Project Title: BladeRunner, Sentiment analysis on Twitter tweets.
(Project Proposal for CS 175, Winter 2015)

List of Team Members:
David Chung, 82757831, chungd4@uci.edu
Bo Hyun Chung, 13619644, chungbh@uci.edu
Michael Rodriguez 21653651, merodri2@uci.edu

1. Project Summary

The project's ultimate goal is being able to do a sentiment analysis on tweets and classify them as positive or negative. For this specific problems, there exists many algorithms such as Bernoulli Naive Bayes which is provided in the NLTK framework. Additionally, Logistic Regression classification algorithm will be developed to compare and contrast the performance. For the final phase of the project, a simple python application will be made. The application will consist of keyword inputs such as company names and gather all tweets related to it. The algorithm will then be applied and the results will be plotted to a graph.

2. Problem Description and Background

With 500 million tweets a day and 284 million active users, twitter has become a successful and incredibly popular micro-blogging service. As such, we will be conducting automatic sentiment analysis on tweets, classifying each tweet as either positive or negative—and all with respect terms or words provided by the user. This can serve as a useful tool for both consumers and companies that desire to observe public sentiment on a given product or service. To further this aim, we will compare multiple classifiers and pre-processing techniques to that of the baseline of a unigram model with a Naive-Bayes classifier.

3. Data Sets

Our data set will be labeled and unlabeled English tweets for training and testing respectively. Tweets are notoriously noisy data sets embedded with user names, links, and casual language that is common in microblogging services like Twitter. However, the collected test and training data do not contain any emoticons. We plan to look into this data in a variety of ways, using the baseline of a sentiment keyword dictionary to define a vocabulary. We will also observe the impact of using unigrams, bigrams and trigrams. With such a limited document size(only 140 characters) the way we choose to pre process data may very well impact our findings.

Links to Data Set:

- http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip
- http://www.sananalytics.com/lab/twitter-sentiment/
- http://socialcomputing.asu.edu/datasets/Twitter

4. Proposed Technical Approach

Our approach will be compared to the baselines of the following: Unigram model (Bag of words) with a Bernoulli Naive Bayes classifier. We will use the "bag of words" approach to collect relevant data for our algorithm. After collected, we plan to train our algorithm using different features models such as presence of specific words to analyze the best set of features to accomplish our goal.

Furthermore, we plan to use a logistic regression classifier compare the accuracy with respect to different n-grams. We will further analyze the effects of adding a third class between positive and negative entitled neutral. Neutral tweets are tweets that don't really shift the sentiment about a given topic in any one way. We can then compare how the classifiers differ in a 3 class problem, if there is any difference. We will train and test our algorithm on multiple datasets and normalize the results in order to perform cross-validation.

5. Experiments and Evaluation

Cross validation techniques will be applied to our algorithms for checking overall results. The data set provided by Stanford University is divided to training and test data which will be used to confirm the accuracy of cross validation. Additionally, random tweets can be gathered in real-time for test purposes. http://www.datasift.com is a website that provides such service. Datasift provides a Python library that allows collecting tweets. These data can be used as an additional test data.

6. Software

For the project, Python will be used as the programming language. Additionally, Natural Language Toolkit (NLTK) and scikit-learn will be used to to develop algorithm as well as collect data and pre-process them. For our baseline, the Naive Bayes Classifier provided from the NLTK Framework will be used. We will further use python and helper functions from both NLTK and scikit-learn to develop the multinomial Naive Bayes and Maximum entropy. We will also use other already implemented systems to compare how our implementation measures up. Some of these tools are already online that demonstrate similar purposes.

7. Milestones

Week 5-6:

- Formatting collected data into a single or multiple text file.
    - Collected data format: .csv
    - Output data format: .txt
    - Every line will begin with a digit, the sentiment value
    - From the next word to the end of line, it will contain a tweet from one user
- Import data into Python program
    - Make sure the System reads in tweets and the sentiment score correctly
    - Define helper functions such as gathering random tweets for the given parameter
        - e.g. n_random_tweets(n)
- Begin classifying training and testing data using Naive Bayes Classifier
    - We will begin the construction of our baseline
    - Make sure we have at least 60%+ result on test data or cross validation
    - Keep repeating with different combination of features and pre processing techniques

Week 7-8:

- Implement more helper functions
    - Word Histogram building function
    - This might need to be done on week 6 if time is not enough
- Begin implementation of Logistic Regression
    - Do more research on the technique
    - Optimize weights using log loss function and gradient decent
    - do cross_validation
    - Iterate the process with different set of features
- Compare to the Naive Bayes Classifier
    - Test cases may vary in performance depending on the distribution of the data
- Optimize algorithms to create faster runtimes

    - as large amounts of data will continue to be processed, fast runtimes are important

Week 9-10:

- Research algorithms for further improvements
- Implement methods to test the algorithms for practical applications
    - A query system that could search based on key terms
    - Print the data in a legible manner
- Test implementation for comparison to other types of sentiment analysis
    - Comparing the implementation to things like movie ratings
- Analyze results
    - Create visual representations of results to make differences apparent
    - Document the results in a legible format

8. Individual Student Responsibilities

Michael Rodriguez:

- Responsible for providing at least one dataset of tweets
- Will write helper functions for the second phase
- Brainstorm together to come up with good set of features
- Will implement Part of the logistic regression along with Bo Hyun

David Chung:

- Responsible for providing at least one dataset of tweets
- Will write the helper functions required for the first phase.
- Test the naive Bayes Algorithm from the NLTK framework
- Brainstorm together to come up with good set of features
- Help part of the Logistic Regression algorithm and the final application of the project

Bo Hyun Chung:

- Responsible for providing at least one dataset of tweets
- Will implement Logistic Regression along with Michael
- Brainstorm together to come up with good set of features
- Will do research on optimizing algorithm