Xi Wang, xwang76@jhu.edu
Yining Wang, wangyining1988@gmail.com

Final Project
Sentiment Analysis of Twitter Posts About Group Emotion Change Pattern
600.466 Spring 2011

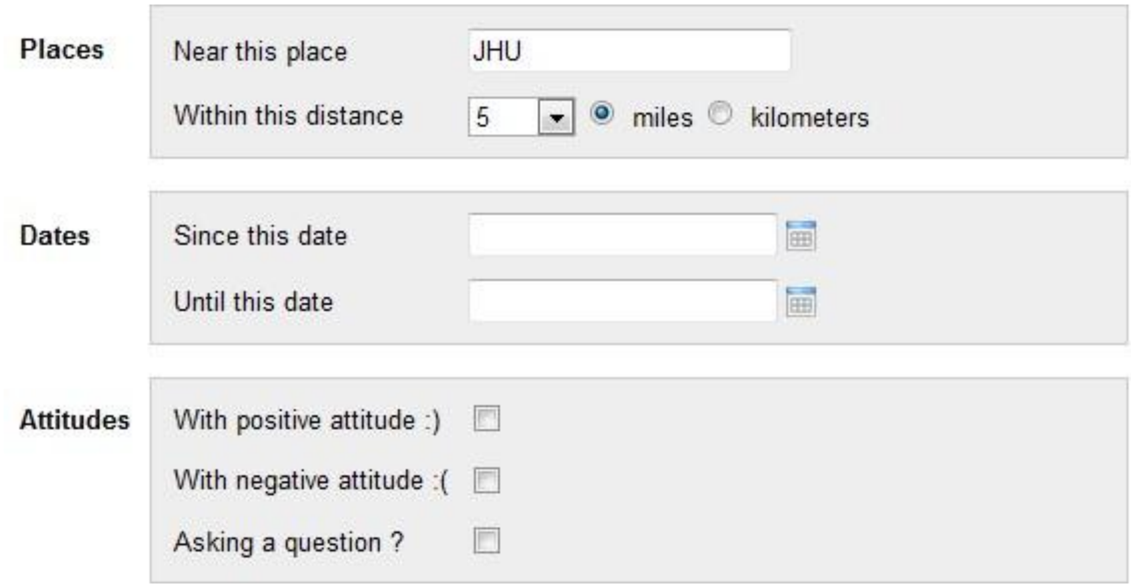
Overview:

Wikipedia defines sentiment analysis as the process that “aims to determine the attitude of a speaker or a writer with respect to some topic.” Automated sentiment analysis is the process of training a computer to identify sentiment within content through Natural Language Processing. Various sentiment measurement platforms employ different techniques and statistical methodologies to evaluate sentiment across the web.

One of the most popular microbloging platforms is Twitter. Twitter has become a melting pot for all - ordinary individuals, celebrities, politicians, companies, activists, etc and especially for college students. When people post on Twitter, reply to other’s posts, or retweet news posts, it is possible that they can express their sentiment along with what they are posting, retweeting or replying to. The interest of this thesis is in how Hopkins student’s emotion changes in the first day before the final exam week and to determine how latest news and incidence affect the emotion change.

Data Collection:

We encounter difficulty in adapting the existing web robot of class directory on hops to parse the Twitter website to gain data. Thus we use the website http://www.search.twitter.com instead. The website has an interface with advanced search filed like below:



Places

Near this place: JHU

Within this distance: 5 ▾ ● miles ○ kilometers

Dates

Since this date:

Until this date:

Attitudes

With positive attitude :) □

With negative attitude :(□

Asking a question ? □

Testing Data

- We use make a HTTP::Request that supports sending information as a POST to the server of search.twitter.com to get back tweets and label them with number presenting the time those tweets are published.

- We only take tweets written from a geographic scale within a radius of five miles of JHU and eventually got 500 tweets during 24 hours at the frequency of retrieving 50 tweets/ 6 minutes.
- The reason we use the discrete data is to avoid the repetition of posting same tweets in a short interval.
- The dataset has the format of XXXX\t | "time" where XXXX indicates the content of the tweets and the time figure indicates the time we retrieve the tweets.
- We trim the data of symbols such as @ (usually follows with a username), RT(represents "retweet") , http address and punctuations.
- Each tweet is equivalent to a "document". The starts of header (e.g T."time") in the vector also separate different tweet from each other. For example,

```
T.16
Waiting
to
meet
the
Deftones
T.16
It's
so
true
……
```

- We segment the text using the replace regular expression function embedded in the software Editplus.

Training Data

- We find more than 10 thousand tweets separately of both positive and negative attitude using the above website as our training database.
- The dataset has the format of XXXX\t |NEG(POS) where XXXX indicates the content of the tweets and the "NEG/POS" indicates the attitude of the tweets.

Sentiment Word Lexicons:

The main task is the construction of sentiment discriminatory-word lexicons that indicate a particular class such as positive class or negative class. The polarity of the words in the lexicon is determined prior to the sentiment analysis work. In our case, we use the publicly available discriminatory word lexicons for use in sentiment analysis from http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

- The appearance of an opinion word in a sentence does not necessarily mean that the sentence expresses a positive or negative opinion.
- There're many misspelled words in the list. They are not mistakes. They are included as these misspelled words appear frequently in social media content.

Design:

Step 1: We compare the words of training data against the sentiment word list and get the raw term frequency of each term of the word list in the corpus of tweets. Because people can only twitter 140 words at maximum, thus the term frequency in one sentence is usually not more than 1.

Step 2: We use the raw term frequency as the weight for each term in the word list. For terms appeared in the positive tweets training set we assign them with positive scores while for terms appeared in the negative tweets training set we assign them negative scores.

Step 3: We then compare the words of testing data against the same sentiment word list and calculate each tweet's lexicon score by the weight we got in the previous step.

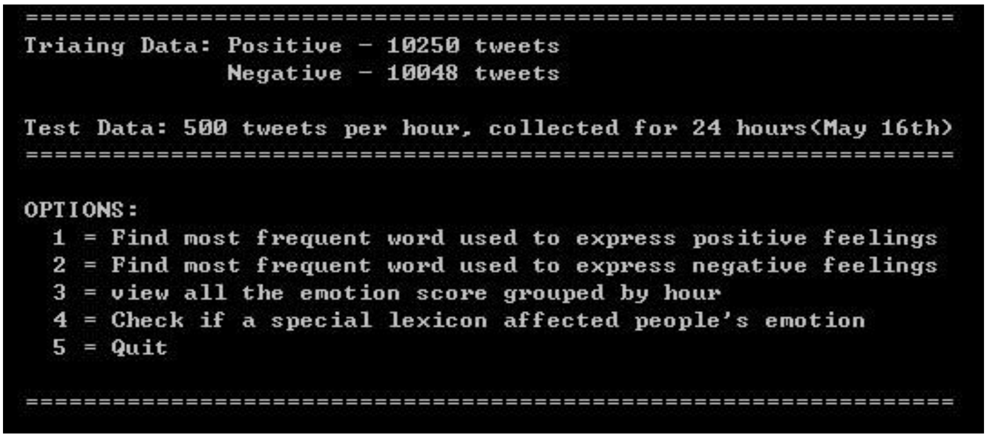
Step 4: We add up all the tweets' lexicon scores of the same time period and divide it by the total document number to normalize the score. We use the normalized score to indicate the overall emotional level at that time. We compare the score with zero, if the score is below zero then we define the attitude as negative, if it's around zero then we define it as neutral, and otherwise it is positive.

Step 5: We then label each tweet sentence as positive/neutral/negative to see how many people are classified into each category.

Step 6: User can enter the latest events/news in the query. The program will parse through all the testing tweet sentences and compare the news words lexicons with the labeled tweets to figure out how people react with different events and news.

Results:

- User Interface:

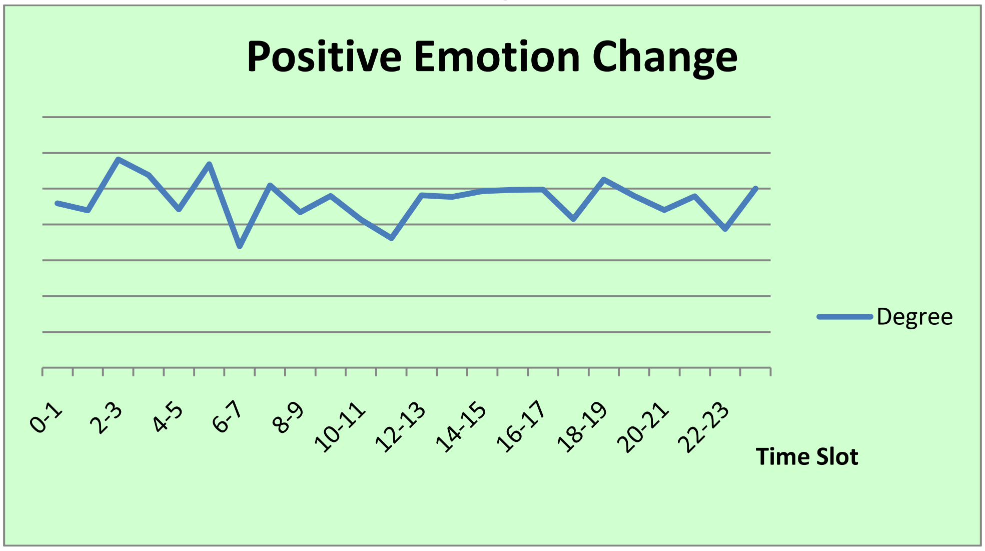


```
==============================================================
Triaing Data: Positive - 10250 tweets
              Negative - 10048 tweets

Test Data: 500 tweets per hour, collected for 24 hours(May 16th)
==============================================================

OPTIONS:
  1 = Find most frequent word used to express positive feelings
  2 = Find most frequent word used to express negative feelings
  3 = view all the emotion score grouped by hour
  4 = Check if a special lexicon affected people's emotion
  5 = Quit

==============================================================
```

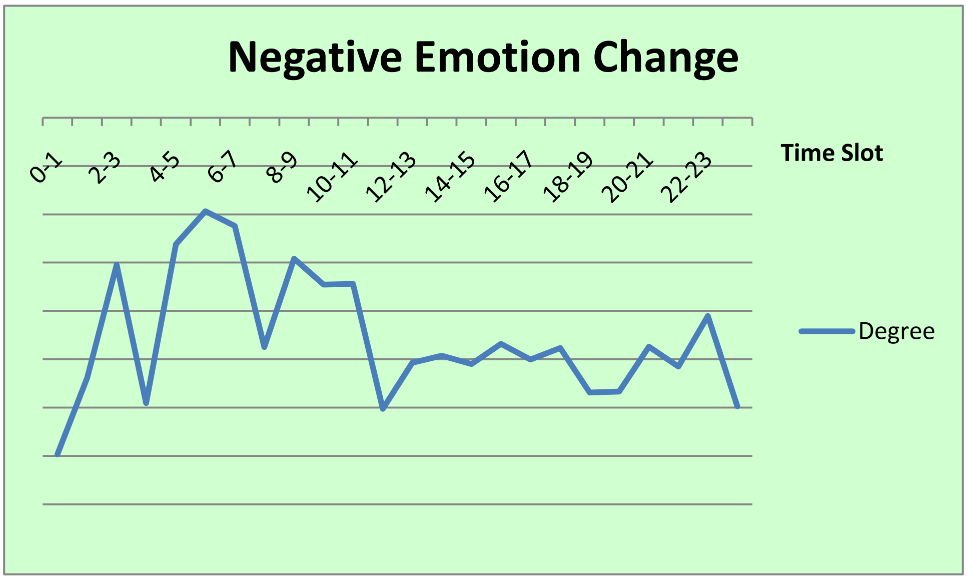
- We output the most frequent word used to express positive and negative feelings together with the raw frequency.

Top 20 Positive Word	Top 20 Negative Word
love	too
just	miss
like	need
good	sad
will	sorry
back	bad
please	hate
want	sick
well	down
happy	long
hope	mean
right	missed
great	little
thank	hurts
best	sucks
better	ill
nice	tired
awesome	shit
work	trying
yes	damn

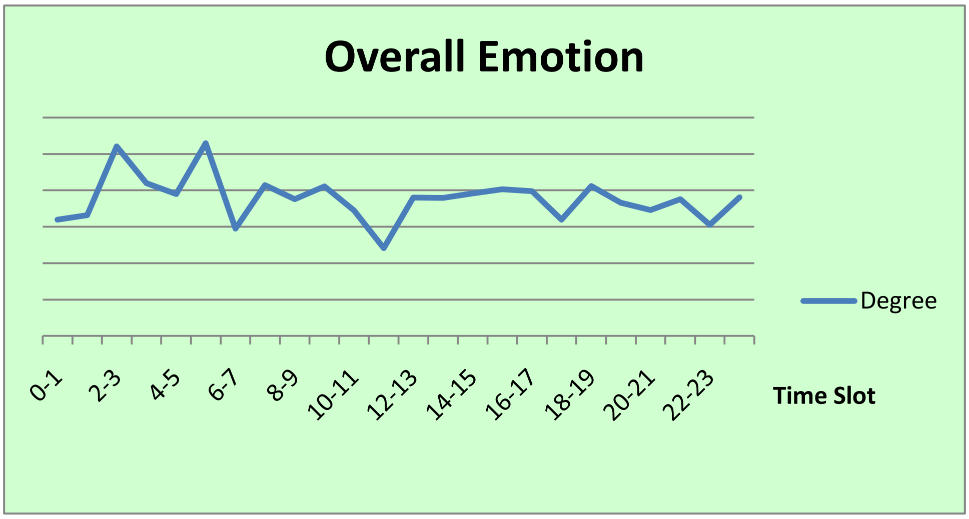
- **People's overall positive level and negative level**



Student's positive emotional level are quite stable on May.16th, the first day of the final exam week, with the emotion level fall down to the lowest point at 6-7 am in the morning the time they have to get up and start working.

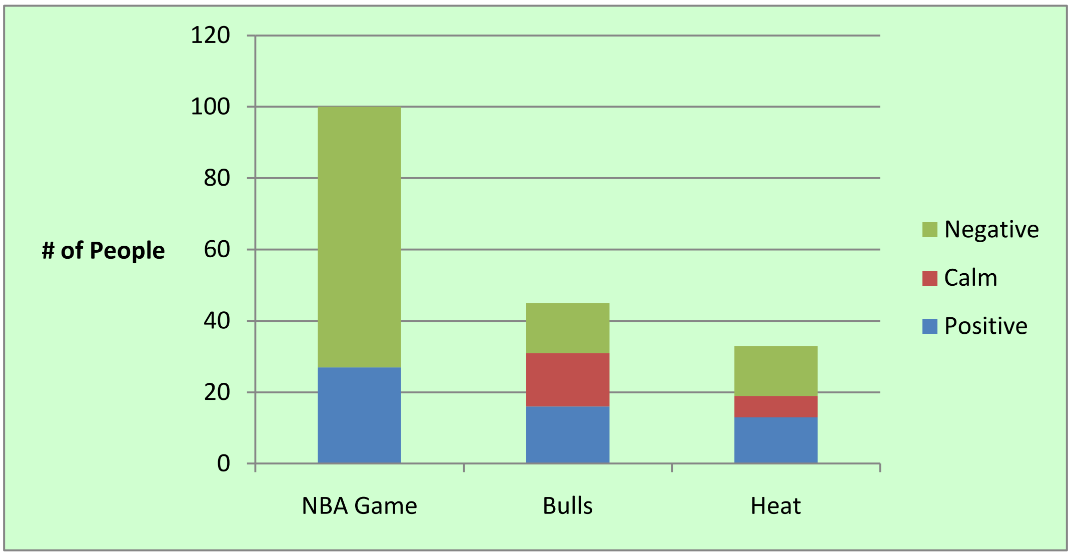


Student's negative emotional level has more ups and downs during the day, especially the trend can be observed from the graph that students are getting more and more negative as time goes on.

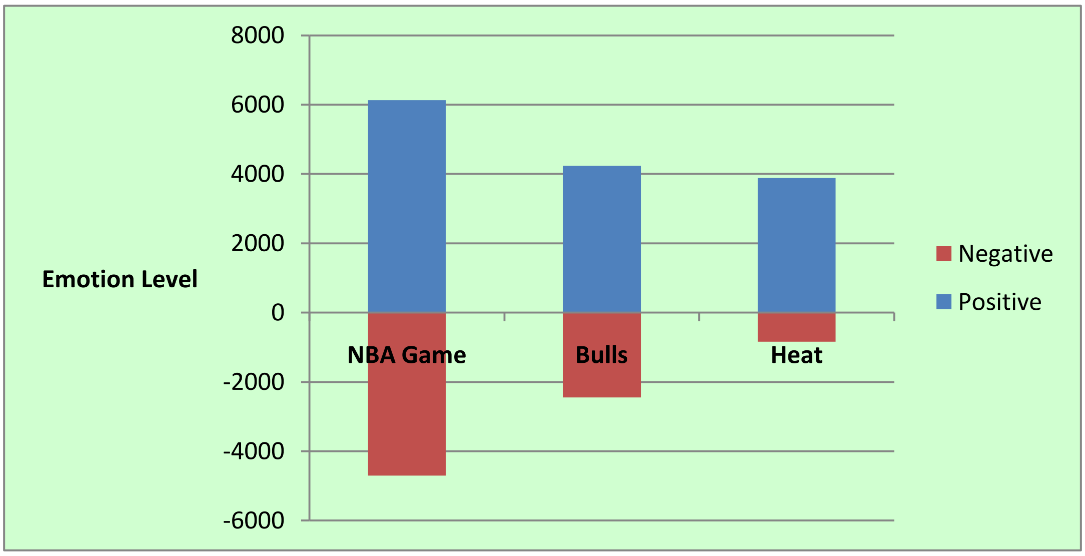


However the student's overall emotion level is quite stable and positive during the whole day.

- **People's attitude towards certain events or latest news.** Here we take the NBA game hold on Sunday as an example. On Sunday May.15, the Bulls manhandled the Miami Heat en route to a 103-82 win to take a 1-0 series lead on the road to the NBA Finals. By typing in query of NBA game, we find 27 people express their positive feeling while 73 people express their feeling of disappointed. It seems that people in overall are not satisfactory with the result that Bulls won over Heat. When typing in query of Bulls, we find 16 people are supportive of the team, while 14 people have negative attitude. The number is 13 positive and 14 negative for the query of Heat. 15 people feel neutral for Bulls, while 6 people feel neutral for Heat.

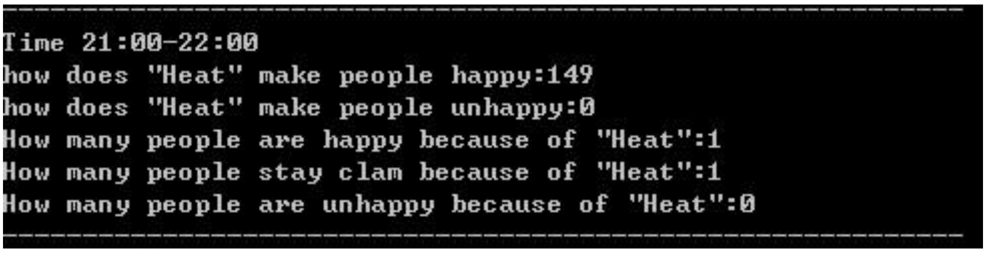


Although the numbers of people who feel negative and people who feel positive are nearly the same for the team Heat, we further analysis their emotional level, we find that the positive emotional level for Heat is way higher than negative level. This can be explained that people who support Heat support the team with the largest effort to put the best word they have to encourage the team. While those who express their negative attitude toward the team are mainly people who feel disappointed about the result thus they only use slightest "bad" word to describe their feeling.

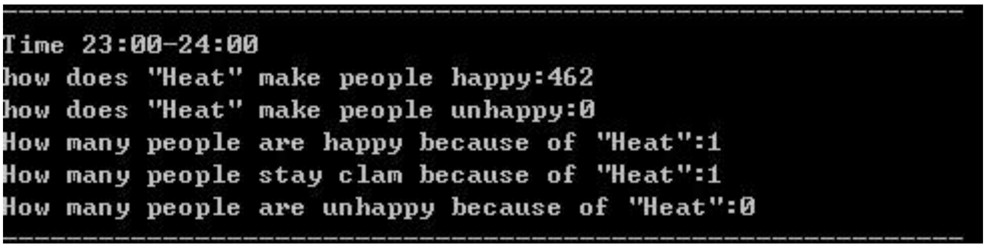


- **Different tweet pattern for different people.**

The result also shows another interesting phenomenon that people usually use the best words they can find to express their positive feeling.



```
------------------------------------------------------------
Time 21:00-22:00
how does "Heat" make people happy:149
how does "Heat" make people unhappy:0
How many people are happy because of "Heat":1
How many people stay clam because of "Heat":1
How many people are unhappy because of "Heat":0
------------------------------------------------------------
```

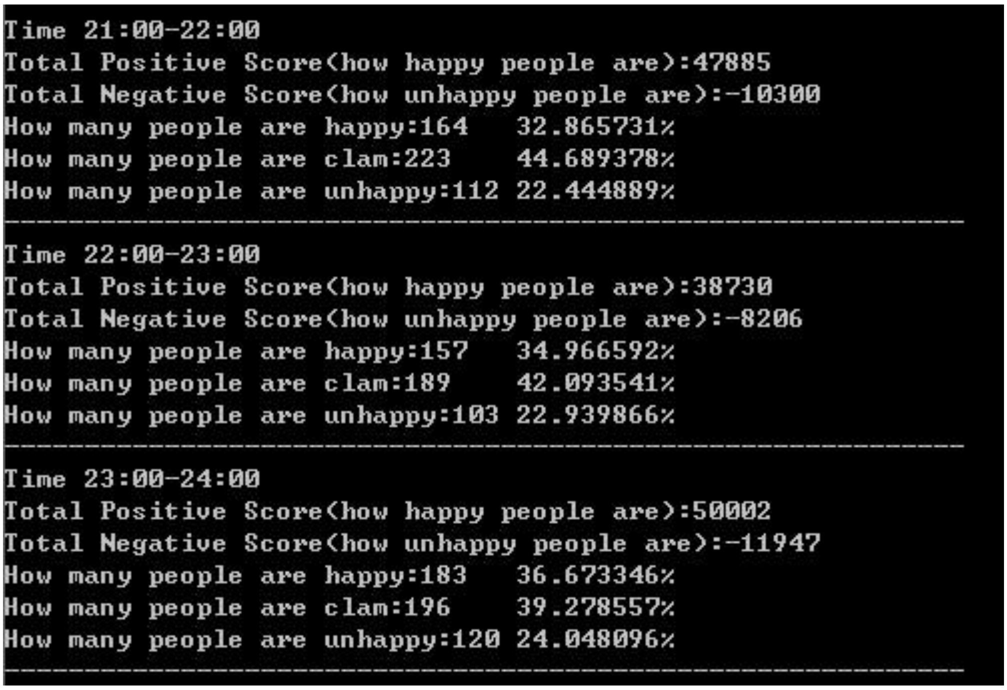


```
------------------------------------------------------------
Time 23:00-24:00
how does "Heat" make people happy:462
how does "Heat" make people unhappy:0
How many people are happy because of "Heat":1
How many people stay clam because of "Heat":1
How many people are unhappy because of "Heat":0
------------------------------------------------------------
```

In the data we collect above we find that although the number of people who express their positive feeling and negative feeling are the same, the emotional level of the two are quite different which shows that some people use more “strong” words, the words that ranked high in the positive word list, to express their feeling and attitude.

Discussion and further application:

1. **Real-time Feedback.** As currently we retrieve our data on May.16 the day after the game ends, thus we can just analysis student’s attitude toward the game result on from an overall aspect. If we retrieve the data during the time the NBA game plays, then we can do real time analysis on how people’s emotion change along the change of the game score. This function can be further applied to TV series and shows census to see people’s reaction on a specific person or performance.



```
Time 21:00-22:00
Total Positive Score(how happy people are):47885
Total Negative Score(how unhappy people are):-10300
How many people are happy:164   32.865731%
How many people are clam:223    44.689378%
How many people are unhappy:112 22.444889%
------------------------------------------------------------
Time 22:00-23:00
Total Positive Score(how happy people are):38730
Total Negative Score(how unhappy people are):-8206
How many people are happy:157   34.966592%
How many people are clam:189    42.093541%
How many people are unhappy:103 22.939866%
------------------------------------------------------------
Time 23:00-24:00
Total Positive Score(how happy people are):50002
Total Negative Score(how unhappy people are):-11947
How many people are happy:183   36.673346%
How many people are clam:196    39.278557%
How many people are unhappy:120 24.048096%
------------------------------------------------------------
```

Program List

analysis.pl	The main program to analysis tweets
robot.pl	The robot to get testing data of tweets from twitter
TrainingData.pl	This program get tweets from search.twitter.com and with labels indicating their attitudes (positive/negative) we use those tweets as our original training data
TrainNeg.txt	Training data labeled negative
TrainPos.txt	Training data labeled positive
TwiData.txt	500 testing data retrieved from twitter on May.15
TrimedTwiData	Trimmed testing data
negative-words.txt	Sentiment Word Lexicons(Negative)
positive-words.txt	Sentiment Word Lexicons(Positive)
FrequentPositive.txt	Output: The most frequent word people use for expressing positive feeling
FrequentNegative.txt	Output: The most frequent word people use for expressing negative feeling

References:

1. Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." ,Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA,

2. Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

3. Bing Liu. "Sentiment Analysis and Subjectivity." An chapter in ; Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.