Tim Powell

CS 410

Final Project Progress Report

1) Which tasks have been completed?

For my project thus far I've completed a majority of the initial planning/setup steps. I've identified the Python Reddit API Wrapper (PRAW). In order to use PRAW, I successfully generated an id and key to use the API from my application. In my application I explored PRAW's different functions in order to see what data is readily available from reddit. I also spent time looking into how this data was formatted and how I'll use it for future project steps. After playing with PRAW, I spent time dissecting the data I was able to retrieve from reddit. I was able to determine which words were occurring most frequently in each post and in the entire subreddit forum (collection of all the posts). Once finished, I researched how to preprocess the data to use the most relevant data for my classification task. I've used Regular Expressions to remove unnecessary content such as URLs from the data retrieved from reddit. In addition to this I've also identified the Natural Language Toolkit (NLTK). It's a suite of libraries that provides natural language processing functions. Thus far I've used this toolkit to further pre-process my data retrieved from reddit. I've been working on using NLTK to return all of the verbs and adjectives from my data. I'll be using all of the verbs and adjectives from the post to determine if it's negative or positive since the verbs and adjectives will best describe the post's nature. After determining how to retrieve and preprocess my data, I began to look into what model I'll use for my classification task. I found the scikit-learn library which provides access to a logistic regression model. I'll use this model to complete the sentiment analysis that my project is based

on. I've also identified a dataset at the following website, https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages, which I'll use as training data for my model.

2) Which tasks are pending?

Tasks still pending include trying to implement the logistic regression model from sci-kit learn. I'll need to train the model with the training data that I found, feed the data that I've retrieved from reddit into the model, and investigate the performance of the model. Afterwards, I'll need to analyze the results. If they aren't sufficient I'll need to backtrack and determine which steps need to be revisited. I'll also attempt to use another type of model from sci-kit learn and compare its performance to that of the logistic regression model. In the end, I'd like to work on displaying my findings in a way that is very easy for our peers to understand.

3) Are you facing any challenges?

The main challenge I'm facing is determining how to find the subject of each post, instead of using the title of the post. I'd like to find the subject so that I can categorize each of the posts and determine which team/player is being spoken of most frequently, and then use my sentiment analysis to determine if they're being spoken of in a positive or negative way.