Question 1 Understand R library "wordcloud"

Following R libraries are imported into R runtime to attend question 1.

- Wordcloud Ver2.6[dependencies : function {'RColorBrewer'}]
- tm Ver0.7-7 [dependencies : packages {'NLP', 'slam', 'xml2', 'BH'}]

1(a) Carry out the commands shown in the practical notes

> library(wordcloud)
> library(tm)
Loading required package: NLP
> wordcloud("May our children and our children's children to a thousand generations, continue to enjoy the benefits conferred on us by 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)
Warning messages:
1: In tm_map.SimpleCorpus(corpus, tm::removePunctuation) : transformation drops documents
2: In tm_map.SimpleCorpus(corpus, function(x) tm:: removeWords(x, tm:: stopwords ())) : transformation drops documents

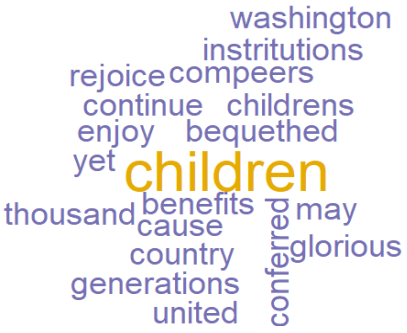


Figure 1 output of 1(a)

1(b) 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 some are excluded and others included?

Words in wordcloud : [**'may',** 'children', ***'childrens',*** 'thousand', 'generations', 'continue', 'enjoy', 'benefits', 'conferred', 'united', 'country', 'cause', 'yet', 'rejoice', 'glorious', 'institutions', 'bequeathed', 'washington', 'compeers']
Words excluded in wordcloud : ['our', 'and', 'our', ***'children's',*** 'to', 'a', 'to', 'the', 'on', 'us', 'by', 'and', 'have', 'to', 'under', 'those', 'us', 'by', 'and', 'his']
Hypothesis : As indicated in Warning message during execution, punctuations and stop words in TM::NLP package are dropped. Default list of stop words is from English [1]. Word children is repeated multiple times hence emphasized with bigger font and central position. Sequence of operation is possibly:

- **Drop punctuations** : Word "children's" becomes "childrens"; which is treated as a whole word. In reality, "children" should have an "NNS" tag and "'s" should have a "POS" tag and treated separately.
- **Assign Importance** : First word "May" is a modal verb and listed as a stop word [1]. But, it is assigned more importance due to its context or position in sentence and preserved. If wordcloud algorithm internally considers PoS tagging as basis of importance, then there is possibility that modal verb "may" is erroneously PoS tagged as noun – the month of "May"; hence preserved.
- **Normalize by Lowercasing** : Lowercasing must be after importance tagging, else a lowercase word "may" is highly likely to be treated as a stop word. stop word. Also, this sequence of operation would preserve semantic nature of proper nouns like "Washington".
- **Drop stop words** : stop words in list , tm:: stopwords (<default: "en">) are dropped

1(c) Check hypothesis about why the wordcloud package is including some words and excluding others. Put your own word-list together (of 10- 20 words, try to repeat some words a few times) and check what wordcloud includes and excludes? Report whether your initial hypothesis was right or wrong and why?

Sample Text = "Hear me the rejoice! Children and saved by Thanos of Titan. Smile.. for even in the death, you are rejoice children the Thanos!"
Words in wordcloud : ['hear', 'rejoice', 'children', 'saved', 'thanos', 'titan', 'smile', 'even', 'death']
Words excluded in wordcloud : ['me', 'the', 'and', 'are', 'by', 'of', 'for', 'in', 'you', 'are']

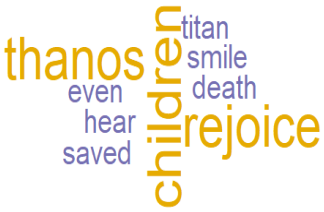


Figure 2 output of 1(c)

Hypothesis made for 1(b) partially stands. In the sample set of words, total 4 words are repeated viz. 'children', 'rejoice', 'thanos'. These words are marked as high frequency words in wordcloud output. "The" is included in wordcloud in spite of being a stop word. Other stop words are correctly omitted.

1(d) See what happens when using word-list with more repeated words (use the same word multiple times)

Sample Text = "Bitty Butter bought butter, but the butter was bitter. Bitty Butter bought another butter to make bitter butter better."
Words in wordcloud : [butter ']
Words excluded in wordcloud : ['bitty', 'bought', 'but', 'the', 'was', 'bitter', 'bitty', 'bought', 'another', 'to', 'make', 'bitter', 'better']

Figure 3 output of 1(d)_1

Sample text contains multiple repeated words like 'bitty', 'butter', 'bought', 'bitter'. Here, Algorithm drops punctuations and removes stop words. But, since "min.freq" parameter is not included [2]; wordcloud picked up most relevant word butter and missed other repetitive Nouns, Adjectives and Verbs like 'bitty', 'bought', 'bitter', ''better' etc**.** [refer Figure3]

- After setting "min.freq = 1" parameter, results improve and all words of importance are included in wordcloud. [refer Figure4]
- Also, words are conditionally treated as stop words. Hence, when "was" and "the" are capitalized as "Was" and "The"; they are considered as part of word cloud. [refer Figure5]

Figure 4 output with min.freq=1

Figure 5 capitalized stop words "The,

Question 2 Find the Google Ngram Viewer online and perform the following operations

Google NGram Viewer is an online search engine that charts the frequencies of any set of search strings using a yearly count of n-grams found in sources printed between 1500 and 2019 in Google's text corpora in English, Chinese, French, German, Hebrew, Italian, Russian, or Spanish.

2(1) Put in "Mark Keane" as a search term and look at the graph shown. You can trace the books that refer to this name. Can you explain the peaks that appear in the graph over time.

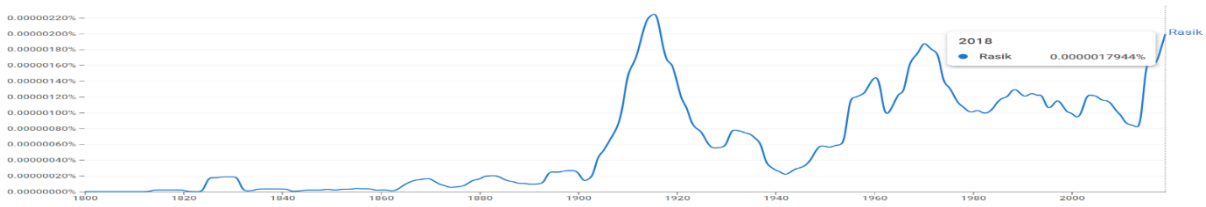
NGram shows a single mention of for name Mark Keane, in a "Report of the Society for Promoting the Education of the Poor in Ireland" during 1920s. After nearly 130 years, the graph shows a prominent spike during decade 1960-1970. These citations refer to Mark Keane who was Executive Director, International City Management Association and his exert opinion is quoted in multiple documents of US government bodies. Peak in 1955 also belong to Mr. Kean, when he was Village Manager in Oak Park, Illinois [3]. The

presence of citations drop post 1075 sharply. Post mid-1980s, Professor Mark Keane's work starts surfacing, which rises from 1990s till 2002; then has steady decline in citations till 2010 and again a steady rise till now. Some of top results returned by significance are - Cognitive Psychology: A Student's Handbook, Computational Approaches to Analogical Reasoning: Current Trends, Machine learning and Data mining in pattern recognition etc. [4].

2(b) Put your own name in and describe what happens, explaining where the hits are coming from. If there are none, then change your name until it starts to produce hits for it.

When I searched NGram for my full name [Rasik Kane], I am not able to find any results; which suggests that combination for my [name Surname] combination has not been used in any published material. When I searched for NGram of my name "Rasik"; it returned results from 1814 till today. 3 humps are observed in decades of 1860s, 1880s and 1890s. Prominent peak 1920s. Majority off these mentions are from "Bengal law reports" and "The modern vernacular of Hindustan"; both during British Raj in India. In the 1915, maxima in the observations is recorded, due to Calcutta University chemistry scholar Rasik Lal Datta, who patented Sulphur retrieval process during the time. After a trough in 1940, NGram starts getting more

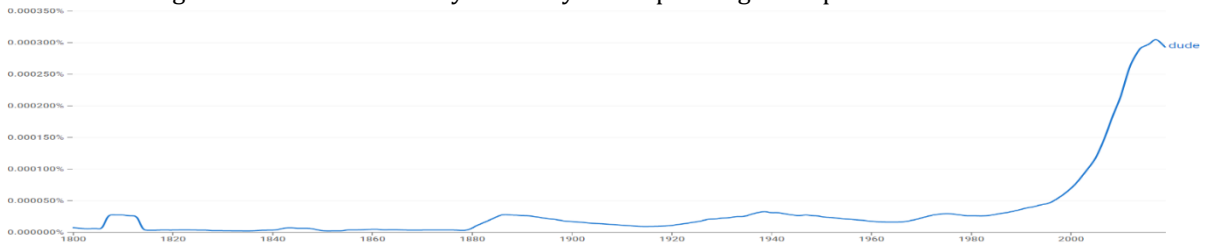
hits in due to documented courtroom records and use in English literature. This cumulates into a peak in 1970; supplemented by mention of a place "Rasik" in US's official Standard name gazetteer on USSR.



Rasik [Sanskrit] literally means a connoisseur [of arts and music]. Post 1980s, NGram experiences up and downs with sharp increase over decade 1910-20. It is mentioned in literature, in context of a 1956 song "Rasik Balama" and Indian classical music – wherever audience are addressed.

2(c) Pick a word that you think is a recent introduction into the English language (like "exit strategy") and plot its emergence. Explain reason If it actually emerges before you thought.

I assumed that "Dude" must be a post WW2 collegiate term originated in The US and might have been colloquialized across globe after American cultural influence in internet age. However, it is old American word dating back to 1870s meant for a person who is conspicuous citified fashionable person visiting a rural location. Yet, the later part of my thought process sustains as use of the word takes a rapid rise since 1980s. The usage of word dude is mainly fueled by books painting concepts of modern manhood.



2(d) Describe some of the effects of smoothening these graphs with different values?

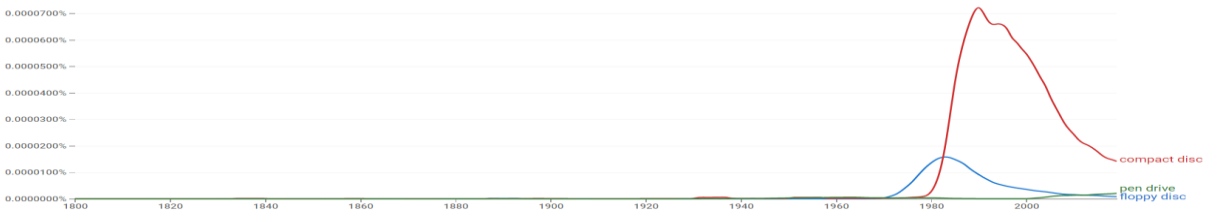
The Smoothening option in NGram Viewer is averaging results. By default, the value selected here in 3, and a value of 0 can be set for precise results.

Mathematically, for smoothing by factor 'n', result of each year is averaged including n previous years and n following years. Corpus is made by randomly dropping and connecting results. So, for year 2010 with smoothening value = 3; results from 2007 → 2012 are taken and the observations are averaged.

Smoothening reduces spikes and leaps in result.

2(e) Do a comparison between 3 or more related terms to see how their relative frequencies have changed over time (e.g., winter, summer, autumn, spring; do not use these ones). Is there anything surprising about how these terms differ in their frequency and, if so, why? Why do you think the frequencies vary in these patterns?

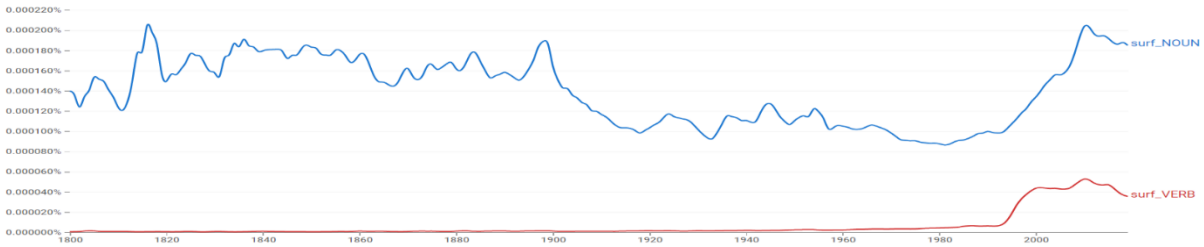
Related Terms chosen for comparison are: Floppy Disc, Compact disc and pen drives – which were prominent offline data sharing mediums since arrival of mini computers. Usage of these terms is governed by the development of technology. Rightly, term "floppy discs" was in use since 1970 and attained a peak in 1981. Era of compact discs overtook usage of floppy discs, which peaked around 1991. Usage of Compact discs diminished thereafter in 90s; which was due to increasing availability of internet, web browsers and search engines. After USB implementers Forum standardized USB protocol, Pen drives became a very hyped product in 21st century and consequently, usage of words floppy discs and compact discs is decreased.



2(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. You will need to research how to do this in Ngram viewer.

Google NGram searches words head to head by spelling. To assign syntactic meaning, we can attach part of speech to a searched word. On google NGram viewer, words can be inputted in syntax – [word]_[Part of Speech]. I looked up for word SURF in following context :

- Surf_NOUN : flakes of surf on the tides of ocean [Noun]

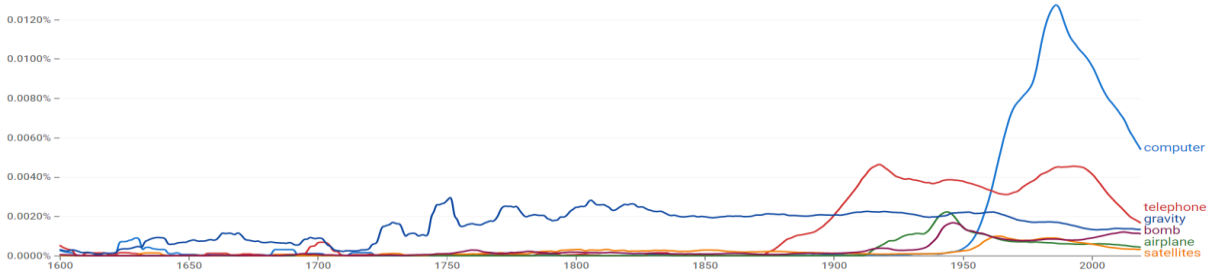


- Surf_VERB : pastime game of surfing on the tides with a float board [Verb]

Term surf is used as a noun extensively in 19th century and after a trough in 19th century, it is steadily inclining. Verb surf is associated with game of surfing since 18th century. Naturally, with increasing popularity of game [as it is introduced in coming Olympics], usage of verb is on the rise since 1990s. Act of "Internet" surfing or browsing also adds up to recent usage of "surf" as a verb.

2(g) Think of some major cultural change that has happened over the last 500 years and some words that could denote this event(s). Check these words for the relevant time-period. Report finding.

- Term "Telephone" shows a sharp spike 1880s, which coincide with invention of the telephone by Alexander Graham Bell. Rightly, it spikes around 1917 WW1 event, after which it became a mainstay technology for military and urgent communications. Usage of keyword is rightly declining after cellular mobiles are replacing telephones. A very similar graph is seen for the "satellites". Associatively, graph of keyword "Computer" sharply from 1950s [conceptualization by Von Neumann] till 1990s, after which the use is in decline due to other forms of computers like laptop, cloud, distributed systems etc.
- Queries about physics phenomena gravity has steady rising graph. Rise around year 1748 is mainly due to Le Sage's theory of gravitation. Surprisingly, there is no significant spike around mathematical formulation of gravity by Sir Isaac newton in 1680s.
- The terms airplane show a sudden increase around world war 2 in 1942. It is interesting to see that word "Bomb" also experiences similar spike but unlike decline in frequency of term Airplane post world war, frequency of term "Bomb" increases given cold war and nuclear race etc.



Question 3 **Open an Excel spreadsheet set up a table with list of 10 words which should be the rows in the spreadsheet and give each a made-up frequency between 0 and 2000 for each of five years (2010,….2014; in the columns). Compute two things: (i) find the large-N, a count of all words across all years (i.e., the sum of all words in the set) (ii) find the small-n, for all the words in each year (e.g., sum of all words in each year, for each year; i.e., column totals). Perform two normalizations**

3(a), 3(b) Overall Normalization and Yearly Normalization

For the question, 10 words are assigned random frequencies for years 2010 through 2014 each using excel function "=RANDBETWEEN(0,2000)" and stored in a csv file. The data is normalized in python script:

- Sum of all words across all years [Method 1] **N = 51969**
- Sum of the words in each year [Method 2] **n = {'2010': 11326, '2011': 10004, '2012': 10840, '2013': 10848, '2014': 8951}**

In the graph for year 2012, the orange line shows global normalizing [Method 1] and the green line indicates year-wise normalizing [Method 2]. Blue line indicates original data for year 2012; which is scaled

Keyword	2010	2011	2012	2013	2014
Neumann	464	1387	1785	984	710
Babbage	1901	134	1627	1104	809
Einstein	1368	147	875	1250	1073
Edington	1095	1355	1130	647	1010
Newton	1407	327	1677	1572	1664
Faraday	1212	1334	311	1352	454
Maxwell	992	901	214	944	347
Rutherford	1426	1187	420	642	1387
Planck	139	1418	1446	1572	19
Dirack	1322	1814	1355	781	1478

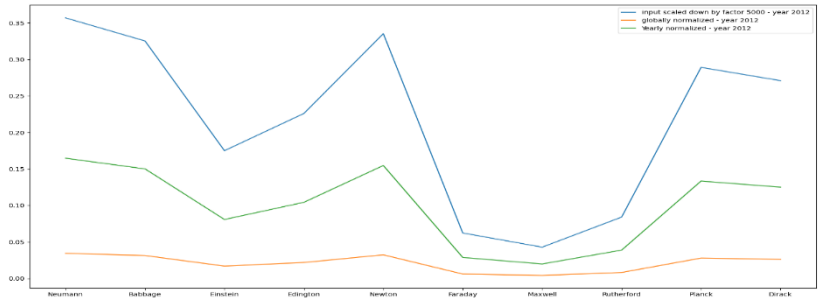


Figure 3 Input data and its values for year 2012 after normalizations

by factor of 5000 only for plotting a neat graph. N and n correlate with central tendency "mean" for their sample spaces. It appears that normalized datasets are covariant with each other as well as with original data, since their crests and troughs vary simultaneously.

3(c) Does normalizing these different ways make a big difference to the scores produced? Graph the differences you find in a histogram and comment on it.

- With overall and yearly normalization methods, choice of a method is dependent on further application. To assess effect on frequency distribution after normalization; I created a new feature in each table, which takes sum of values for each row of "original data" – "N normalized data" – "n normalized data".

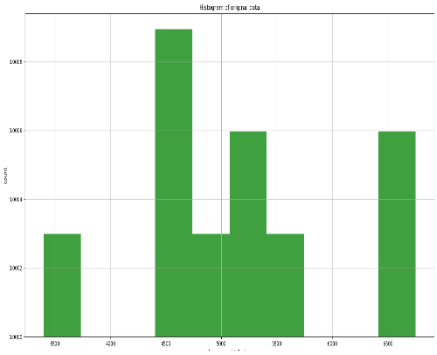


Figure 4 Original Data

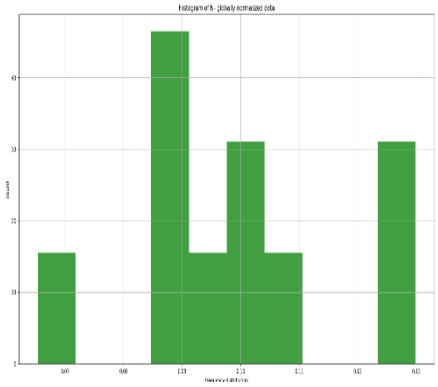


Figure 5 globally normalized (N)

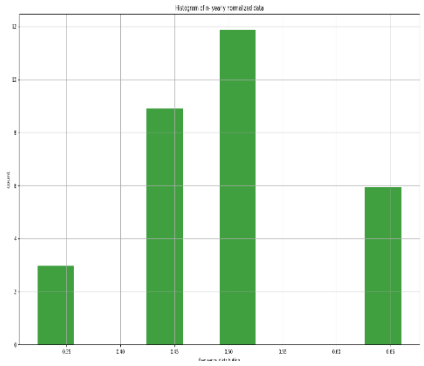


Figure 6 Year-wise normalization (n)

- As shown in figures 4, 5 and 6; histograms for sums with 10 bins shows that, Method 1 better approximates the original data distribution. This is natural as Method 1 essentially scales whole data with a constant N; which is representative central tendency of all frequencies in dataset. Whereas, yearly normalization considers frequencies for particular year.
- Also, as shown in figure 7, ranges of normalized values is not same for both normalizations. These magnitudes do not match with original magnitudes. Also, normalization magnitudes are specific to corpora for normalization [which is "entire data" for N, and "data for particular year" for n].
- Example : Frequency for word "Planck" is only 19 for year 2014. This oddity is normalized in overall Norm (N), and histograms for each word follows similar distribution of original data. Whereas, value for Year Norm (n) for 2014 drastically drops compared to other years due to small frequency of word "Planck". Due to this small normalization factor; value of frequency for other words is positively boosted. Observe distribution of "Einstein" and "Newton"; where 2014 frequency is larger than original distribution.

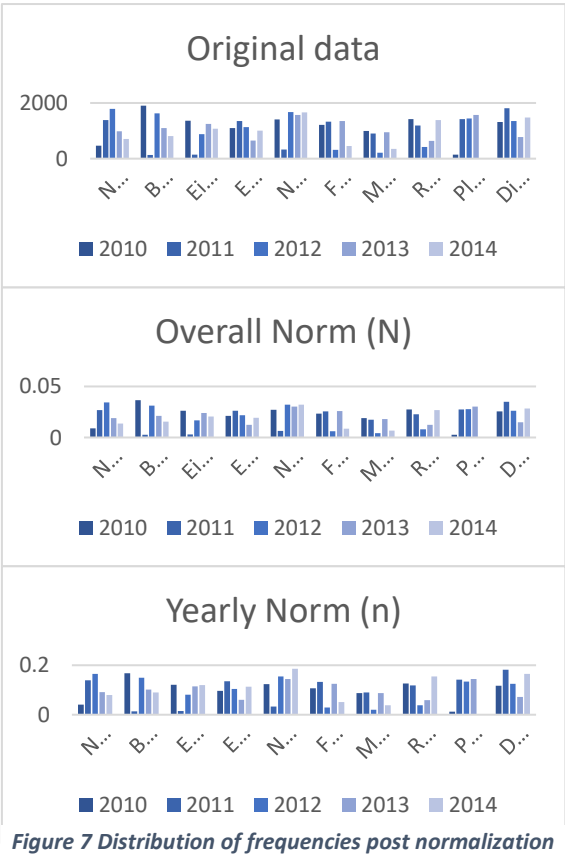


Figure 7 Distribution of frequencies post normalization

References:
[1] Removing stop words
[2] Wordcloud documentation
[3] Remembering mark e keane
[4] Mark Keane NGram for books book