remaining query inputs because we can’t get all possible results.

3. ***Which programming language do you plan to use?***

   I plan on using Python and then using the Dash framework to create a dashboard which will essentially be used as an interface to get user input and display results accordingly.

4. ***Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.***

   This project will definitely take me at the very least 20 hours. I plan on dividing the project into 3 parts.

   Part 1) This will involve exploring the API and the structure of its results, then processing those results to create some kind of a usable dataset that consists of a list of tweets and their corresponding geographic regions. In the case of a tweet, we will use the geo-tagging feature, but for the google query we will use the location used in the query itself (5 hours). This step alone is pretty time consuming and tedious because it requires a thorough understanding of the different fields that are available in the API and choosing which ones to fully leverage.

   Part 2) This will involve doing the actual sentiment analysis on the datasets aggregated by geographical regions. I plan on doing a geographical analysis of positive and negative sentiments on the vaccine or any news around Covid-19. (12 hours)

   Part 3) This will involve displaying the results, for this I will use the Dash framework. (10 hours)

   Part 4) I will test various queries and make sure the results are displayed properly. (2 hours)