Text Analytics: Practical 2
(for Lecture 2: De Data)

1) Try to get used to using then **nltk** package. Create a text file with, 50 words in it (written as a collection of sentences) and make sure it has items that challenge the tokeniser (e.g., I.B.M. and other weird things). But, you need to make these up yourself, so you think about what it is doing.

    a. Load the file in and use nltk.word_tokenizer() on it. Report the list of tokens that are produced from it and note any oddities that arise. Comment on these oddities and how they might be handled.
    b. Now, take this output from the tokenizer and do the normalization step
    c. Now, take the output from normalization step and run it through a pos-tagger. Report this output as your answer and highlight any inaccuracies that occur at this stage.

2) Now, take a second, new-text-file do the following:

    a. Tokenize the new-text-file (50 words) and the stem it using Porter Stemming. Report your outputs and some of weird things that Porter Stemming does.
    b. Tokenize this the new-text-file and then lemmatize it using WordNet Lemmatizer; note you may have to pos-tag the sentences first and then convert the tags to make this work. Report the result of these steps and point out some of the things that look wrong.
    c. Compare the outputs from Porter Stemming and the Lemmatisation of the same file. Which do you think is the best to use and why?