Text Mining for the Social Sciences
Assignment 2

Please work in same groups as formed for first problem set. Code for Q1 and a write-up of Q2 due on Thursday 30 April.

1. Extend the class introduced in the iPython notebook for pre-processing to include two additional methods:

   (a) *count*(self, dictionary), which counts the number of occurrences of terms in "dictionary" in each document.

   (b) *tf_idf*(self, dictionary), which applies tf-idf weighting as seen in lecture to index documents according to terms in "dictionary".

2. Pick a dictionary (or dictionaries) of your choice from the Harvard IV set, the Loughran-McDonald set, or some other of your choosing that you think may be relevant for the data you collected for Q4 of the previous problem set. Then conduct the following exercise:

   (a) Use the two methods above to score each document in your data.

   (b) Explore whether the scores differ according to the meta data fields you gathered: for example, do different speakers/sources/etc tend to receive a higher score than others?

   (c) Do the answers to the previous question depend on whether tf-idf weighting is applied or not? Why do you think there is (or is not) a difference in your answers?