David Gutierrez
CS410: Text Information Systems
Final Project Proposal
Fall 2020

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.
    a. Individual team
        i. Name: David Gutierrez
        ii. NetID: davidmg4
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?
    a. Topic: Sentiment analysis based on geography in the United States using Twitter. This tweet-based text analysis will break down each geographic region (50 states) into a sentiment analysis based on sub-topics as government (e.g. corruption), weather (e.g. natural disasters), quality of life (e.g. cost of living), pollution (e.g. air quality), lifestyle (e.g. traffic) to help people make informed decisions about where to live without the potential biases and pitfalls present in survey data. I expect that this will line up with publicly available information, but may present some interesting challenges inherent to social media such as the use of irony and sarcasm. I plan to use Twitter's built-in API and library along with the NLTK library to train the classifier and implement the analysis. Once this data is collected, I plan to use numpy's libraries to perform a statistical analysis to measure the significance of each sentiment and create a threshold value above which the positive or negative value will be considered significant on a per-state basis. Then they will be compared against other states to generate an interactive grid with which user will be able to sort by state and sub-topic as well as apply filters to narrow down to only specific states that they care about
3. Which programming language do you plan to use?
    a. Python

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.
    a. Download Twitter tweet training database/information retrieval via web-scraping: 5 hours
    b. Develop Codebase to parse both geography and sub-topic information: 10 hours
    c. Analysis of data : 5 hours
    d. Presentation in web-based intuitive and interactive data visualization: 5 hours