Section 4 – Sentiment Analysis

There are a few new things in this script, but it's not much more complicated than the last set.

Ignoring the print statements, this script does the following:

First we load a twitter account, as before. We are then using a variable to decide which number tweet to select from our list.

Next we split the tweet into individual words - rather than one long list of characters with spaces in between, we now have a list of single words. You can see the original tweet and the split version in the console output.

We then load the two large word lists in negative-words.txt and positive-words.txt into the lists negativeWordList and positiveWordList.

This time our for loop is going through each word in the tweet, rather than a list of tweets.

We then have an if statement on line 24 - the following two lines (25 and 26 - notice how they are tabbed in further than the rest of the loop) are only executed when the if statement is true.

Python is quite nice in that a lot of python statements describe exactly what they are doing - this is true for our if statement.

Lines 25 and 26 are executed IF the word that our loop is looking at is IN the big list of positive words.

Coding Challenge: Edit this script so that it also calculates the number of negative words in the tweet in question. The simplest way to do this is with a second copy of all the code between lines 20 and 30 (copy and paste is your friend!) with some minor edits. This is simple if you understand what the code is doing – if you're not sure, ask us for some help and we can go over it.

Questions to Investigate:

- Calculate the positive and negative word totals for a number of tweets from both accounts. Do they tend to use positive and negative language differently?
- Who tends to be more positive? (This and the next question might be easier to answer with the solution to the Extension Challenge below, which gives you the mean positive and negative words per tweet)
- Add a filter and investigate positive and negative word use for tweets on a particular theme. E.g. if a tweet from Obama contains the word 'republican' does it tend to be positive or negative?
- Look at the words that the script picks out as being positive or negative, do you agree with those choices?
- Why is doing this kind of analysis on human language hard?

Extension Challenge: Write a script to calculate the mean positive and negative words per tweet for an account (you could also calculate this for a single score e.g. number of positive words minus number of negative words).