Text Analytics: Practical 4
(for Lecture 4: Beyond Frequencies)

1) Find 10 short text-items (20-30 words); they could be emails, short docs, tweets or whatever… Make sure they all deal with some common topic of interest; so they have some of the same words
   a. Remove the standard stopwords from them using some standard list, use nltk.
   b. Compute the TF scores for all the remaining words in the texts and use R to show the word-cloud for these words. In your answer provide the matix of TF scores and the word-cloud image.
   c. Now, compute the TF-IDF scores for all the same words in the texts. Construct a set of words that represents the TF-IDF scores you have found, for all the words. Use R to show a word-cloud for these words. Also, provide the matrix of TF-IDF scores and the word-cloud image.

2) Using Python or R, compute the PMI scores for all adjacent pairs of words in your 10-doc corpus (ie the texts after stop-word removal).

   List the top-10 pairs based on the PMI scores found for the pairs.

   Do the results make sense? If not, then introduce a minimal cut-off frequency and re-compute the top-10 until they seem sensible.

3) Entropy has been used to determine whether tweet set is interesting (contains variety) or repetitive (spam).

   Create two sets of 10 made-up tweets:

   a. **spam-set**: where the 10 tweets are very similar containing an advert for a product
   b. **random-set**: where the 10 tweets are very different, chosen at random from Twitter.

   Now, find a Python/R program or package that computes entropy and find the entropy values for (i) spam-set, (ii) random-set, (iii) the two sets combined.

   Report the program you used and its source, the tweet data and the entropy values found.