Henry Garcia

Amr Elsisy

John Almaraz

CS179G Project Proposal

For our project, we decided to work with twitter streaming API. We will download ~5GB worth of tweets, and will do some analysis on them, and show our results on a graph/map on a webpage.

For part1, we will create a crawler that uses Twitter4j, which is a twitter API for java, to download tweet information. We will store all the attributes of the tweets, which will come in handy when we are doing our analysis. For example, we can use the coordinates of a tweet, to show on a map the location where the tweet was posted. We will store the downloaded tweets in files of size ~100MB each. The tweets will be stored one tweet per line, and that is to make it easier to parse/create JSON objects, when doing our analysis. The user will also have the option of pausing the crawler, and resuming it at some other time.

For part2, we will analyze the tweets by determining if they are positive or negative. Using hadoop we will have 2 map reducers. The first map will be in charge of making an assumption of how negative a tweet is, and then splitting the tweets based on their negativity score. It will then pass that score, and the text of the tweet, to the reducer. The first reducer will receive tweets with a certain score, and verify that the score given to them is accurate, by running some more analysis tests. The second map will be in charge of splitting the analyzed data based on the state they originated from. The second reducer will then calculate the average negativity score for all the tweets for each state, and will output that score, and the ratio of negative tweets from a given state. Due to the fact that some tweets are not necessarily negative due to phrasing of sentences we will obtain a method in order to correctly analyze some of these cases. For example, a tweet can say, “I hate racist people.” Although, there is some negativity in the tweet in terms of word choice, the overall message of the tweet is positive. After obtaining our scoring for the tweets, we will then find a relation between the density of negative tweets in a given state, and the high school/college graduation rates in that state.

For part 3, the user will be presented with a map of the United States showing the correlation between the ratio of vulgar tweets and graduation rates. Based on how long the previous parts take us, we might allow the user to press on the state they desire, and will then show the user a more detailed analysis of that state. For ex, we can show the user the percentage of graduation for each of the three main races, (whites, blacks, hispanics).