Text Analytics Practical 3 Sumit Chawan -18200549

Q1.a) Creation of word cloud for given sample data :

The below picture depicts the world cloud formed using the given code snippet in the question

Q1 b)Words included in the process : children ,bequeathed, childrens, may ,continue, compeers, rejoice, may, institutions, upon , united, thousand, glorious, yet, generations, rejoice, conferred, country, cause, benefits, enjoy

Words excluded in the process : our, and, to, a, the, us, by, have, to, under, those, his.

Observations : 1)All the **stop words** were **excluded**.

2)The **words excluded are mostly neutral** that do not convey any useful summary from the text like positiveness ,negativity of the data.

3)The process tried bringing words like **children's to children that is in its root form** .

Q1.c) Sample Text :

Positivity relates to confidence that good and better things will happen. It also means that all the shots and attempts which we make will help take us closer to our goal. It's exactly opposite to negative thinking which is being uncertain, panic, fearful ,restless and unsure of success. Being Optimist and confident about achieving goal is positive thinking it's also a self -belief that our attempts always teach us something. Positive Thinking has sound effects and is important in overall personality development. It is positive thinking that encourages people to make good attempts and efforts to achieve Success.

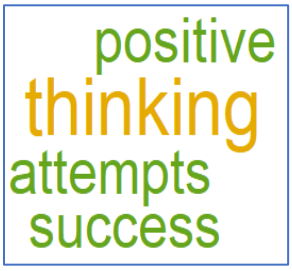
Observations :

Maximum Frequency of word : Thinking = 4

Words Included : Positive ,Thinking, attempts, success.

Words excluded : All Stop words ,Shots, fearful, uncertain, restless, unsure, negative, relates, confidence, good, happen, closer, optimist, achieving, sound, effects, overall, personality, development, *teach*, something, encouraging

Comparison between the initial Observation :



Yes my initial judgement about word Cloud not considering the neutral words is correct from the fact that :

1) It didn't consider words like Shots, sound, effects, overall, personality, something, closer which in overall summary had no positive or negative polarity.
2) Also all the stop words were excluded even in second sample text .
3) But it selects words of maximum occurrence.

Q1 .d) Observations for Word Cloud after choosing a self made sample text, increasing the frequency of particular word and package change: Task 1: Maximum word frequency = 4 for word = success.

Word Included : Success.

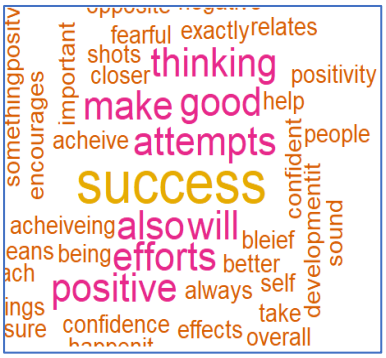
Words Excluded : All words were excluded except success which appeared 4 times in the text.

Important Observation :

1)The Word Cloud Package applies frequencies to each word and words falling with higher frequency ranges are only displayed .

2)The word Cloud package also doesn't consider upper or lowercase of the word and maps it appropriately. For e.g. : Success and success in my text shown above.

Task 2 : Behavior of Word Cloud after changing parameters :



Code snippet :

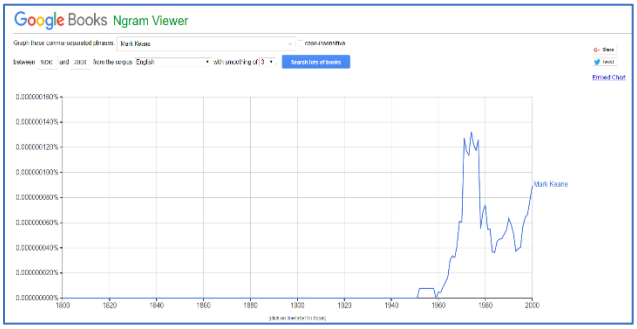
```
> wordcloud("Positivity relates to confidence that good and better things will happen.It als
o means that all the shots and attempts which we make will help take us closer to our Succes
s.Its exactly opposite to negative mindset which is being uncertain,panic, fearful ,restless
 and unsure of success. Being Optimist and confident about acheiveing success is positive th
inking its also a self -bleief that our efforts always teach us something.Positvity has soun
d effects and is important in overall personality development.It is positive thinking that e
ncourages people to make good attempts and efforts to acheive Success.",colors = brewer.pal(
6,"Dark2"),random.order = FALSE,min.freq = 1,max.words = Inf)
```

1)At the first instance, the word Cloud portrayed a normal behavior and included or excluded the words based on their occurrences.

2) After changing parameters :Minimum Frequency:**'min.freq' =1** which implies to include more words

3) Maximum words : **'max.words'=Inf** which signifies to include maximum words, almost all words were included excluding only the stop words.

Q2. A) "Mark Keane" in Google N-gram viewer :



1)The peaks shown in the graph indicate that the word "Mark Keane" was used many times in the period 1951-1980.

The name was noticed for the first time in the "APWA Reporter" in 1966 volumes 33-34 at pages 54,55 and 56 subsequently.

2)This graph was at its peak in 1974 after which there has been an extensive printing of the books which includes 'Mark Keane' like "Proceedings of the Annual Conference of Cognitive Science Society"

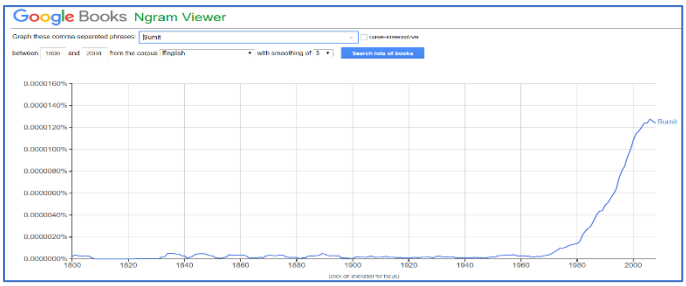
Which had many conferences as shown in the links :

https://www.google.com/search?q=%22mark%20keane%22&tbm=bks&tbs=cdr:1,cd_min:1975,cd_max:1999&lr=lang_en

Q2 b) "Sumit Chawan" as search term

1)The name "Sumit Chawan" has no mention anywhere in the books.

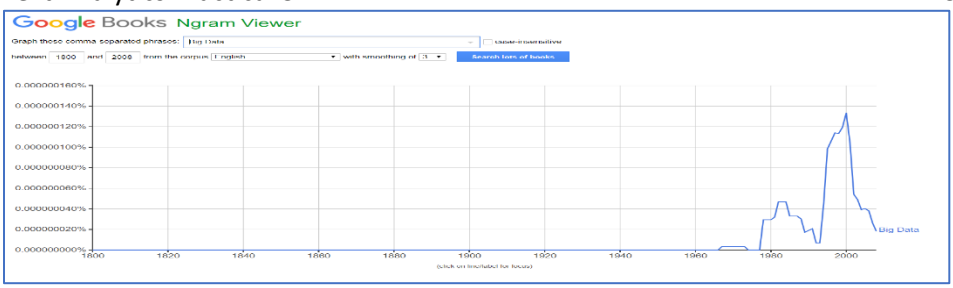
2) When the search term was changed to a **1-Gram word "Sumit"** there is a peak seen in the graph after 1980s.



3)This name Sumit may have been popular after 1980s because of the author "Sumit Talwar" who wrote book "Illustrated accounts Movement for Women" with many copies printed.

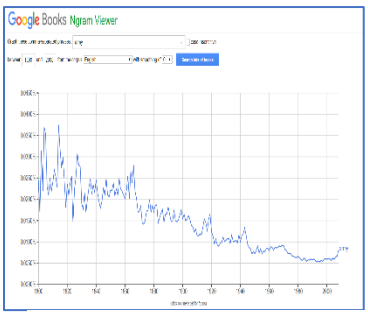
Q2 c) Search Term "Big Data"

1)The 2-Gram word "Big Data" is new introduction to English which can be seen in the graph distinctively, no mention of it anywhere in the books before 1970s

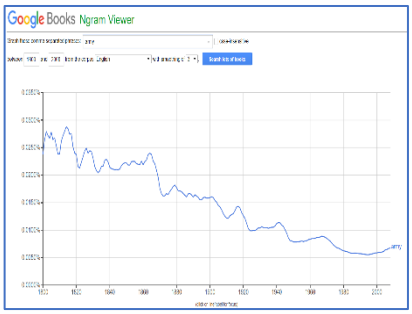


2) The word saw a steep increase from 1980s as millions of Data is getting generated each day and various theories were formulated to make use of this data in many books.

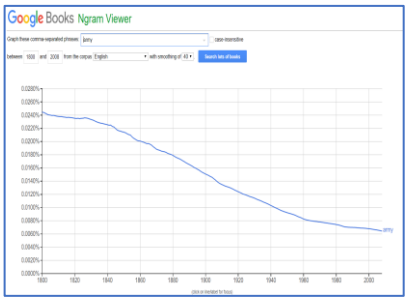
Q2 d) **Smoothening Effect on word "Army"**



Smoothening factor 0 on Army



Smoothening factor 3 on Army



Smoothening factor 40 on Army

Smoothening helps to view data as a moving average making it easy to notice trends in the data.

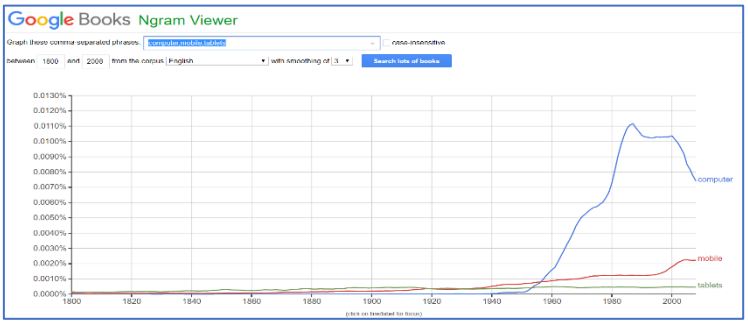
1. Smoothening factor of 0 indicates nothing but the raw data.
2. Smoothening factor of 3 on army for year say 1940 will be average of 7 values i.e. 1937,1938,1939,1940,1941,1942 and 1943.
3. Smoothening factor of 40 for instance will be an average of (40+40+1) on either sides of the selected year and the selected year.

Refrences : https://books.google.com/ngrams/info

Q2 e) **Comparison on words computer, mobile, tablets:**

1)The word Computer has a **steep increase in graph** than the other two devices mobiles and tablets.

2)Probably the **computers have greater frequencies overall** because there has been a constant development in the field of Computers and on which **books have been published to a much greater extent.**



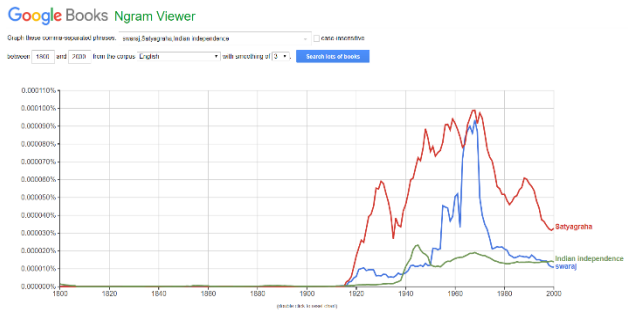
3) The **mobiles** on the other hand have shown a **little peak from 2000 because it was the time when many books and theories about mobiles (handheld devices) were published but the information on tablets was still sparsely available till 2008 and also the use of tablets is less in relation to other two devices which leads to lesser mention of tablets in relation to others.**

Q2 f) Report on syntactically different words(Bow-Noun),(Bow-verb)

1)The Noun form of word Bow was highly used in early period as it means a weapon and hardly its used in 20th century due to availability of many more destructive weapons.

2) As per the graph both the noun and verb form of bow is hardly used in any books as the graphs tends to lean on the negative scale on the y-axis.

Q2 g) Words ' Satyagraha, 'Swaraj',' Indian Independence' signify Time Period difference :



1)The words 'satyagraha' and 'swaraj' meaning "fighting with peace" and "self rule" respectively, started taking peak in India for the Indian Independence. Before 1918 there was no mention of these words .These were the major ideologies on which the fight for the Indian Independence was based

2)India attained independence in 1947 after which the above two words were mentioned in many of the books between 1950 -1980 which is exactly shown by N-gram analysis in the form of peaks.

Q3) a,b) Normalization -a) Method 1 : Normalization using Grand Total

b)Method 2 : Normalization with total words in the same year.

Words and Frequency(Assumption)

Sr.No	Word	2010	2011	2012	Total
1	Hello	500	800	100	1400
2	Welcome	1200	600	200	2000
3	Computer	300	300	600	1200
4	Science	1600	1300	800	3700
5	Technical	400	500	200	1100
6	Education	1400	1200	1000	3600
7	Economics	300	400	500	1200
8	Bow	160	300	940	1400
9	Bark	600	800	900	2300
10	surface	500	700	1000	2200
11	Table	400	600	700	1700
12	Dublin	300	450	750	1500
13	Capital	900	1000	1200	3100
14	Mobile	300	600	1000	1900
15	Homework	900	700	500	2100
Grand Total		9760	10250	10390	30400

Method 1

Sr.No	Word	2010	2011	2012
1	Hello	0.01644737	0.026316	0.003289
2	Welcome	0.03947368	0.019737	0.006579
3	Computer	0.00986842	0.009868	0.019737
4	Science	0.05263158	0.042763	0.026316
5	Technical	0.01315789	0.016447	0.006579
6	Education	0.04605263	0.039474	0.032895
7	Economics	0.00986842	0.013158	0.016447
8	Bow	0.00526316	0.009868	0.030921
9	Bark	0.01973684	0.026316	0.029605
10	surface	0.01644737	0.023026	0.032895
11	Table	0.01315789	0.019737	0.023026
12	Dublin	0.00986842	0.014803	0.024671
13	Capital	0.02960526	0.032895	0.039474
14	Mobile	0.00986842	0.019737	0.032895
15	Homework	0.02960526	0.023026	0.016447

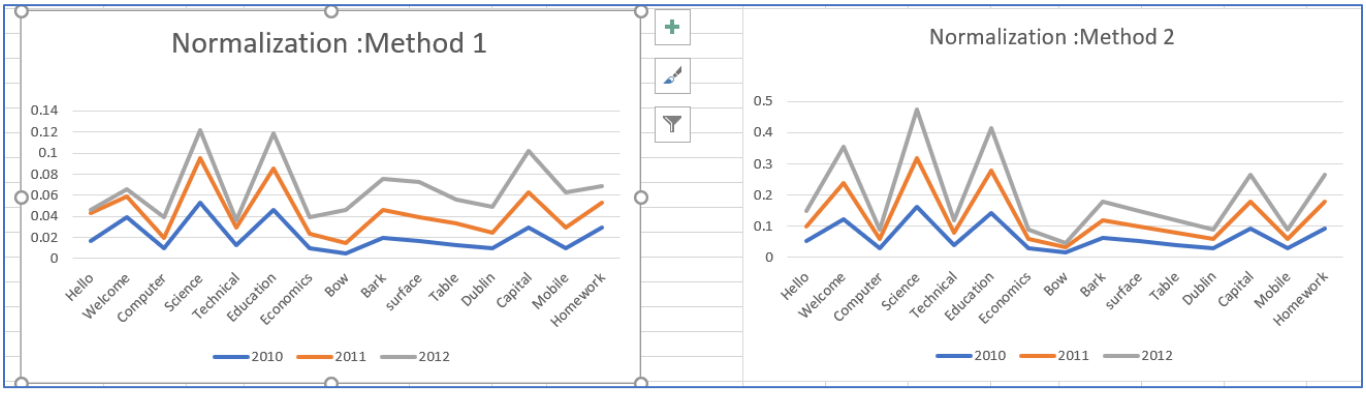
Method 2

Sr.No	Word	2010	2011	2012
1	Hello	0.05123	0.04878	0.048123
2	Welcome	0.122951	0.117073	0.115496
3	Computer	0.030738	0.029268	0.028874
4	Science	0.163934	0.156098	0.153994
5	Technical	0.040984	0.039024	0.038499
6	Education	0.143443	0.136585	0.134745
7	Economics	0.030738	0.029268	0.028874
8	Bow	0.016393	0.01561	0.015399
9	Bark	0.061475	0.058537	0.057748
10	surface	0.05123	0.04878	0.048123
11	Table	0.040984	0.039024	0.038499
12	Dublin	0.030738	0.029268	0.028874
13	Capital	0.092213	0.087805	0.086622
14	Mobile	0.030738	0.029268	0.028874
15	Homework	0.092213	0.087805	0.086622

Observation: The Normalization frequencies are almost the same for both the techniques but there may be a variation in frequencies in method I we choose Grand Total as the Normalization factor and in Method 2 we just choose only the total words in the same year.

Q3)c) The graphs in Method 1 and Method 2 show the same trend only there's a difference in the scaling factor for the plots along y-axis.

For example if we look at the word Dublin the frequencies show a constant increase each year(as assumed in the table). This characteristic is even evident in both the Normalizations plots for method 1 and 2.Hence only there is a difference of scaling factor and no change in the overall trend.



Q4) Ford Prediction Model -2009

The research paper "Predicting the Present with Google Trends" by Hyunyoung Choi and Hal Varian defines how we can use "Google Trends" and "Actual Sales" to predict the future trends. This model acts as a baseline and may be further refined to extract more relevant analysis.

R program Working :

1) I have downloaded two files "googletrends.csv" and "FordSales.csv" attached in zip file from google trends.
2) Renamed the headers in google trends to date, trends and from FordSales.csv to month ,sales respectively, as required by the code snippet in the paper.(variables)

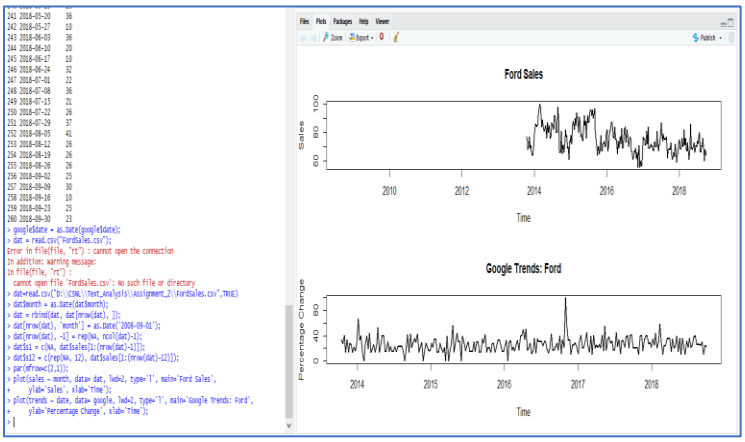
3)Time lag is one of the important functionalities taken into consideration by google trends and its used to define predictors. Both the data sources are further merged .

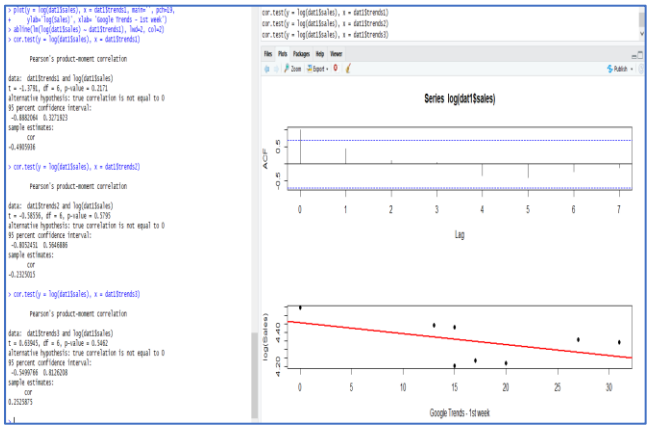
4)In the next step data is split into two parts one for fitting and the other for predicting sales in the next month.

5)In the end a fit model is extracted and summary is specified .

6)As a final step prediction function is used to display the predictions.

Code Snippets at various steps with the corresponding outputs:





Final Prediction :