# Part 1: Get Twitter Data

Completed in week 2

# Part 2: Sentiment analysis of the tweets

1. Examine the file AFINN-111.txt and the file AFINN-README.txt. You will see pairs of words and numeric values in AFINN-111.txt. The values equate with sentiments for a set of words – “5” is strongly positive and “-5” is strongly negative.
2. The python file “tweensentiment.py” will read the sentiment file into a python dictionary. It then read the tweets you saved in the file “output.txt (it parses the tweets) and then compares the words in the tweet against the sentiments. Next it increments a counter for each sentiment value and then prints the sentiment value count.
3. Run the command:

*python tweetsentiment.py AFINN-111.txt output.txt > newoutput.txt*

Look in your file and you should see the *newoutput.*txt file. The output that I got was:

-5 sentiments 2

-4 sentiments 24

-3 sentiments 53

-2 sentiments 91

-1 sentiments 112

0 sentiments 0

1 sentiments 101

2 sentiments 134

3 sentiments 116

4 sentiments 31

5 sentiments 0

1. Consider the following questions
   1. What type of data are the sentiment values i.e., the values {-5, -4, …, 5}?
   2. Compute a mean value for the sentiments.



where is the number of occurrences of the th sentiment.

* 1. What, if anything, does this mean value mean?
  2. In calculating the mean, what additional assumptions about the type of the sentiment data have been made?
  3. Suppose that in AFINN-111.txt that there were different numbers of words that have been rated -5, -4, -3, …, 4, 5. What effect would this have on your analysis? What might be a way do compensate for this problem?