Practical 3

Text Analytics: Simple Frequencies

Otto Hermann
otto.hermann@ucdconnect.ie
16203034

COMP47600

University College Dublin
5th October 2017

1)a.

Command used as requested:

```
wordcloud("May our ... his compeers.", colors=brewer.pal(6,"Dark2"),random.order=FALSE)
```

1)b.

Excluded list

```
['a', 'and', 'by', 'have', 'his', 'our', 'the', 'those', 'to', 'under', 'us']
```

Included list

```
['benefits', 'bequeathed', 'cause', 'children', 'childrens', 'compeers', 'conferred',
    'continue', 'country', 'enjoy', 'generations', 'glorious', 'institutions', 'may',
    'rejoice', 'thousand', 'united', 'upon', 'washington', 'yet']
```

It appears that the excluded words are stop words, which is confirmed by the R wordcloud package: our input is a character vector and, given that we did not specify the $freq$ parameter in our command, standard stop words were removed prior to plotting the word cloud.

1)c.

New text

This is my text. It contains about thirty words, of which I expect roughly half to be excluded as stop words. If I modified the input of the R command, I could change the results.

Excluded list

```
['about', 'as', 'be', 'could', 'i', 'if', 'is', 'it', 'my', 'of', 'r', 'the', 'to',
    'which']
```

Included list

```
['change', 'command', 'contains', 'excluded', 'expect', 'half', 'input', 'modified',
    'results', 'roughly', 'stop', 'text', 'thirty', 'this', 'words']
```

These results are consistent with my original theory and the documentation for the wordcloud package.

1)d.

New text

This is my text. It contains about fourty words, of which I expect roughly half to be excluded as stop words. If I modified the input of the R command, I could change the results. I will repeat two words: donkey, donkey, donkey and is is is.

Excluded list

```
['about', 'and', 'as', 'be', 'change', 'command', 'contains', 'could', 'excluded',
    'expect', 'fourty', 'half', 'i', 'if', 'input', 'is', 'it', 'modified', 'my',
```

```
'of', 'r', 'repeat', 'results', 'roughly', 'stop', 'text', 'the', 'this', 'to',
'two', 'which', 'will']
```

Included list

```
['donkey', 'words']
```

Only *donkey* and *words* have been selected. It seems that in the absence of an explicit value for $min.freq$, wordcloud filters stop words and verbs. It seems there is some basic form of text pre-processing occurring, which would include aspects of normalisation, stemming, and possible lemmatization. Various other inputs (not displayed) corroborate this.

To make wordcloud more inclusive of words in the word-list, make the parameters explicit (see here): enter $freq$ and set the value of $min.freq$ e.g. $min.freq$=1 will include all words used at least once, including standard stop words. Further precision can be achieved by modifying elements of the other packages we are using.

2)a.

Sources: Google NGram Viewer, Google Scholar, UCD Page

Our Professor Mark Keane was born in 1961, so I'll ignore the data prior to 1979 under the assumption he didn't publish anything of note prior to being 18.

- 1987: On retrieving analogues when solving problems, The Quarterly Journal of Experimental Psychology 39 (1), 29-41
- 1988: Analogical problem solving., Halstead Press
- 1988: Incremental Analogical Mapping: A Computational Model of Analogy, Third European Working Session on Learning
- 1998: Adaptation-guided retrieval: questioning the similarity assumption in reasoning, Artificial intelligence 102 (2), 249-293
- 1998: Principle differences in structure mapping, Analogy'98
- 2000: Efficient creativity: Constraint-guided conceptual combination, Cognitive Science 24 (2), 299-349 (this was the biggie)
- 2005: Cognitive Psychology: A Student's Handbook: A Student's Handbook 5th Edition, Psychology Press
- 2005: Retrieval, reuse, revision and retention in case-based reasoning, The Knowledge Engineering Review 20 (3), 215-240

2).b

There are not hits for "Otto James Hermann"; this is because I've yet to find anyone who shares my name, certainly anyone who has published.

2).c

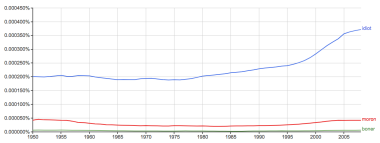
I suspected that "credit crunch" would have had a wider adoption by 2007;

however, it seems that the term was commonly employed as early as 1961, e.g. in Federal Banking Law Service, Volume 37 . Also of interest is that the phrase was relatively more frequently during the US Savings and Loan Crisis, than it was during the more recent banking fiasco (note, however, that the Google NGram Viewer cuts off in 2008 and most of the canonical work on the credit issues of the late 2000's were published from 2009 onwards).

2).d

As the smoothing value s increases, the shape of the graph approaches a straight line; see here. In brief, the smoothing value n is the input in a moving average $f(n)$ such that $2n + 1$ values will be averaged, with n years before a given year, n years after, and the year itself.

2)e.



I searched for synonyms referring unkindly to a simpleton: "idiot", "moron", and "boner". I ran my search from 1950 until the final data in 2008. I'm not surprised by the dominance of "idiot", as it has more uses and definitions than the other two terms, but I am surprised that from 1955 onwards the shape of the graphs for each of the three phrases is closely and consistently positively correlated. As their relative frequencies are also fairly constant, I find it rather interesting that the three words have maintained a similar volume of use over time.

I would speculate that these three phrases exhibit this behavior for myriads reasons, but principally that "moron" is a type of "idiot", and "boner" is a type of "moron", while their non-related uses are random in their textual appearance e.g. "boner" referring to tumescence.

2)f.

sex_NOUN has many results, whereas sex_VERB has none. I find this odd as, the 1933 paper by Masui and Hashimoto was widely cited, and subsequent work continues in sexing chickens. Perhaps there is a naive morality filter in the Google Ngram Viewer. But no, because fuck_NOUN and fuck_VERB are well documented; it seems that sex_VERB does not occur enough in the corpus to generate a search result.

2)g.

I chose "equality", "suffrage", and "liberation" as they all related to widepsread rise of democratic government and individual rights that have been on the rise, on average, for the past 500 years.

Libery is the most frequently used and seems to spike whenever one of the other two spikes. Equality is about twice as frequent as suffrage and has the most consistent upward trend of use. The use of liberty has fallen precipitouslys since 1791, whereas equality has risen at a steady rate, to the extent they are almost equal now. I would suspect that other synonyms, like freedom, would have risen (they have).

3)c.

Method 1 is consistently larger than Method 2. The denominator in Method 1 has only three elements, whereas the denominator in Method 2 has 15, including an overlapping element. $m_1 = r_1/(r_1 + r_2 + r_3)$, $m_2 = r_1/r_1 + c_2 + c_3 + ... + c_15$, so $m_1 < m_2 \rightarrow r_2 + r_3 > c_2 + ... + c_15$, i.e. r_2 and r_3 would need to be on average more than seven times larger than the elements c_2, c_3, ..., c_15.

For $m_1 < m_2$ on average, at least eight rows would need to conform to the above argument. But then 2 out of 3 columns in each row would be at least seven times as large as the values in the other column, which would mean each of those columns have a higher sum than the rows, which means m_1 is not less than m_2 on average. See here for a spreadsheet example.

4)

The article relies on two primary sets of data: a googletrends.csv and a ford-sales.csv. The former is readily available, but the latter might be more challenging to obtain.

A seasonal autoregressive model is easy enough to implement (and the paper provides explicit code for R), so the challenge in replicating this comes from the data. It may be possible to obtain the same Google Trends data as was used, but using the website it's not clear if the data now available would match the trend data used eight years ago, even if it's ostensibly the same data.//
The Ford sales data might not be easy to obtain: the figures would need to be sourced from Automotive News, access to which might be problematic. The data could also have been revised since the initial publication, so the actual results might be different even if the methods and model were perfectly replicated.