Zachary Shorts

LING 360

Final Project Proposal

1. I want to use the Twitter api in order to figure out what type of content is being tweeted out each day. I specifically want to try to figure out how to perceive the emotions of the tweet based on the text content. Are there more happy tweets or are there more sad tweets? Are there more angry rants or are there more curious tweets? I will also be analyzing when these certain types of tweets are sent. We can hopefully analyze a trend between when certain emotional tweets are sent out. This will help me develop the programming skills that I might need in the future because I tend to analyze large quantities of data. Having a project like this that tackles mixing emotion with data would surely provide a challenge and would help me grow in my skills.
2. I plan to analyze tweets from twitter. I am debating how much data I would like to analyze. The more data, the better the results. But at a certain point it might be too computationally difficult for the computer to process in a reasonable amount of time. I will probably know better once we dive into the twitter api how far back in time I should go. In the get request from the twitter api, there will be the tweet body as well as the time stamp for when that tweet was sent out.
3. My plan to carry out this project is to use a lot of the NLTK python library, a fair share of regular expressions, and a lot of tweets. I will get the data through the twitter api for as many users as is reasonable. There might be limits on how many tweets I can get back at a certain request. If that is the case, then I will need to perform many different requests and piece the data together. In order to ensure and measure the accuracy, I will have to take a small sample and run my program against it. I will find a way to compare the results from personal analysis of the tweets with the results that are returned from my program. This will ensure that the program gets the right emotion out of the tweet. That is the main thing that I am going to need to test. I am also going to be analyzing the times that each type of emotion tweet is sent out. I predict that the later into the night it gets, the more negative the tweets are going to be and the early the tweets are sent out, the more positive the tweets are going to be. As for the curious and angry emotions, I am not sure yet. I will accomplish the programming tasks by generating as many files of tweets in JSON format as I can. I will save these as my own form of database and use it for all of the testing. Once I have parsed all of the body of the tweet and the actual time that it was sent, I need to apply my program to try to categorize the tweet into a certain emotion bucket. These buckets will be compared to each other and the results will be shown. To categorize them, I will need to use certain regular expressions to look for key words and phrases and probably some aspects of the NLTK library.
4. Again, I will use regular expression and python libraries. I will need to use variables, lists, and dictionaries in this project to store and retrieve data from. Since we will be getting the data through HTTP requests, I might need to learn how handle asynchronous tasks in python. In JavaScript I have used promises and I wonder if it might be necessary to use a similar aspect of python for this project.