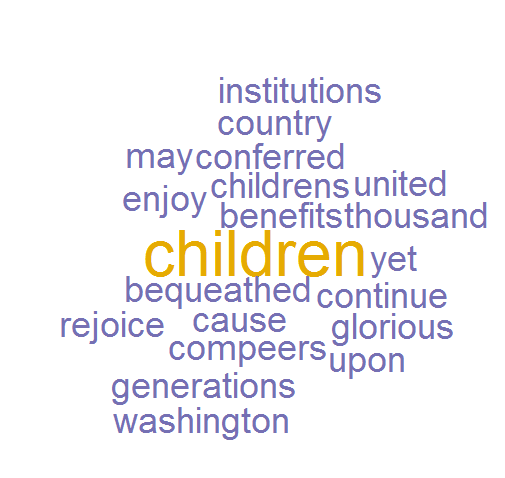
Practical 3

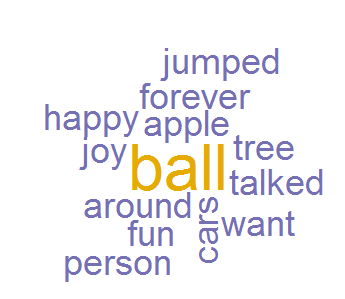
**Question 1:**

**b.)** The wordcloud produced is seen to the right. These are the words included. The words excluded are as follows:

* our
* and
* to
* a
* have
* under
* those
* us
* by
* his

The above sets of words are all either pronouns (our, us, his), prepositions (to, by, under), demonstrative adjectives (those), conjunctions (and), articles (a), or auxiliary (aka helping) verbs (have). All of these parts of speech are very common in English phrases and can often be omitted without losing too much meaning. Additionally, the frequent necessity to use these words in English could skew any word counter, as these types of words will almost always be the most frequent, but carry little meaning, which could defeat the purpose of a wordcloud. As a side note, it is interesting that everything was made lowercase and characters, such as apostrophes, were dropped from the original text and then processed (as exemplified by the wordcloud presence of the word “childrens” which was, originally, “children’s”.

**C.)** To test this theory the following word list will be used:

I go around up down in out and all I want to do is jumped talked and have fun forever He has a ball He has the ball and an apple and those cars that person is so happy, joy tree

The word cloud seen right was produced.

As predicted, all prepositions (around, up, down, in, out, to), auxiliary verbs (be, do, have), pronouns (I, he), conjunctions (and), articles (a, an, the), demonstrative adjectives (that, those), and adverbs (so), have been excluded from the wordcloud. It is also interesting to see that the verb “go” was also eliminated, despite it being an action verb which carries more immediate meaning than auxiliary verbs. It is also interesting that the word “that” was included in the wordcloud when it was accidentally capitalized in the middle of the text sample, but upon correcting this error, it disappeared. I can only assume that the program is build to check for capitalized words and assume that they carry some importance but doesn’t do it for pronouns like ‘he’(although it should be noted that the capitalized words are still made lowercase in the wordcloud, thus hinting at some normalization going on behind the scenes). As to whether or not my theory was correct, I definitely believe I am on the right track – none of the parts of speech I initialized guessed would be eliminated were present. Additionally, I tested whether or not it eliminated verb endings (such as past tense ‘ed’) like stemming and lemmatizing, but this is not the case. I am sure I haven’t gleaned every rule, but I think it must have something to do with very common words and their use in sentences, specifically as to how much meaning they carry.

**d.)** The list has been changed to repeat more words:

go go around up down want want and all I want to do is jumped talked talked fun forever forever He has a ball He has the ball and an apple and a ball and those cars cars person is so happy happy joy joy joy tree tree

Interestingly, the addition of new repeats of words yeilded the wordcloud seen right, which only consists of three words. Words previously included and still present in the list, and even more oft repeated, are now excluded. I believe that this is because the program counts the frequency of the most frequent word and judges what is worth including based on that number. So in the resulting tree, these are the words that are repeated three times, and the ones only repeated twice have been dropped.

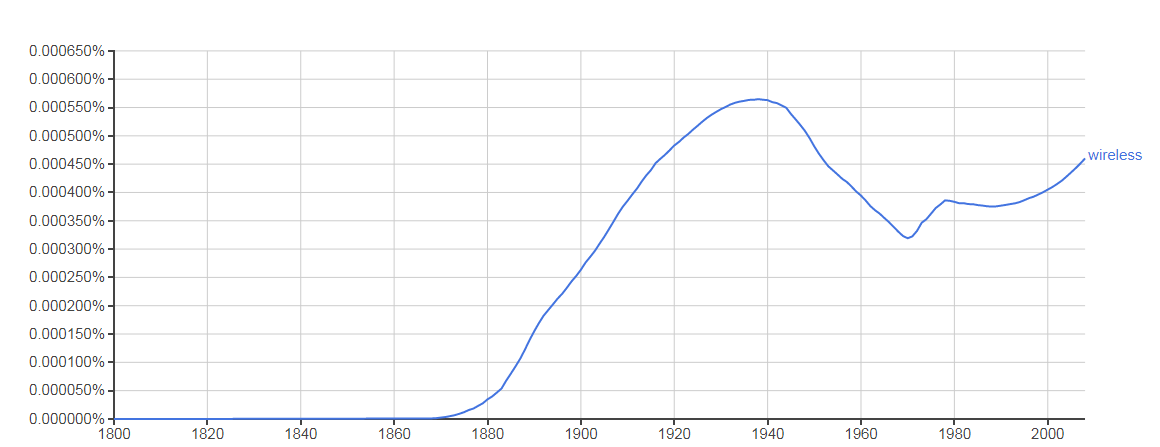
The package and arguments used to build the wordcloud can be changed to include more words. In the arguments, setting the min frequency can change the appearance of some words. In this example, setting the minimum frequency to 2 yeilds the wordcloud to the right , which includes the words that had been previously dropped. If it is set to 1, more words would be included.

Additionally, more words could be removed or included if tm\_map from the tm library is used (Kodali, 2015). For example, the command tm\_map(jeopCorpus, removeWords, c('the', 'this', stopwords('english'))) removes the words ‘the’ and ‘this’ (as selected by the user) as well as English stopwords (such as I or he) from the file jeopCorpus. This gives the user more control over the wordcloud display.

**Question 2:**

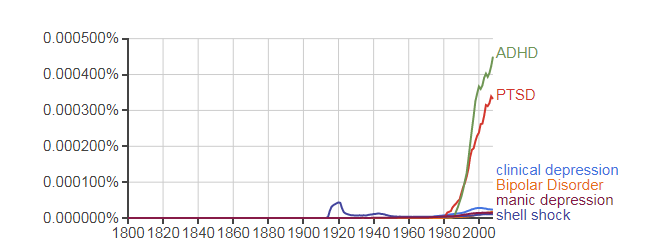
1. The first appearance of the phrase “Mark Keane” appears in 1952 and stays at the same frequency (.0000000078%) until 1959 where it goes back to zero. However, in 1960, it appears again, this time sharply increasing until peaking in 1969, dipping a bit in the following years, then reaching an all-time high in 1974 , dipping in ’75 and ’76 then peaking again in ’77 wherein after it started to decrease again, having a smaller rise in 1980, hitting a low in ’83 then rising sharply in ’89, dipping in the ’90 but then sharply increasing during all years between 1992 and 2000 hitting a peak number as high as it had been in ’74 in 2001. A decrease has been shown since then.
2. The first appearance of the phrase “Katherine Campbell” went from zero to 0000012888% in 1806 and remained at that level until 1812 where it promptly plummets back to zero. It sees a small presence between 1831 and 1843 and again, completely disappears until between 1869 and 1876 where it peaks at its highest. In 1880 it plummets back down and by 1887 is at zero again and never fully recovers, climbing a bit in 1927, and going back down. It should be noted though that since the ‘60s it has been steadily increasing.
3. I searched for the word “wireless” and was shocked to see it first really peak in 1918, dip again, and then have another smaller appearance in 1938-1940 until finally, as initially hypothesized, skyrocketing from 1990 onward (wireless housephones, cell phones, speakers, wireless networks, etc.). However, after researching, it was revealed that the field of “wireless telegraphy” was an increadibly popular field between 1914 and 1918 due to its becoming an integral part of communication during ground warfare (The Australian War Memorial, 2017, March 20). Considering the dates of the early peaks of this graph, and also considering the dates of the two World Wars, things start to make sense. I had also forgotten that radios, becoming popular in the first half of the 20th century, are also wireless technologies.
4. With a smoothing value of 3, the graph for the word “wireless” is as such:

A smoothing value of 10 reveals this graph:

And a smoothing value of 30 renders this one:

The graphs look completely different. Firstly, the smaller changes have disappeared, literally “smoothing” the graph. Additionally, the values on the y-axis change. Finally, as can be seen in the last graph, even the shape and general trend of the line can change depending on the smoothing value. This is because the smothing is the average raw count of a year plus the values of x years on either side, where x is equal to your smoothing value (Google Books, 2013).

1. The words tested were related to psychiatric disorders: “ADHD, PTSD, shell shock, clinical depression, manic depression, Bipolar Disorder. This was found in the English corpus with a smoothing value of 3. The following graph was rendered:

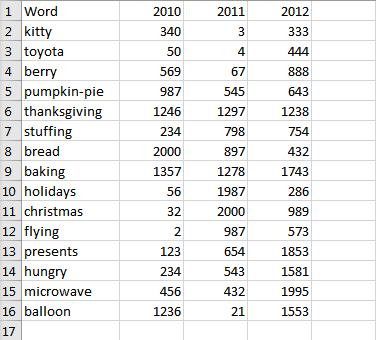


As can be seen, the older terminology “shell shock” was used, predictably, earlier, experiencing a small peak between 1918 and 1920 (years corresponding to WW1). The term PTSD (post-traumatic stress disorder) was not invented to replace the phrase “shell shock” until later, with its use only appearing in 1943 (corresponding to WW2 years) but not overtaking its predecessor until 1980, where the use of the term PTSD began to sharply rise until currently where it is the second most commonly found term among the searched phrases. The term ADHD also does not appear until 1943 and remaining with incredibly low frequency until 1985 ’86 and ’87 where it sharply increases in use and continues to do so, becoming the currently most frequently used of the searched phrases. The term ‘Bipolar Disorder’ was the last of the tested phrases to appear, showing up as late as 1968, whereas its other, more dated terminology, ‘manic depression’ appears as early as 1905, although not with much frequency. It is the second oldest of the tested terminologies, only younger than ‘shell shock’ which first appeared in 1861 (also corresponding with the first year of the American civil war). Considering historical dates, and the evolution of the field of psychology, I did not find these changes in frequency particularly surprising, except for the consistently low frequencies of the terms ‘manic depression’ and ‘bipolar disorder’ as there has been extensive research done on the subject. The relative scarcity of the term ‘clinical depression’ also really surprised me, as that is a prevalent and relatively non-taboo subject to write about, at least in more modern years. Finally, initially I was also surprised at the survival of archaic terms such as ‘shell shock’ but considering that all books are being searched, historical accounts could explain the present-day use of this term.

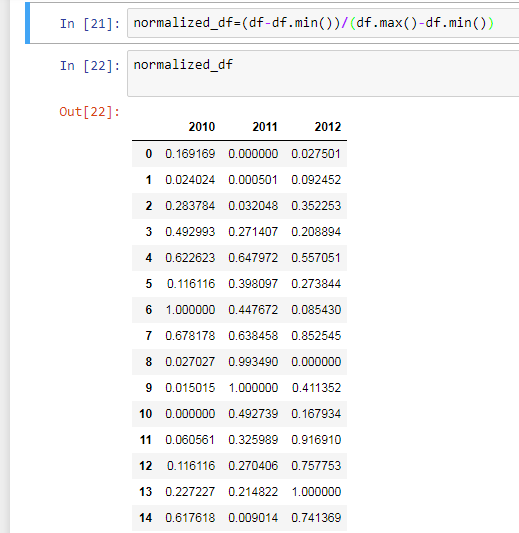
1. The two different words tested were peer (verb) and peer (noun). It is clear that both phrases have been in use since 1880, but that the noun use has always been much more common, although use of the verb has been steadily increasing. Interestingly, the noun use actually decreased from 1800 to 1941 and began to steadily rise again (although it must be noted that it never had a lower frequency than the verb form).
2. The cultural change explored was germ theory. Reflecting this subject are the phrases/words “hand washing” “germ” “germs” “bacteria” “cleanliness” “virus” and “viruses”. Germ theory was initially proposed in 1546 and then expanded upon in 1762, although it did not catch on and was held in disdain until the 1850s with Pasteur and Koch providing evidence. Viruses were then discovered in the 1890s. Although the word germ and germs were found as early as 1509, clearly it doesn’t relate to germ theory. The idea of virus started around 1667 although its meaning of virus as “an agent that causes infectious diseases” was first recorded in 1728 and indeed, from 1739 we see the use of the word significantly beginning to climb (although its prevalent use from the 1990s onwards probably has something to do with its referring to computers as well. The words germ and germs have a bump in the 1800s, really beginning to rise in the 1850s (aligning perfectly with the experiments and revelations of Pasteur and Koch). The word bacteria first appears in books in 1756 which is after their first discovery in 1676 (by Antonie van Leeuwenhoek) but before the word “bacterium” was coined in the 1820s so this is interesting to me. This word makes a strong appearance in the 1820s and then from 1847 sharply rises from therein peaking from 1916 to 1918. The height of papers written about bacteria and sickness in 1918 makes perfect sense considering that 1918 was the year of the deadly Spanish Flu. The word ‘germ’ also hits its highest point during this timeframe.

**Question 3:**

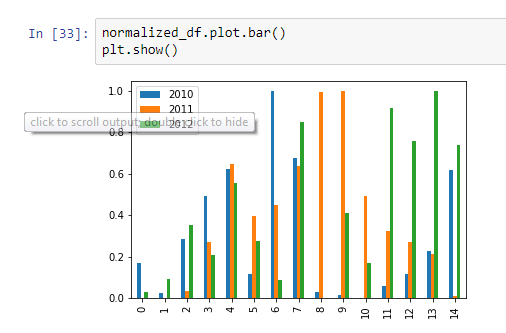
The CSV used for this is as follows:



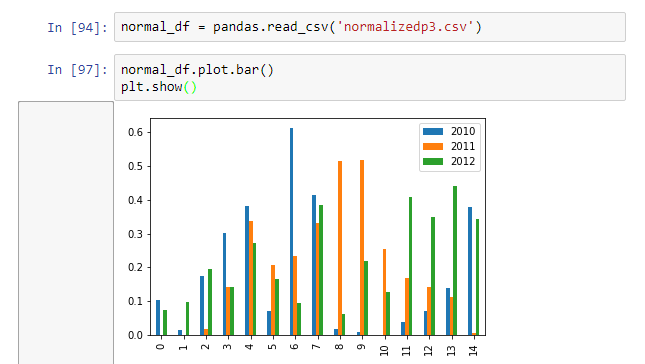
1. Method 1 – normalize using total words over all the years – this was accomplished by dropping the “Words” column and then computing the total minus the minimum over the standard deviation.



1. Method 2 – normalize using N words in a given year – I used scikitlearn.preprocessing.normalize on each column
2. Graphs of the values of the two different normalizations are seen



Above: method 1



Above method 2.

According to the graphs above, on this particular data set, the two different normalization methods didn’t have very many differences but word number 1 shows a much bigger gap between 2010 and 2012 values and number 4 is interesting in that method 1 has the 2011 value as the highest and method 2 has it as the middle value. This is because in terms of the single year and the count of all years, there is a significant difference in what the minimum and maximum value might be. This changes the values yielded by normalization. It is all a matter of comparison. The method best to use is dependent on the data.

**Question 4**

After reading and studying the R code used by Choi & Varian (2012), the first line is to get the data which is read from the googletrends.csv for whatever example they’re exploring. This can come from the google trends website. They also needed Ford sales data CSV for the same timeframe as the google trend. They procured theirs from automotive news. Unfortunately, following that path only resulted in a site that said the csv is available to paying members only. So I decided to see if even some of the program could be run if I found any corresponding data. For our exploratory purposes, let’s say it’s travel to the USA. To get it I had to go to the google trends website and find the trend for this search term and download the csv. I then had to find the data of arrivals to the USA in 2017. I read in these csv as data and attempted to compare. However, the dates for the international arrivals to the USA did not correspond to the dates of the google trends csv. Allowing more time, however, I think I could get this to work, and I understand the underlaying principle that plotting sales and google trends data and finding the lag between the two allows for the making of future predictions of sales based on google searches.

Works Cited

The Australian War Memorial. (2017, March 20). Technology, Weaponry and Communications in 1918. Retrieved October 04, 2017, from <https://www.awm.gov.au/visit/exhibitions/1918/technology>

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88(s1), 2-9.

Google Books. (2013). Google Ngram Viewer. Retrieved October 04, 2017, from https://books.google.com/ngrams/info

Kodali, Teja. “Building Wordclouds in R.” R-Bloggers, 28 Aug. 2015, www.r-bloggers.com/building-wordclouds-in-r/. Accessed 3 Oct. 2017.