

## Data description

Being an avid shopper at Costco, I personally like the range of varieties it offers in all kinds of shopping. I usually suggest all my friends and family to shop at Costco. With the help of this data analysis, I can show them real statistics based on the reviews I searched. I conducted data analysis on Costco to check how often people shop at Costco and what they feel about the range of products at Costco offers.

## Structured data after text processing

The RStudio interface displays a data frame 'df' with 23 rows of text reviews. The Environment pane shows the 'Large LDA\_VEM' model with various parameters and row totals.

1	ever participated challenge
2	polygamist euphemism somebody picks sample
3	even okay take bake since pizza yeah foo
4	told dillydally
5	miss phineas wholesale full set yakisoba im
6	tbh ive never pizza im sorry frien
7	killary get back ur favorite toilet paper isle amp push ur...
8	ate much food jesu
9	elder scrolls online morrowind psxb
10	heres thought pizza excellent take bake
11	growing spending leaving proud self control
12	shout polygamists
13	guess membership yep adult bitch
14	aye sometimes got blessed us pack
15	secrets shopping
16	walking yes hello vincent price circa please
17	man like well hit every sample station didnt fuck
18	saw
19	wow havent chicken bake years huge craving id lick toe ...
20	th birthday love much
21	thank nina ended getting amp recycle program good kn...
22	like reasoning lacks logic s economy

Showing 1 to 23 of 6,818 entries

Environment:

- Global Environment
- Large LDA\_VEM (2.2 Mb)
- max: 0.203865412698103
- max\_pos: num [1:34090] 0.205 0.211 0.203 0.202 0.202 ...
- neg: chr [1:4783] "2-faced" "2-faces" "abnormal" "aboli...
- pos: chr [1:2006] "a+" "abound" "abounds" "abundance" "..."
- rowTotals: Named num [1:6817] 3 5 8 2 6 6 15 4 5 6 ...
- v: Named num [1:7271] 6239 6037 596 420 385 ...
- x: 1L

## Topic numbers

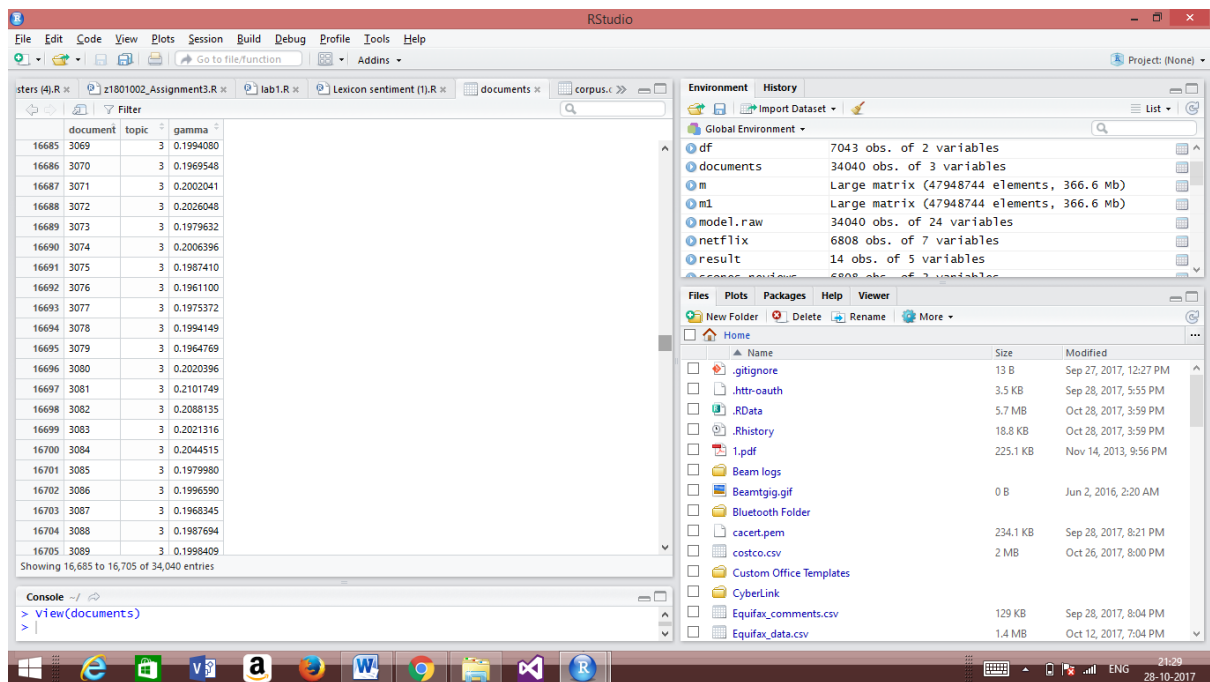
The RStudio interface displays a data frame 'topics' with 21 rows of topic numbers. The Environment pane shows the 'df' data frame with 7043 observations and 2 variables.

	topic	term	beta
1	1	challeng	1.714608e-04
2	2	challeng	2.048019e-04
3	3	challeng	3.590293e-04
4	4	challeng	3.186876e-05
5	5	challeng	6.760161e-05
6	1	ever	9.741882e-05
7	2	ever	2.398444e-03
8	3	ever	2.428463e-03
9	4	ever	1.580934e-03
10	5	ever	3.272976e-03
11	1	particip	8.724672e-05
12	2	particip	4.664462e-05
13	3	particip	4.775501e-05
14	4	particip	1.508367e-04
15	5	particip	2.509003e-05
16	1	euphem	2.980099e-05
17	2	euphem	1.872338e-05
18	3	euphem	3.321310e-05
19	4	euphem	2.370405e-05
20	5	euphem	1.378254e-05

Showing 1 to 21 of 35,215 entries

Environment:

- Global Environment
- df: 7043 obs. of 2 variables
- documents: 34040 obs. of 3 variables
- m: Large matrix (47948744 elements, 366.6 Mb)
- m1: Large matrix (47948744 elements, 366.6 Mb)
- model.raw: 34040 obs. of 24 variables
- netflix: 6808 obs. of 7 variables
- result: 14 obs. of 5 variables
- result.raw: 6808 obs. of 3 variables



```

      document topic      gamma
      <chr>  <int>    <dbl>
1         1      1 0.1988803
2         2      1 0.1990961
3         3      1 0.2030079
4         4      1 0.2017868
5         5      1 0.2024302
6         6      1 0.1963707
7         7      1 0.1928946
8         8      1 0.1985521
9         9      1 0.1956651
10        10      1 0.1983563
# ... with 34,030 more rows

```

## Sentiment Label

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Midexam 670.R MID FINALLR scores\_reviews netflix Topic clusters (4).R z180

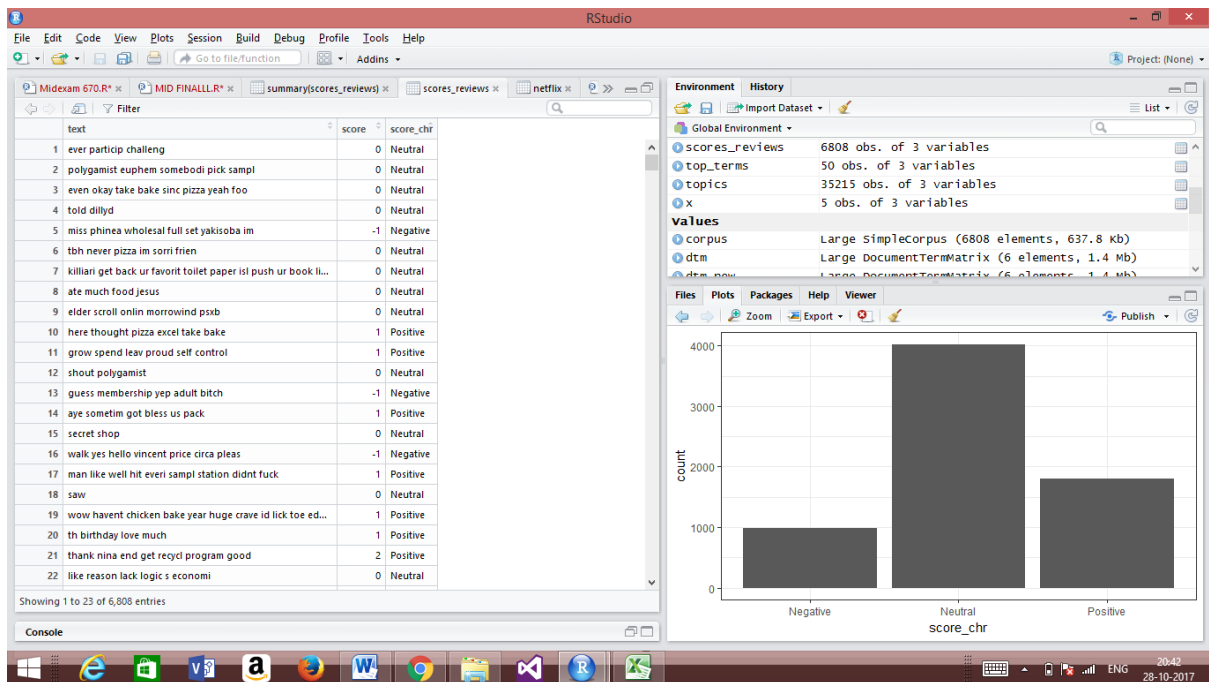
Filter

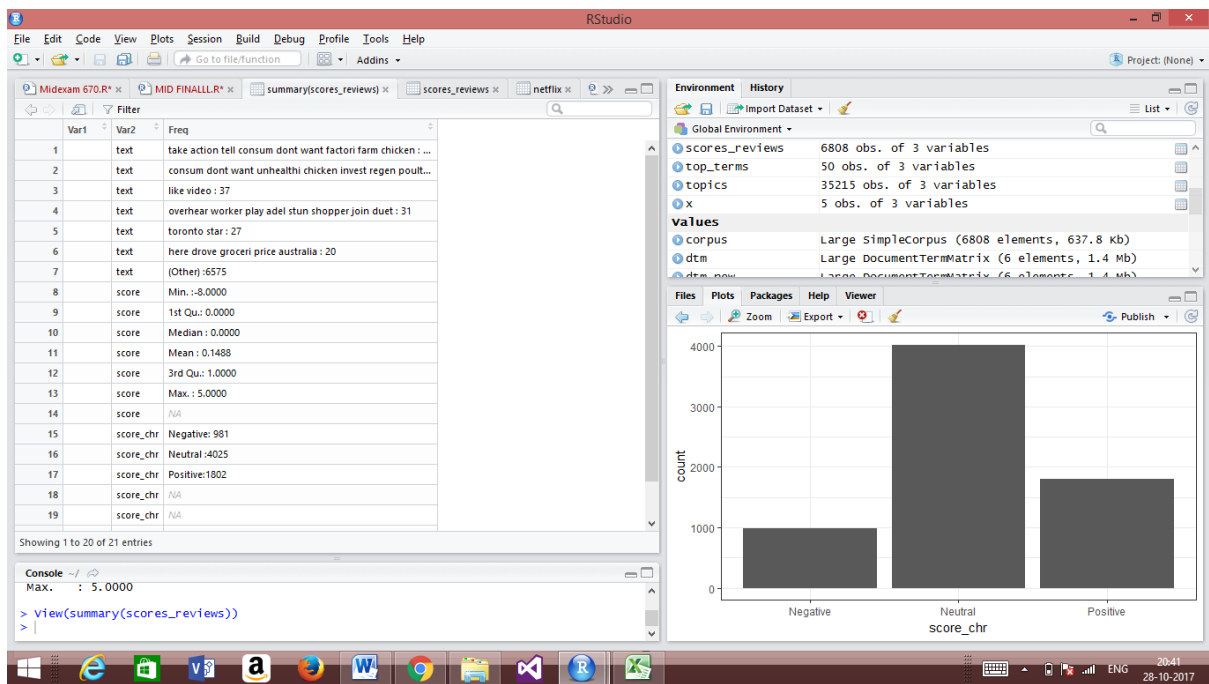
	text	score
1	ever particip challeng	0
2	polygamist euphem somebodi pick sampl	0
3	even okay take bake sinc pizza yeah foo	0
4	told dillyd	0
5	miss phinea wholesal full set yakisoba im	-1
6	tbh never pizza im sorr frien	0
7	killari get back ur favorit toilet paper isi push ur book li...	0
8	ate much food Jesus	0
9	elder scroll onlin morrowind pssb	0
10	here thought pizza excel take bake	1
11	grow spend leav proud self control	1
12	shout polygamist	0
13	guess membership yep adult bitch	-1
14	aye sometim got bless us pack	1
15	secret shop	0
16	walk yes hello vincent price circa pleas	-1
17	man like well hit everi sampl station didnt fuck	1
18	saw	0
19	wow havent chicken bake year huge crave id lick toe ed...	1
20	th birthday love much	1
21	thank nina end get recycl program good	2
22	like reason lack logic s economi	0

Showing 1 to 23 of 6,808 entries

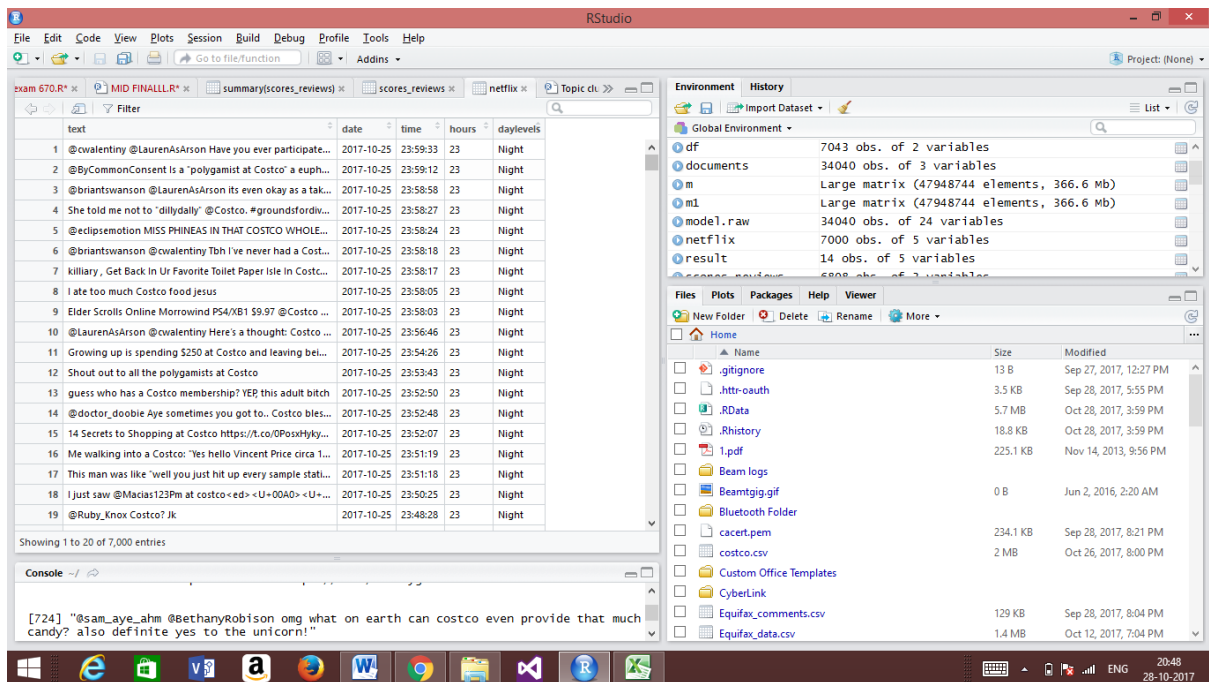
Console

## Sentiment Score

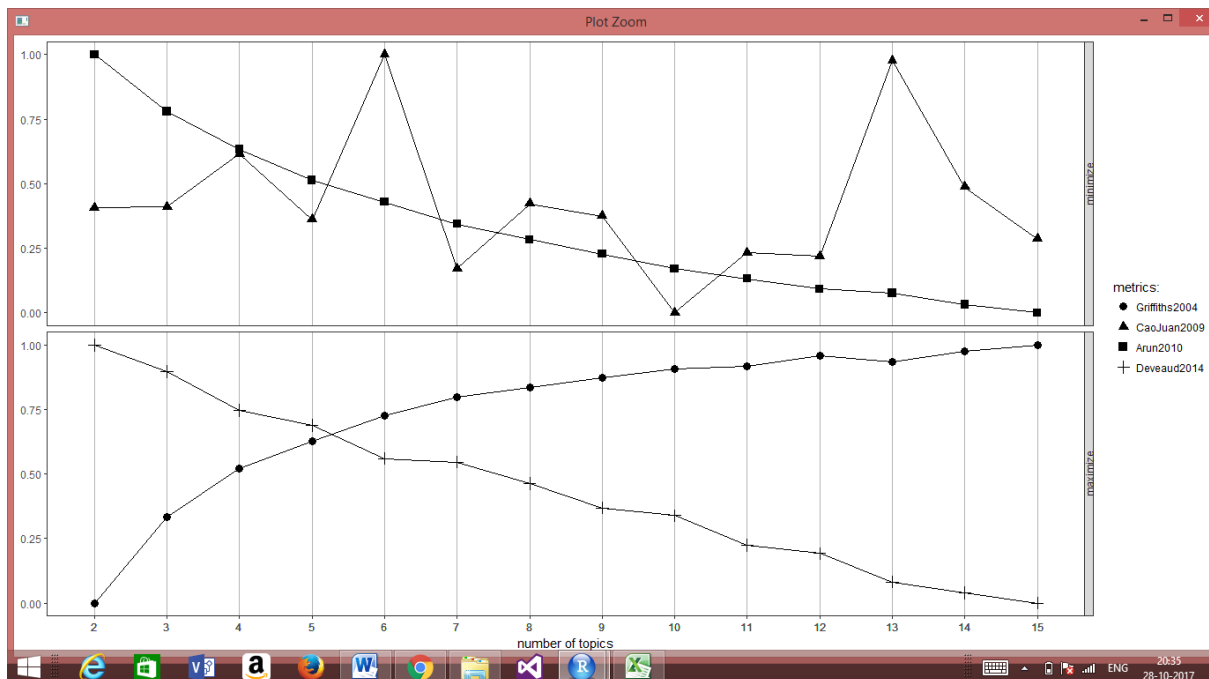
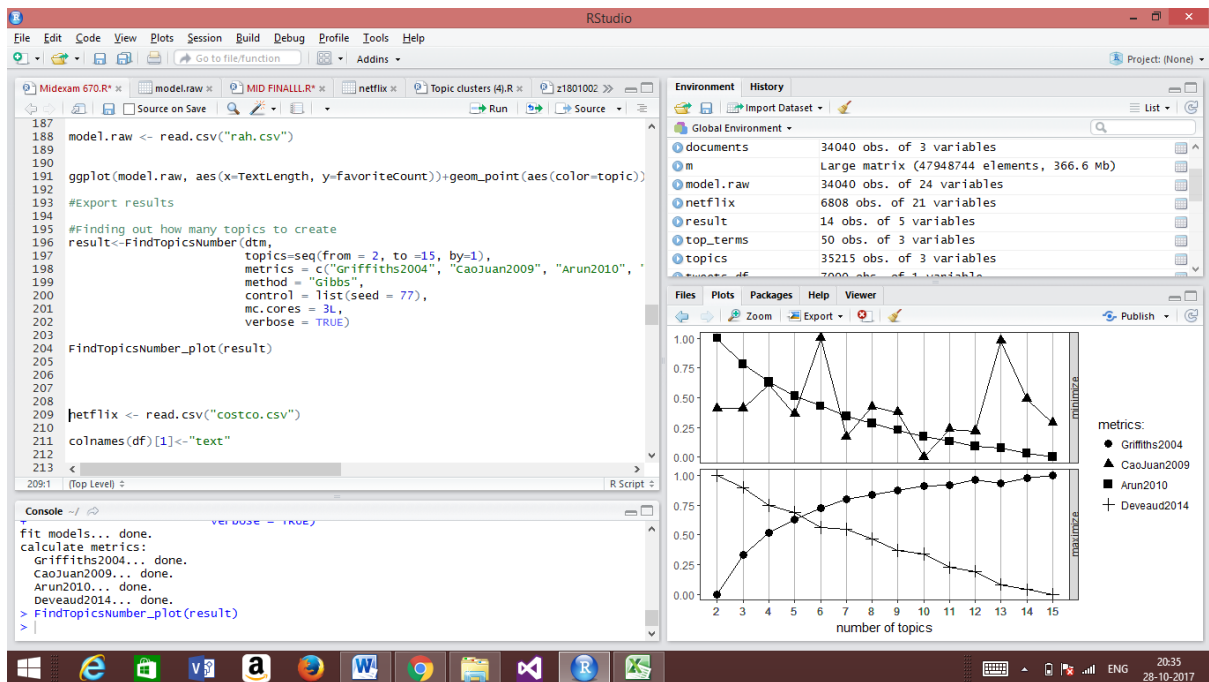




## Hour of the day



## Test deciding on the number of topics generated



Simple approach is to analyze the metrics to find extremum.

minimization:

Arun2010

CaoJuan2009

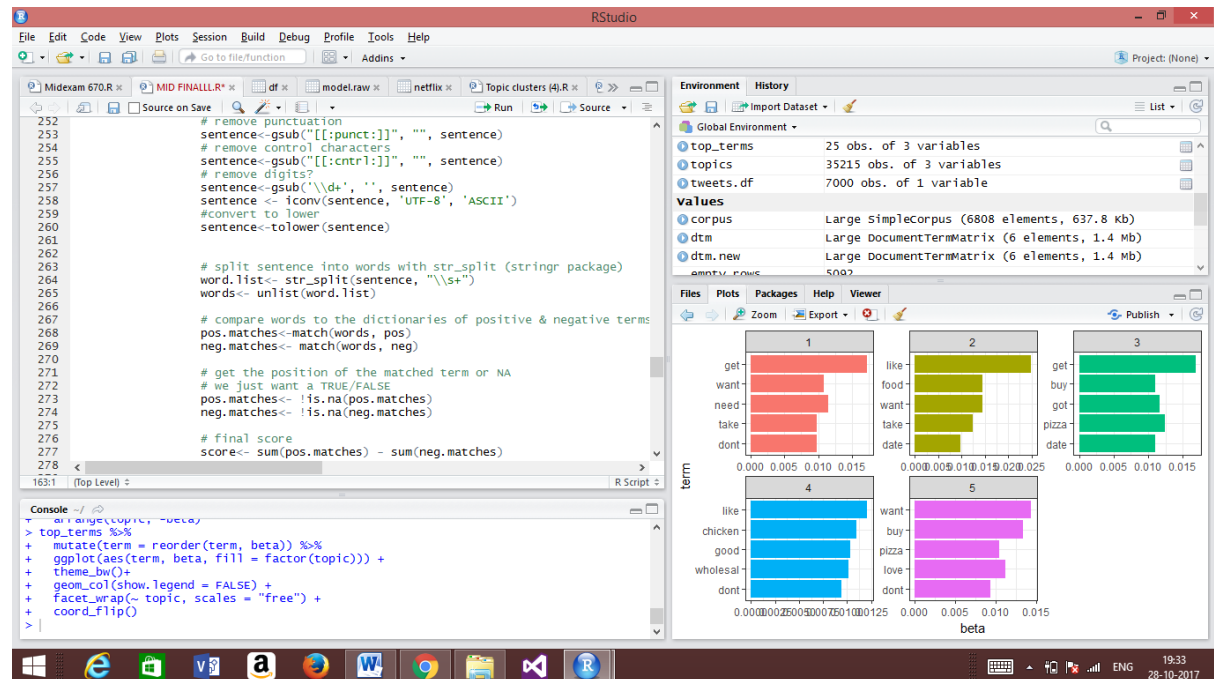
maximization:

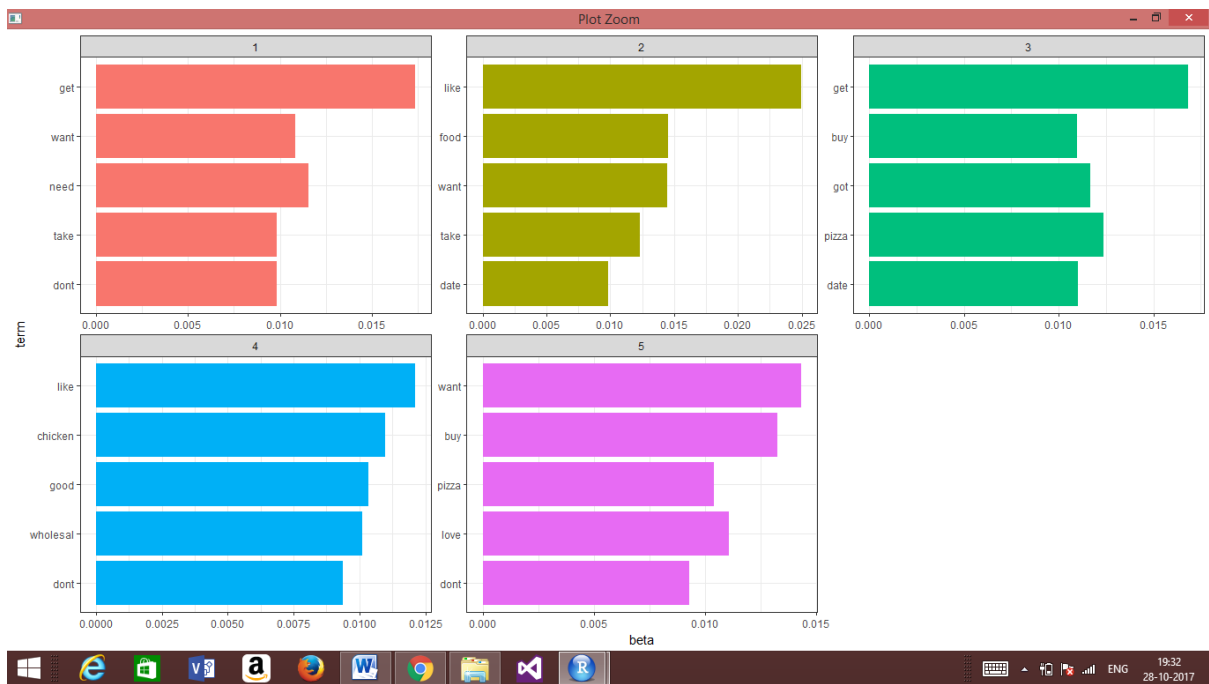
Deveaud2014

Griffiths2004

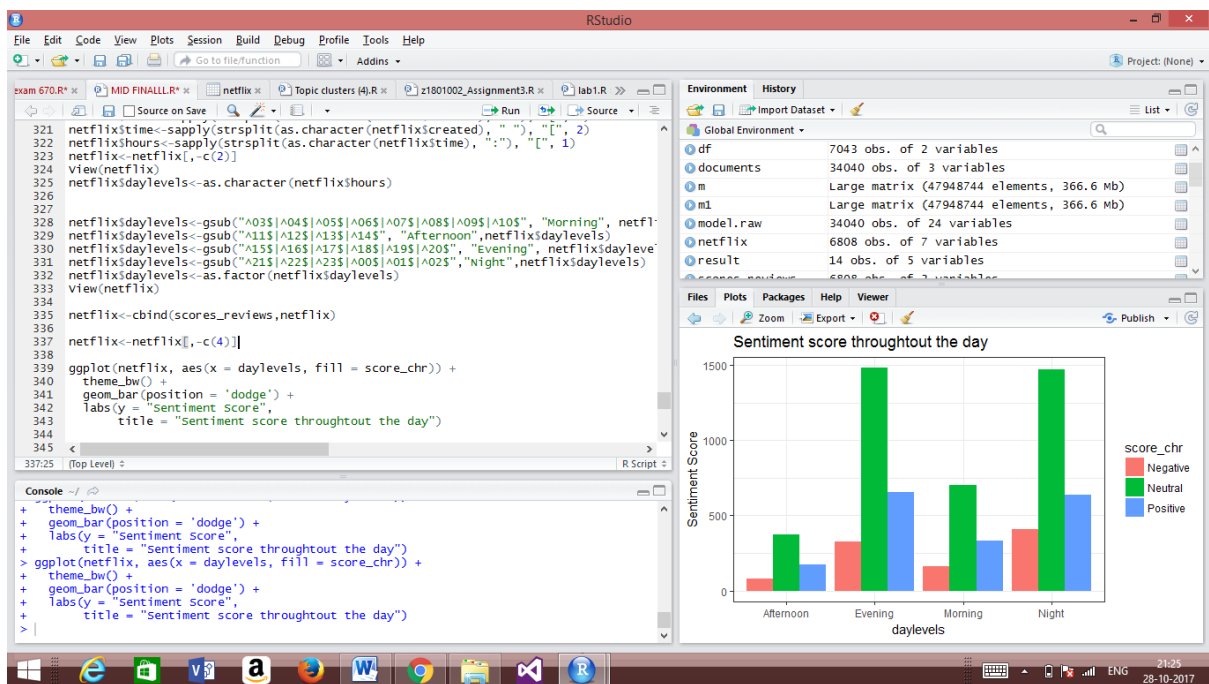
From this plot we made conclusion that optimal number of topics is in range 7-10.

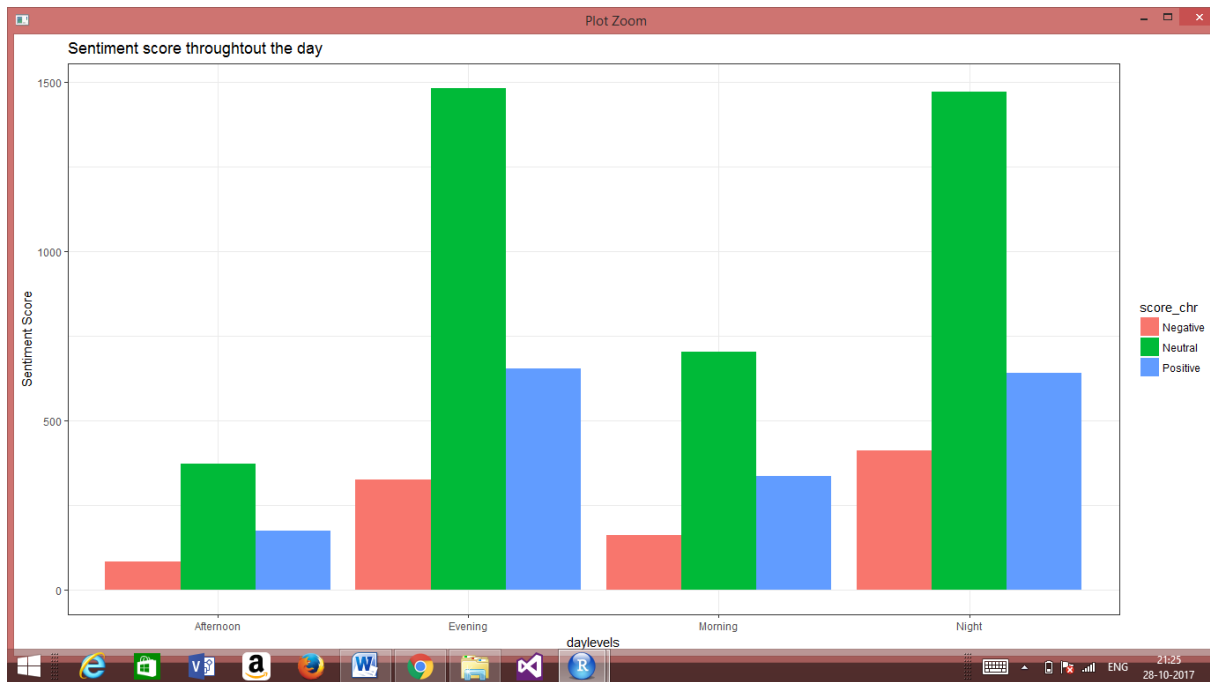
## Top 5 words for each topic generated



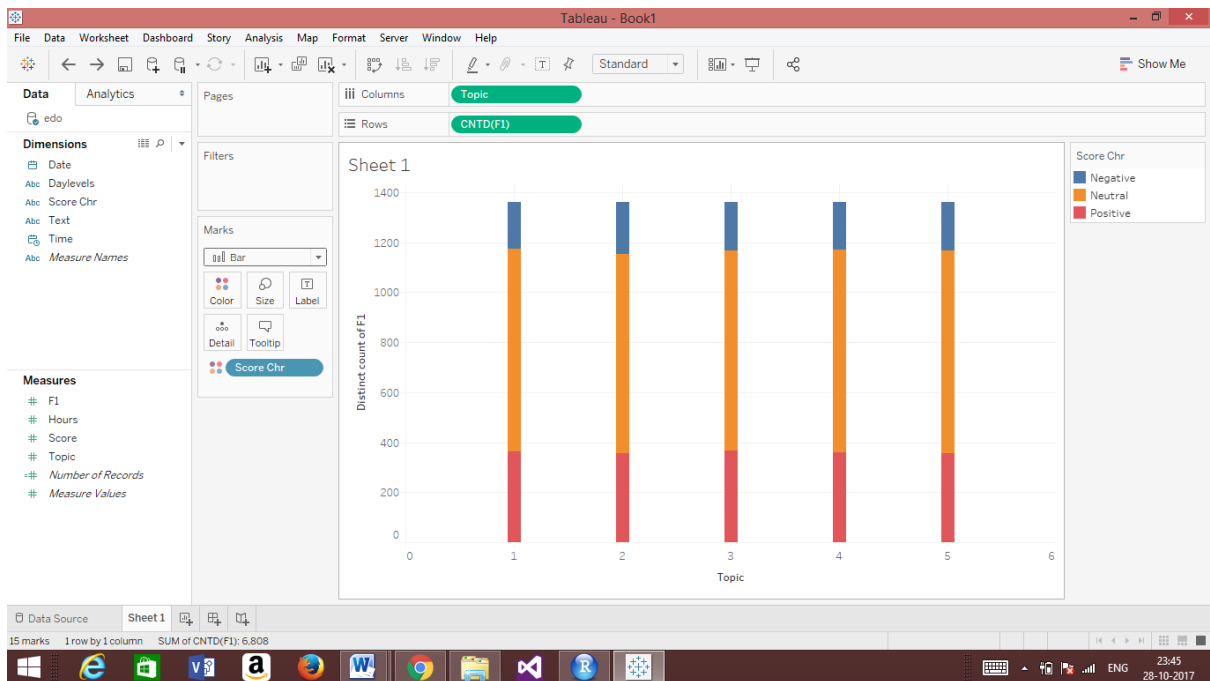


Sentiment scores are varying over the time range

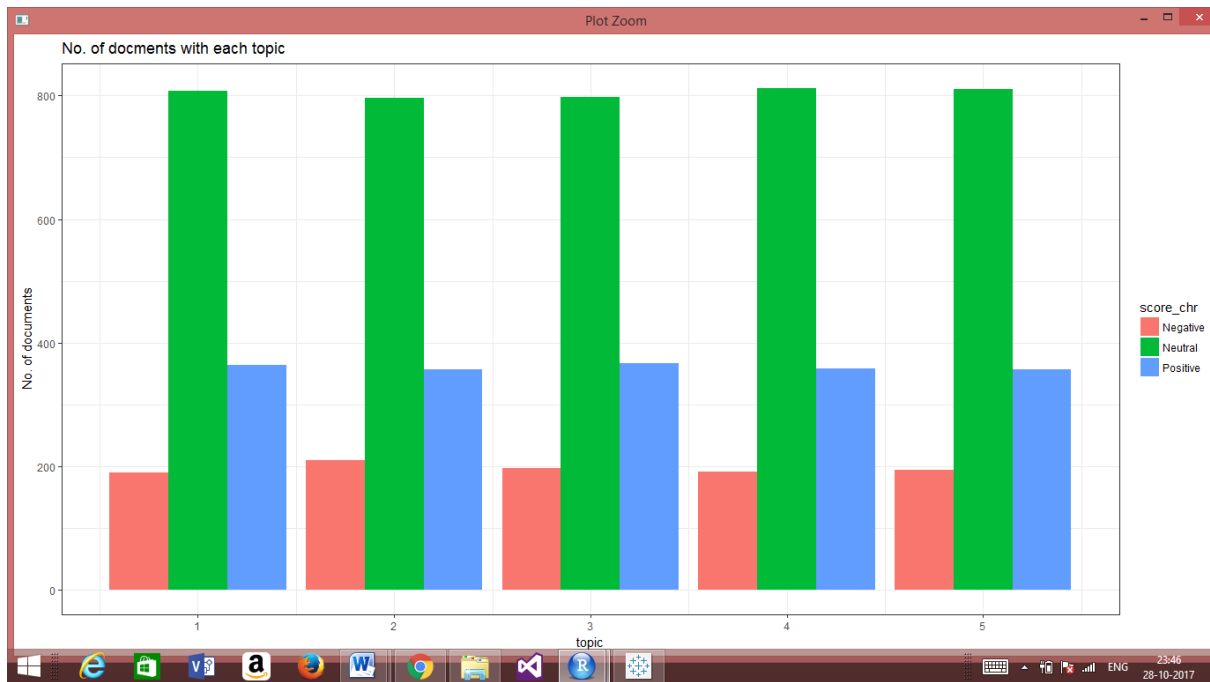




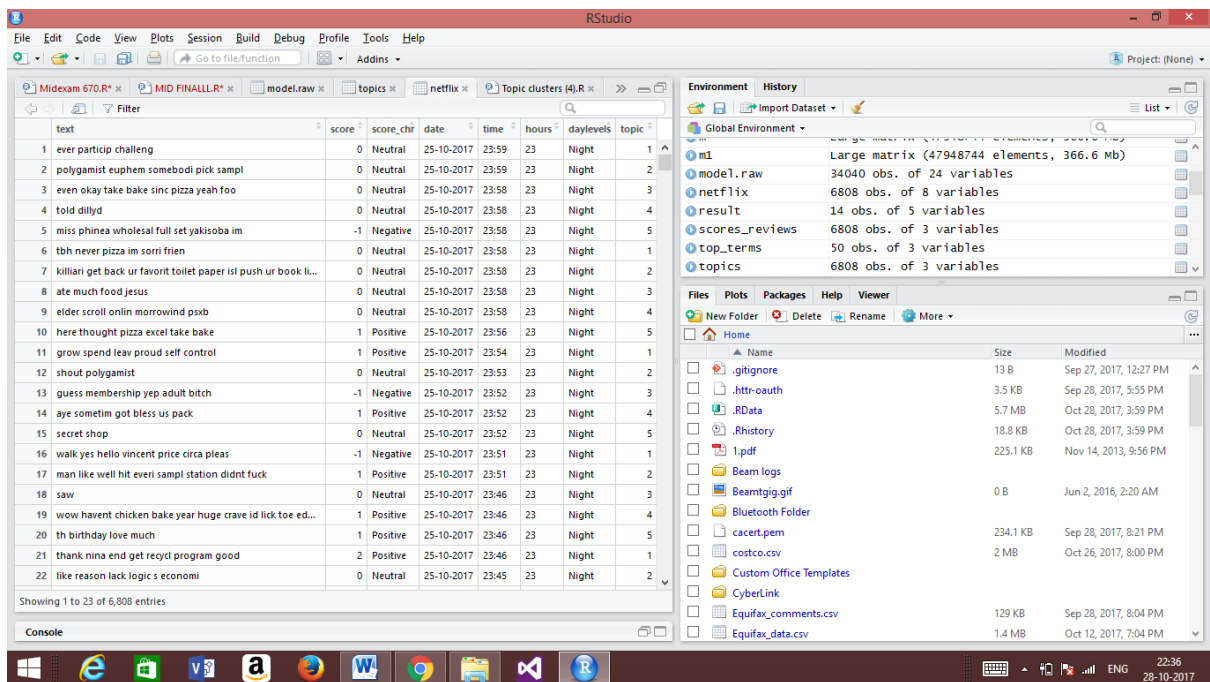
## Distribution of topic numbers segmented by sentiment labels







## New features added with master file



## Insights on second visualization of section 4

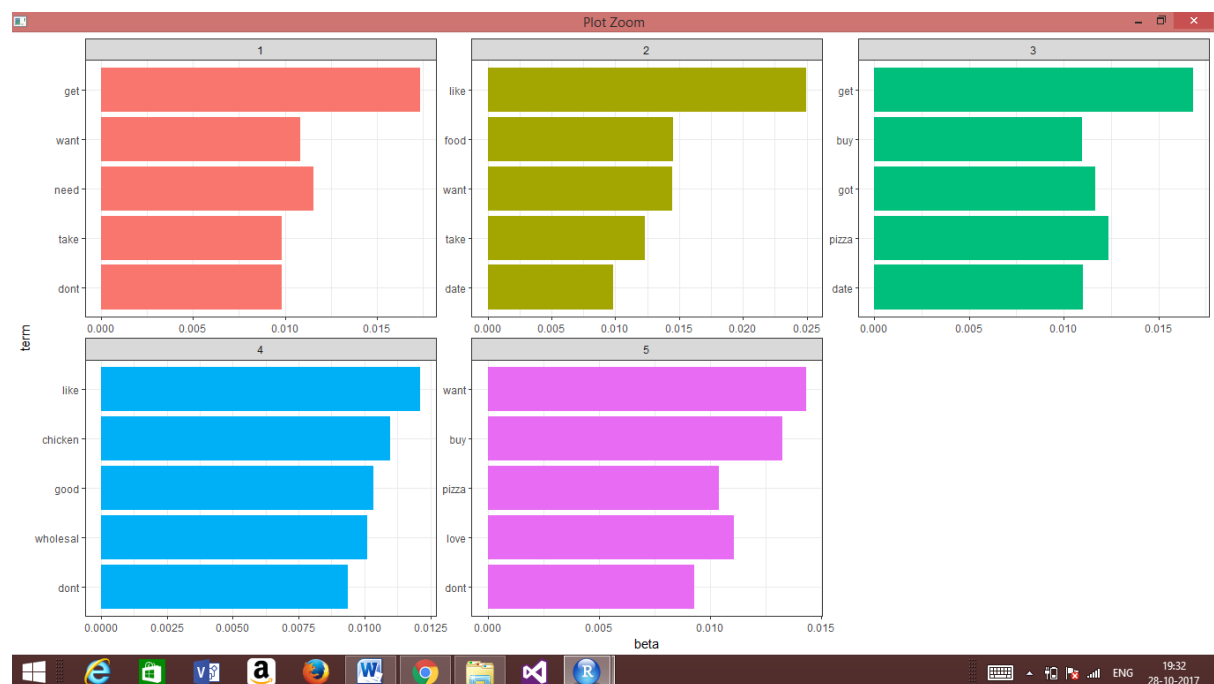
In the first topic, we observe more positive words such as get, want, need, and take. And we just have one negative word don't. We can infer that most product are suitable for purchase through this topics.

In the second topic, we observe very positive words such as like, food, want, and date. This implies that Costco has a wide range of foods and people prefer buying groceries at Costco.

In the third topic, we can infer that amongst the food varieties pizza is most purchased one at Costco.

In the fourth topic, we can observe that Costco is a wholesale chain with chicken and pizza foods most purchased and liked.

In the fifth topic, we can see that pizza word is repeated in this topic. So we can infer that people prefer buying pizza at Costco amongst the wide food varieties.



#### Insights on third and fourth visualization of section 4

In the third visualization, the most neutral scores were during the evenings and nights. The most positive scores were observed during evenings and nights too. The most negative scores were observed during evenings and nights. Also we can infer that most of the comments were generated during evenings and nights. However, most comments on Costco are more neutral than positive in the graph.

From the fourth visualization, we calculate the number of comments per topic and also segregate them based on sentiment label. We can observe that in topic 1, 4, and 5 the neutral comments are slightly more than the rest. In topics 1 and 3 the positive comments are slightly higher. Whereas, the negative comments in the 2<sup>nd</sup> topic is more than the rest.

