Project Proposal: Classifying Party of Political Tweets

Melissa Lynn

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.
Melissa Lynn, NetID 655811658. I will be completing the project individually.

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 project, I will scrape political tweets from Twitter, build a dataset of tweets labeled by political party, and train a classification algorithm to classify tweets by political party. This project is interesting because it will involve analyzing the content of tweets by politicians, giving insight into the priorities of politicians from each party and how they communicate.
The project will involve the following steps:

(1) Scrape tweets from the Twitter accounts of politicians. There are various blog articles available online showing how to scrape tweets. Twitter's API, Tweepy, is one possibility, though it looks like that has some limitations.
(2) Clean and label gathered tweet data, to create a dataset where each line is a tweet, and each tweet is a sequence of words without any extra punctuation. This will include tokenization and eliminating stop words, to simplify the dataset and make it more suitable for the next steps. Each tweet will be labeled with the political party of the politician who posted the tweet.
(3) Perform some exploratory data analysis on the dataset, to see if there are interesting patterns in which words are commonly used by politicians from each party.
(4) Train classification algorithms on the dataset, to create a classification model that classifies tweets according to the political party of the poster. I have not yet gone through that part of the course, so I am not sure about the details here yet.

My project will be successful if I am successfully able to construct the dataset of tweets, and if I am able to train a classifier that is somewhat successful at classifying political tweets. I think that I should certainly be able to achieve an accuracy over 50%, however I expect that I won't be able to get close to perfect accuracy, given the nature of tweets. My guess is that somewhere around 60% to 70% would be good performance.

3. Which programming language do you plan to use?
I will use 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.
Here is my anticipated breakdown of time spent on this project:

(1) Scraping tweets: ≈ 4 hours. Since I haven't done Twitter scraping before, and this course was the first time I've done web scraping, I expect that it will take me some time to figure out how to use the API, and gather tweets from various politician's Twitter accounts.
(2) Cleaning and labeling tweets: ≈ 8 hours. I expect that it will take a significant amount of time to clean up the data and labeling the tweets.
(3) Exploratory data analysis: ≈ 3 hours. Since I think that the dataset itself will be interesting, I want to dedicate significant time to exploring the data, and trying to identify any interesting patterns.
(4) Training classification algorithms: ≈ 5 hours. I will experiment with different classification algorithms and parameters, to try to produce the most accurate classifiers that I can. I think it would be particularly interesting to look at decision trees and linear models, where I can easily interpret how the models are determining a classification.