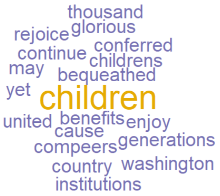
Jason Ballantyne | Text Analytics COMP47600 | 09/10/2021
13432788 | Practical Number 3

Q1.

a)

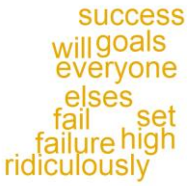


b) Please see the list of the words from the original quote that are included in the word cloud and those that are not in the table below.

Words Included from Quote	Words not Included from Quote
thousand, rejoice, glorious, continue, conferred, may, childrens, bequeathed, yet, children, united, benefits, enjoy, cause, generations, compeers, country, washington, institutions	our, and, to, the, upon, us, by, have, under, those, his

It appears from the table above that the words; "**our, and, to, the, upon, us, by, have, under, those, his"** have been excluded from the word cloud. I believe this is likely do to these being primarily stop-words. These words hold less weight and would therefore provide little insight into the quote. As a result, these less important words were removed from the word cloud.

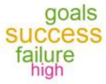
c) The text I used is a quote from James Cameron and can be seen below:
"If you set your goals ridiculously high and it's a failure, you will fail above everyone else's success".
The corresponding word cloud can be seen below:



Words Included from Quote	Words not Included from Quote
success, will, goals, everyone, elses, fail, set, failure, high, ridiculously	if, you, your, and, its, above

It appears my initial hypothesis still holds true. Looking at the table above we can see the words not included from the quote are; "**if, you, your, and, its, above"**. These would still be considered stop-words and the more essential words still do appear in the word cloud.

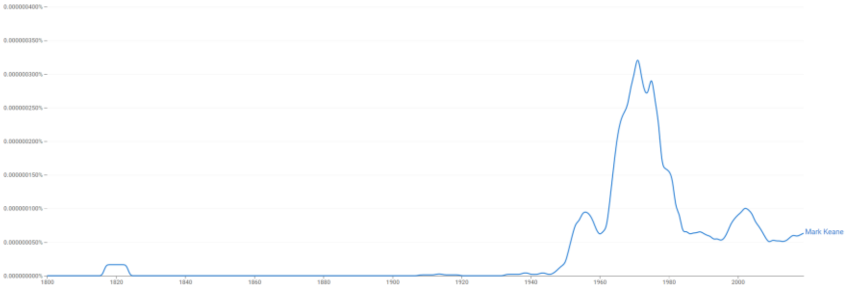
d) I've added more repeated words to the list, the sentence is no longer coherent but this is in order to illustrate the differences in the word cloud. The new text now reads;
"If you set your goals goals goals goals ridiculously high high high and it's a failure failure failure failure, you will fail above everyone else's success success success success success." The corresponding word cloud can be seen below:



In this scenario, I've added "**goals, success, failure, high**" multiple times to the text. It appears that only the repeated words are displayed in the word cloud. The words are only appearing in the word cloud if they hit the default min.freq which is equal to 3. It is possible to adjust this min.freq to see more of the text if we desire.

Q2:

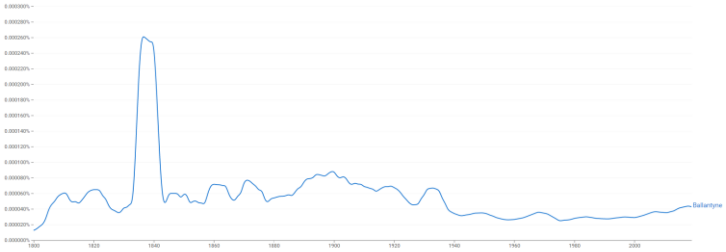
a)



"Mark Keane" appears in various books, reports and articles. Please see below table for an explanation of the peaks and when they occurred.

Peaks	Explanation for the Peaks
1800 - 1957	1. Selection of Reports and Papers of the House of Commons. (1836) 2. Bibliography of the Athapascan Languages - Volumes 14-19 - Page 123 (1892) 3. Report of the Society for Promoting the Education of the Poor (1892)
1958 - 1972	1. Nominations of D.C. Commissioner, Assistant to Commissioner, and Nine City Council Members (1967) 2. Problems of Air Pollution in the District of Columbia. (1967)
1973 - 1974	1. National Policy and Priorities for Science and Technology Act, 1974:
1975 - 2013	Several Psychology articles have reference to Mark Keane during this period. An example of a few are: 1. Cultural, Psychological, and Typological Issues in Cognitive Linguistics (1995) 2. Cognitive Psychology: A Student's Handbook (2005)
2014 - 2019	1. Design Studies in the Geometry of Frank Lloyd Wright (2017) 2. Metaphor: A Computational Perspective (2016)

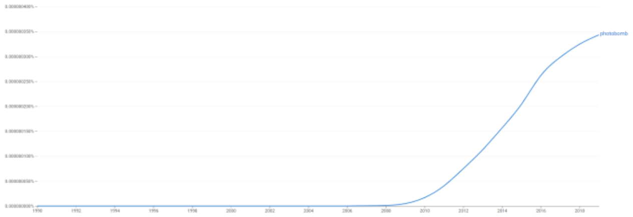
b) Unfortunately, there are no appearances for my full name, "Jason Ballantyne". However, I do get some hits for my surname alone, "Ballantyne".



There is one major spike in hits around 1837, then two smaller hits around 1900 and 1932. Please see below table that outlines where these came from.

Peaks	Explanation for the Peaks
1837	The majority of these hits come from the minutes of evidence before select committee that appear in the “Parliamentary Papers, Volume 12 By Great Britain. Parliament. House of Commons. (1837)
1900	1. The author of the book “Madman and the Pirate” – R. M. Ballantyne 2. “Snowflakes amd sunbeams” – Robert Michael Ballantyne
1932	“Ballantyne's Folly” – A novel

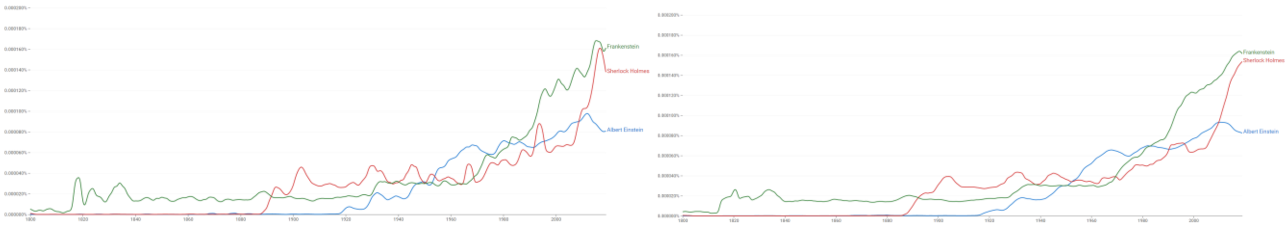
c) The word I have picked is “**photobomb**” which refers to the action of spoiling a photograph by unexpectedly appearing in the camera's field of view.



As expected, the emergence of the word photobomb roughly coincides with the increased popularity of camera phones in 2007. The earliest recorded uses of the word appeared to be used in the following:

1. The Chicago Manual of Style (2003)
2. Time - Volumes 173-174 (2009)
3. Geektionary: From Anime to Zettabyte, An A to Z Guide to All Things Geek (2010)

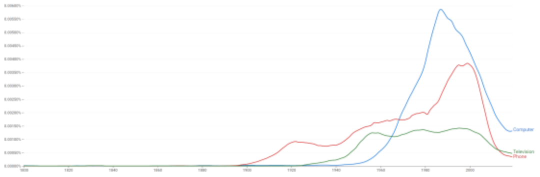
d) Smoothing helps to make the graph more legible and easier to analyse. It averages out values over a range of years. For example, all the graphs displayed to date have a smoothing factor of 3 which averages out the values over a 3-year period rather than just 1. I will now compare the same graph with a smoothing of 1 vs a smoothing of 3.



(Albert Einstein, Sherlock Holmes, Frankenstein with a smoothing of 1 vs a smoothing of 3)

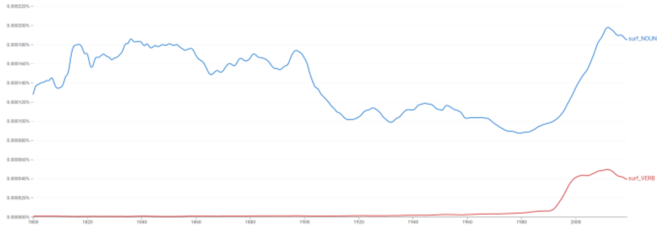
As you can see from the graphs above, a smoothing of 1 is much closer to the raw data. The spikes are harsher, and it is harder to identify general trends without the use of smoothing.

e) The three words I have chosen are **Computer, Phone, Television**.

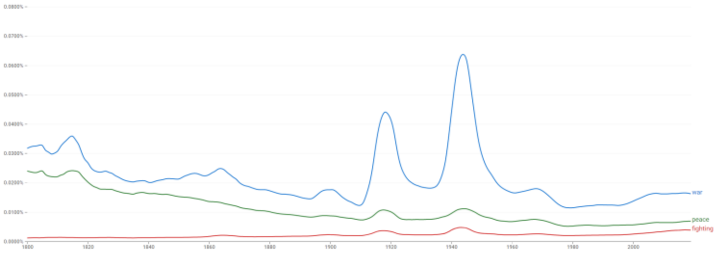


What is quite surprising from this graph is that pre 1967 the “Phone” had the highest index out of the three categories and now, it is ranked the lowest of the three terms. This is quite surprising but is likely more attributed to the rise and popularity of computers and televisions. Another surprising point of note is that “Phone” had a severe drop in frequency between 2004 and 2020, similarly the "Computer" had an equally astonishing drop between 1988 and 2015. Throughout all of this, "Television" remained relatively stable with no noticeably major spikes. This is likely due to the hype that surrounded computers and phones, meanwhile television did not acquire the same amount of attention. Aside from the addition of colour, televisions improvement has been slow and gradual in comparison to both computers and phones.

f) The word surf was used as both a noun and a verb for this section. As illustrated in the graph above is it quite clear that surf as a noun is used much more frequently than its verb counterpart. Surf as a verb has begun to gain popularity in 1991 and has been steadily growing since.



g) War has caused major cultural changes; it destroys communities and families and often disrupts the development of the social and economic fabric of nations. The relevant words I have chosen that pertain to this are "**war, fighting, peace**".



All three words appear to follow the same trend with major spikes happening in 1918 and 1943, which are, by no coincidence, right in the midst of World War 1 and World War 2 respectively. The next closest spike in frequency appears in 1814, which is also during the War of 1812, between the United States and its Indigenous allies. This is then followed by a spike in 1864 during the American Civil War.

It is fascinating to see the frequency of these words lining up exactly to the most notable war related cultural changes in the past 200 years.

Q3:

Words	**Frequency**				
	2010	**2011**	**2012**	**2013**	**2014**
engine	1884	471	58	622	1333
tired	1586	1024	384	626	336
homely	967	595	1195	406	1204
cattle	766	338	1729	1498	498
top	704	527	1808	772	1862
flock	930	1095	927	1485	723
blood	479	21	902	1617	1939
screeching	1433	1281	463	1954	443
root	591	547	380	449	1080
fallacious	317	904	156	283	1887

(i) The large-N for the data for the table above is 45479.

(ii) The small-N for each year is as follows:

Year	**Small-N**
2010	9657
2011	6803
2012	8002
2013	9712
2014	11305

Jason Ballantyne Text Analytics COMP47600 09/10/2021
13432788 Practical Number 3

a) & b)

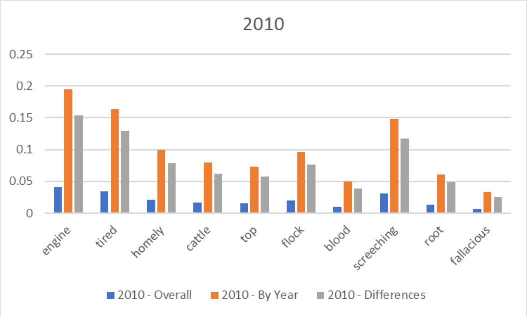
Using large-N						Using small-N					
	2010	2011	2012	2013	2014		2010	2011	2012	2013	2014
engine	0.041426	0.010356	0.001275	0.013677	0.02931	engine	0.195092	0.069234	0.007248	0.064044	0.117912
tired	0.034873	0.022516	0.008443	0.013765	0.007388	tired	0.164233	0.150522	0.047988	0.064456	0.029721
homely	0.021263	0.013083	0.026276	0.008927	0.026474	homely	0.100135	0.087461	0.149338	0.041804	0.106502
cattle	0.016843	0.007432	0.038018	0.032938	0.01095	cattle	0.079321	0.049684	0.216071	0.154242	0.044051
top	0.01548	0.011588	0.039755	0.016975	0.040942	top	0.0729	0.077466	0.225944	0.079489	0.164706
flock	0.020449	0.024077	0.020383	0.032652	0.015897	flock	0.096303	0.160958	0.115846	0.152904	0.063954
blood	0.010532	0.000462	0.019833	0.035555	0.042635	blood	0.049601	0.003087	0.112722	0.166495	0.171517
screeching	0.031509	0.028167	0.010181	0.042965	0.009741	screeching	0.14839	0.188299	0.057861	0.201194	0.039186
root	0.012995	0.012028	0.008356	0.009873	0.023747	root	0.061199	0.080406	0.047488	0.046231	0.095533
fallacious	0.00697	0.019877	0.00343	0.006223	0.041492	fallacious	0.032826	0.132883	0.019495	0.029139	0.166917

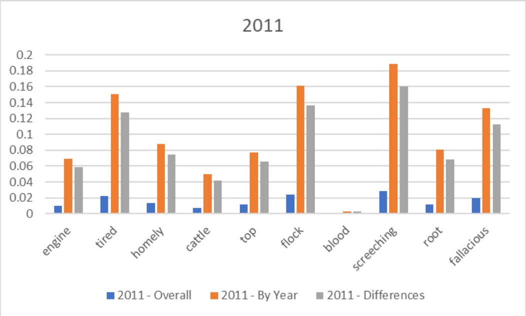
Overall Normalization was done by dividing by the large-N for each word. The corresponding large-N value for the table above is 45479.

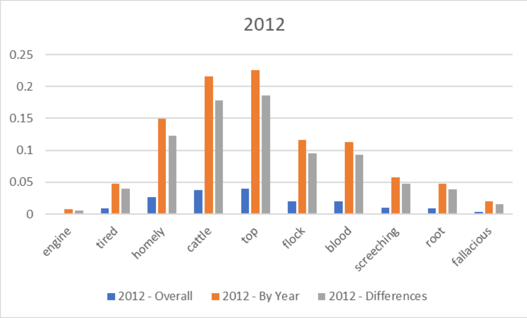
By year Normalization was done by diving by the small-N for each word in its respective year. Please see table below for corresponding small-n values.

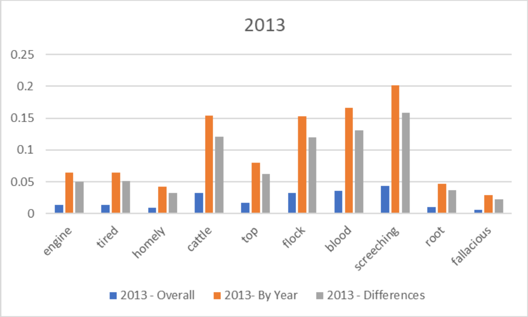
	2010	2011	2012	2013	2014
Small-N	9657	6803	8002	9712	11305

c) Normalizing these in different ways does make a big difference to the scores produced. Overall normalization deals with the entire corpus whereas by-year normalization deals with only per year corpus for calculating frequencies. As a result, scores produced are much smaller using the overall normalization method as opposed to the by-year normalization. Please see the results of the tables and their differences in the graphs below.

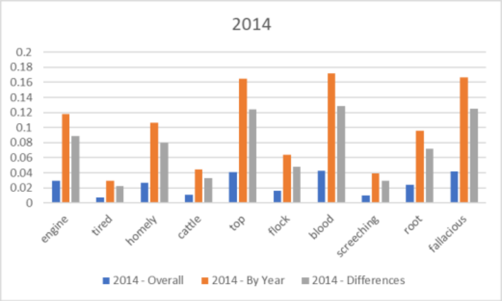








As mentioned, the overall normalization scores are much lower than the by-year normalization scores and are made even more evident by the graphs. For this example, it makes more sense to normalize the data using the by-year method as we have the frequencies segregated in years. Depending on when the normalization needs to be performed, indicates whether we should use the overall or by-year normalization method.


engine
tired
homely
cattle
top
flock
blood
screeching
root
fallacious
2014 - Overall
2014 - By Year
2014 - Differences