Q1

Text used:

This is a sentence based on which a wordcloud should be displayed to my right
Sentence lacks length but makes up in content such that R can properly display it
in the plot panel. It seems that only one word has been repeated so far.
worclouds are fun to look at especially when there are multiple words that are repeated in the text

Result:

Fig 1.

Fig2. Repetition of 'text' with min freq=1

Analysis:

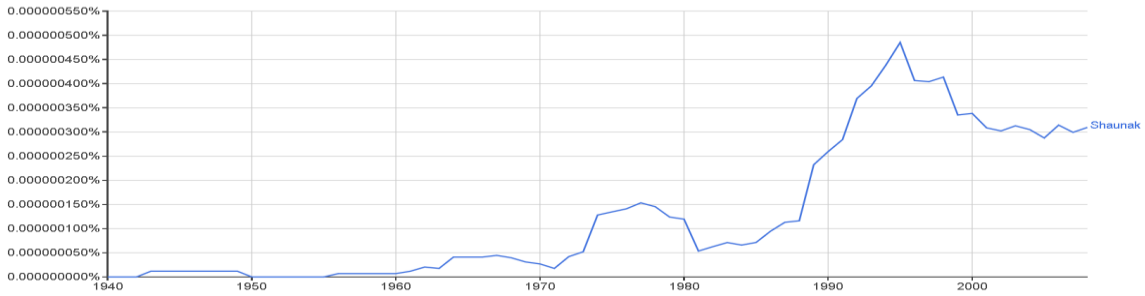
- Certain words like "is, on, a, in, at, to" aren't displayed in wordcloud. This is because otherwise only those words would pollute the wordcloud and they provide no extra information to the wordcloud.(see fig 1.)
- Adding too many same words leads to the wordcloud containing only that particular word.
- After adjusting wordcloud's parameter of min.freq the word cloud appropriately makes the other words smaller. I repeated text string to see the effect.(see fig 2)

Q2

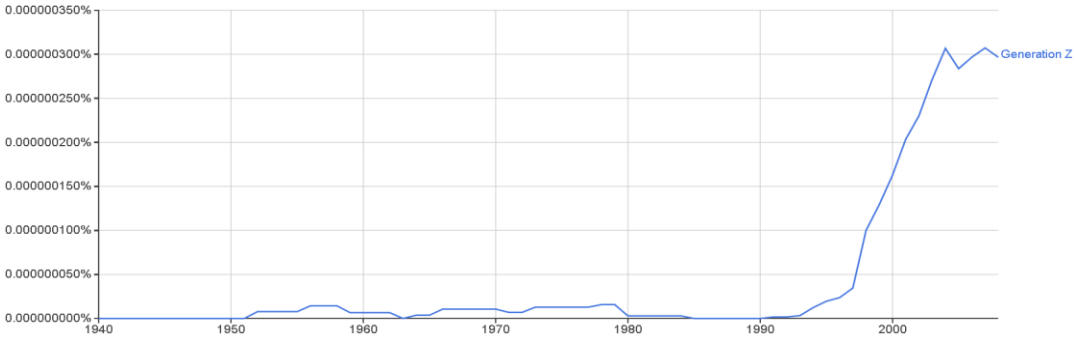
- Putting Mark Keane in google ngram viewer results in a cliff like structure around the 1970s.

According to me this seems to be due to the multiple publications in conferences. One of them is called "Proceedings of the 17th Annual Conference of Cognitive Science Society"

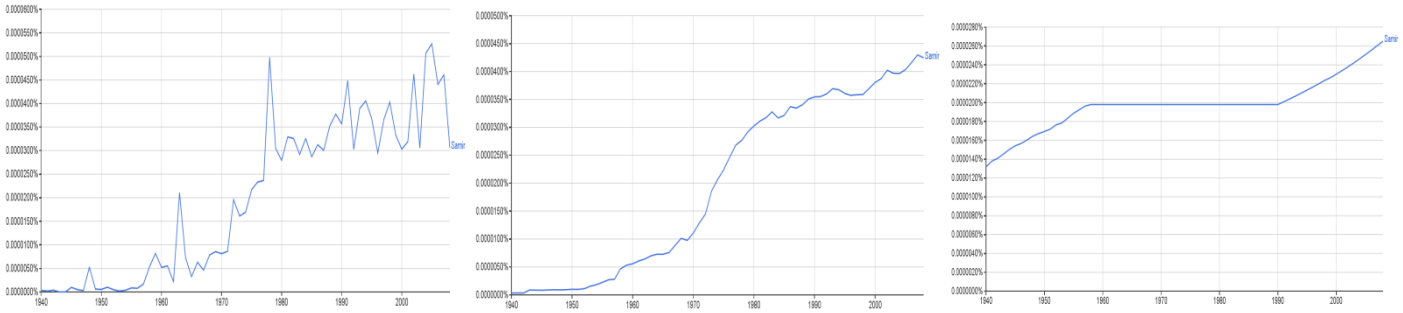
There is another peak around 2002. This seems to be due to a Hurler with the same name. The frequency of the name also increases with more books around that era



- Putting my name in the ngram viewer sadly didn't leave any hits. I think this is because the ngram viewer only reaches till 2008 and I was only 12 years old and my name is quite unique hence there are not hits. For my first name the graph follows a rising pattern with several peaks with the major one around 1996. I think this is because the name really started getting popular around then. Overall it's still a bit niche.

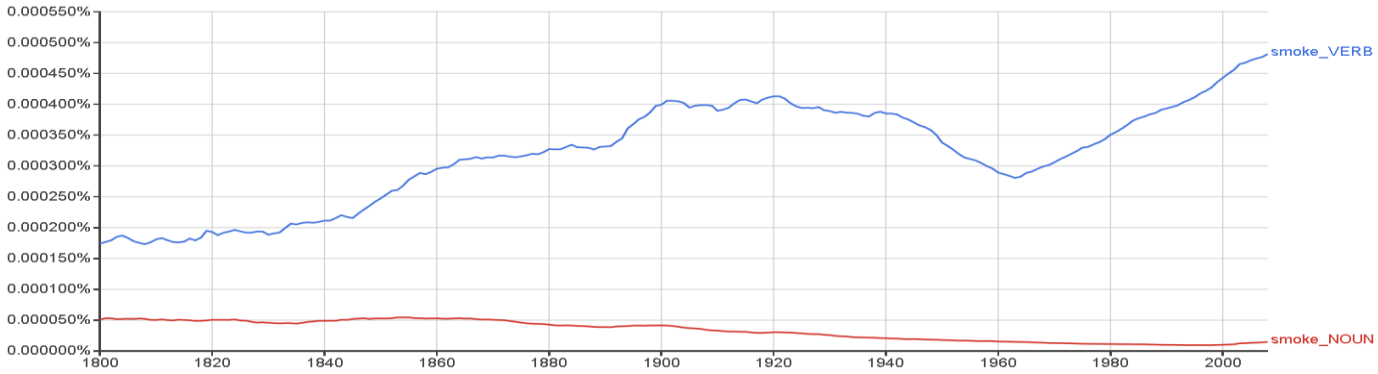


- Generation Z is a new word I tested in ngram. I expected it to start around 2000s. It started in 1995 but there were still a few hits before that. After investigating I found that generation Z was used before rarely for scientific purposes in a different context. Mostly as a generation of new nodes with the label Z. It was used mostly in scientific research concerning biology.

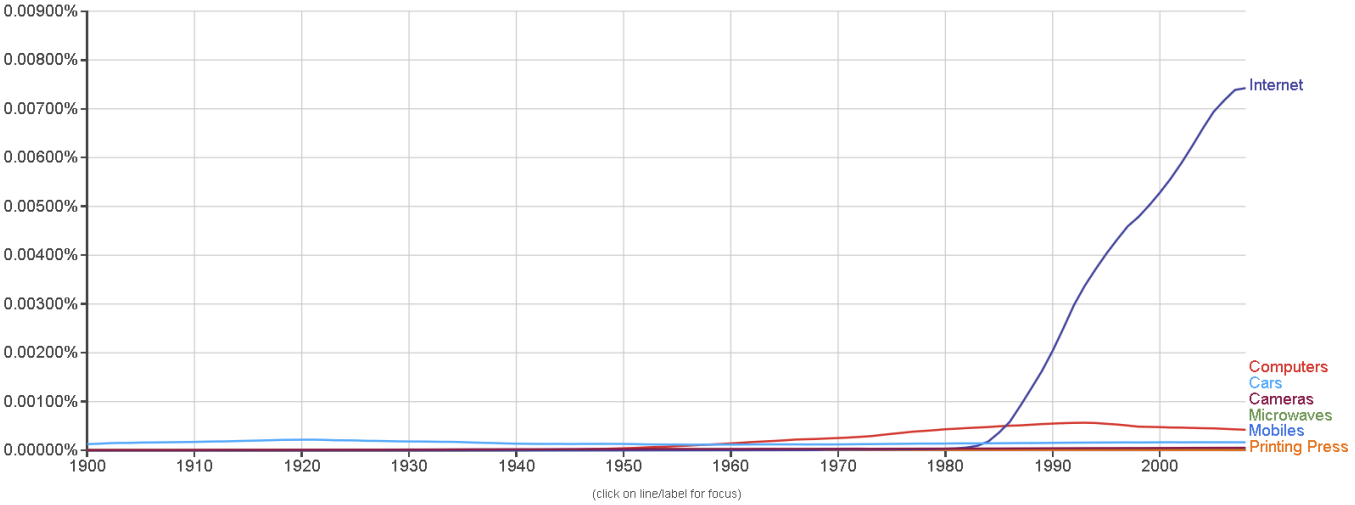


0 smoothing 5 smoothing 50 smoothing

- I tested a word 'Samir' a popular name in India. With no smoothing it showed random hits and a very jagged graph. With smoothing around 5 the graph made finding trends easier. With smoothing 50 it showed an almost smooth curve with a line. The general trend was overly summarized in this method. It also eliminated a lot of small and big peaks and instead showed as straight line through the peaks implying there were mentions in years where there were no mentions. In some cases it also showed a grossly bigger frequency (1940) and also showed an almost straight line (1960-90).



- I tested the terms smoke as a noun and a verb. I expected the smoke noun to rise in the 90s because of the problems of industrial revolution but it was mostly stable. Actually it was gradually decreasing. Smoking as a verb however had more mentions .I believe this is because of the research on the effects of smoking and pollution on the whole. This was more prevalent in books in that era.



- I used the terms Mobiles, Computers, Microwaves, Printing Press, Internet to demonstrated new technologies in 1900s. In the graph its apparent how significant the fields of Computers (red) and the Internet (violet) are in academics compared to other similar technologies. Adding any number of other technologies we take for granted nowadays like Air Conditioners, Cameras even general words like Cars ;all are dwarfed by internet and computer mentions. Even revolutionary discoveries like electricity (not pictured) couldn't compare to the mentions of internet.

Q3

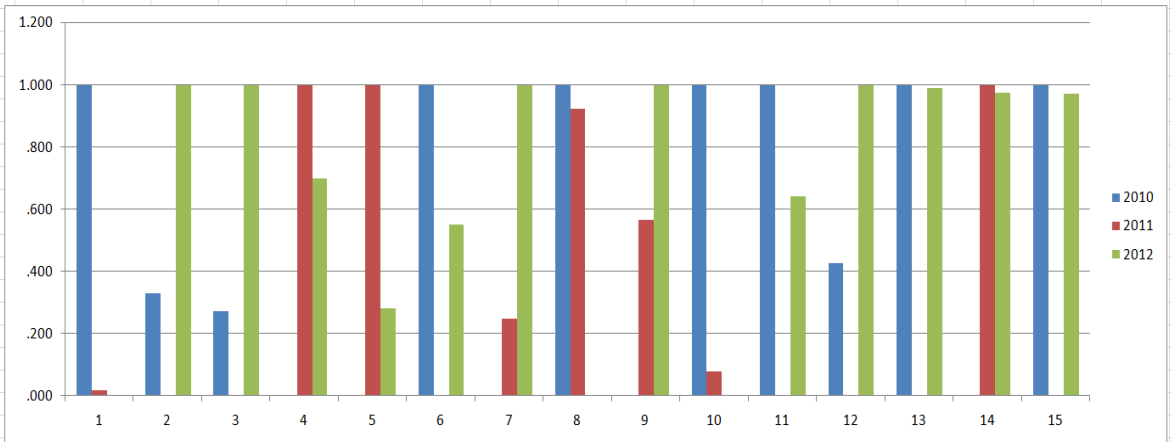
Dataset used for analysis:

Sales	Coke	Pepsi	Sprite	Nuka Cola	7up	Budweiser	Kingfisher	Soda	Lemonade	Mountain dew	Red Bull	Dr Pepper	Fanta	Mirinda	bisleri
2010	999	345	287	134	194	765	124	900	344	765	871	457	328	526	910
2011	100	120	74	589	653	148	329	852	645	480	201	235	109	783	453
2012	84	799	856	452	324	489	952	268	874	456	632	753	326	777	898

Frequency of soft drinks/drinks mentions in 2010-12

Normalizing frequencies using range transformation (0 to 1) across years results in:

Year	Coke	Pepsi	Sprite	Nuka Cola	7up	Budweiser	Kingfisher	Soda	Lemonade	Mountain	Red Bull	Dr Pepper	Fanta	Mirinda	bisleri
2010	1.000	.331	.272	.000	.000	1.000	.000	1.000	.000	1.000	1.000	.429	1.000	.000	1.000
2011	.017	.000	.000	1.000	1.000	.000	.248	.924	.568	.078	.000	.000	.000	1.000	.000
2012	.000	1.000	1.000	.699	.283	.553	1.000	.000	1.000	.000	.643	1.000	.991	.977	.974



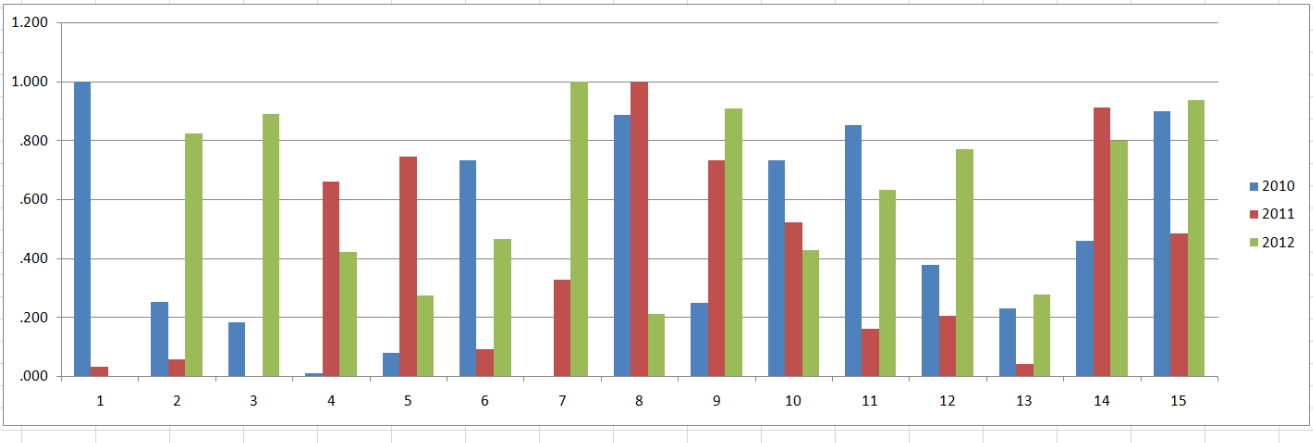
(the indexes 1-15 refer to the attributes in the data set)

We have basically normalized each attribute with no relation to other attributes.

In this technique it results in each attribute having a 0 or 1 value. The next year has a value in the middle. This tells us which year had the biggest /lowest value in the 3 years. The third value lies between 0 and 1 and it tells us the proportion of how much the value is compared to the bigger and lower values. This method is good to compare the changes of other companies over the time period. But we can't find relation between the highest lowest value because they are all normalized to 1/0 . A lot of the data shown is redundant because of this. Also we can't accurately estimate how much a loss/gain is for a given company because it completely depends on lower or higher values. For example: for attribute 1 year 2010 has highest mentions while 2012 has lowest. But we can only measure the comparison between year 2010 and 2011 and not the no. of changes as a whole.

Normalizing frequencies based on a given year for a given attribute:

Year	Coke	Pepsi	Sprite	Nuka Cola	7up	Budweiser	Kingfisher	Soda	Lemonade	Mountain	Red Bull	Dr Pepper	Fanta	Mirinda	bisleri
2010	1.000	.253	.186	.011	.080	.733	.000	.887	.251	.733	.854	.381	.233	.459	.898
2011	.033	.059	.000	.662	.744	.095	.328	1.000	.734	.522	.163	.207	.045	.911	.487
2012	.000	.824	.889	.424	.276	.467	1.000	.212	.910	.429	.631	.771	.279	.798	.938



This technique enables us to actually compare between given attributes for the 3 years separately. This results

in a good uniform representation to properly compare among attributes. Example :For the year 2010(blue) first attribute has highest normalized frequency while seventh attribute has lowest normalized frequency. Attribute 15 also has a significant frequency of almost 1, while attribute 4 has similar frequency to attribute 7(0).

By aggregating among the years further one can also roughly tell which attribute has most mentions for all the years. Attributes 8, 9, 14, 15 had the most frequency in all the years. Attribute 4, 13 had the lowest frequency respectively.

Q4

The model tries to predict the present values by simply looking at Google queries for the given/related subjects during that time for some given dimensions. The paper demonstrates what queries can be used for determining some dimensions.

For house sales it demonstrates that counting google searches for real estate services is a good predictor for the no. of sales.

It tries to measure the no. of tourists to Hong Kong from various countries. It uses various predictors like the value of Hong Kong currency, the effect of Beijing Olympics and searches for 'Hong Kong'. The model predicts the data remarkably well.

In the future perhaps it can be used to peer a little in the future too as these models get more accurate.

The googletrends.csv and Fordsales.csv are missing. These are vital for the program because Google has its own indexing format and according to the indices the models are plotted by R.