Beyond Frequency

Practical 4: Simple Frequencies

TF-IDF
Q1

Prac1 Q1(a) select texts

- Find 10 short text-items (20-30 words); 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
- Remove the standard stopwords from them using some standard list using **nltk**

Prac1 Q1(b): Compute TF

- Use R show the word-cloud for the words; NB this will be based on the TF of the words
- In your answer, also provide the matrix of TF scores and the wordcloud image

Prac1 Q1(c): Compute TF-IDF

- 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

PMI
Q2

Prac4 Q2 Compute PMI

- 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, retaining the original order)
- List the top-10 pairs based on the PMI scores found for 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.

Entropy
Q1

Prac4 Q3 Entropy

- Entropy has been used to determine whether tweet set is interesting (contains variety) or repetitive (spam)
- Create two sets of 10 made-up tweets:
  - **spam-set**: where the 10 tweets are very similar containing an advert for a product
  - **random-set**: where the 10 tweets are very different, chosen at random from Twitter
- Now, find a python program or package that computes entropy and find the entropy values for (i) spam-set, (ii) random-set, (iii) the two sets combined