1. Assuming you have R installed (if not install it). Load up the various packages you need for using the wordcloud packages:

a. Carry out the commands shown in the practical notes:

> library(wordcloud)

> library(tm)

> wordcloud("May our children and our children's children to a thousand generations, continue to enjoy the benefits conferred upon us by a 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?

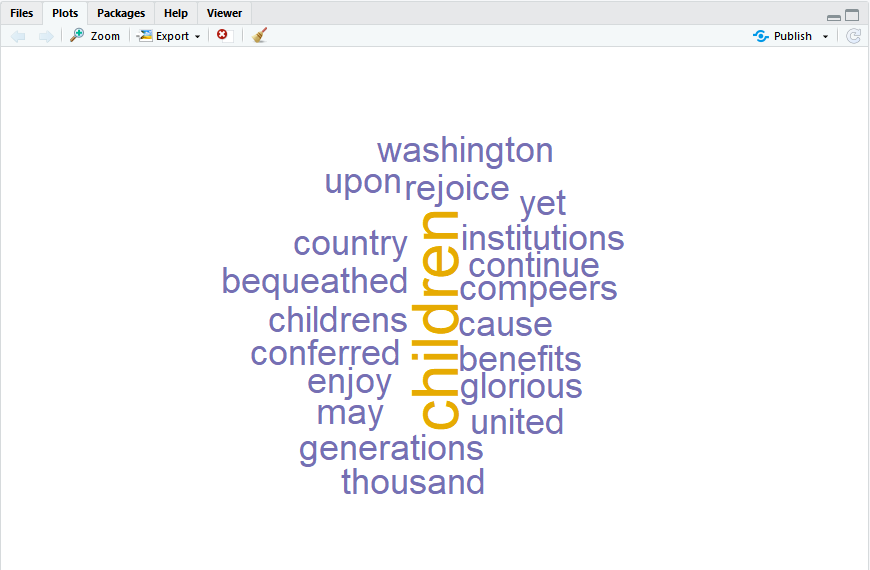
c. Now, check your theory about what the wordcloud package included and excluded. Put in your own word-list together (30-50 words) and check what wordcloud includes and excludes? Report whether your initial theory was right or wrong and why?

d. Again, using your word-list add more repeated words and see what happens? Can you change the package’s to make it more inclusive of the words in the word-list?

**Ans1a.** Following packages are installed:

* wordcloud package: It is required to generate the visual representation of the text data which is known as “WordCloud”.
* tm package: It is required to analyse the text (text mining).
* Rcpp package: It provides both R methods and C++ classes to collaborate both R & C++.
* RColorBrewer package: It is required for the color schemes of the visual representation.
* slam package: It is used for data structures and the algorithms for sparse matrices and arrays.

The two packages ‘wordcloud’ and ‘tm’ are loaded using library() method. Further wordcloud() is executed with the text and a wordcloud is generated depicting the term ‘children‘ as most important word with highest frequency in the text.

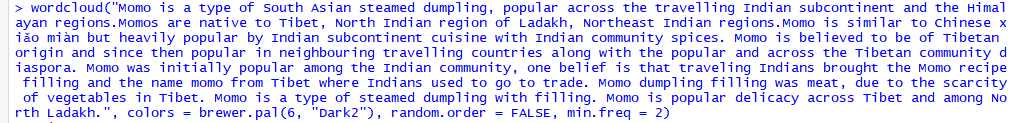


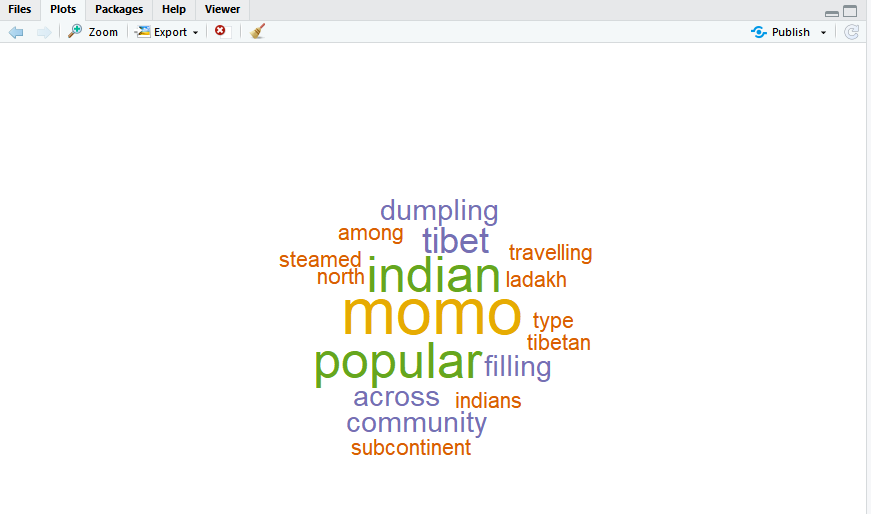
**Ans1b.** A list of few words included and excluded from the wordcloud are listed below:

|  |  |  |
| --- | --- | --- |
| Sr. No | INCLUDED | EXCLUDED |
| 1. | children | our |
| 2. | washington | and |
| 3. | rejoice | to |
| 4. | benefits | a |
| 5. | glorious | the |
| 6. | generations | us |
| 7. | institutions | have |

The internal functioning the wordcloud() is such that it invokes the text mining package (tm) to analyse the text, generate the collection of words and remove the stopwords from the string of words given as the input. Also the default value for the property min.freq is 3, every word with freq < 3 will not be displayed in the wordcloud. That’s why some words are excluded from the wordcloud visual representation.

**Ans1c.** wordcloud() is given a string of my own words.





A list of few words included and excluded from the wordcloud are listed below:

|  |  |  |
| --- | --- | --- |
| Sr. No | INCLUDED | EXCLUDED |
| 1. | dumpling | type |
| 2. | indian | belief |
| 3. | momo | Northeast |
| 4. | popular | delicacy |
| 5. | tibet | among |
| 6. | filling | is |
| 7. | community | similar |

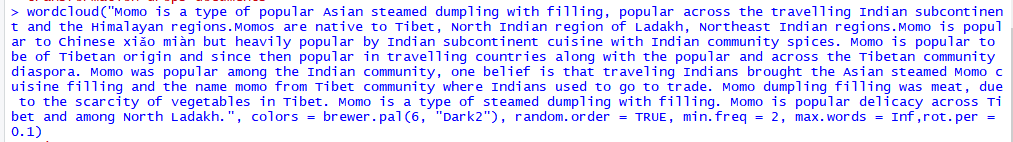
There’s slight difference in the initial theory, It not only removes the stopwords from the text but also doesn’t plot those words with frequency less than that specified by the feature ‘min.freq’.

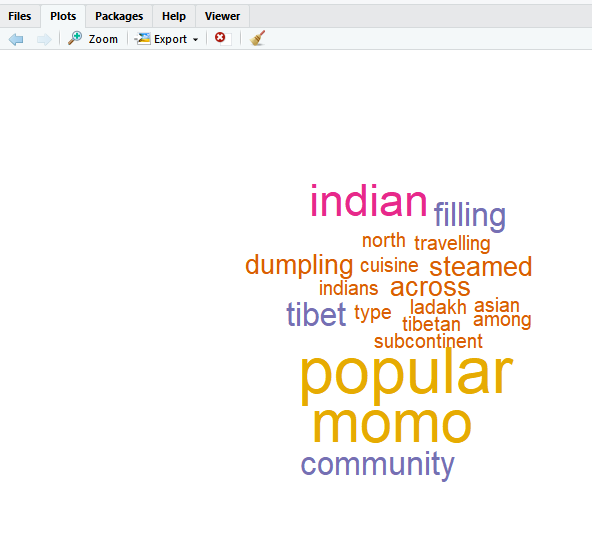
This property depicts the minimum frequency of the words to be plotted in the wordcloud. The initial theory was right that it does remove the stopwords from the text in addition to the words with frequency lower than min.freq.

**Ans1d.** More repetitive words are inserted in the string and passed to the wordcloud() method to plot the tagcloud. Here there are two words with similar and most frequency.The height of each word shows the frequency of occurrence of the word in the entire text.

Package properties that can be used to include more words from the input text are:

* min.freq: If it is set to 1, tagcloud will plot every word but of varying size.
* max.word: If it is set to a huge number, wordcloud can plot upto that number of words (ie. All the words from the input text).
* excludeWords: It is a vector containing the dictionary of own stopwords to be excluded from the text. If it is set to NULL, no user’s stopwords will be eliminated from the text.





==================================================================================

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

a. Put in “Mark Keane” as a search term and 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.

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 smoothing 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\*. Is there anything surprising about how these terms differ in their frequency and, if so, why? Why do you think the frequencies vary in the way they do.

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.

g. Think of some major cultural change that has happened over the last 500 years and some words that could denote to ths event/events. Check these words of the relevant time period. Report what you find.

**Ans2a. Ngram**: It is defined as contiguous sequence of n words from a sample of text. It is used in text mining and NLP, Can be used to make Next word prediction, spelling corrections. Example: “San Francisco” -> 2 gram.

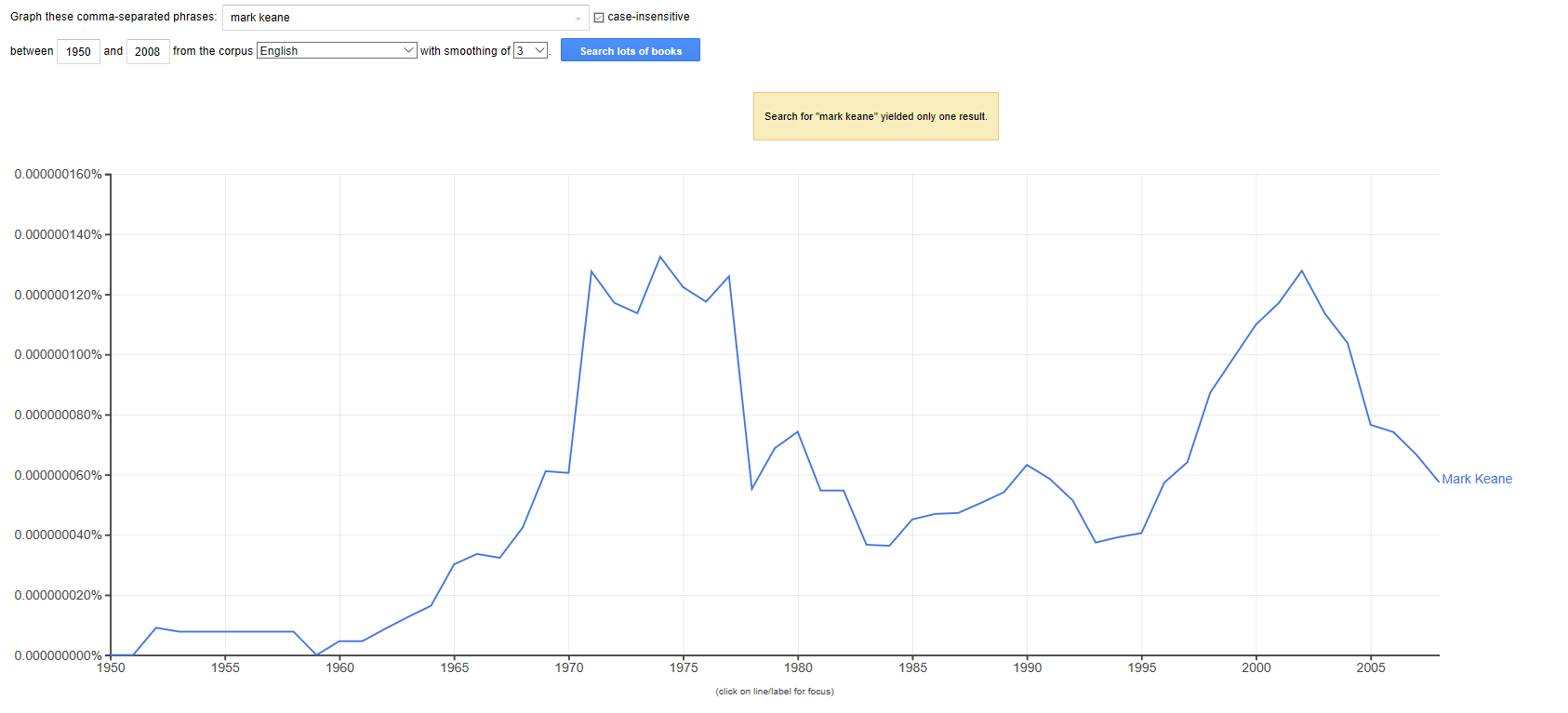
**Google Ngram Viewer**: An online search engine that displays graph indicating the frequency of any set of comma-seperated search strings on the basis of yearly count of n-grams, searched in books during the time period of 1500 and 2008 in Google's text collection written in English, language.

It shows how frequently the phrases have occurred in a collection of text over the selected years.

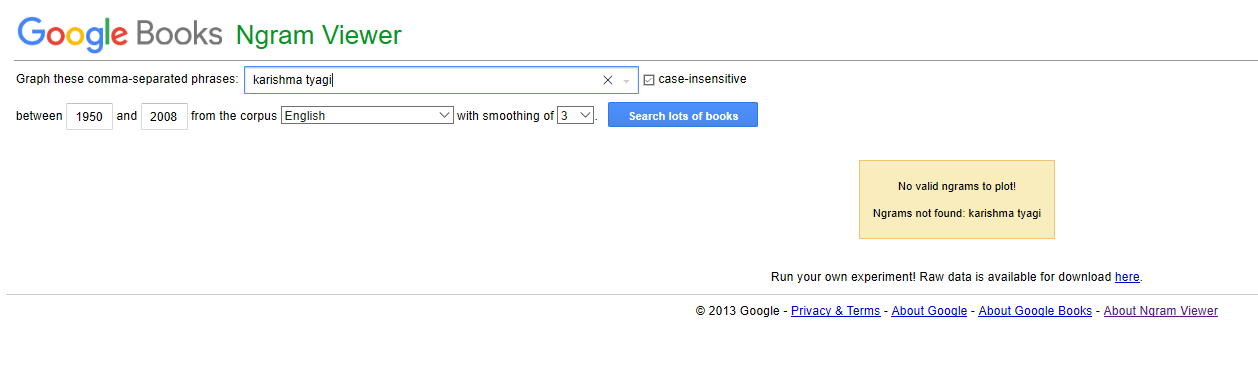
- When the search term “Mark Keane” is fed to the Google Ngram Viewer, following trend is displayed:

**The graph depicts the elevating trend during the time period of 1960-1971, then in early 1974s. Further it hits the peak during 1976. It then decreases drastically in 1977. Similarly it gradually increases during the period 1984 to 1990. In the end it elevates from 1995 and hits the peak at 2002.**

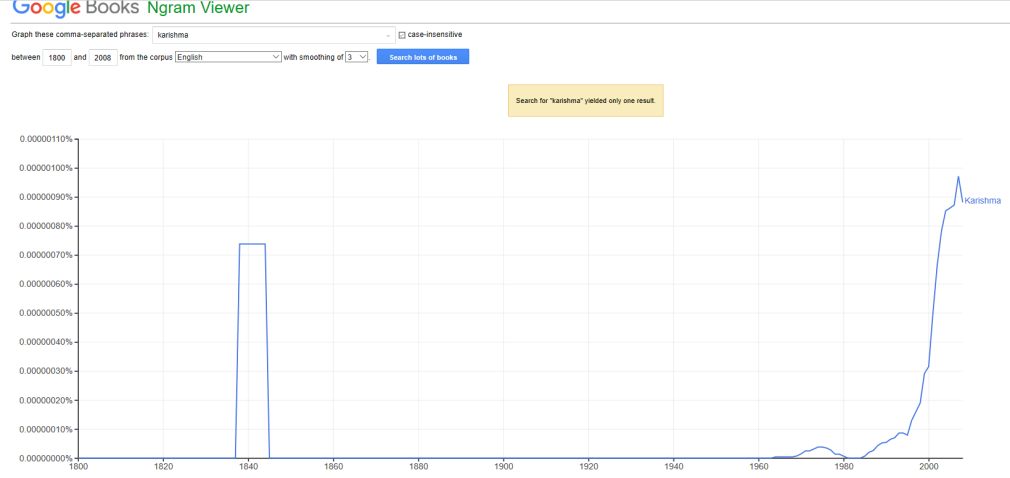
- The peaks indicates that during this period of time, frequency of the search term “mark keane” is the maximum. It is due to the corpus including number of **books(cognitive psychology), hearing(Problems of Air pollution in the district of Columbia), reports(Committee on district of Columbia),(Annual City Manager School) meetings, City Manager’s newsletter, Conference Proceedings of the International City Managers** published, printed during that time where the search term appears the most. For some years, the data is not visible publicly due to their removal under the data protection law. In **1974, Wheeling Herald Newspaper Archives** is published where this search term ‘mark keane’ can be found frequently. Further for the period of 1975- 2002, a large number of books named “Cultural, Psychological and Typological Issues in Cognitive Linguistics” , “Cognitive Psychology: A student’s handbook”etc published during that period of time with the frequency of 0.000000130%.



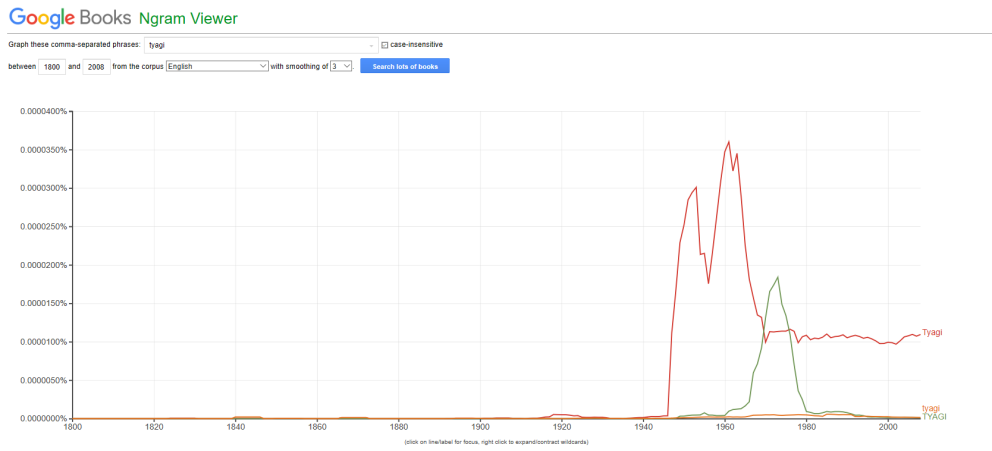
**Ans2b.** I searched my name “karishma tyagi” but no results were showed for the bi gram.



But if I gave the input of my first name (uni gram) as the search term, it displayed the following graph depicting its gradual increase in the time period 1995-2006 hitting the peak at 2 values (0.00000075% in 1840s and 0.00000100% in 2006 period of time). The reason of this frequency is due to number of books(Key Competencies in Brief Dynamics Psychotherapy,Tourist Guide to Goa), blogs, Bollywood news articles(National survey of prison health care), Case studies etc.

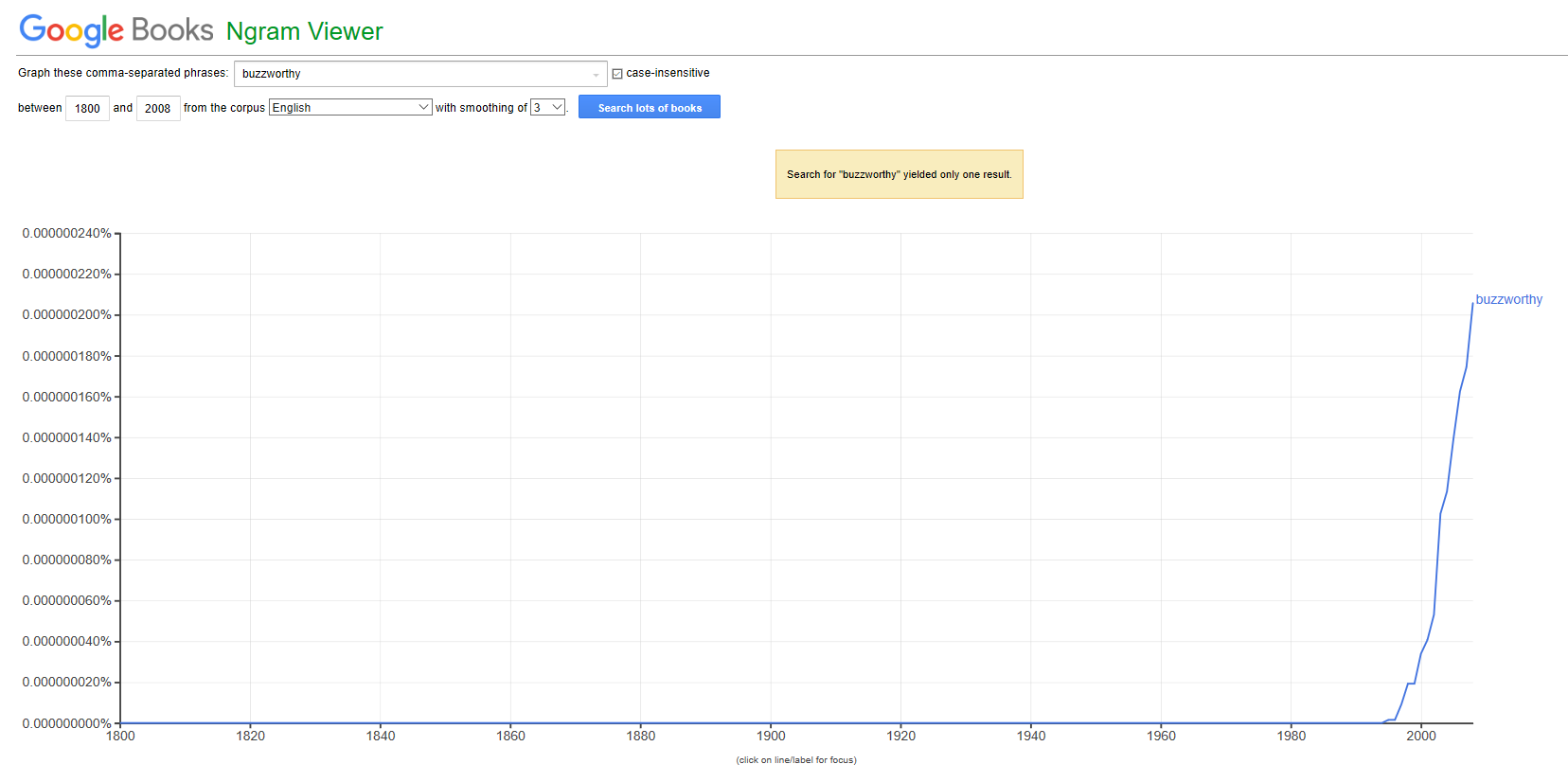


Similarly when the last name(uni gram) is fed to the Google Ngram Viewer, it plots the following 3 graphs which shows the gradual elevation during 1950s to late 1970s and further a steady rate of 0.0000120% from 1980s to 2000s. Reason being the number of books(Advances in Pharmacology and Chemotherapy), articles(Foundation of Software technology and Theoretical Computer Sciences, Consumer Behavior), annual reports(Annual Report - Central Soil Salinity Research Institute.), bibliographies(Bibliography of agriculture) etc during the time period.



**Ans2c.** A new English word recently added to the dictionary **“buzzworthy” [1]** (Mental Floss, *mentalfloss.com/article/31363/35-modern-words-recently-added-dictionary*) is given as the input to the Google Ngram Viewer and it depicts the below graph.

According to my assumptions, it should have been added during the 2010 and so on but as the graph displays it occurs most frequently during the year 2000 and so on. It is because of the news(Maroon 5 Announce Their First National Headline Tour; 'Harder to Breathe' 'Buzzworthy' at MTV in the year 2003), books(Connected Marketing: The Viral, Buzz and Word of Mouth Revolution during 2005-2006, Poetic Arts and new Media in 2007, Beautiful Liars , etc in 2008 ), Notebook Journals published majorly in 2019World Records,Trending topics etc.

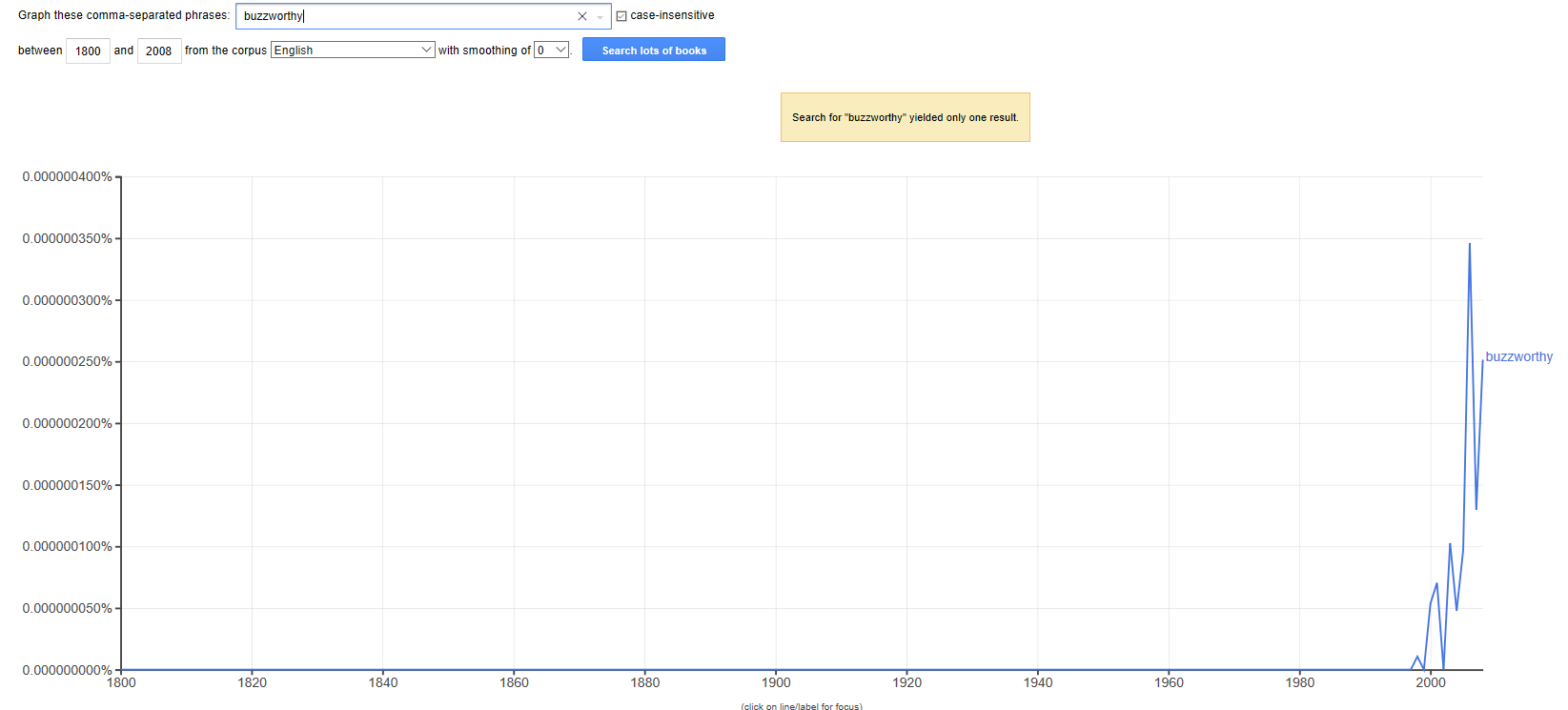


**Ans2d.** Smoothing means how smooth the graph is at the ends. The most accurate representation depicts a smoothing level of 0, but difficult to read. The default value is 3. Example: Smoothing=1 means that the data shown for x year will be an average of the raw count for (x) plus 1 value on either side: ("count for (x-1)" + "count for x" + "count for x+1"), divided by 3.

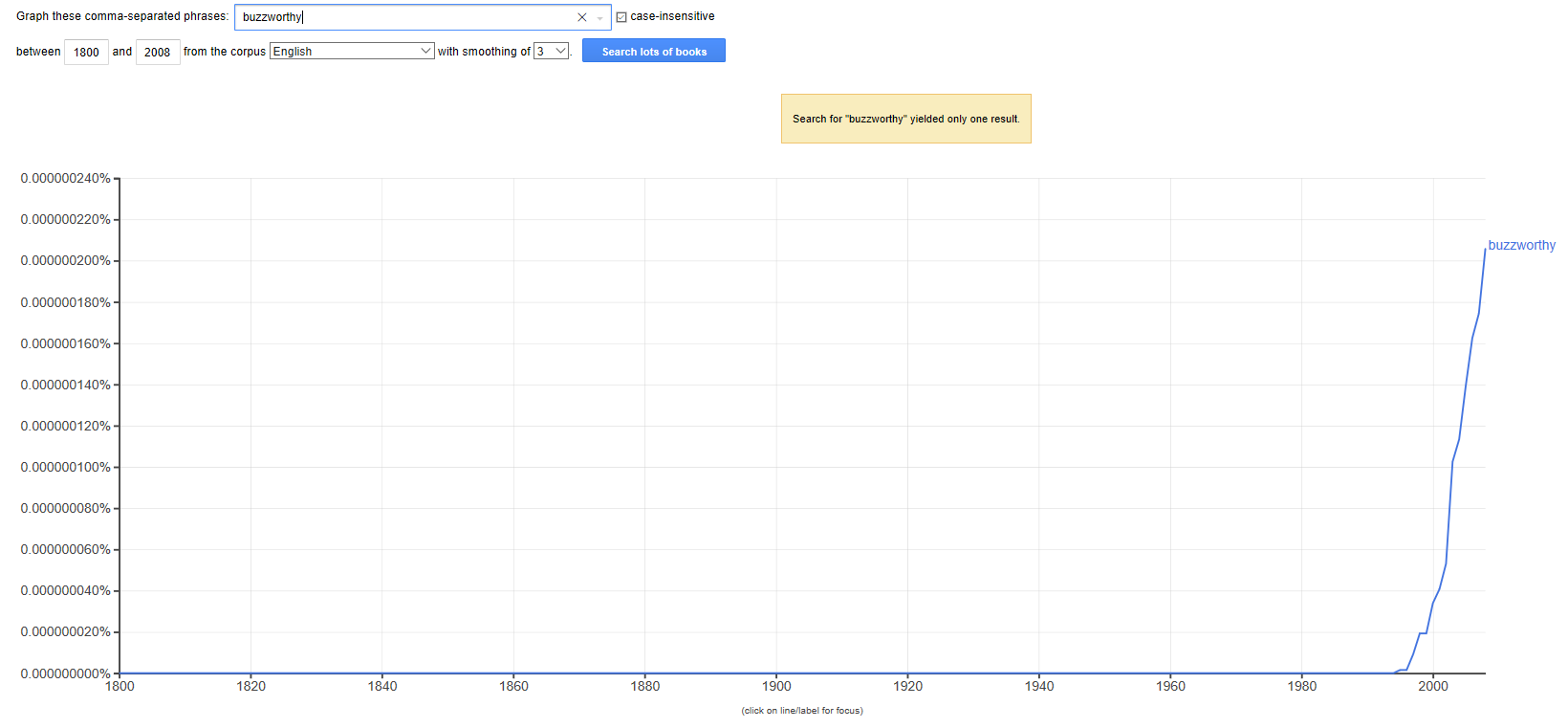
**Smoothing of 0, 3 and 30 is displayed below respectively.**

**Smoothing of 0 depicts the complete raw data. It shows the peak values but it is difficult to read whereas the smoothing of 30 depicts more clear curvy graph between the years as it will average total 61(Left 30 + itself + Right 30) values and it is easier to read them but no clear peaks can be found in this.**

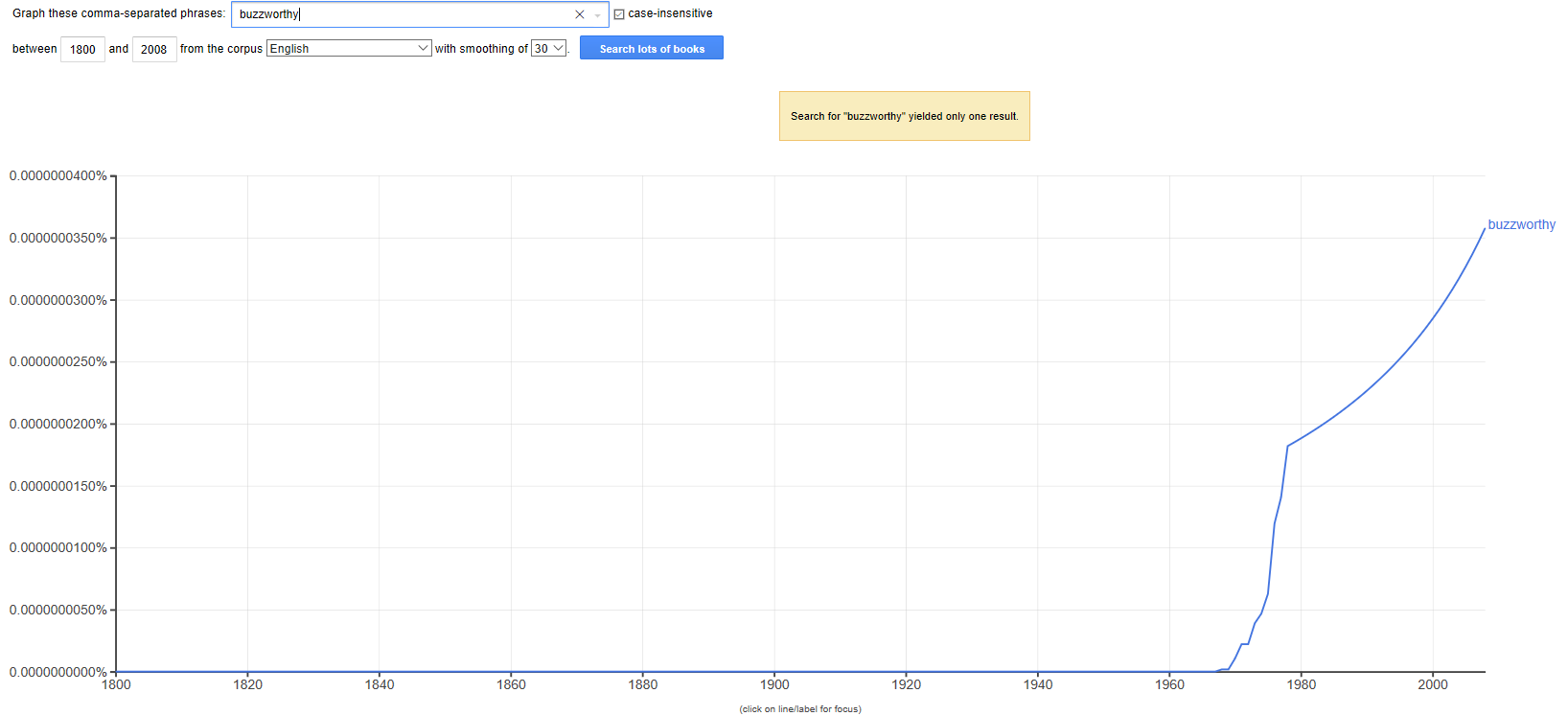
Smoothing = 0



Smoothing =3



Smoothing = 30



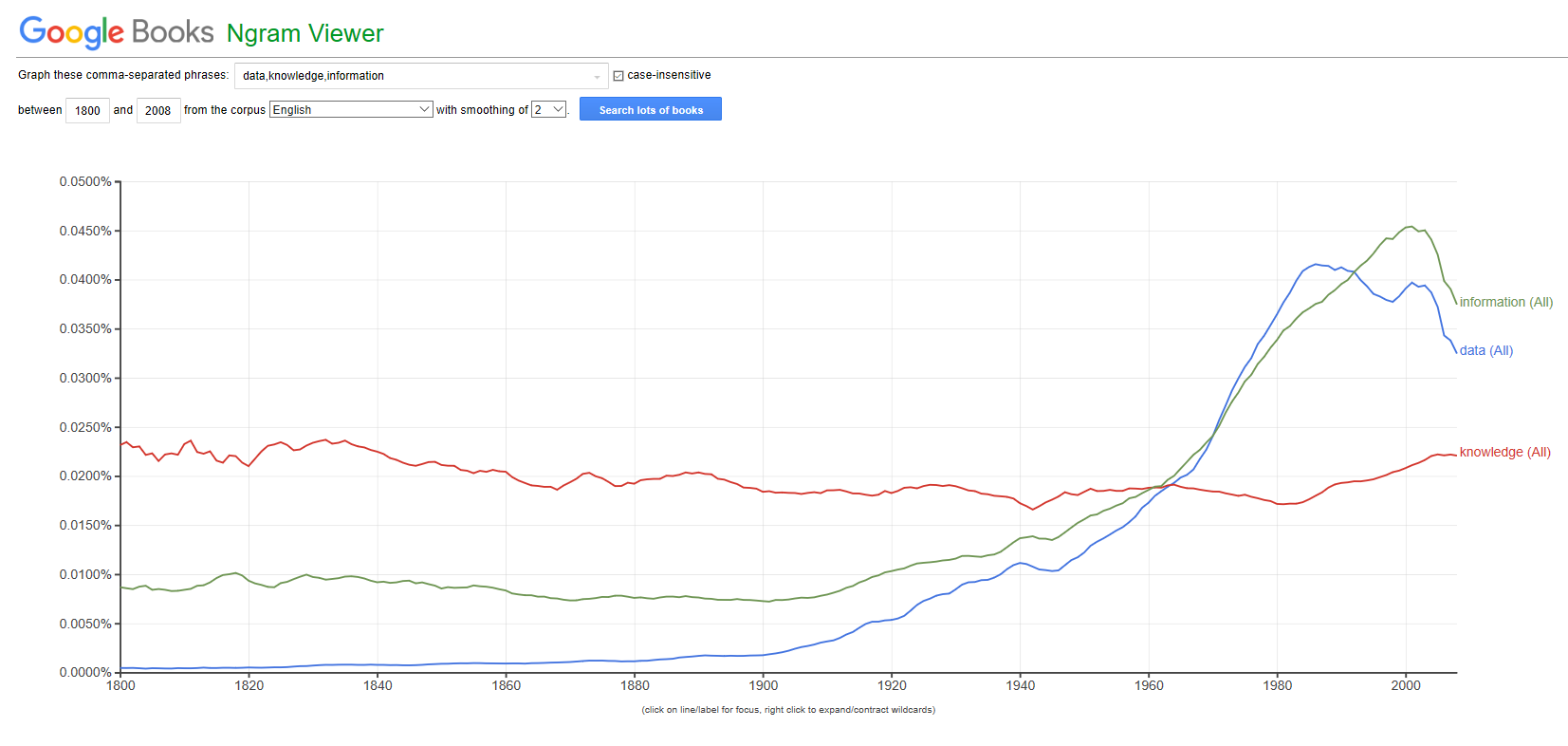
**Ans2e.** The 3 related terms (data, knowledge,information) between 1800 and 2008 with smoothing 2 are visualized by Google Ngram Viewer.

**Till 1900s, ‘data‘ was least searched/frequent. Thereafter it has a continuous growth hitting its peak in 1980s and 2000s.**

**Further ‘knowledge’ was the most frequent in 1800s and has a steady rate of growth throughout.**

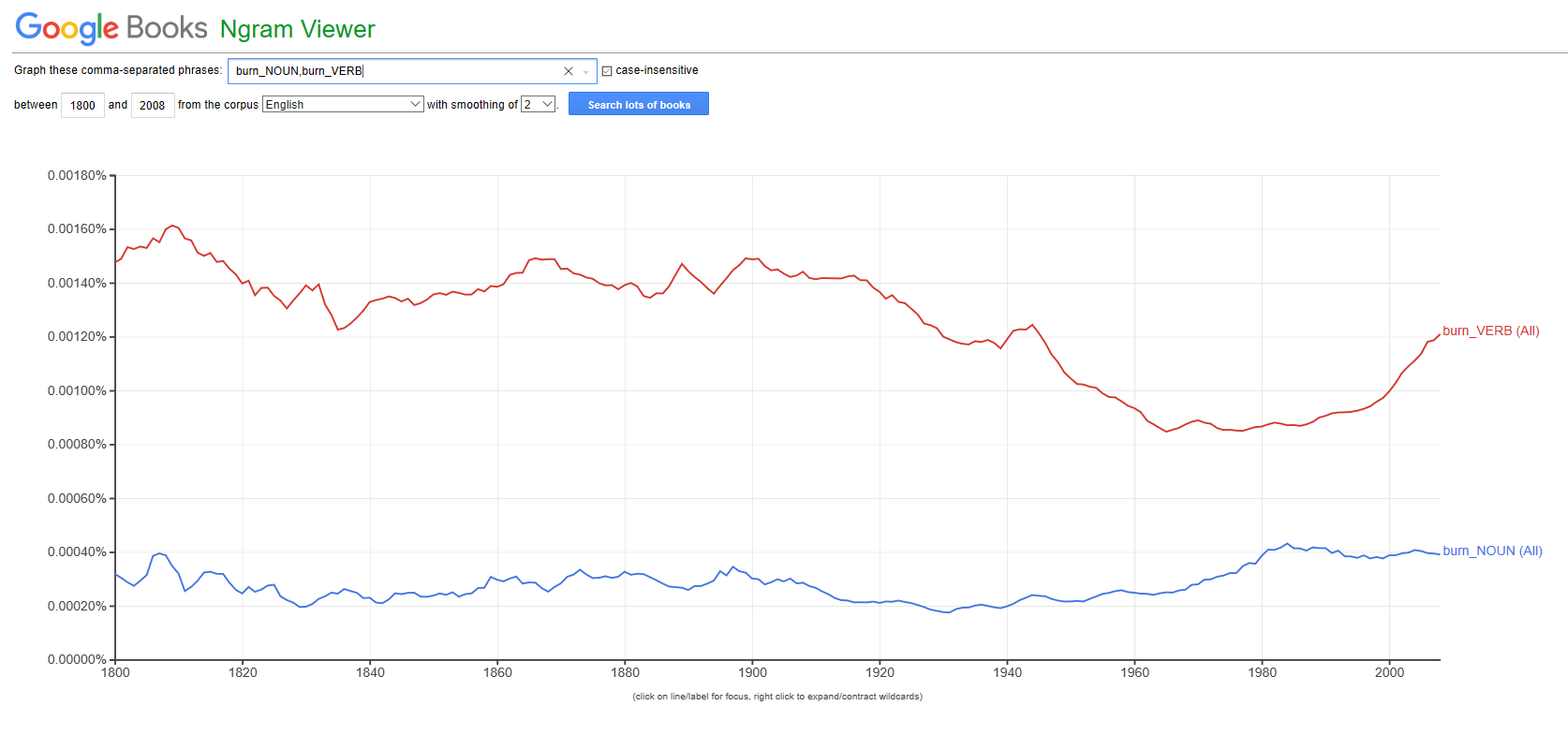
**Last ‘information’ has the medium frequency among 3 and after 1940s, it took a shoot hitting the peak in2000s becoming the most frequent in the recent times.**

All the 3 terms are clashing during the period of 1960s. Surprising thing is: ‘data’ which is least frequent for a long duration of time (1800-1920) becomes most frequent in1980s beating knowledge and information,although knowledge was the most frequent throughout. Over the period of time, processed data(information) becomes more searched/frequent term due to number of books, articles(Principles of information Security,Handbook of information resource management) in the era of 1980s till now. In last few decades, everything revolves around Information and Information Systems were boom throughout.



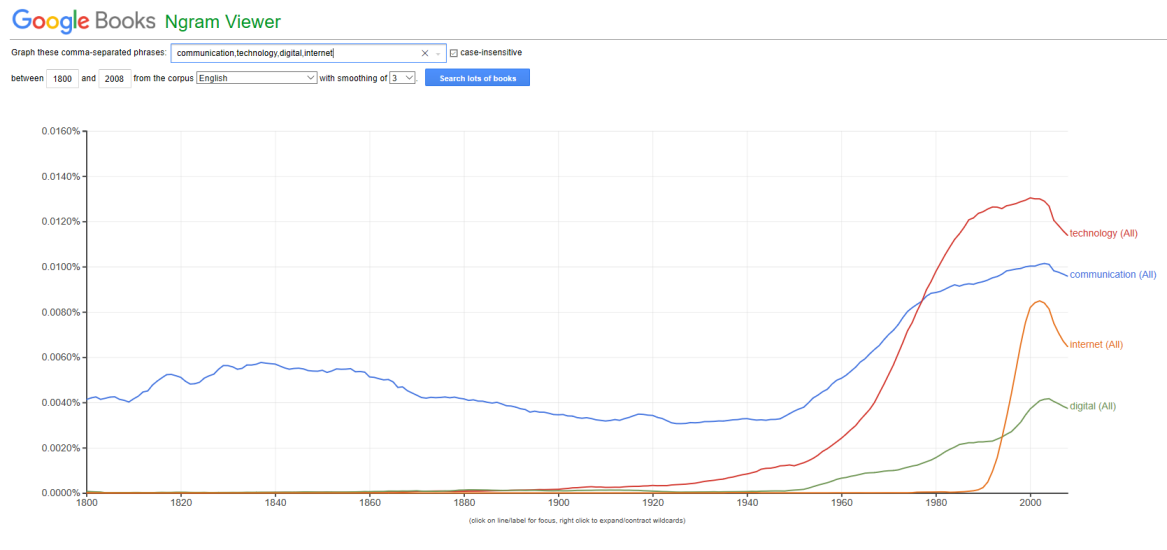
**Ans2f.** The term both noun and verb taken is **‘burn’**. The graph is displayed below:

The ‘burn\_NOUN’ has a steady growth rather low frequency throughout as it is less used comparatively whereas ‘burn’ as a VERB is mostly used and that’s why it is most frequent among the 2 throughout.



**Ans2g. ‘The Internet’** is the cultural change happened over last 500 years majorly last 50 years and the related terms denoting this event are: **Communication, Technology, Digital, Internet.**

Graph is depicted below:



Here, **Technology** increases gradually from 1920s continuously till now hitting the peak in 2000 but the frequency of Technology is almost negligible from 1800 till 1920s.

**Communication** from the beginning has a steady growth till 1950 then it shoots hitting the peak in 2000.(Example: Social Media)

**Internet** becomes the buzz from late 90s, growing popularly till now.Example: Huge digital database of answers for User queries.

The term **Digital** has negligible growth till 1950 and then growing but still lowest frequent among all (when things started digitizing rather then manual work) Example: Mails over Letters

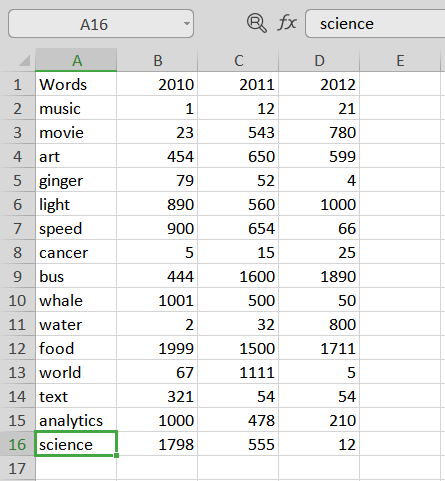
Q3. Using an Excel spreadsheet set up your own list of 15 words and give each a made-up frequency between 0 and 2000 for each of three years (2010, 2011, 2012). Now perform two different normalization on them:

a. Method1: produce a normalized frequency for each word in each year, using the total N of words over all the years (i.e., Grand Total)

b. Method2: produce a normalized frequency for each word in each year, using the total n of words in a given year

c. Does normalizing by method1 or method2 make a big difference to the scores produced? Graph the difference and comment on it.

**Ans3.** Snippet of 15 words with made up frequency between (0-2000) across 2010,2011,2012 is displayed:

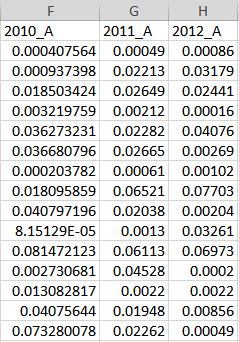


**Ans3a. Method 1 Normalization:**

Here the grand total is the total sum of all the frequency across all the 3 years (2010,2011,2012).

Further Normalized Frequency of each word using the grand total = freq(word)/grand\_total.

Output of Method1 is attached below:

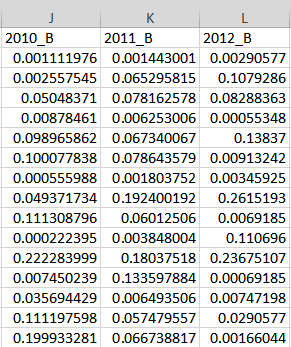


**Ans3b. Method 2 Normalization:**

The total sum of the frequency of each year 2010/2011/2012 = sum of frequency of year 2010/2011/2012.

Further Normalized Frequency of each word in each year = freq(word)/sum\_freq(particular year).

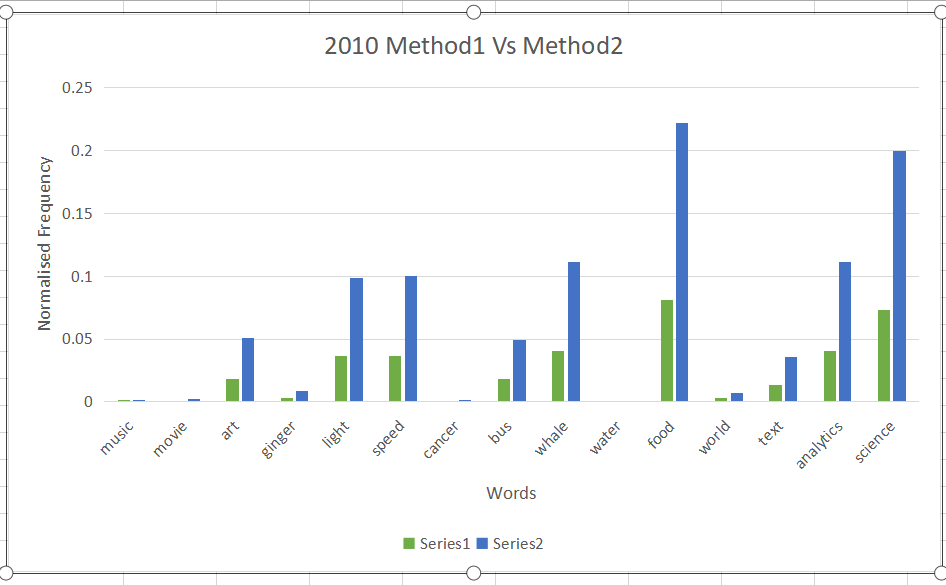
Output of Method2 is attached below:



**Ans3c. Differences between method 1 and 2 over the years 2010,2011,2012.**

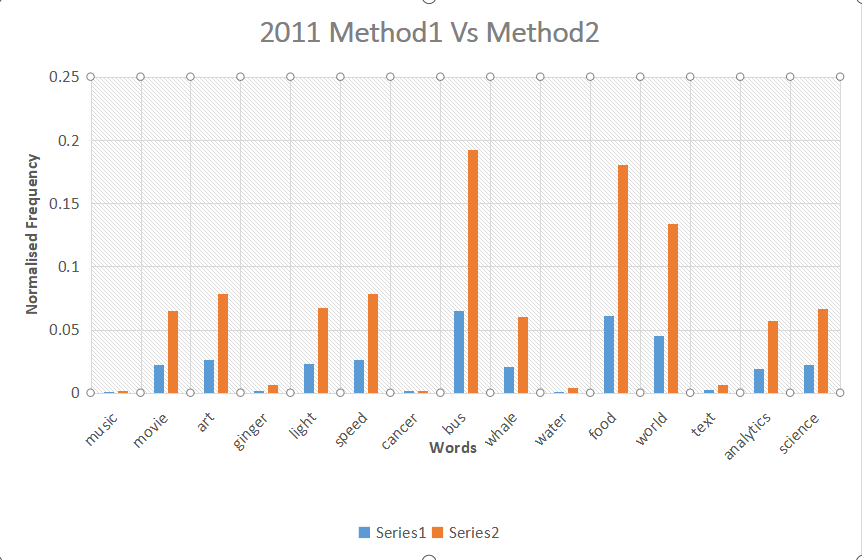
**2010 Method1 Vs Method2 is shown below:**

Here Series1 depicts the Method 1 2010 and Series 2 depicts Method 2 2010. Example: food having freq 1999 in 2010 gets normalized to 0.08 by Method 1 and 0.22 by Method 2.



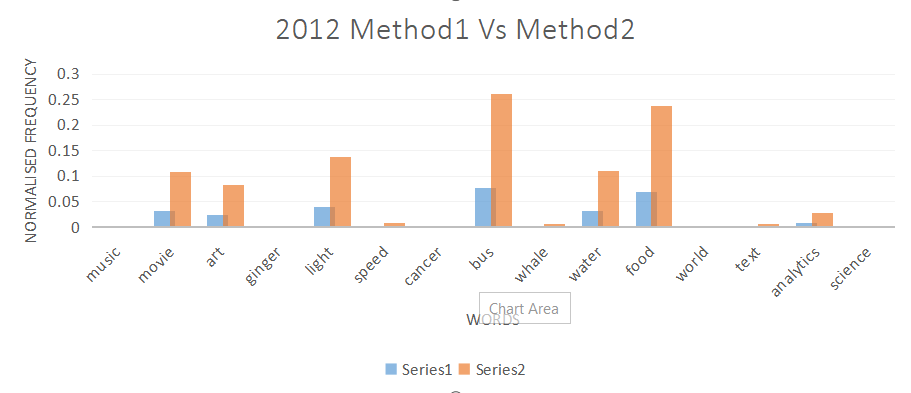
**2011 Method1 Vs Method2 is shown below:**

Here Series1 depicts the Method 1 2011 and Series 2 depicts Method 2 2011. Example: bus having freq 1600 in 2011 gets normalized to 0.065 by Method 1 and 0.192 by Method 2.



**2012 Method1 Vs Method2 is shown below:**

Here Series1 depicts the Method 1 2012 and Series 2 depicts Method 2 2012. Example: bus having freq 1890 in 2012 gets normalized to 0.077 by Method 1 and 0.26 by Method 2.



**Differences:**

- Method 1 does normalize frequency using the total sum of all the frequencies across the years(2010,2011,2012).

- Method 2 does normalize frequency using the sum of frequency of an individual year(say 2010).

- It depends on the condition that if it is required to normalize the word frequency on the (total corpus based across all years) total summation basis, then Method 1 can be used.

- And if it is required to normalize the word frequency using n of words of a (corpus based on individual year) particular year, then Method 2 can be used.

- Method 2 is better over Method 1 as it calculates the total of each year and thus it has less bias and less noise. Also it generates a common summation of frequency against which all the frequencies are compared.

- Method 2 normalized the word frequencies in between 0 and 1(broader scope) whereas Method 1 normalized the word frequencies to a lower scope between (0 and 0.1).

- Method 2 has impact over Method 1 as it deals with specific frequency of corpus instead of whole corpus.

**==========================================================================================**

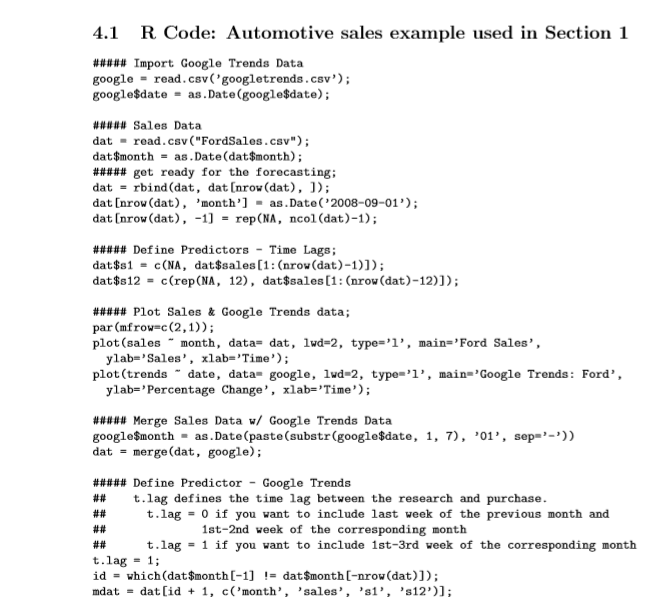
Q4 Find the article by Choi & Varian (2009/2011/2012) and find the R program they give for their Ford prediction model. What do you need to do to run this program? Can you do this?

**Ans4.** The article by Choi & Varian(2009/2011/2012) is found at:[2] (Apr 10 2009, Hyunyoung Choi,Hal Varian,Predicting the Present with Google Trends [https://static.googleusercontent.com/media/www.google.com/en//googleblogs/pdfs/google\_predicting\_the\_present.pdf](https://static.googleusercontent.com/media/www.google.com/en/googleblogs/pdfs/google_predicting_the_present.pdf) )

- The R program for Ford prediction model is also found out within the same paper: Predicting the Present with Google Trends (2009/2011/2012) is found at: (Apr 10 2009, Hyunyoung Choi, Hal Varian, Predicting the Present with Google Trends [https://static.googleusercontent.com/media/www.google.com/en//googleblogs/pdfs/google\_predicting\_the\_present.pdf](https://static.googleusercontent.com/media/www.google.com/en/googleblogs/pdfs/google_predicting_the_present.pdf) )

A dataset of Ford Sales (FordSales.csv) is required to run this program but not available with us. So, could not execute the program.

Snippet of the code is attached below:



**REFERENCES**

**[1]** (Mental Floss, *mentalfloss.com/article/31363/35-modern-words-recently-added-dictionary*)

**[2]** (Apr 10 2009, Hyunyoung Choi,Hal Varian,Predicting the Present with Google Trends [https://static.googleusercontent.com/media/www.google.com/en//googleblogs/pdfs/google\_predicting\_the\_present.pdf](https://static.googleusercontent.com/media/www.google.com/en/googleblogs/pdfs/google_predicting_the_present.pdf) )