**CS195 Final Project Proposal**

Max Fowler, Naveed Jooma, Mohsan Elahi

April 15, 2014

**Overview**

Our group is interested in analyzing users’ texting data. For the purposes of this assignment, we would like to write a data munger, in the form of a script which runs on your desktop (probably in python) which accesses an iPhone’s SMS database. This data can then be used to conduct various types of analysis described below.

**Sentiment Analysis**

Using the text conversations, we would like to conduct sentiment analysis on the data. This can produce several interesting results. Here are some of the things that we intend to explore:

1. Calculate the sentiment of your conversations per person with different people.
2. Calculate the aggregate sentiment of all your conversations.

The above two ideas can be expanded to include more types of analysis as well. For example, it is possible to calculate a sentiment for each hour of the week. This will allow people to analyze what time and days they are the happiest on. This information will be visualized in the form of a heatmap as was done in Assignment 4.

**Word Frequency Analysis**

We will use word frequency to compare users to each other and to visualise individual’s vocabulary and how it changes over time. By calculating the frequency of user’s word usage in their text message we specifically intend to visualize:

1. What words does an individual use most frequently? What words do they use most frequently compared to another user or to the average frequencies of all users in our data set?

2. How large is a user’s vocabulary compared to another/average?

3. What words has a user’s frequency had the greatest volatility over time?

**Category Analysis**

We will attempt to separate texts into different categories (e.g. sports, academics, activities, etc.). The training data will consist of pre-categorized texts, and then the test data will be predicted. Analyses of these categories could yield some interesting information, like

1. What categories your interpersonal relationships fall in
2. What category you spend most time talking about
3. If categories are time specific (e.g. you talk about sports on Sundays)

Furthermore, you could cluster your contacts based on what categories your conversations with them fall in, and even provide a “friend recommendation service” where people who are in the same cluster could become friends.

**Milestones**

1. Create a script to import SMS data as JSON
2. Create a cloud database so we can all access a single data source
3. Run the analyses described above
4. Create a visual representation of the analyses

**Success**

Success will be defined as determining relationships between people based on texts, as well as finding interesting correlations between sentiments, categories, and vocabulary.