Kshitija Shelar

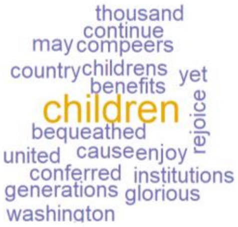
Text Analytics COMP4760

Date: 18/10/20

20200155

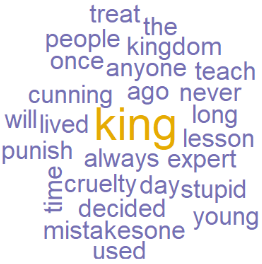
Practical 3

1) b.

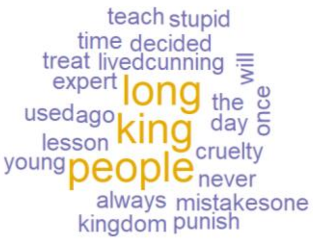


Words listed	Words not listed
Thousand, continue, may, compeers, country, childrens, children, benefits, yet, rejoice, bequeathed, united, cause, enjoy, conferred, institutions, generations, glorious, washington	our, and, to, a, the, by, have, under, those, us

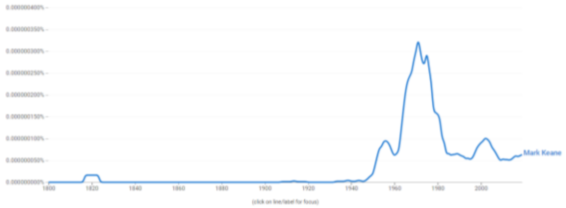
1) c. As hypothesised, the stop words were removed again. And depending upon the frequency, for ex here King word was used twice, hence its in a bigger font. So the hypothesis was correct.



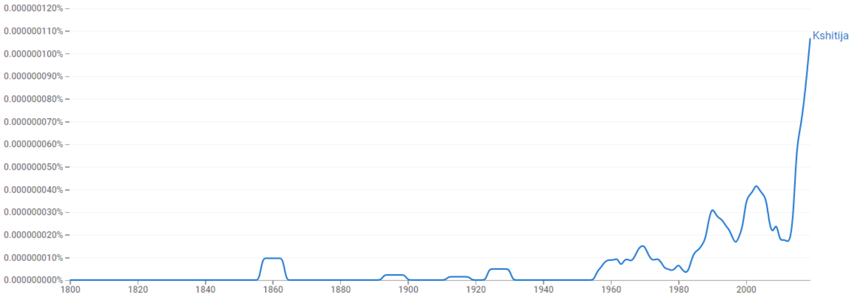
1) d. after adding the frequency of other words, they appear as same color and same font as king's word. That means, internally every word is assigned a fixed font and color depending on it's frequency and it is placed on the UI randomly. One more observation, if the frequency of one word is increased a lot, all other words disappear and only that word appears. See below figure. This happens because there is a minimum frequency cut off in that case which doesn't let words with low frequency appear on the wordcloud.



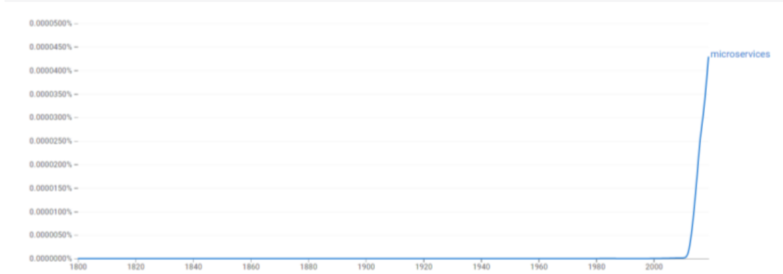
2) a. Searching Mark Keane in google N gram gives the following graph. N gram basically means N consequent words from a corpus. From the graph we can observe major hits in the year 1955, 1971, 1975, and 2001. After searching, we can find many books in these years where MARK KEANE appears. For ex. In 1971, Mark Keane, the executive director of ICMA, was mentioned in many books related to United States Congress Senate. In early 2000s, we can observe another peak, this is because of Professor Mark Keane's books. There has been a lot of mention of his books in multiple conferences.

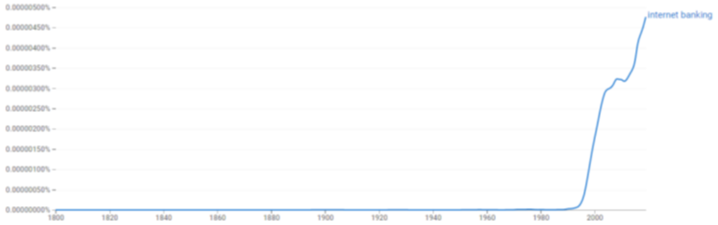


2) b. Searching Kshitija produces the following graph. The major hits are observed 1968, 1990, 2002 and 2019. In the late 1960's there was a book written about Hindu Astronomy which mentions Kshitija in it many times which explains the hits in early 2000s. The name is also mentioned in a book of baby names. The name appears in many astronomical related books which explains the random peaks.



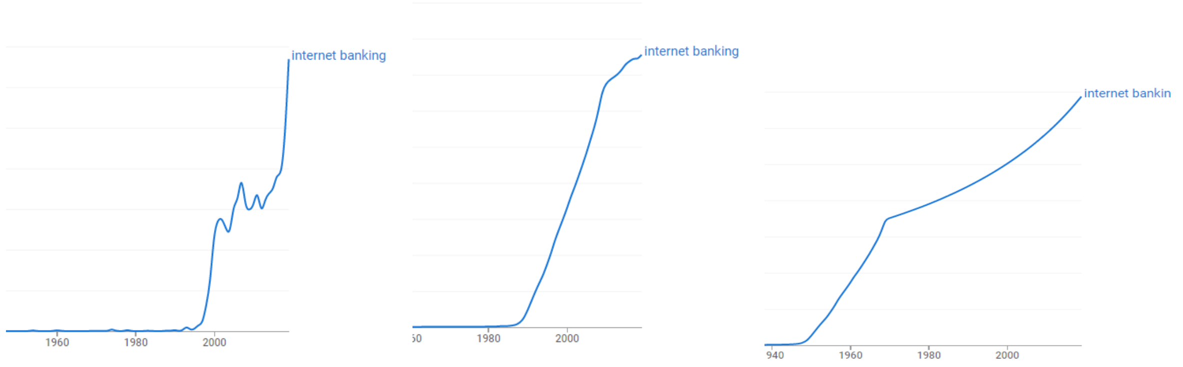
2 c. The term microservices started emerging after 2010





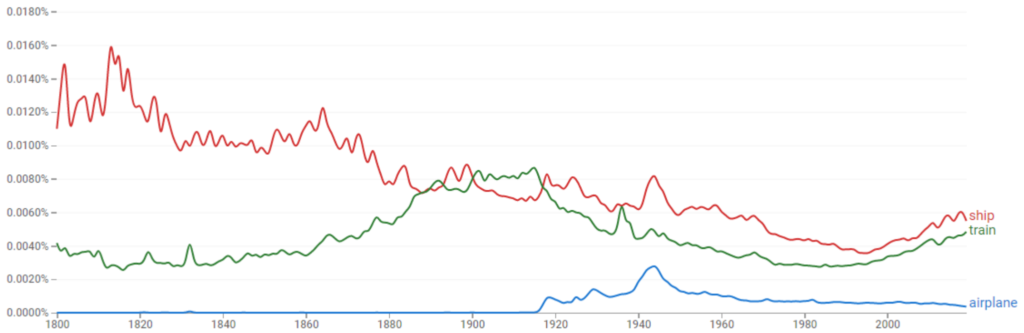
Whereas the term internet banking can be observed in late 1980's. Internet banking started in 1996. The fact that this word appeared before it actually started is because there were a lot of debates and discussion around this topic in United States Congress regarding the security and other risks associated with internet banking.

2 d. Smoothing is a process of broadening the graph. By increasing the smoothing number, the graph gets flatter. Smoothing is performed to prevent 0 probability from getting assigned to test data that does not exist in the trained N-gram. If the n-gram is too sparse, smoothing is necessary so the graph can be analysed clearly.

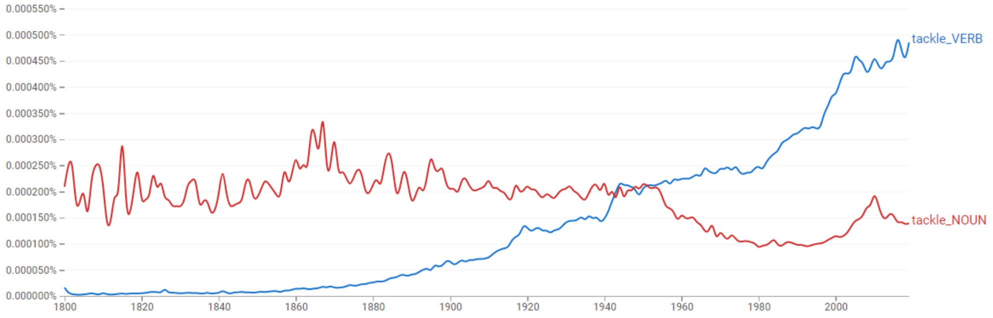


N gram with 0 smoothing | Same N gram with soothing of 10 | Same N gram with smoothing of 50

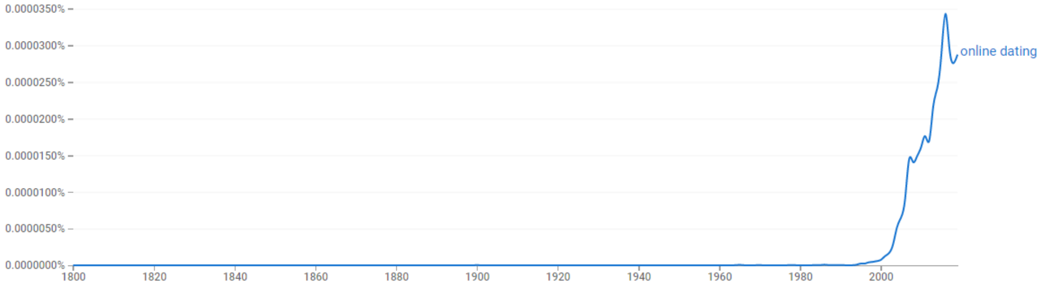
2. e. Related words like Airplane, ship, and train have the following frequency pattern. We can see that ships have always been on a higher frequency. It is because the ships came into use a very long time ago. It was the only medium of travel across countries/continents. Early dynasties used to expand their kingdom and do trading using ships and boats. Hence we see a very high frequency of ships in the graph. Trains were invented shortly after, but because it has many limitations like the tracks routes are fixed, etc,. It is not as popular as ships. As we all know, airplanes were the most recent invention, hence, it emerges out of the graph very late. As airplanes too have limitations of goods it can carry because of pressure and weight restrictions, in today's day as well, ships have the highest frequency in the graph. As most of the goods are still transported from ships. We can see a sharp rise in the train graph after 1880s. That is because it was then when electric trains were invented and it gained popularity, even surpassing ships shortly after. However, because of its limitaions, ships took over again.



2. f. Syntatcically similar words are tackel noun and tackel verb. The graph looks like follows. To search words using pos tag we write it as word_POSTAG ex. Tackle_VERB



2. g. Terms like online dating started only after the boom of internet. From the graph below we can observe the peak around 2015 which is approximately the year dating apps like Tinder and Hinge were released and gaining popularity. Before online dating, people used to meet in pubs etc.



3. i. N = 5130

3. ii. n2010= 525, n2011=790, n2012=905, n2013=1240, n2014=1670

words	2010	2011	2012	2013	2014
education	100	150	50	90	100
empowerment	50	90	100	110	30
development	10	30	50	40	70
music	50	100	150	200	400
literature	10	20	25	40	60
Arts	100	90	70	80	60
Science	80	100	120	150	200
Computers	50	100	150	250	350
Technology	25	50	100	200	300
Media	50	60	90	80	100
Total	525	790	905	1240	1670

3 a. Overall Normalization: Dividing all column data by the total number of words i.e N = 5130 and multiplying it by 1000

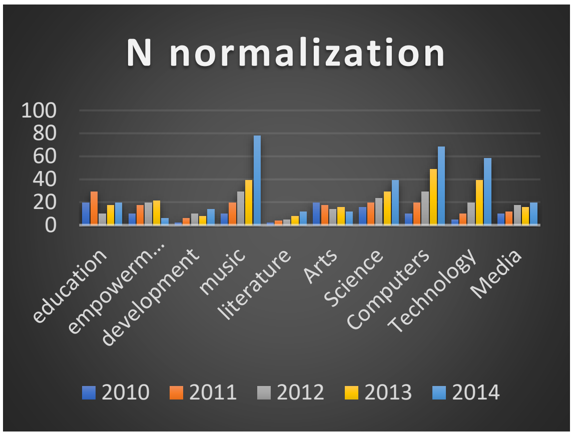
words	2010	2011	2012	2013	2014
education	19.49	29.2	9.74	17.54	19.4
empowerment	9.74	17.54	19.4	21.44	5.84
development	1.94	5.84	9.74	7.79	13.64
music	9.74	19.4	29.2	38.98	77.9
literature	1.94	3.89	4.87	7.79	11.69
Arts	19.4	17.54	13.64	15.59	11.69
Science	15.59	19.4	23.39	29.2	38.98
Computers	9.74	19.4	29.2	48.73	68.3
Technology	4.87	9.74	19.49	38.98	58.4
Media	9.74	11.69	17.54	15.59	19.4

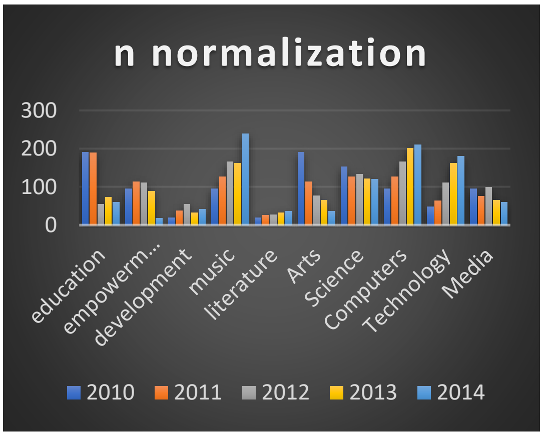
3 b. By- year Normalization: dividing each element by the total number of words in that particular year. Example all values in 2010 are divided by 525, all values in 2014 are divided by 1670 etc. and multiplied by 1000.

words	2010	2011	2012	2013	2014

education	190.4	189.87	55.24	72.58	59.88
empowerment	95.23	113.92	110.49	88.7	17.96
development	19.04	37.97	55.24	32.25	41.91
music	95.23	126.58	165.74	161.29	239.52
literature	19.04	25.31	27.62	32.25	35.92
Arts	190.4	113.92	77.34	64.51	35.92
Science	152.38	126.58	132.59	120.96	119.76
Computers	95.23	126.58	165.74	201.61	209.58
Technology	47.61	63.29	110.49	161.29	179.64
Media	95.23	75.94	99.44	64.51	59.88

3 c.





- N normalization is the normalization done on all the words in all 5 years. i.e suppose computer in 2010 has 50 wordcount, total number of words in all 5 years is 5130. So after normalization it will become 9.74.
- “n” Normalization is the normalization done on all words of that particular year. i.e suppose word count of computer in 2010 is 50 and total number of words in 2010 is 525 then after normalization computer will become 95.23
- The education bar of 2010 and 2011 in N normalization have a lot of difference. Whereas, the education bar of 2010 and 2011 in n normalization are almost of the same height. This is because, the total number of words in 2010 are less compared to number of words in 2011. Therefore, even though the number of time education word appeared is slightly less in 2010, after normalization, both bars appear to be same. Education in 2010 appears 100 times and in 2011 appears 150 times. The total number of words in 2010 is 525 whereas in 2011 it is 790 words. Because of this the ratio are coming almost similar i.e 100/525 *1000= 190.4 and 150/790 * 1000 = 189.8.
- Literature word has similar kind of progression in both normalizations i.e it’s increasing steadily.
- The 2010 Media column is less than 2011 in N normalization and greater than 2011 in n normalization. This is again because of number of words in 2010 and 2011 vary significantly.