Text Analytics: Practical 3
(for Lecture 3: Simple Frequencies)

1) Assuming you have R installed (if not install it). Load up the various packages you need for using the wordcloud packages:

   a. Carry out the commands shown in the practical notes:

```
> library(wordcloud)
> library(tm)
> wordcloud("May our children and our children's
children to a thousand generations, continue to enjoy
the benefits conferred upon us by a united country,
and have cause yet to rejoice under those glorious
institutions bequeathed us by Washington and his
compeers.",
colors=brewer.pal(6,"Dark2"),random.order=FALSE)
```

   b. When you have done this, report the list of the words from the original quote that are included in the wordcloud and the list of those that are not. Report why do you think some are excluded and others included?

   c. Now, check your theory about what the wordcloud package included and excluded. Put in your own word-list together (30-50 words) and check what wordcloud includes and excludes? Report whether your initial theory was right or wrong and why?

   d. Again, using your word-list add more repeated words and see what happens? Can you change the package's to make it more inclusive of the words in the word-list?

2) Find the Google Ngram Viewer online and do the following with it:

   a. Put in "Mark Keane" as a search term and explain the peaks that appear in the graph over time.

   b. Put your own name in and describe what happens, explaining where the hits are coming from.

   c. Pick a word that you think is a recent introduction into the English language (like "exit strategy") and plot its emergence, showing the graphs. If it actually emerges before you thought, explain why?

   d. Describe some of the effects of smoothening these graphs with different values?

   e. Do a comparison between 3 or more related terms to see how their relative frequencies have changed over time*. Is there anything surprising about how these terms differ in their frequency and, if so, why?

Why do you think the frequencies vary in the way they do.

f. Use the syntactic tags in a search for two words that are the same but syntactically different (e.g., fish-verb, fish-noun; *do not use fish*) and report what you find.

g. Think of some major cultural change that has happened over the last 500 years and some words that could denote to ths event/events. Check these words of the relevant timeperiod. Report what you find.

3) Using an Excel spreadsheet set up your own list of 15 words and give each a made-up frequency between 0 and 2000 for each of three years (2010, 2011, 2012). Now perform two different normalisations on them:

   a. **Method1**: produce a normalised frequency for each word in each year, using the total *N* of words over all the years (i.e., Grand Total)
   b. **Method2:** produce a normalised frequency for each word in each year, using the total *n* of words in a given year
   c. Does normalising by **method1** or **method2** make a big difference to the scores produced? Graph the difference and comment on it.

4) Find the article by Choi & Varian (2009/2011/2012) and find the R program they give for their Ford prediction model. What do you need to do to run this program? Can you do this?