Text Analytics Tutorial 3

Question 1 a

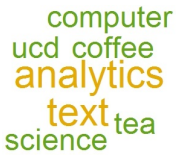
Carrying out the commands in the practical notes resulted in the above word cloud when I plotted it in R.

Question 1 b

The words which were visualised were institutions, united, enjoy, glorious, continue, upon, county, cause, benefits, yet, conferred, children, bequeathed, thousand, childrens, compeer, may, generations, rejoice and washington. Those not shown (with their respective frequency) are *have*: 1, *our*: 2, *under*: 1, *to*: 3, *those*: 1, *a*: 2, *his*: 1, *the*: 1, *us*: 2 and finally *by*: 2.

The words which are omitted appear to be those which are most commonly used in the English language and bear little meaning to the text's meaning overall. The frequency of the words has little or nothing to do with it because *to* has as many occurrences as *children* but is still omitted. The examples which could be considered to contradict this are *yet* and *may* (which do appear in the plot), so there could be a random element to its selection as these words carry little meaning also. Another possibility (more likely) is that the visualisation tool simply has a set list of words which it always omits.

Question 1 c



I made a sample of text with very little variety in words to examine this closer. I again included all the words which were omitted from the first example and increased their frequency more than any other word this time, none of them made it to the final plot. This solidifies my theory that there are probably key words which are left out of the plot by default, this makes sense since these words almost always convey little or no meaning and would just serve to clutter up the plot. For certain those words which have higher relative frequency always stand out on the plot by being bigger and/or having a different colour (depending on your parameters used).

Question 1 d

I added more frequency of the word *text* and the plot completely changed. All words (even analytics) grew smaller and only *text* stood out. This new information clearly shows that the size of the words is

relative to the dataset. This again makes perfect sense when you consider that the size of each body of words which you may wish to plot is almost always going to be (sometimes radically) different.

If you change the parameter *min.freq* in the package you can make R plot only those words which are at or above a certain frequency. *min.freq = 1* will include all words for example whilst *min.freq = 2* will only include those which appear two or more times.

Question 2 a

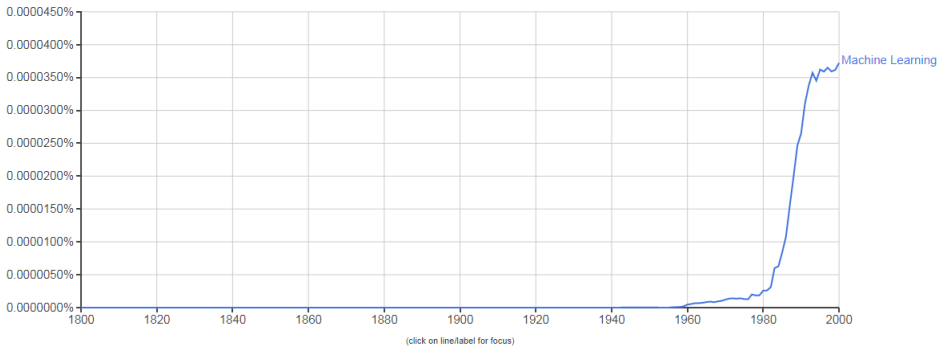
Google's Ngram Viewer shows the relative frequency of certain words which appear in that order (if there's more than one) each year. Frequency in this case is N of an Ngram in a given year divided by the total N of words in the corpus for that year. It is important to divide by the total number of books because the amount will change each year and it is better to "smooth" out the graphs.

The peaks in the graph when "Mark Keane" is searched for represent years of relatively high frequency for the Ngram. The more recent spikes in 2000 can be attributed to the Professor of the same name at UCD. The older spikes in the 1970's can be attributed to academic Mark A. Keane from the Department of Chemical Engineering at The University of Leeds by being cited in or publishing papers.

Question 2 b

There are not any hits for my name "Eoin Kenny". This is because there is no occurrences of this Ngram in the entire Google corpus.

Question 2 c



As expected the Ngram *Machine Learning* is a recent introduction to the English language. There are no major spikes before the 1990's when the emergence of cheap computational power, large datasets and more sophisticated ML methods such as *Deep Learning* became available.

Question 2 d

Using the "smoothing" functionality simply averages the curves in the graphs more. A smoothing of 1 means that the data shown for 1950 will be an average of the raw count for 1950 plus 1 value on either side. A smoothing score of 1 makes the "Machine Learning" search extremely jagged and perhaps more difficult to see a trend, the graph above has a score of 3 and shows the trend (perhaps) more clearly. A score of 50 however could be argued to do more harm than good because (for example) the year 1953 is shown as having a high frequency when it reality it should have 0%

Question 2 e

I compared the Ngrams *Machine Learning, Statistical Modeling* and *Artificial Intelligence*. There is actually nothing surprising in how these terms differ in frequency, *Statistical Modeling* is quite low relatively speaking but with a small peak towards the year 2000, probably because of the rise of big data and the ability to analyse it effectively. *Machine Learning* is several times higher in frequency than the former, again because of the rise in big data, new ML methods and cheaper computational power. The reason I think that the latter is higher in frequency is because the term *Machine Learning* is simply more interesting and exciting to people, so it was chosen as the preferred way to describe this field. *Artificial Intelligence* was at one time (1990) twenty times higher in frequency than even machine learning, however it dropped significantly in the 1990's with a turnaround in 2000. The reason this term is so much more popular than the others is likely because everyone knows this phrase and use it often (in academic & popular fiction), even since the 1940's which is roughly when it became famous. The reason for the dip in the 1990's could possibly be attributed to the infamous AI Winter.

Question 2 f

I searched for the word *love*. As a noun the word has dropped in frequency by around one third since 1800. As a verb it is much less frequently used, it starts off in 1800 being used on third as much as its noun equivalent but remains relatively unchanged up to 2000 when it is only slightly less than half the frequency of love as a noun.

Question 2 g

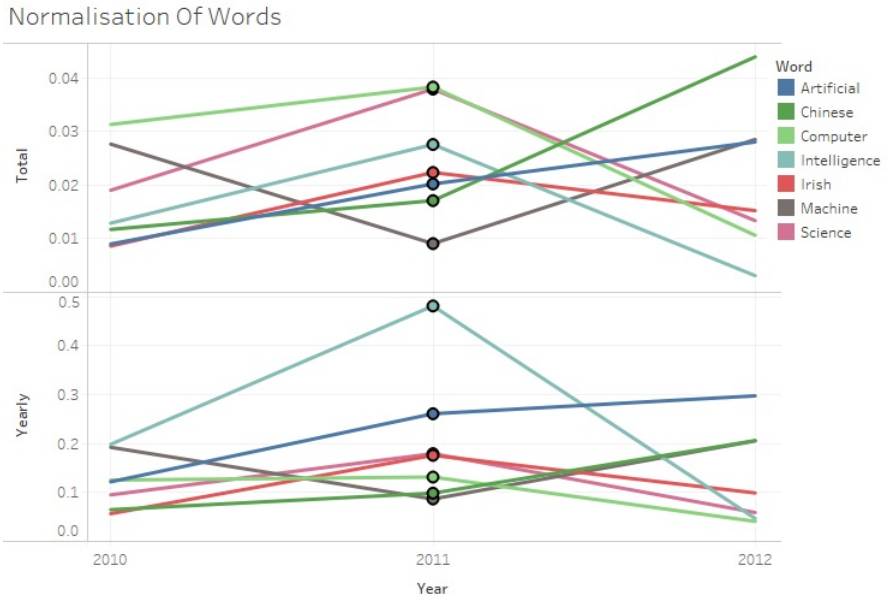
I searched for *Prohibition* in the Ngram viewer. As expected the word was relatively unused until the 1900's when there was a huge spike in the 1920's correlating with the American prohibition which ran from 1920 to 1933. The term has since leveled out between these two frequencies and been relatively consistent for over the last sixty years.

Question 3 a & b

	2010		2011		2012	
Word	Total	Yearly	Total	Yearly	Total	Yearly
Hat	0.009	0.031	0.033	0.090	0.032	0.096
Cat	0.029	0.102	0.026	0.078	0.039	0.128
Bat	0.009	0.035	0.018	0.057	0.003	0.011
Computer	0.031	0.125	0.038	0.131	0.011	0.041
Bag	0.019	0.085	0.041	0.164	0.023	0.092
Science	0.019	0.095	0.038	0.179	0.013	0.059
Chinese	0.012	0.065	0.017	0.099	0.044	0.206
Food	0.017	0.100	0.029	0.187	0.015	0.091
Irish	0.009	0.057	0.022	0.176	0.015	0.099
Machine	0.028	0.192	0.009	0.087	0.029	0.205
Learning	0.042	0.362	0.018	0.190	0.016	0.145
Artificial	0.009	0.122	0.020	0.261	0.028	0.297
Intelligence	0.013	0.198	0.028	0.481	0.003	0.047
Father	0.021	0.393	0.022	0.754	0.036	0.565
Mother	0.032	0.999	0.007	0.991	0.027	0.998

The above excel chart shows the results of questions 3a/b. The column *Total* refers to using the total N of words over all the years (method1 part a), whilst *Yearly* refers to a normalised frequency for each word in each year, using the total n of words in a given year (method2 part b).

Question 3 c

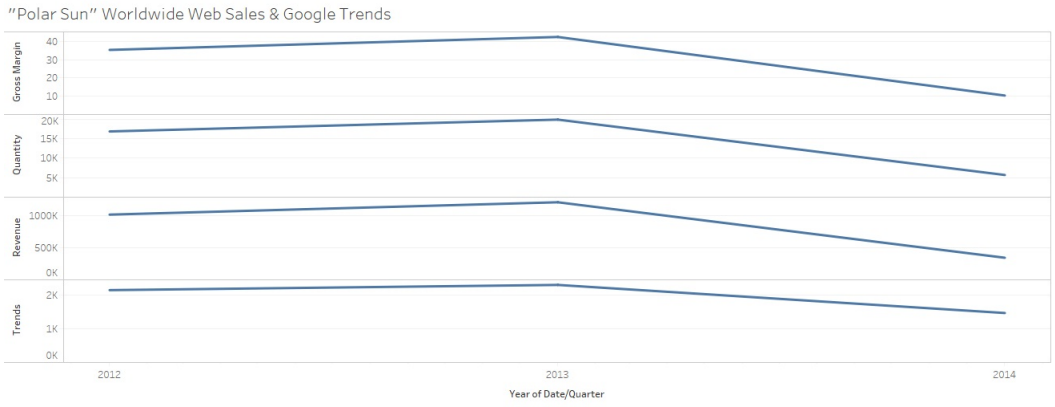


Normalising by method1 or method2 makes a big difference to the scores produced. The above graphs show the magnitude of this with *Total* referring to method1 and *Yearly* method2. *Yearly* generally shows all frequencies to be around one decimal place bigger than *Total*. I only showed a sample of the words to avoid cluttering the graph. The best example which illustrates the difference between both methods is the word *Intelligence* which has the highest frequency when normalising by each year but more of a mid-range frequency when doing so by the total N across three years.

Question 4

Here the data used by Choi & Varian is not available for use so I cannot run their program. However I have decided to use some initiative and attempt this question. Firstly, data for sales of *Polar Sun* sunglasses was obtained from IBM at *https://www.ibm.com/communities/analytics/watson-analytics-blog/sales-products-sample-data/*, then Google trends data for the Ngram *Polar Sun* was collected in CSV format for analysis. With this data there was enough information to attempt this question in a manner similar to Choi Varian.

Much pre-processing was done using excel to clean and join the datasets into one CSV file. After this a preliminary investigation was done using Tableau to observe if a trend was present, the results of which can be seen below.



The plot above clearly indicates a decline in *Google Trends* could accurately predict a company's profit margins for a specific product. After this an attempt to copy the R code of Choi & Varian yielded the following results.

```
# Make a train test split
dat1 = df[1: (nrow(df) -4), ];
dat2 = df[(nrow(df) -4): nrow(df), ];

# Code omitted for space (can be viewed in original paper)

##### Fit Model;
fit = lm(log(Trends) ~ log(Revenue) + log(Quantity), data=dat1);
summary(fit)

##### Diagnostic Plot
par(mfrow=c(2,2));
plot(fit)

#### Prediction for the next month;
predict.fit = predict(fit, newdata=dat2, se.fit=TRUE);
```

Listing 1: Code For Question 1 a

Coefficients:				
—	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-34.099	87.473	-0.390	0.713
log(Revenue)	9.639	21.348	0.451	0.671
log(Quantity)	-9.538	21.362	-0.447	0.674
Residual standard error:	0.2057 on 5 degrees of freedom			
Multiple R-squared:	0.22, Adjusted R-squared: -0.092			
F-statistic:	0.7051 on 2 and 5 DF, p-value: 0.5373			

I was able to run the R code of Choi & Varian but the plots produced in R were omitted for space proposes. The correlation between Google Trends and the product's online sales showed very obviously in Tableau when rolling up to years, but the results from the R code modelling were disappointing. However, I am almost certain the reason for the low correlation scores is simply because there was not enough data in the final CSV file to allow for proper training.