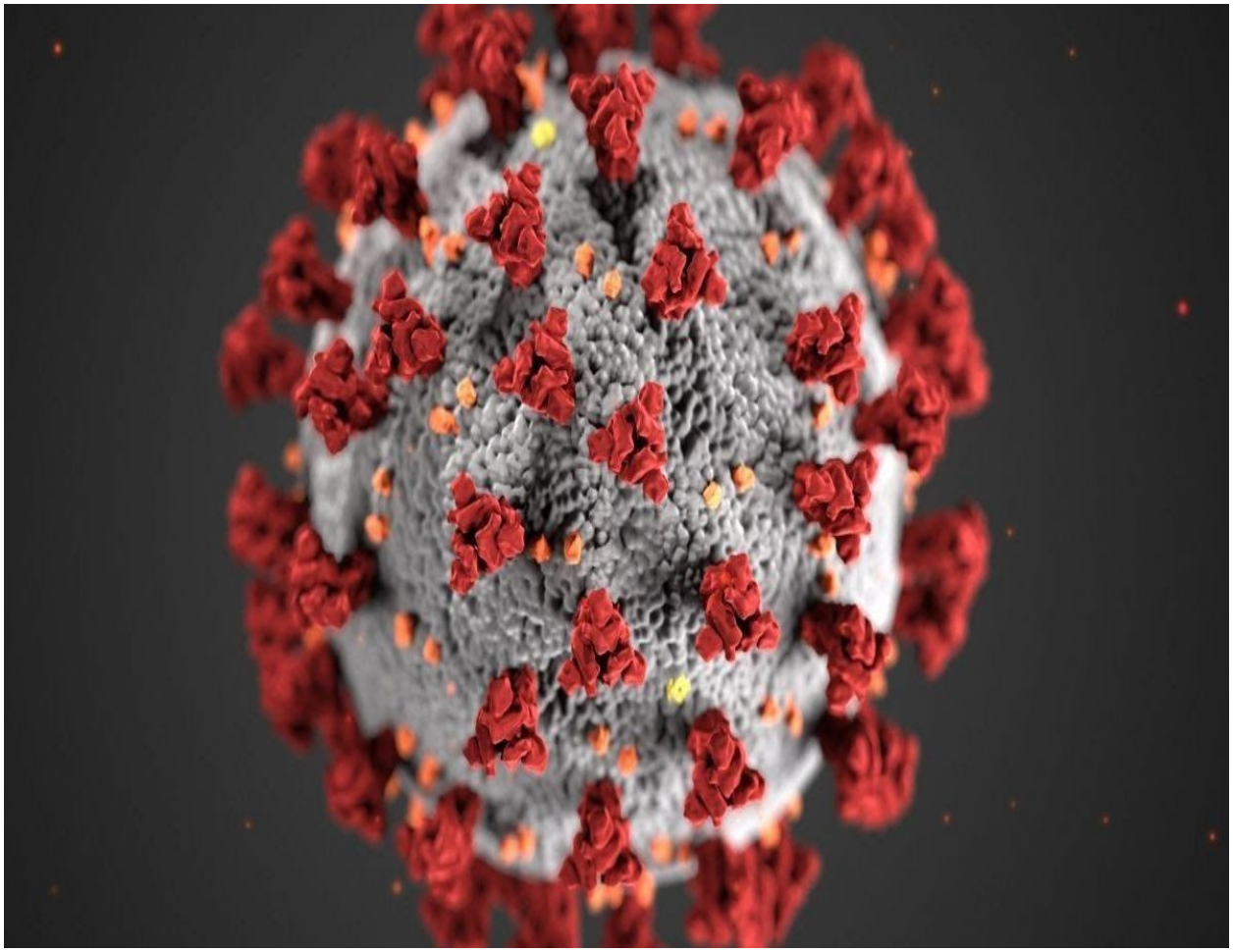
ANALYSIS OF TWEETS ABOUT "CORONA" REPORT

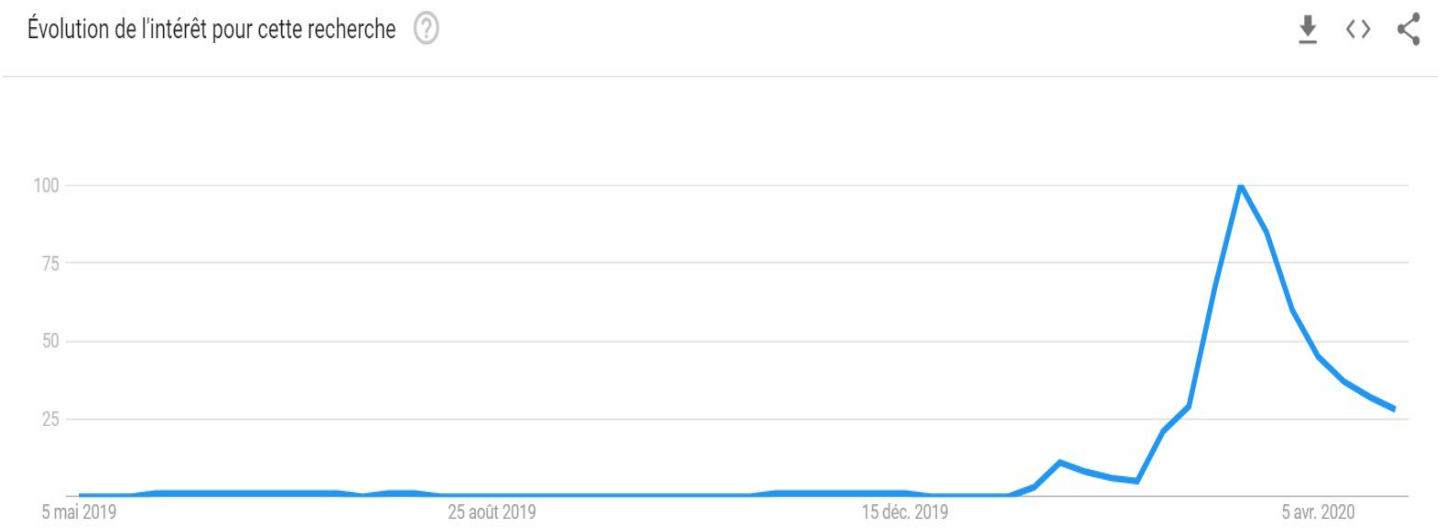


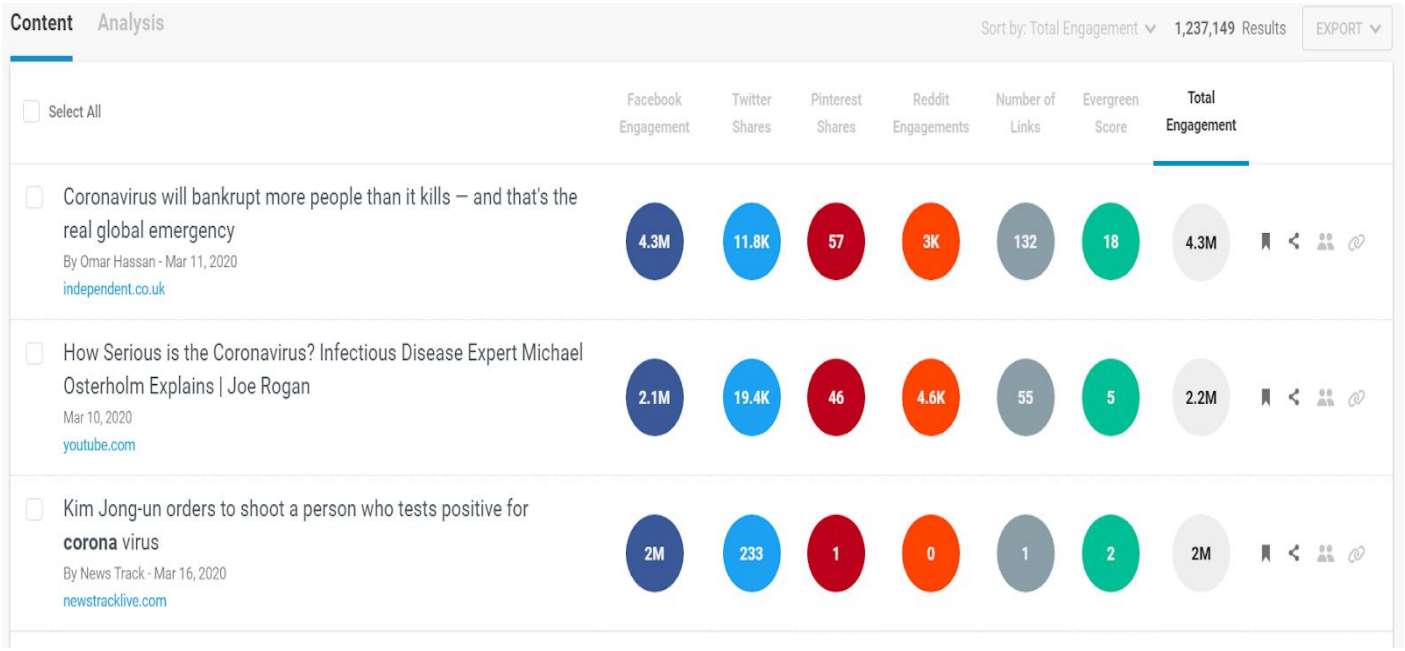
Ahmed Guouader

Background

It’s obvious and clear that corona or covid-19 is one of our main topics everyday and it involves everyone all over the world because our situation became dependent on the evolution of this virus.

My query is corona, i want to download as maximum as possible of tweets to get more representative results.





Content Analysis

Sort by: Total Engagement 1,237,149 Results EXPORT

Select All	Facebook Engagement	Twitter Shares	Pinterest Shares	Reddit Engagements	Number of Links	Evergreen Score	Total Engagement
Coronavirus will bankrupt more people than it kills – and that's the real global emergency By Omar Hassan - Mar 11, 2020 independent.co.uk	4.3M	11.8K	57	3K	132	18	4.3M
How Serious is the Coronavirus? Infectious Disease Expert Michael Osterholm Explains | Joe Rogan Mar 10, 2020 youtube.com	2.1M	19.4K	46	4.6K	55	5	2.2M
Kim Jong-un orders to shoot a person who tests positive for **corona** virus By News Track - Mar 16, 2020 newstracklive.com	2M	233	1	0	1	2	2M

As we can see from these charts that corona has got a lot of interest from people and we can see that in the chart it start increasing from february 2020 to achieve a pic in 5 April and start decreasing but we can see from the second graph the highest number of tweets about this topic which is about 1237,149 results.

RELATED #CORONA HASHTAGS BY INSTAGRAM:

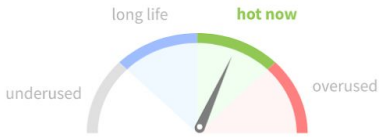
The hashtags that are advised by Instagram. Use them to get similar tags.

#meme	107.72 m	#china	33.16 m	#stayhome	28.02 m
#coronavirus	19.24 m	#quarantine	16.6 m	#covid19	14.62 m
#staysafe	13.23 m	#sick	11.06 m	#covid_19	9.74 m
#socialdistancing	7.25 m	#lockdown	6.39 m	#covid	5.3 m
#stupid	5.14 m	#who	3.23 m	#virus	2.93 m
#irvine	1.37 m	#coronamemes	1.34 m	#covid19	1.33 m
#washyourhands	1.29 m	#covid19	1.22 m	#3dmodeling	1.09 m
#coronavírus	1.09 m	#prevention	1.06 m	#korona	851.76 k
#coronado	817.4 k	#wuhan	728.39 k	#disease	672.41 k
#ranchocucamonga	664.79 k	#coronavirüsü	632.13 k	#becareful	611.58 k
#3drender	559.28 k	#3dartist	540.54 k	#viruscorona	522.95 k

We can conclude from this graph that corona and its related hashtags such as #staysafe #socialdistancing #quarantine #stayhome and its synonymous such as #covid19 are the main used hashtags in instagram these days

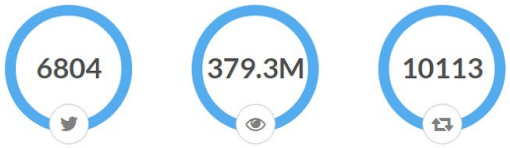
#coronavirus Twitter Hashtag Analytics

FREE Last 24 Hours Stats



RiteTag Recommendation:
Use this hashtag to get seen now

✔ #coronavirus is not banned on Instagram

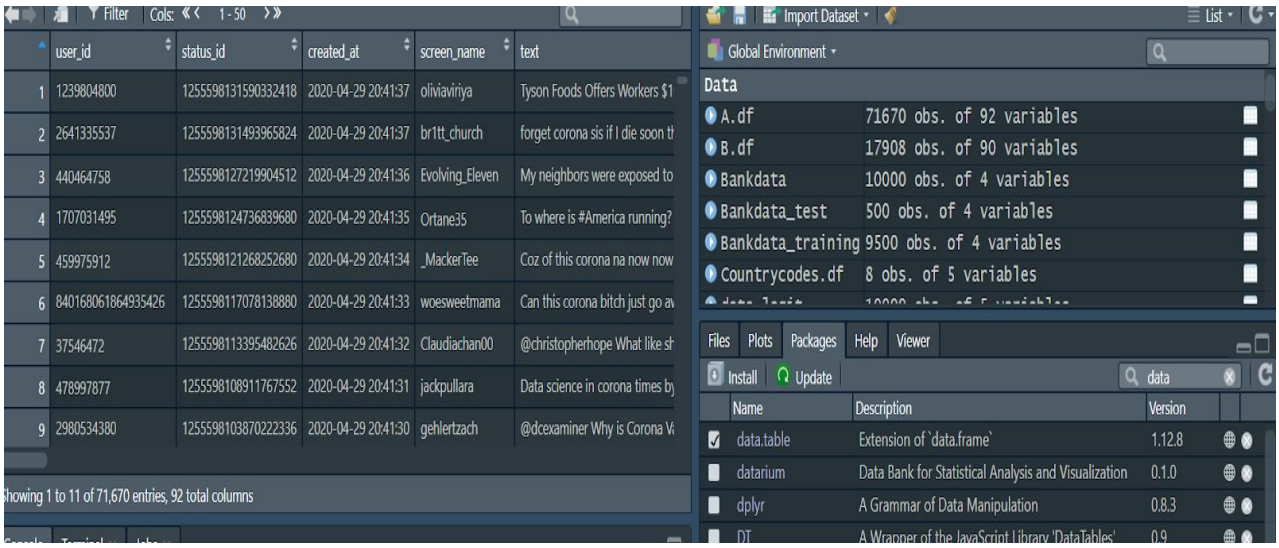


As we can see that coronavirus is considered hot topic now which has 6804 tweets per hour and 10113 retweets per hour and the total of 379.3 M for the hashtag exposure.

HYPOTHESIS

My hypothesis is to see if people are optimistic or not when it's related to coronavirus especially after the decrease of new infected cases in europe, that's why i want to do some marketing analytics in order to see if people are having hope or not regarding the current situation.

ANALYSIS AND INFOGRAPHICS



	user_id	status_id	created_at	screen_name	text
1	1239804800	1255598131590332418	2020-04-29 20:41:37	oliviaviriya	Tyson Foods Offers Workers $1
2	2641335537	1255598131493965824	2020-04-29 20:41:37	br1tt_church	forget corona sis if I die soon th
3	440464758	1255598127219904512	2020-04-29 20:41:36	Evolving_Eleven	My neighbors were exposed to
4	1707031495	1255598124736839680	2020-04-29 20:41:35	Ortane35	To where is #America running?
5	459975912	1255598121268252680	2020-04-29 20:41:34	_MackerTee	Coz of this corona na now now
6	840168061864935426	1255598117078138880	2020-04-29 20:41:33	woesweetmama	Can this corona bitch just go av
7	37546472	1255598113395482626	2020-04-29 20:41:32	Claudiachan00	@christopherhope What like sh
8	478997877	1255598108911767552	2020-04-29 20:41:31	jackpullara	Data science in corona times by
9	2980534380	1255598103870222336	2020-04-29 20:41:30	gehlertzach	@dcexaminer Why is Corona V

Showing 1 to 11 of 71,670 entries, 92 total columns

Global Environment

Data

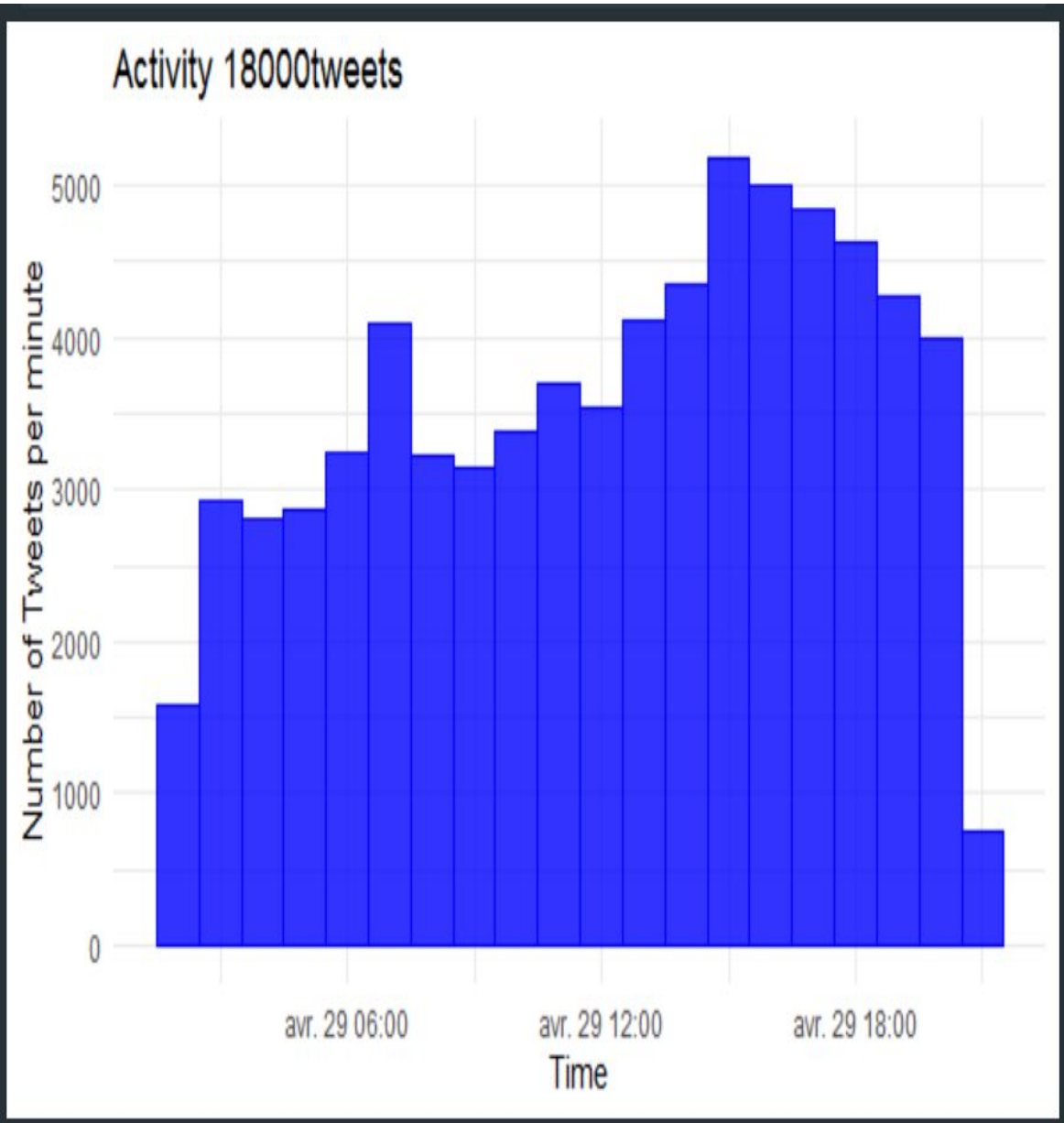
A.df	71670 obs. of 92 variables
B.df	17908 obs. of 90 variables
Bankdata	10000 obs. of 4 variables
Bankdata_test	500 obs. of 4 variables
Bankdata_training	9500 obs. of 4 variables
Countrycodes.df	8 obs. of 5 variables

Name	Description	Version
data.table	Extension of `data.frame`	1.12.8
datarium	Data Bank for Statistical Analysis and Visualization	0.1.0
dplyr	A Grammar of Data Manipulation	0.8.3
DT	A Wrapper of the JavaScript Library 'DataTables'	0.9

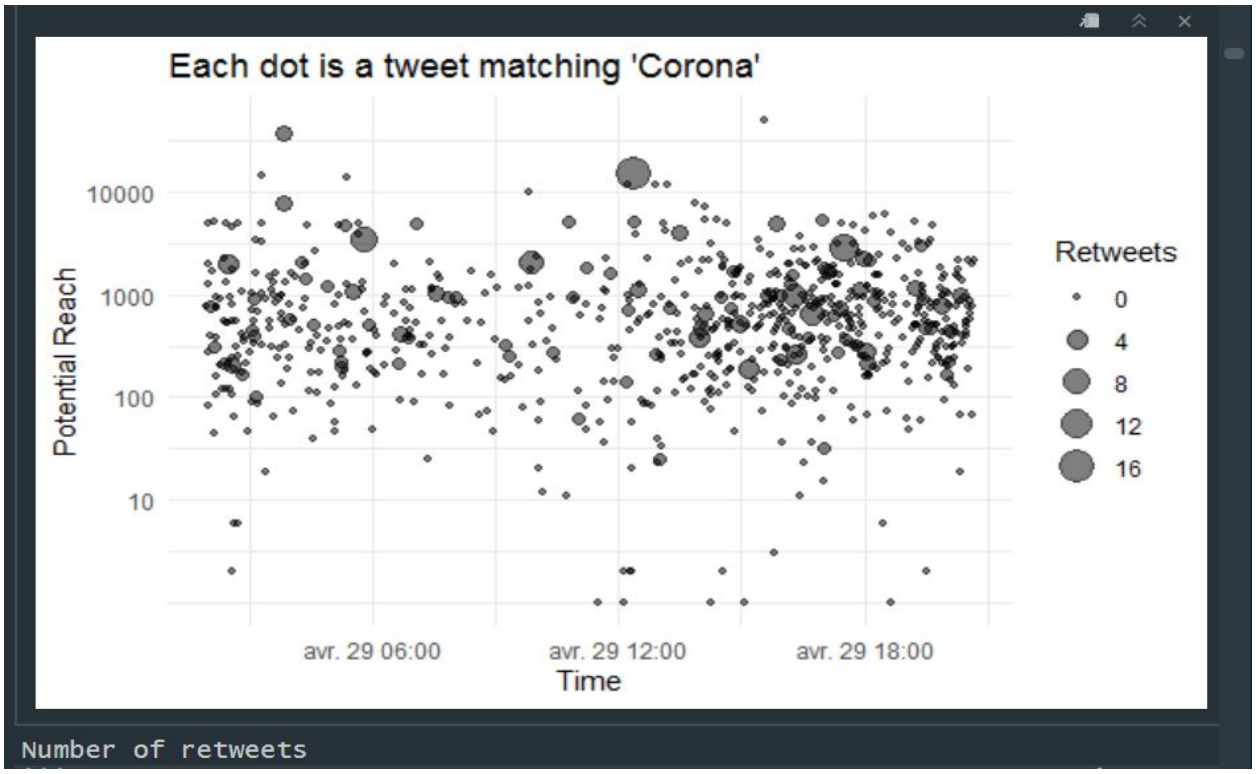
As you can see, I downloaded 71670 Observations which is about 92 total columns.

Here are some examples of columns that we get :

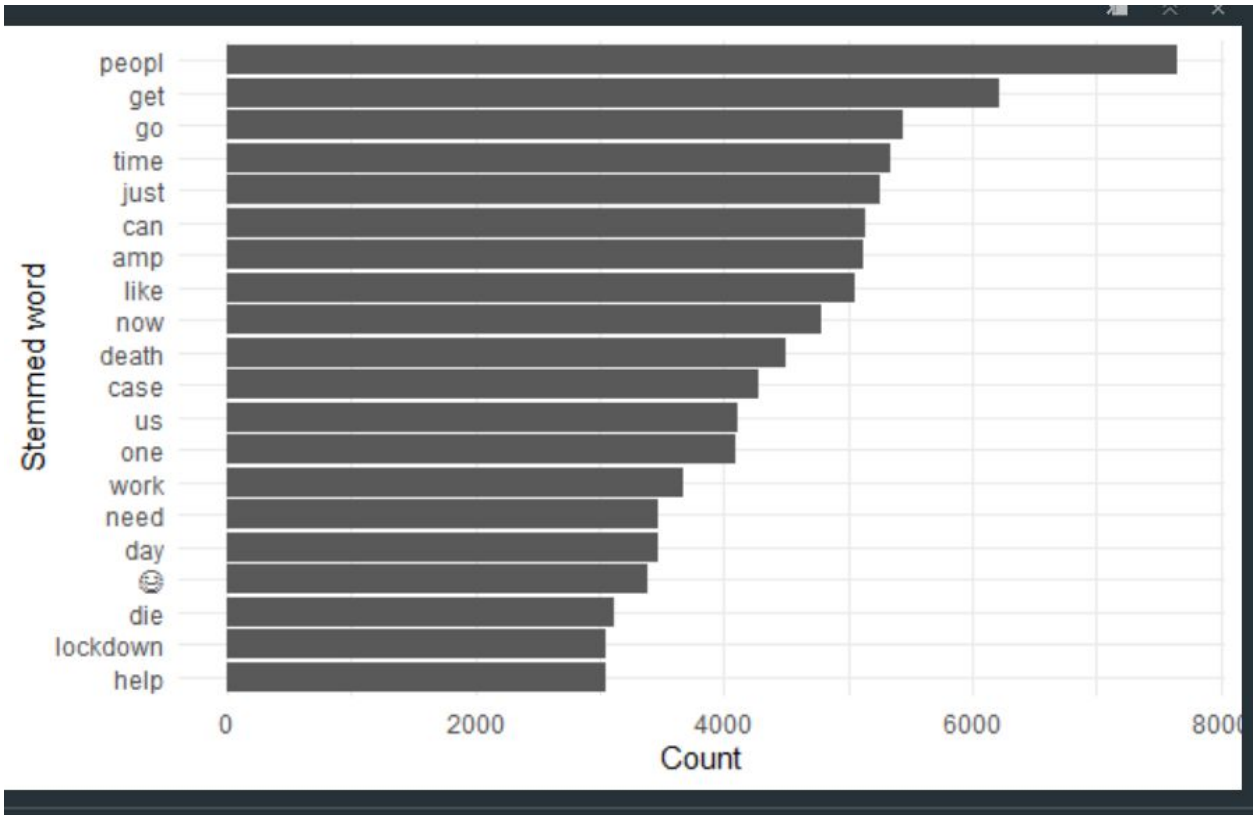
```
 [1] "user_id"              "status_id"            "created_at"
 [4] "screen_name"          "text"                 "source"
 [7] "display_text_width"   "reply_to_status_id"   "reply_to_user_id"
[10] "reply_to_screen_name" "is_quote"             "is_retweet"
[13] "favorite_count"       "retweet_count"        "quote_count"
[16] "reply_count"          "hashtags"             "symbols"
```



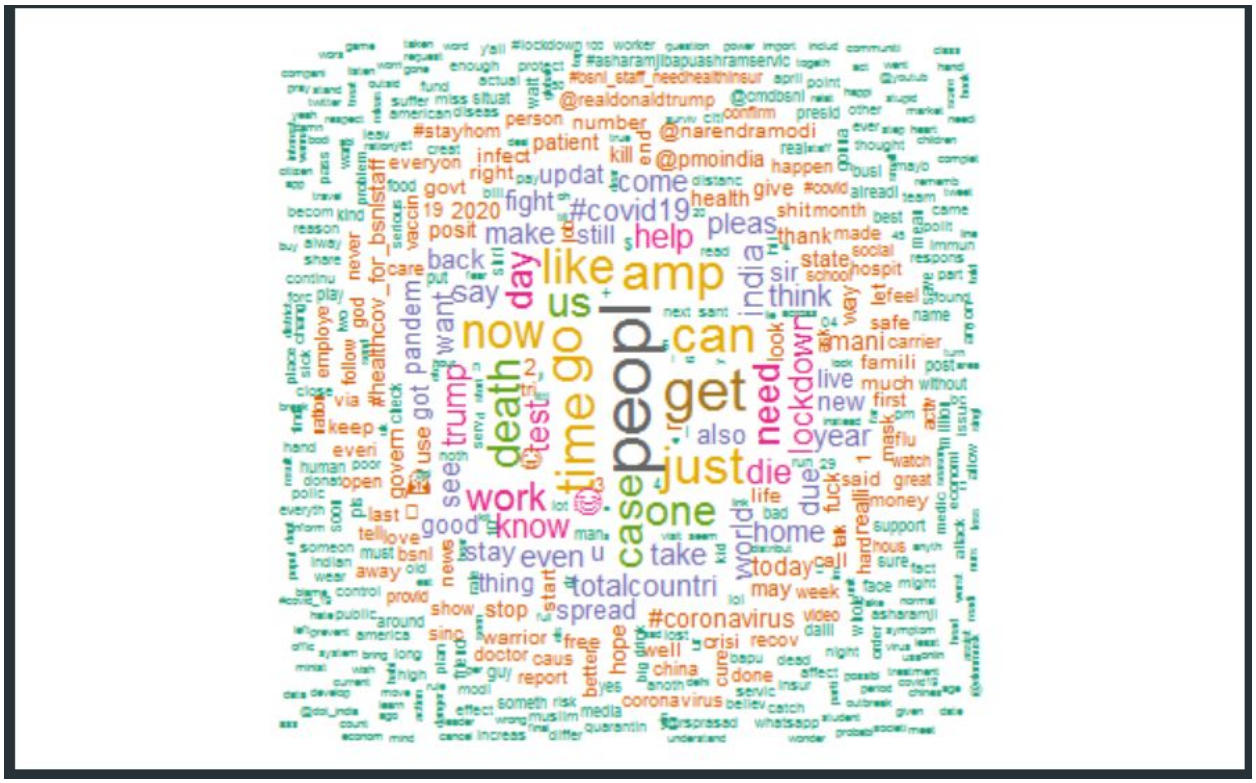

```
The number of tweets vary a lot in the day time that in the early morning
number of tweets is about 2750 tweet per hour, after 6am, number of tweets
increases dramatically and attend a pic of 4000 tweet per minute, then it
keep increasing until it attend other pic of 5000 tweet at 16:00 then it
decreases.The average number of tweets per hour is about 3000 tweet.
```



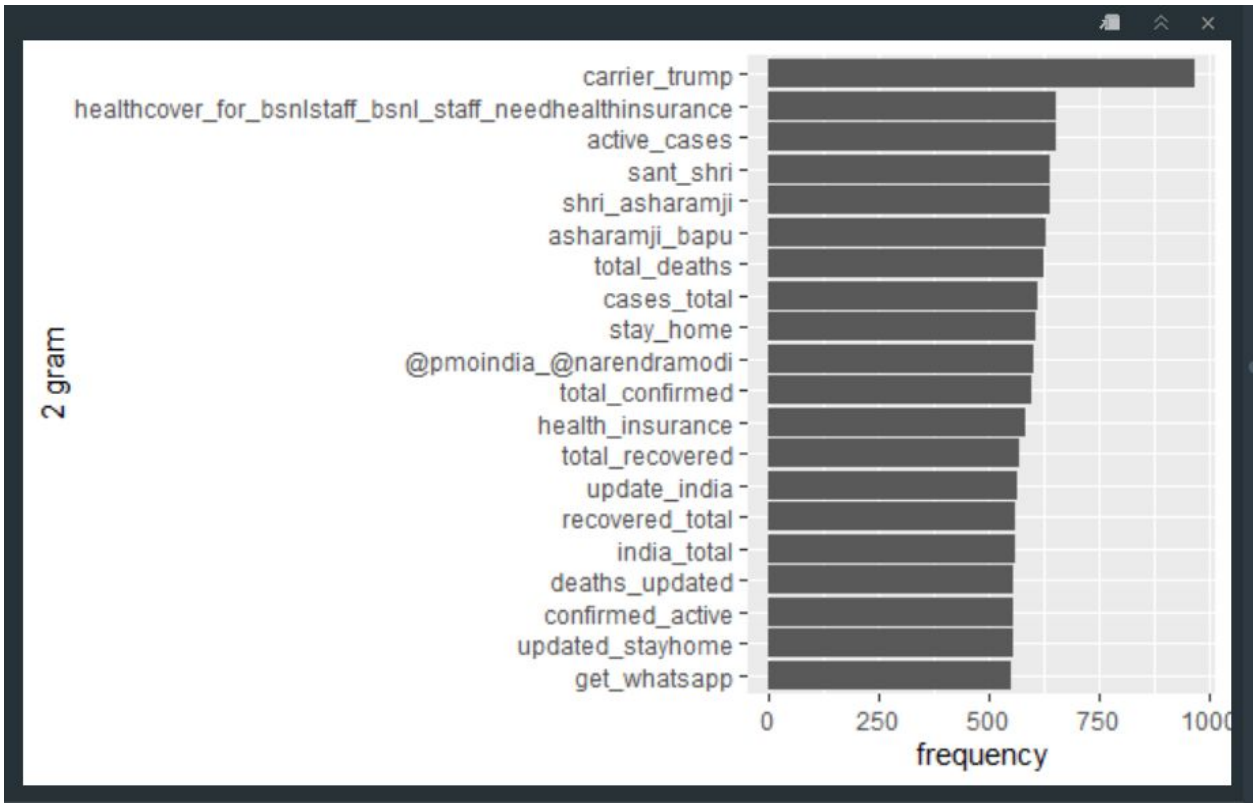
What we can observe from this graph which is about seeing the number of retweets in USA especially and what we can notice is that it vary from 0 to 16 and retweets are especially available from 12:00 to 20:00



After excluding some words that are more likely to be synonymous to corona such as covid_19, covid, @corona, virus, we get this graph which is showing us the most used and interesting words related to our hypothesis are “death”, “case”, “die”, “work”..



From this word frequencies graph which is more general and the size of the word is related to its frequency of appearances. From here we can see some important words in my analysis such as : stop, love, spread, death, care, kill, patient, feel, great,free : these words are important in our sentiment analysis later because it will give us an insight about people’s feelings.

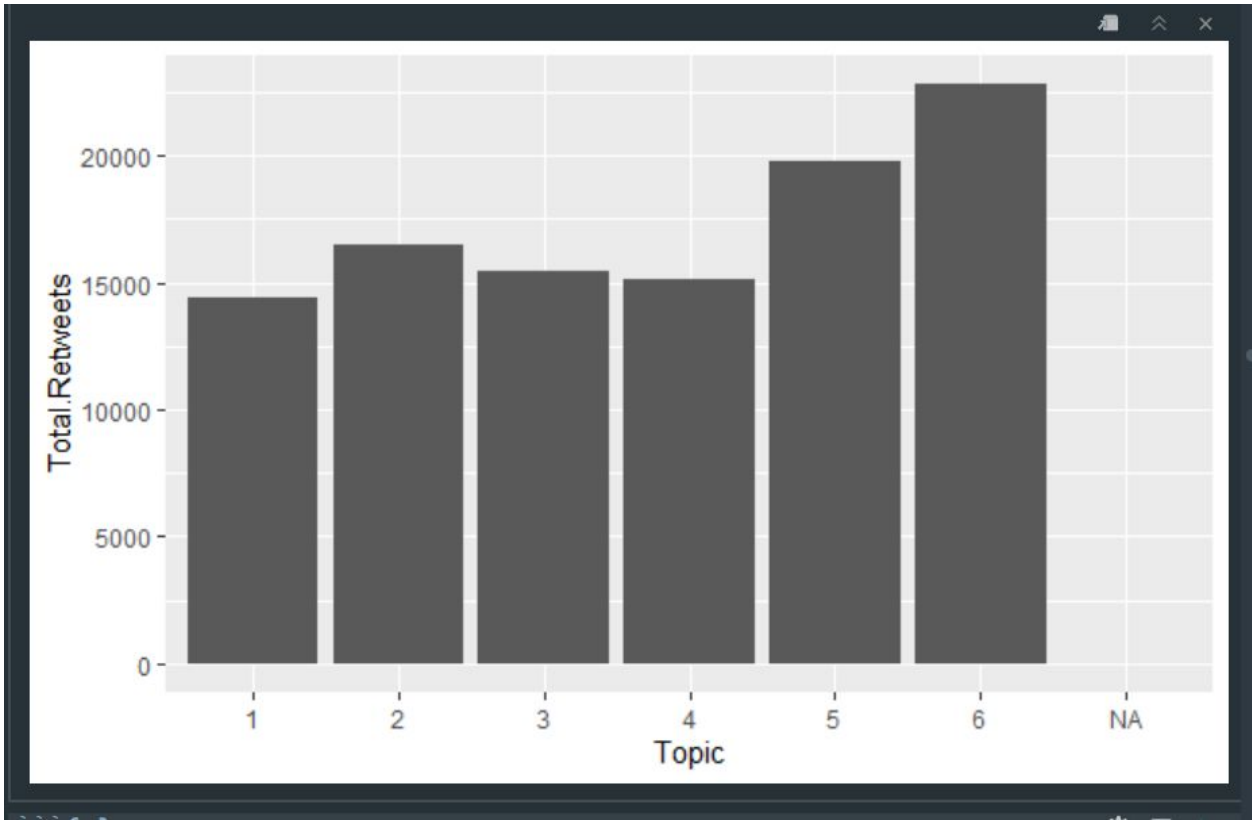


Here i chose n=2 grams to see the most used tokens in my dataset of tweets and the number is about carrier which is a company saved by Trump which is really buzzy these days because of his speechs and current decisions but it's not that important to my topic. For instance, total_deaths which is about 600 times. Also, total recovered is important which has appeared 550 times which is also representing hope of good future.

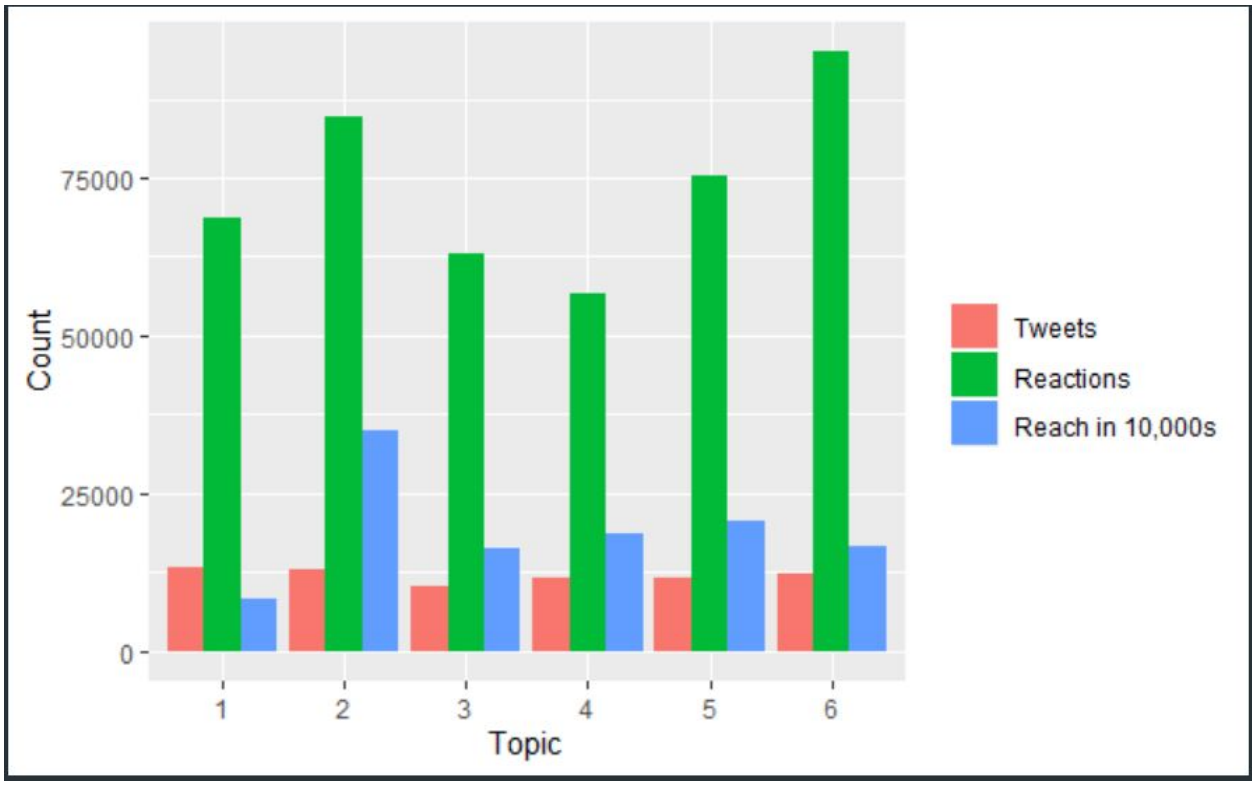
Topic Modeling :

```
     Topic 1   Topic 2    Topic 3      Topic 4 Topic 5 Topic 6
[1,] "peopl"   "get"      "total"      "peopl" "get"   "peopl"
[2,] "test"    "us"       "case"       "day"   "amp"   "now"
[3,] "time"    "time"     "death"      "case"  "one"   "like"
[4,] "now"     "amp"      "go"         "just"  "can"   "\U0001f64f"
[5,] "just"    "like"     "\U0001f602" "can"   "need"  "think"
[6,] "can"     "lockdown" "updat"      "work"  "help"  "sir"
[7,] "countri" "infect"   "amp"        "time"  "bsnl"  "can"
[8,] "die"     "mani"     "india"      "u"     "work"  "day"
```

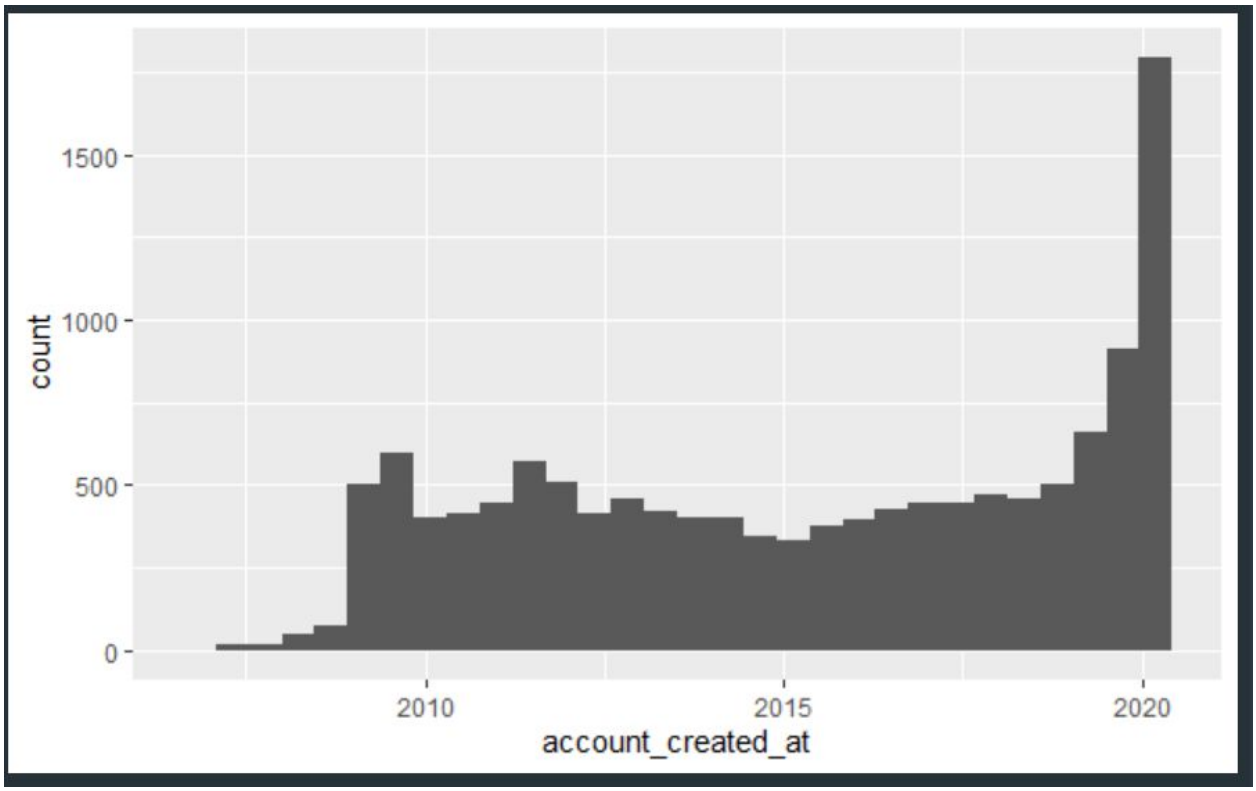
Here are my six Topics.



From this graph, we can observe that Topic 6 has the most number of retweets so maybe later in the sentiment analysis it can be more representative than number 1 as an example.



This graph shows us that in topic 6 we have the largest number of reactions which is about 90,000 reactions while topic 3 has the lowest number of reactions which is equal to 55,000 but for the number of tweets in all topics is likely to be the same.

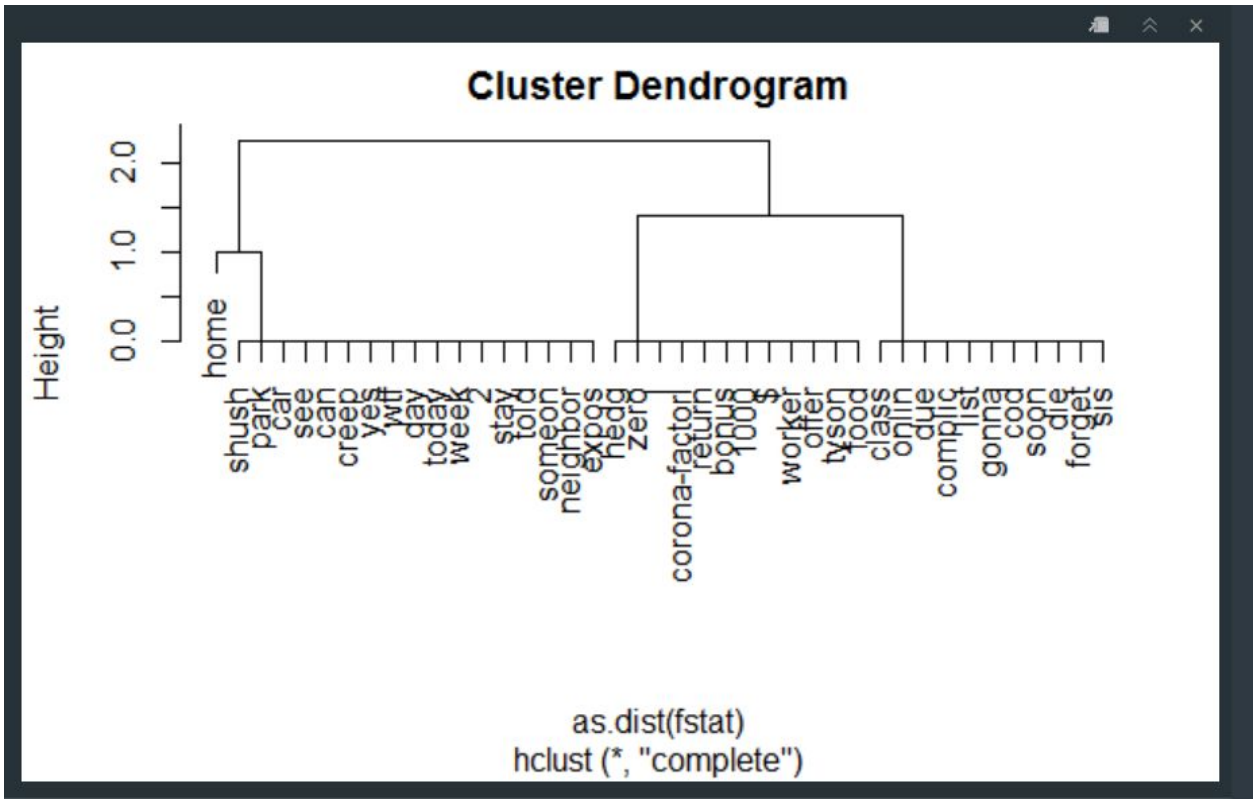


We can notice form this graph that huge number of new accounts is newly created which will may us think of the availability of some spam in our tweets.

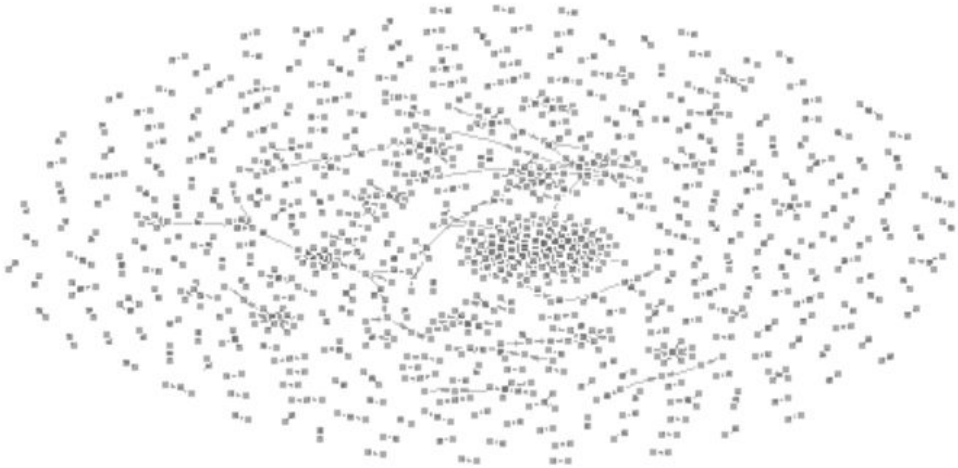
Here is the reorder of each term by associating its beta which represent its appearance in each topic. We can observe that some negative terms such as death, die, infect tend to more appear in topic 1 and especially topic 3 but for the others some neutral words take place which are not really representative yet of what's people thinking and writing about corona these days.

Hierarchical clustering :

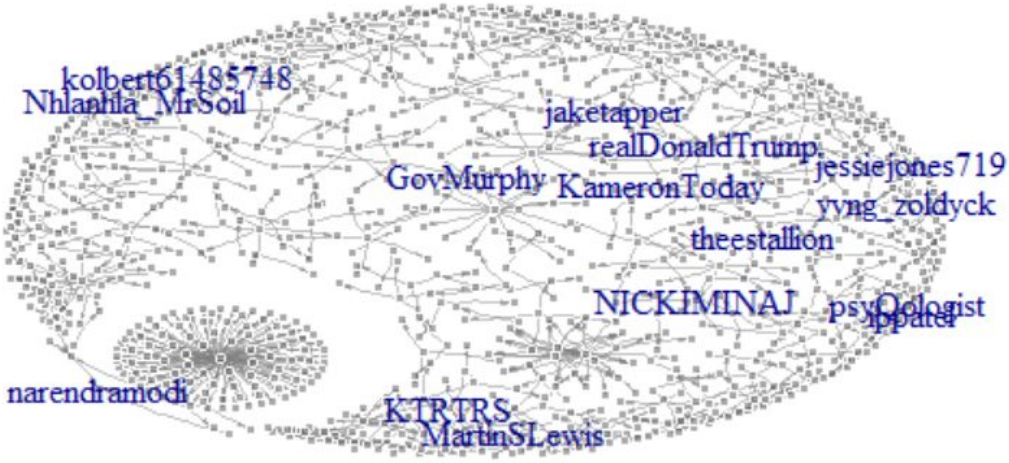
Cluster Dendrogram :



According to this dendrogram, at a height of 2, we get two main clusters, one which is on left is related to normal life context : home, day, neighbor, car, park. The second cluster is more about corona effects there we have : online ,class, corona factory, die..



Here is the Network of users that i get from my data which is represent the network of all the users (actors) mentioned (by @username) in a corpus of tweets.



Here we have the highest ranked users :

```{r}
top.ranked.users(actorGraph.simplyfied)[1:15]
```

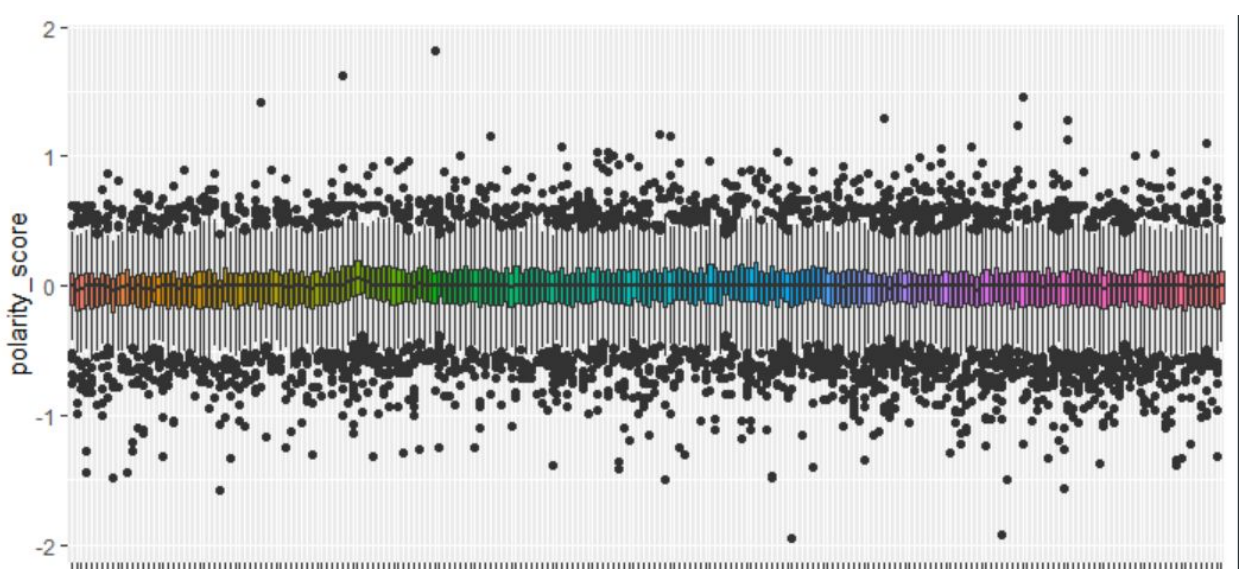
```
 [1] "kolbert61485748" "yvng_zoldyck"    "Nhlanhla_MrSoil"
 [4] "jessiejones719"  "jaketapper"      "realDonaldTrump"
 [7] "psyQologist"     "GovMurphy"       "narendramodi"
[10] "NICKIMINAJ"      "theestallion"    "KTRTRS"
[13] "MartinSLewis"    "KameronToday"    "ippatel"
```

Sentiment Analysis :

```
        Welch Two Sample t-test

data:  sentiment_by_tweet[Topic == 1, ave_sentiment] and sentiment_by_tweet[Topic == 2, ave_sentiment]
t = -4.5498, df = 26126, p-value = 5.393e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.018032187 -0.007173575
sample estimates:
  mean of x   mean of y
-0.04154275 -0.02893987
```

As we can see after doing the T-test p_value is equal to 5.393e-06 which means that the difference is significant.



What we can conclude from these box plots is the existence of a lot of outliers but the mean is likely to equal 0 and the boxplots are being identical by having the same shape, same mean, same min and max

. <fctr>	Freq <int>
virus	15953
trump	2878
pandemic	2321
death	1652
government	1485
fight	1471
crisis	1458
shit	1417
stop	1295
die	1227

1-10 of 10 rows

Here are the 10 most negative terms in the tweets and we can observe that death appears 1652 time, crisis appears 1458 time, die tend to appear 1227 time.

. <fctr>	Freq <int>
like	898
please	443
sir	404
work	355
good	351
care	347
new	290
free	279
money	273
safe	248

1-10 of 10 rows

Here are the 10 top positive terms which appear on our tweets. We can see that free appear just 279 time, safe tend to appear 248 time, good is used in tweets just 351 time.

Conclusion : The negative terms tend to appear more than the good terms.

```
Each time `extract_sentiment` is run it has to do sentence boundary disambiguation when a
raw `character` vector is passed to `text.var`. This may be costly of time and
memory.  It is highly recommended that the user first runs the raw `character`
vector through the `get_sentences` function.Each time `extract_sentiment` is run it has to do sentence
boundary disambiguation when a
raw `character` vector is passed to `text.var`. This may be costly of time and
memory.  It is highly recommended that the user first runs the raw `character`
vector through the `get_sentences` function.[1] "Topic 1"
pos
    like   please positive    right     well    money     good
     833      480      471      400      354      349      334
    work     hope      big
     307      289      213
neg
   virus      flu    trump      die    death     died     shit
    3315      731      529      496      466      433      384
    stop    fight pandemic
     336      273      272
[1] "-----------------------------------------------"
```

```
vector through the `get_sentences` function.[1] "Topic 6"
pos
    like   please      sir     good      new     work     free
     909      623      560      375      329      274      267
    food inspired     safe
     240      229      228
neg
   virus    trump pandemic      bad     fuck     shit    fight
    2732      830      507      355      353      339      293
    lost   crisis     poor
     260      259      250
[1] "-----------------------------------------------"
```

Here we have the frequency of positive and negative words in topic 6 and 1. I choose these two topics, because topic 1 has the lowest number of tweets and topic 6 has the highest number of reactions. As we can see here that deep negative words such die, death, bad, lost tend to appear more than real positive words such as free, inspired,

safe.

CONCLUSION

After doing all of these analytics, i figured out that people when they think about coronavirus they tend to be more pessimistic than what i thought because these days the number of new cases tend to decrease in many european countries but it could be that the situation in USA is affecting the analysis due to the high spread of the virus, it could be also related to the unexistence of a vaccine until this moment.

REFERENCES

1. https://app.buzzsumo.com/content/web?q=corona
2. https://trends.google.com/trends/explore?q=corona%20virus%20update
3. https://ritetag.com/hashtag-stats/coronavirus