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 wordcloud (e.g., wordcloud, tm; as you try to use them they will tell you about missing packages):

   a. Carry out the commands shown in the practical notes (n.b., type these in by hand, do not cut and paste as errors can arise:

```
> library(wordcloud)
> library(tm)
> wordcloud("May our children and our children's
children to a thousand generations, continue to enjoy
the benefits conferred on us by 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 hypothesis about why the wordcloud package is including some words and excluding others. Put your own word-list together (of 10-20 words, try to repeat some words a few times) and check what wordcloud includes and excludes? Report whether your initial hypothesis was right or wrong and why?

   d. Again, using your own word-list add more repeated words (i.e., use the same word multiple times) and see what happens?

      [You should copy the images of these wordclouds in your report]

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

   a. Put in "Mark Keane" as a search term and look at the graph shown. You can trace the books that refer to this name. Can you 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. If there are none, then change your name until it starts to produce hits for it.

   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 (e.g., winter, summer, autumn, spring; *do not use these ones*). Is there anything surprising about how these terms differ in their frequency and, if so, why? Why do you think the frequencies vary in these patterns?

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. You will need to research how to do this in the Ngram viewer.

g. Think of some major cultural change that has happened over the last 500 years and some words that could denote this event(s). Check these words for the relevant time-period. Report what you find.

3) Open an Excel spreadsheet set up a table with your own list of 10 words which should be the rows in the spreadsheet and give each a made-up frequency between 0 and 2000 for each of five years (2010, 2011, 2012, 2013, 2014; which should be the columns). Now compute two things: (i) find the large-*N*, a count of all words across all years (i.e., the sum of all words in the set) (ii) find the small-*n*, for all the words in each year (e.g., sum of all words in each year, for each year; i.e., column totals).

   Now do two different normalizations on each word:

   a. **Overall Normalization**: produce an overall normalized frequency for each word in each year in the table, using the large-*N*
   b. **By-year Normalization:** produce a normalized frequency for each word within a year, using the small-*n* for each word
   c. Does normalizing these different ways make a big difference to the scores produced? Graph the differences you find in a histogram and comment on it.