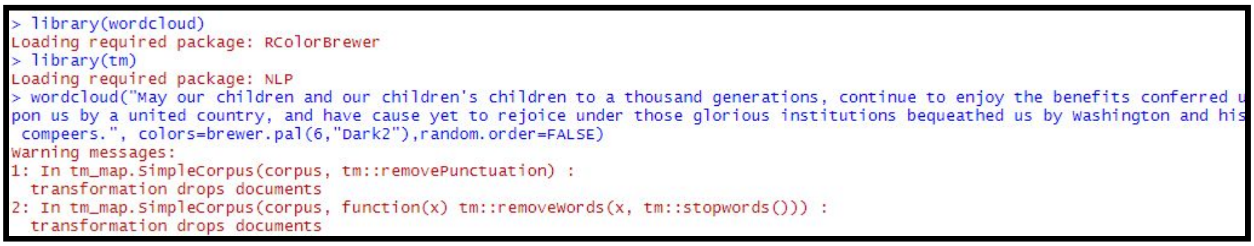
Chirag Shah | **Text Analytics COMP47600** | **06/10/2019**
19200072 | **Practical Number 3**

Q1: Assuming you have R installed (if not install it). Load up the various packages you need for using the wordcloud packages:

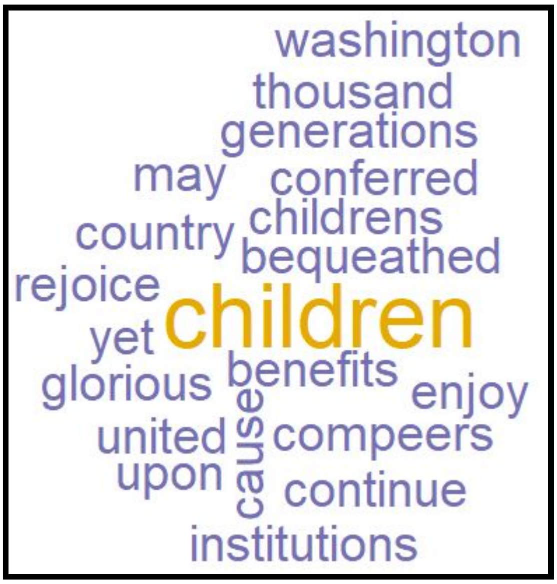
a. Carry out the commands shown in the practical notes:

Code Snippet :



```
> library(wordcloud)
Loading required package: RColorBrewer
> 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 u
pon 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)
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
```

Result:



b. When you have done this, report the list of 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?

From the results, it is observed that words such as - ***<u>to, the, of, our, under, by, his</u>*** etc. are excluded. This can be because these are stop-words and they are not as important as the other words and their frequencies in the raw text for performing text analytics.

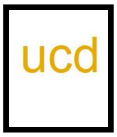
Words included	**Words excluded**
washington, thousand, generations, conferred, childrens, bequeathed, rejoice, yet, glorious, benefits, enjoy, united, upon, cause, continue, institutions, compeers	to, a, the, our, under, by, his

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?

The text used for analysis is taken from UCD website. (About UCD, 2015)

Text: "UCD was created from an idea, UCD is one of Europe's leading research-intensive universities; an environment where undergraduate education, masters, research, innovation and community engagement form dynamic spectrum of activity. Newman's vision, UCD, true enlargement of mind, embodies the aspiration to provide a holistic experience beyond classroom. Interpreted today as Think Bigger it is a rallying call education, for the university to unleash the unique potential of the individual to meet global challenges"

Result:



Word included : ucd
Words excluded :- everything else(including stop-words)

The initial theory that wordcloud package removes the stop-words is wrong because it is evident from the aforementioned results that ***wordcloud removes stop words*** _(an, is, it, of, and, a for, the,to → excluded)_ on custom and ***it also works as per the min.freq attribute*** in the package. That is, by default the _min.freq attribute is set to 3 which means that there is only one word ("ucd") that occurs >= 3 times in the custom text._

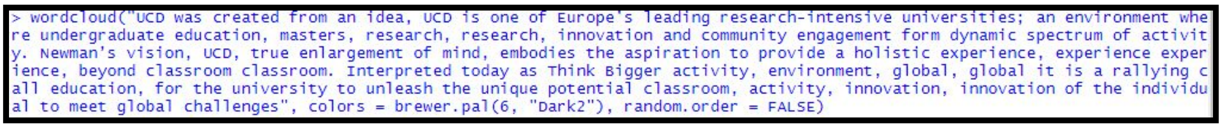
.

Also, the wordcloud package has the _"freq" attribute_. Since this attribute is not set in the afore-mentioned command, it _by default removes the stop words prior to plotting_. Hence, some words are excluded and others included. When the "freq" attribute is set; it doesn't perform stop words removal prior to plotting.

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?

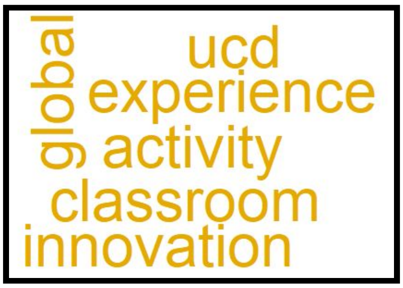
Here, a few words like **global, experience, activity, innovation, classroom, ucd** are *repeatedly added* in the text to get the following results. Hence, these are the aforementioned words that have frequency greater than min.freq = 3 in the text corpus.

Code:



```
> wordcloud("UCD was created from an idea, UCD is one of Europe's leading research-intensive universities; an environment whe
re undergraduate education, masters, research, research, innovation and community engagement form dynamic spectrum of activit
y. Newman's vision, UCD, true enlargement of mind, embodies the aspiration to provide a holistic experience, experience exper
ience, beyond classroom classroom. Interpreted today as Think Bigger activity, environment, global, global it is a rallying c
all education, for the university to unleash the unique potential classroom, activity, innovation, innovation of the individu
al to meet global challenges", colors = brewer.pal(6, "Dark2"), random.order = FALSE)
```

Result:

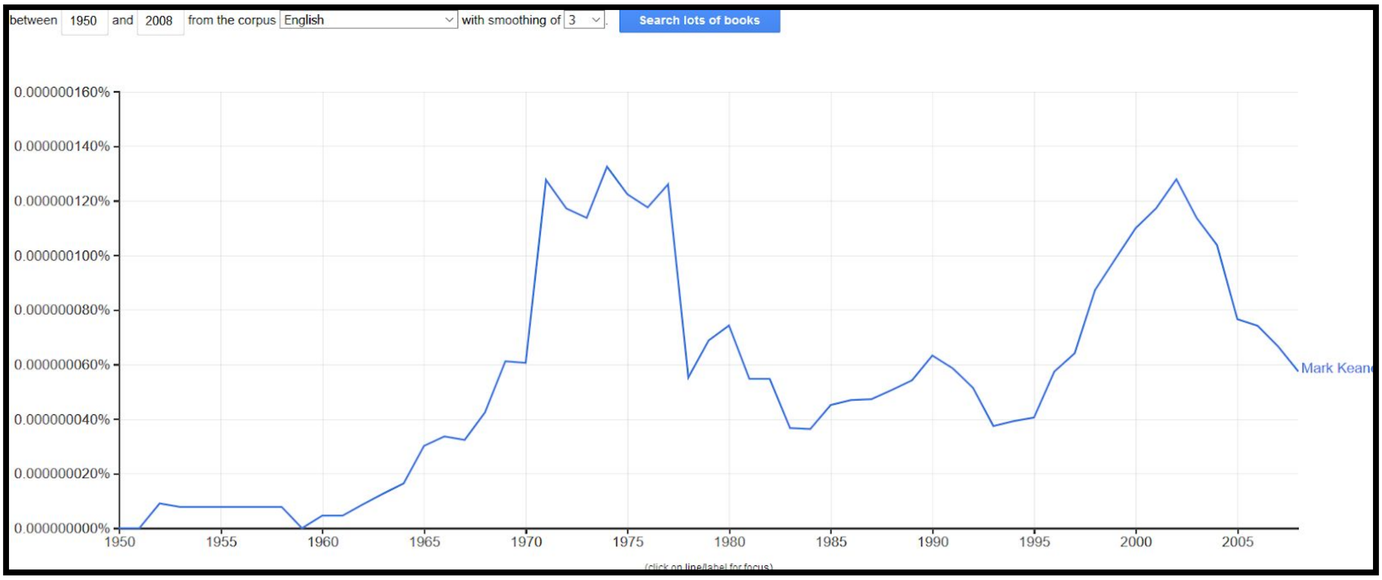


To make more words inclusive of the wordcloud, we can toggle min.freq parameter in the wordcloud package. Following is the result achieved when ***min.freq is set to 1.***

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

Google n-gram viewer displays how the search phrases occur in the corpus of books over chosen years. The most common use of n-grams is to find occurrence count of n-grams in corpus (set of documents) and then use their relative occurrence count. (Google N-gram viewer team, 2013)

a. Put in "Mark Keane" as a search term and explain the peaks that appear in the graph over time.



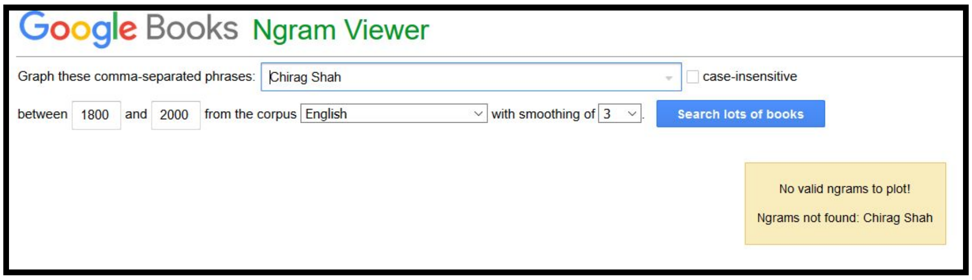
between 1950 and 2008 from the corpus English with smoothing of 3.

There are four major peaks observed which are summarised below. -

Peak period	**Observations**
Till the 1970’s	The search “Mark Keane” occurs in various books/articles/reports such as - **1.** “Hearings, Reports and Prints of the Senate Committee on the District of Columbia” **2.** Nominations of D.C. Commissioner, Assistant to Commissioner, and Nine City … **3.** Problems of Air Pollution in the District of Columbia **4.** The APWA Reporter, Volumes 33-34 etc.
1971 - 1973	There are many articles indexed with the search phrase “Mark Keane” in this time period. But they are not available in public via Google Books
1974	A newspaper article indexed the phrase “Mark Keane” in this time period
1975 - 2002	Many research articles/ papers/ reports indexed the search phrase in this period. Following are a few: **1.** Proceedings of the Nineteenth Annual Conference of the Cognitive Science **2.** Proceedings of the Twenty-Third Annual Conference of the Cognitive Science **3.** Cognitive Psychology: A Student's Handbook **4.** Advances in Case-Based Reasoning: Second European Workshop etc.
2003 - 2008	Following research articles/ papers/ reports included “Mark Keane”. Following are a few:- **1.** Cognitive Psychology: A Student's Handbook (new edition) **2.** The Baby Faced Assasin - The Biography of Manchester United's Ole Gunnar 3. Adaptive Hypermedia and Adaptive Web-Based Systems:

b. Put your own name in and describe what happens, explaining where the hits are coming from

When the name “Chirag Shah” is put as a bi-gram, there are no results observed.



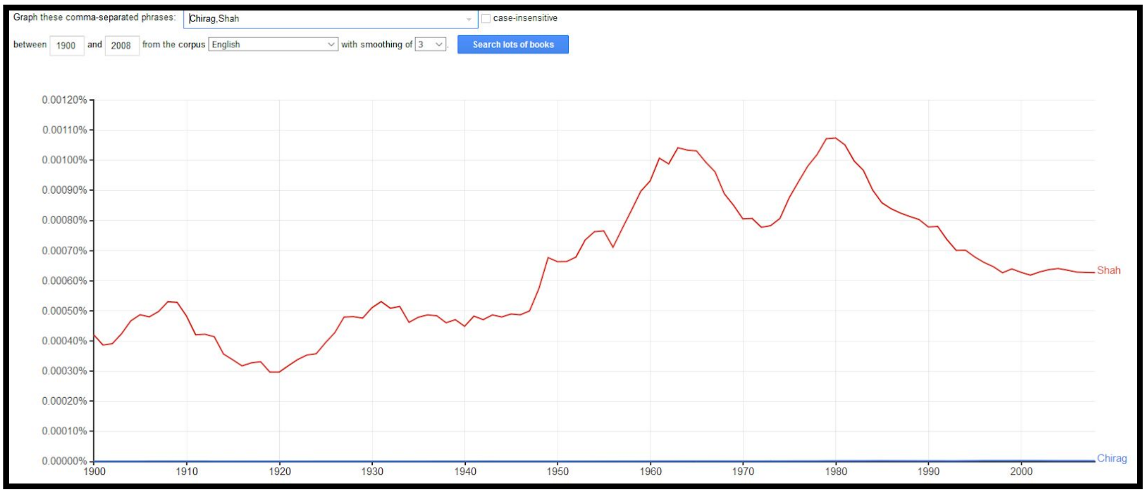
Graph these comma-separated phrases: Chirag Shah □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3

No valid ngrams to plot!

Ngrams not found: Chirag Shah

However, when each first name and last name is put separately as a unigram in the search space, they produce following results -



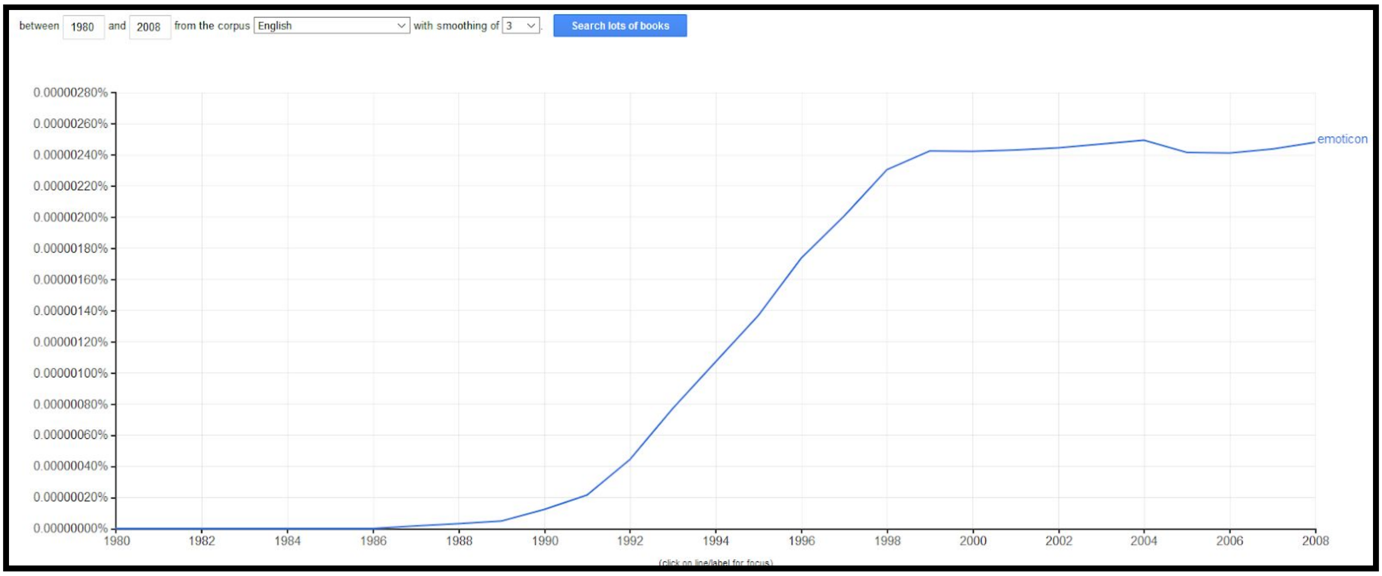
Graph these comma-separated phrases: Chirag,Shah □ case-insensitive

between 1900 and 2008 from the corpus English with smoothing of 3. Search lots of books

Also, the last name has more number of hits than the first name in this observation. These hits are produced from hits observed in books/articles/newspapers/research-papers etc.

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?

The word "emoticon" was introduced in the early 1990's. Following is the graph that shows its emergence.



between 1980 and 2008 from the corpus English with smoothing of 3. Search lots of books

Since the word is comparatively newer than other words, I expected the word to be of early 2000's. However, the word was introduced in 1990's and the early indexes for the word can be found in following books -

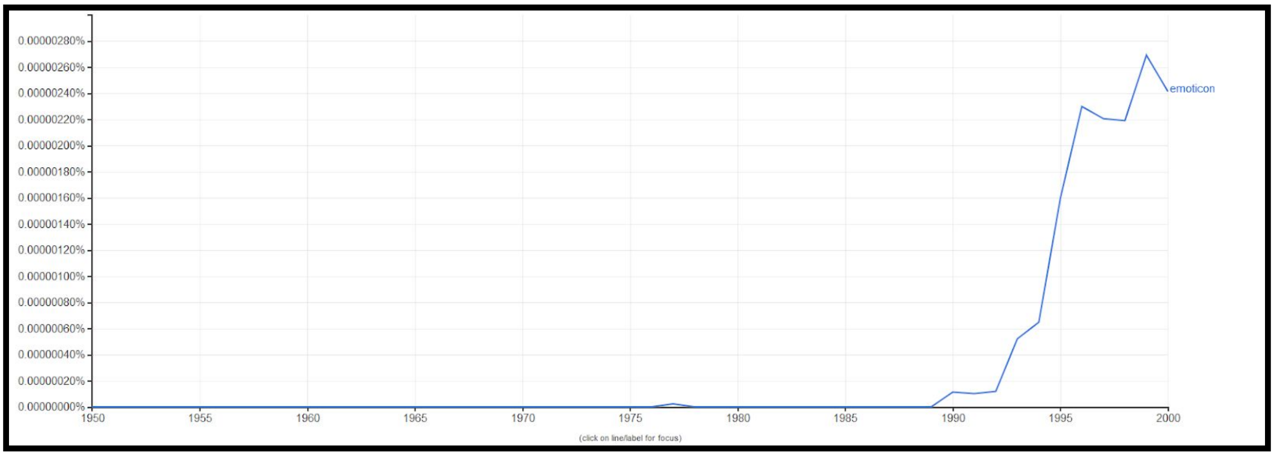
1. The Hackers Dictionary of Computer Jargon
2. The ABC's of SCO UNIX
3. The Handbook of Language, Gender, and Sexuality etc..

d. Describe some of the effects of smoothing these graphs with different values?

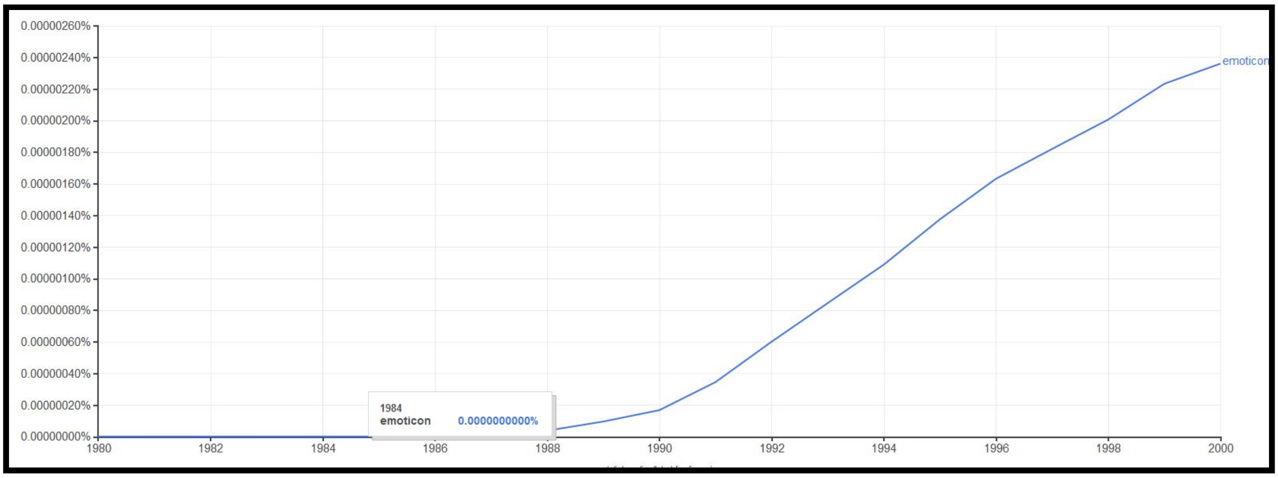
Smoothing is done to see the better patterns and trends in the existing graphical observations. Generally smoothing is performed to smooth out the irregular roughness. For seasonal data, we might smooth out the seasonality so that we can identify the trend. Smoothing doesn't provide us with a model, but it can be a good first step in describing various components in the series.

In practice it is necessary to *smooth* the probability distributions of n-gram hits by also assigning non-zero probabilities to unseen words or *n*-grams. The reason is that models derived directly from the *n*-gram frequency counts have severe problems when confronted with any *n*-grams that have not explicitly been seen before.

Effects of smoothing on the word 'emoticon' when smoothing = 0 :



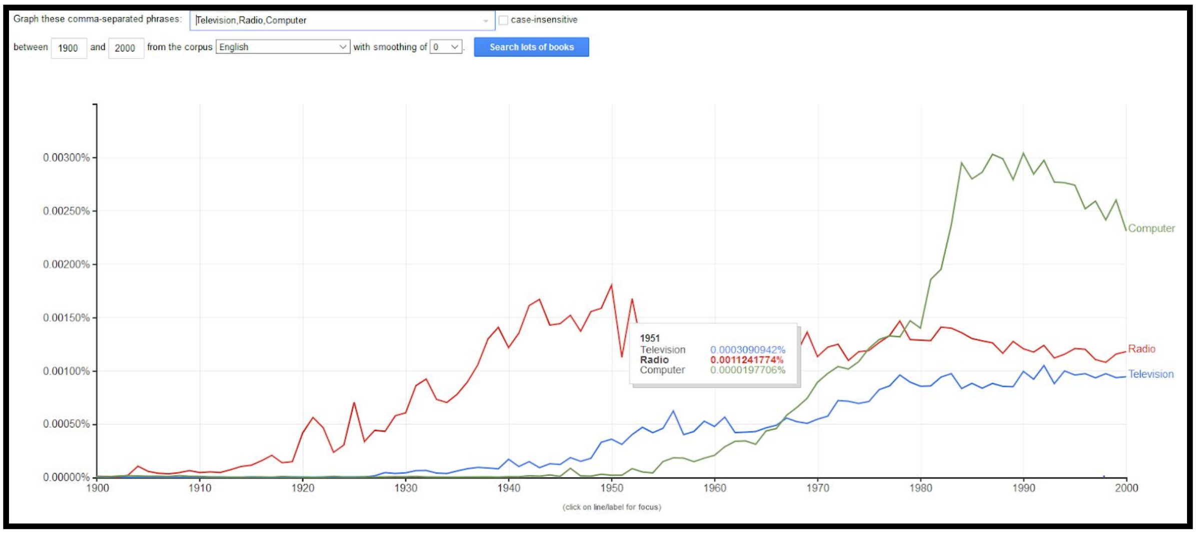
Effects of smoothing on the word 'emoticon' when smoothing = 4:



Here it is clearly observed that as we improve the smoothing from 1 to 4, ***the graph gets more clear without any irregularities. This happens because smoothing refers to moving average and it removes the average from the neighboring values. (The number of neighbors chosen is depended on the smoothing value)***

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.

The three words chosen are **television, radio and computer.**.



Graph these comma-separated phrases: Television,Radio,Computer □ case-insensitive

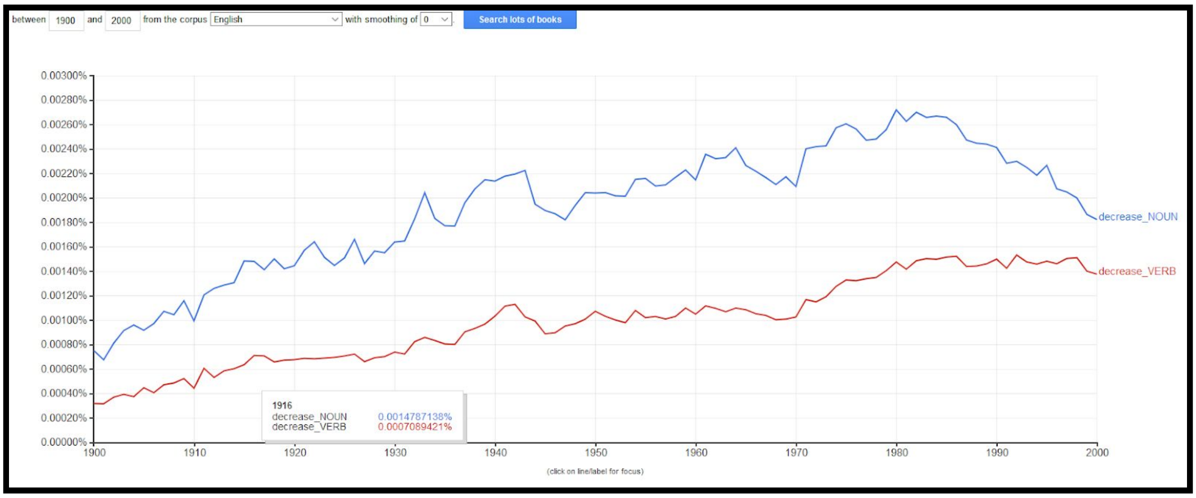
between 1900 and 2000 from the corpus English with smoothing of 0 Search lots of books

1951
Television 0.0003090942%
Radio 0.0011241774%
Computer 0.0000197706%

The surprising fact is that the word radio is indexed heavily and it is always more than television(in terms of frequency)

The frequencies vary because of the change in indexing of each term as per change in time. It is clearly observed that these terms behave as per the changes in time. Radio was the word that was most indexed before the 1980's period. This trend slowly changed and computers gained popularity post 1980's period and their indexing sky-rocketed during the early 2000's. On similar grounds, we can see that television saw a few spikes during the launch of color television in the early 1980's.

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.

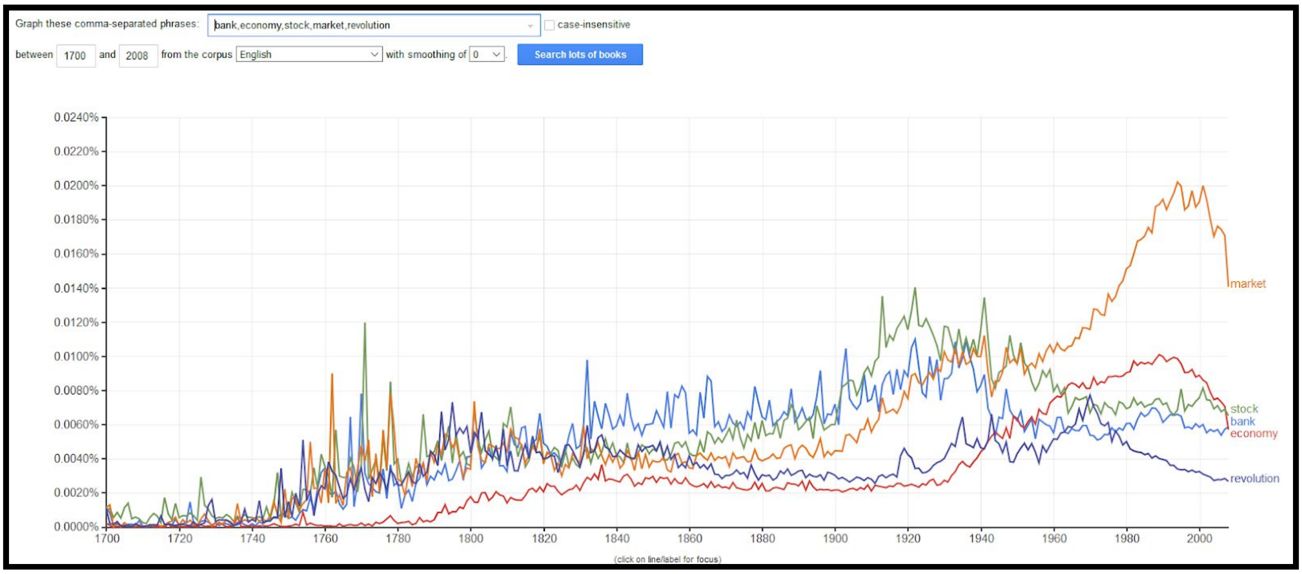


between 1900 and 2000 from the corpus English with smoothing of 0 Search lots of books

1916
decrease_NOUN 0.0014787138%
decrease_VERB 0.0007089421%

The word decrease is used as a noun and verb for the analysis. It is clearly observed that the use of decrease as a noun is heavily indexed than the use of decrease as a verb.

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



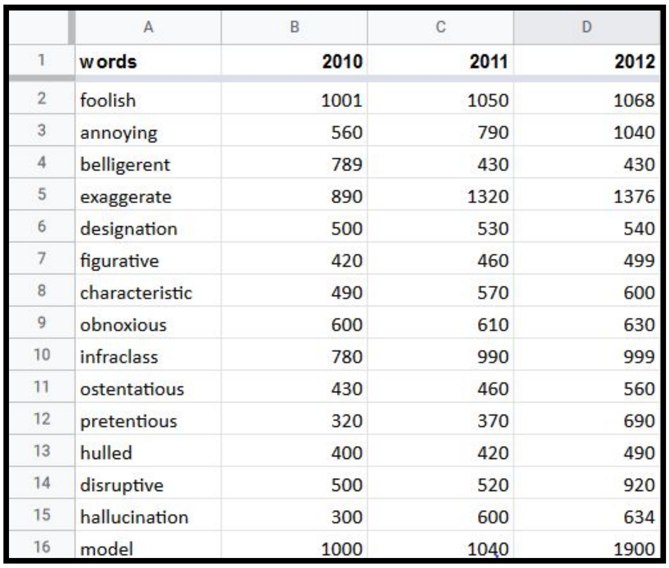
Graph these comma-separated phrases: bank,economy,stock,market,revolution □ case-insensitive

between 1700 and 2008 from the corpus English with smoothing of 0 . Search lots of books

Finance revolution is something that has caused a major impact in the last generation. The relevant words chosen pertaining to the revolution are - bank, economy, stock, market, revolution.
The indexing of words market has strongly increased in the 19th century which clearly shows the right trend of positive things happening in the finance revolution Also the word revolution has seen a downside trend post 1950's because the major events pertaining to the financial revolution took place in the early 1900's and 1800's. The terms 'stock' and 'bank' have been in a steady shape throughout the graph. There are some events like - **Worldwide great depression of 1929**. This is clearly evident from the fact that there is a steep decline in the market, stock and economy curve during the period of 1929. Also, during 2008, we can see a steep decline in all the curves because of the **housing bubble crash of Financial markets** in 2008.

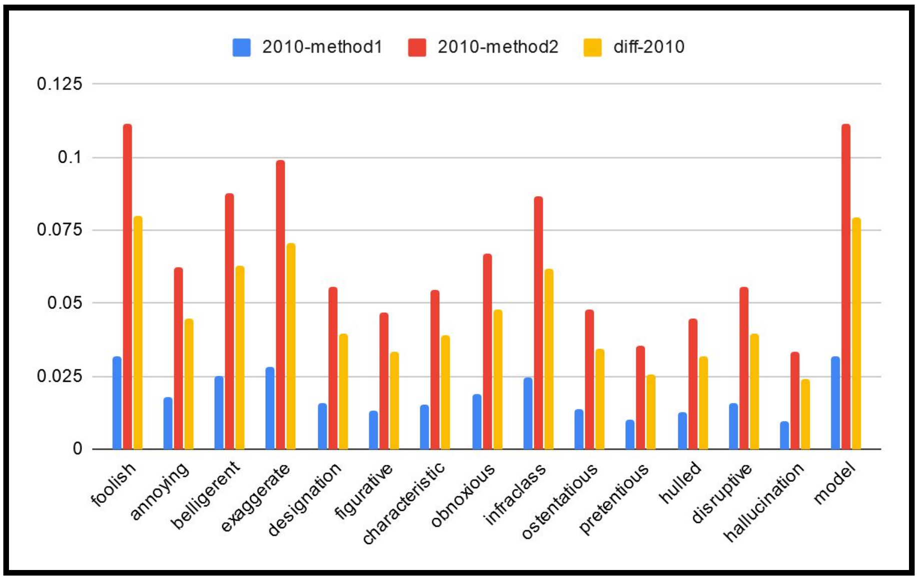
Q3: Using an Excel spreadsheet set up your own list of 15 words and give each a made-up frequency between 0 and 2000 for each of three years (2010, 2011, 2012). Now perform two different normalisations on them:

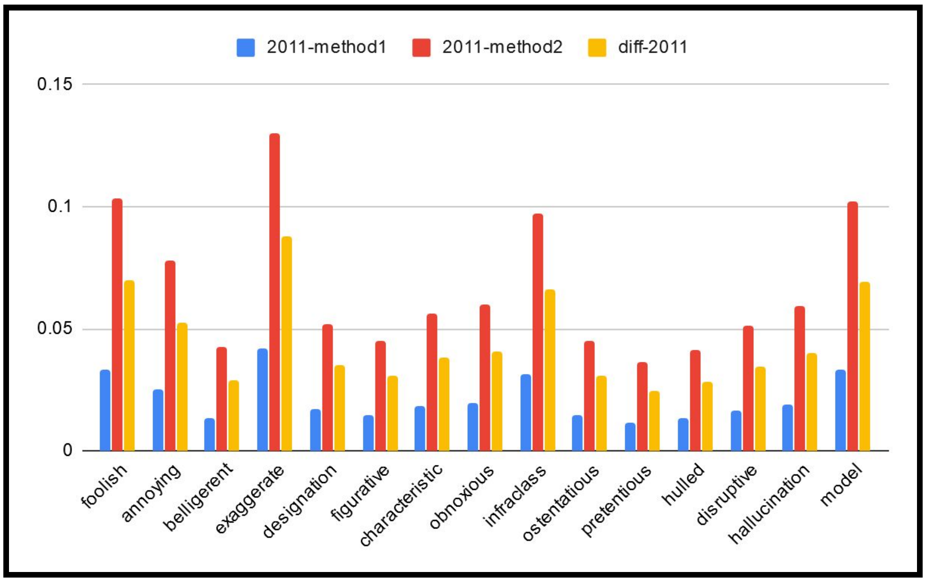
Following is the data-

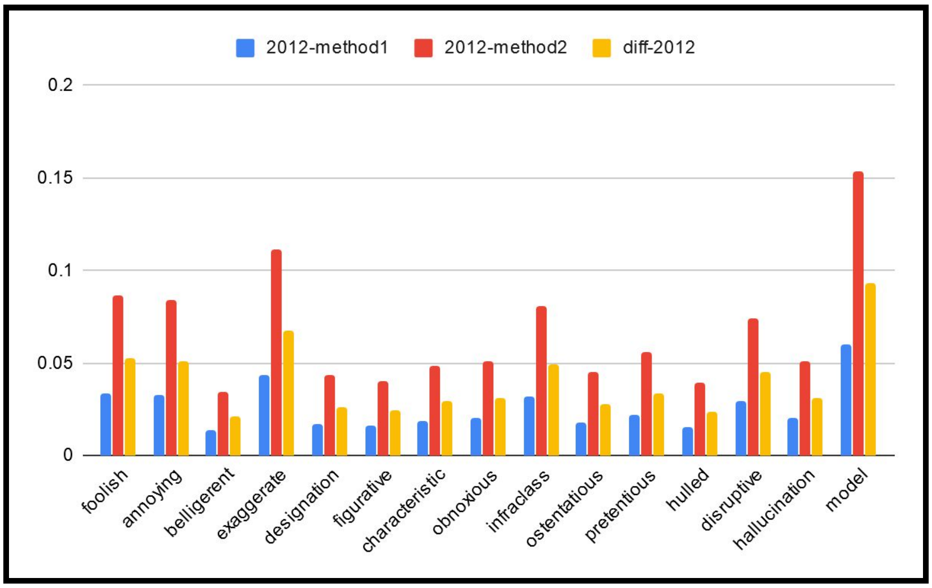


	A	B	C	D
1	words	2010	2011	2012
2	foolish	1001	1050	1068
3	annoying	560	790	1040
4	belligerent	789	430	430
5	exaggerate	890	1320	1376
6	designation	500	530	540
7	figurative	420	460	499
8	characteristic	490	570	600
9	obnoxious	600	610	630
10	infraclass	780	990	999
11	ostentatious	430	460	560
12	pretentious	320	370	690
13	hulled	400	420	490
14	disruptive	500	520	920
15	hallucination	300	600	634
16	model	1000	1040	1900

Following are the charts showcasing words on x-axis for each year and the normalised frequencies on y-axis







In each of the afore-mentioned graphs,

Blue barplot → method 1 for respective year

Red barplot → method 2 for respective year

Yellow barplot → absolute difference between freq(method 1) and freq(method 2)

→ **Method 1:-**

For each word, the corresponding normalised frequency in each year is calculated by taking the sum of frequencies from the entire corpus for 2010, 2011, 2012.

For the word foolish:
normalised_freq(foolish) → 2010 = [freq(foolish) → 2010] / sum(all_frequencies)

→ **Method 2:-**

For each word, the corresponding normalised frequency in each year is calculated by taking the sum of frequencies per year

For the word foolish:
normalised_freq(foolish) → 2010 = [freq(foolish) → 2010] / sum(all_frequencies → 2010)

→ **Difference between method 1 and method 2 :-**
Method 1 deals with the entire corpus for calculating frequencies and method 2 deals with subset of corpus(per year) for calculating frequencies.

→ **Impact of difference:**
Method 2 normalizes data in better manner according to yearly basis.This is more impactful and better as we are dealing with yearly data. On the contrary, method 1 deals with entire corpus which makes less sense when we have the frequencies segregated as per years.

→ **When to use what :**
It depends on how the normalisation has to be performed. If we are dealing with year-by-year data, method 2 has to be preferred. However, if we are dealing with the entire data in corpus, method 1 looks better.

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?

The article by Choi & Varian is available at -
https://static.googleusercontent.com/media/www.google.com/en//googleblogs/pdfs/google_predicting_the_present.pdf

The code available for the paper requires the Ford Sales dataset. SInce the dataset is not available, it is not possible to run the R program and perform predictions.

References

About UCD, (2015), *University College Dublin - About US*. Available at: http://www.ucd.ie/about-ucd/

Google Ngram Viewer info, (2013). Available at: https://books.google.com/ngrams/info#