Pranav Shah

Data Mining CS 6140

**Intermediate Report**

1. What progress you have made towards your proposed goal?

I used two approaches to estimate the polarity of a tweet text which can be later used to draw more conclusions from the tweets.

1. If you tried some basic approaches: what worked well and what did not?

I used regular expressions to preprocess the data and removed all the stop words, punctuation marks and the words which have length less than 3.

The two approaches which I used are as follows:

* 1. Using nltk to calculate polarity: In this approach we used NLTK library to calculate polarity of each tweet text and used matplotlib to visualize it. This seemed to be too straight forward approach, so we tried the method b.
  2. Using a corpus of positive and negative words: After pre-processing the data, all the words in a text are stemmed using Porter Stemmer. Later, these words are checked if they belong to the positive wordlist or negative wordlist and then the polarity of each tweet text is calculated.

Approach “a” worked well than approach “b”.

1. What could be done to improve the basic approaches?

Approach “b” did not work well because the size of wordlist used to calculate polarity was very small. We are still trying to find better wordlist/corpus of positive and negative words to improve the working of approach “b.” Or else, we will stick to approach “a” and use data mining methods over the results of approach “a” to draw better conclusions.

1. What experiments have you run and are you planning to run to demonstrate the effectiveness?

We have run the sentiment analysis using two approaches, that is with and without using library. We will try to club this sentiment analysis with other Data clustering approaches to draw better results.