**Objective of the project  :**

The aim of this objective is to analyze the tweets related to the ropic Corona virus. In fact , I took a query named ‘COVID\_\_19’ and download about 71 000 tweets.

The main goal of this objective is to create data and text visualization charts from the tweets downloaded.

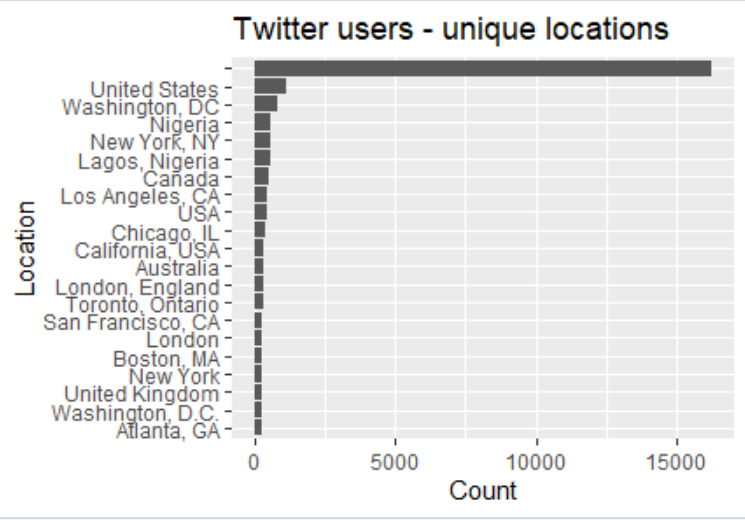
What is expected from this report is to find more positive insights than negative since we are in a period of deconfinement in many countries in the world.

**Data visualization :**

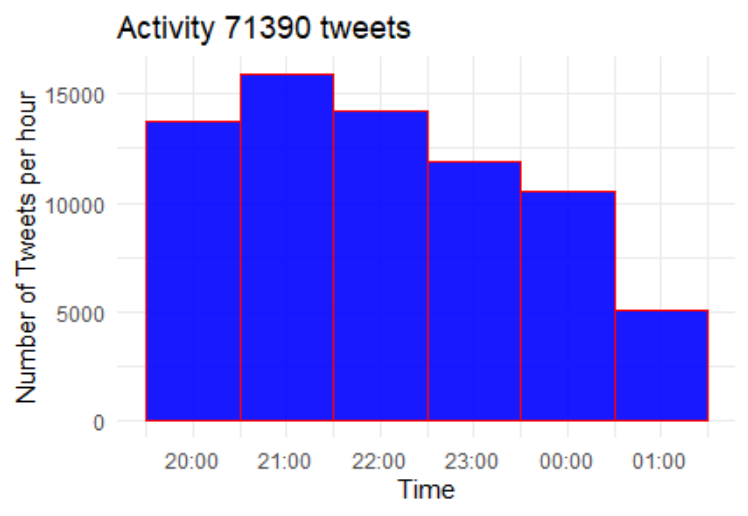
Our dataset contains 71682 rows and 90 columns

Twitter users per location :

The majority of users are from United States ( more than 150 000 users ) then , Washington and Nigeria come after ( less then 3000 users )

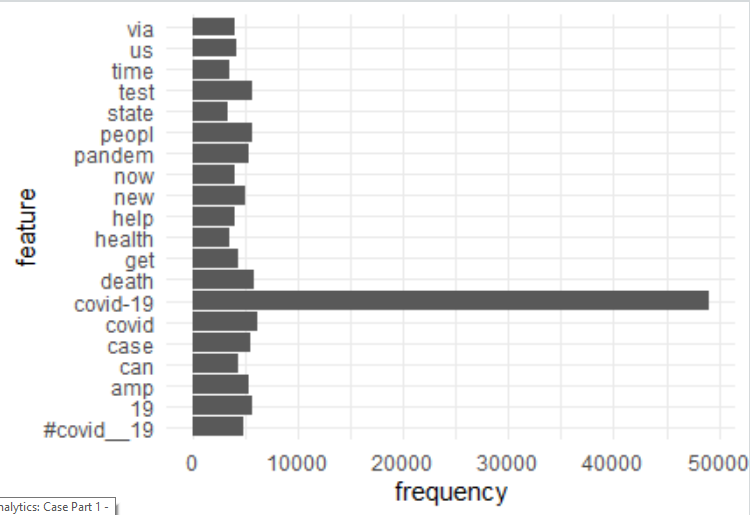


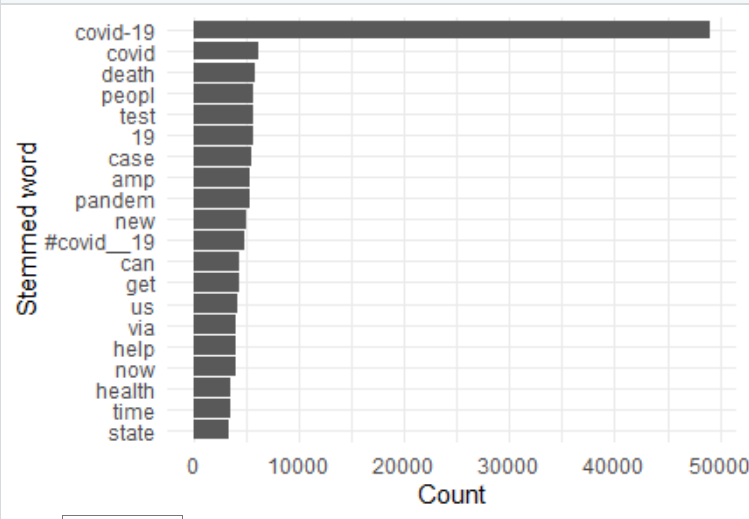
As we can see , the distribution of the tweets activity in the last 5 hours is skewed to the left ( at 21PM and 22PM we have the highest number of tweets ) . At 21PM we have 150 000 tweets and at 1AM we have only 5000 tweets



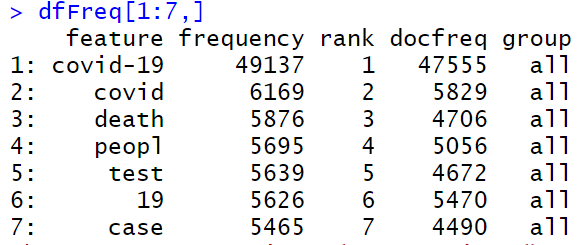
**Word frequencies :**

Covid-19 is the most used word in the 71 thousands tweets . It was used almost 50 000 times. Then it comes ‘covid ‘ and death == ) People are using negative words more then positive ones like ‘help.

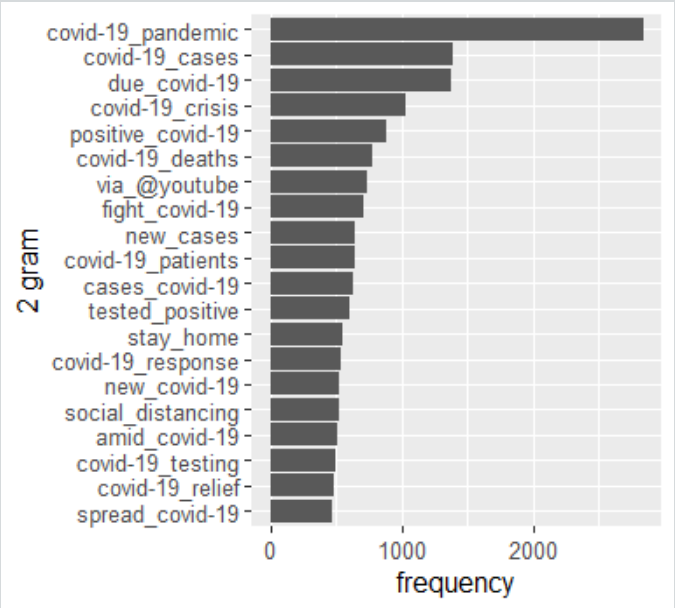




As we can see , covid-19 is presented 49 137 times in 47 555 documents which is a huge number comparing to other features ( All of the other six features don’t reach the frequency of covid-19 alone)

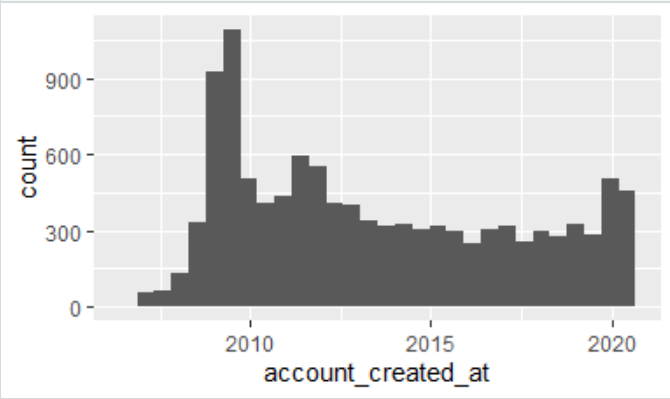


Using the 2 Grams method , we find that ‘covid-19’ and ‘pandemic’ , covid-19 , crisis ‘ , ‘covid-19 , cases ‘ are used more than 6000 ==) People are associating the virus with Negative words and negative insights ( it’s a pandemic , many cases , causes crisis …)



Time distribution of the account creation :

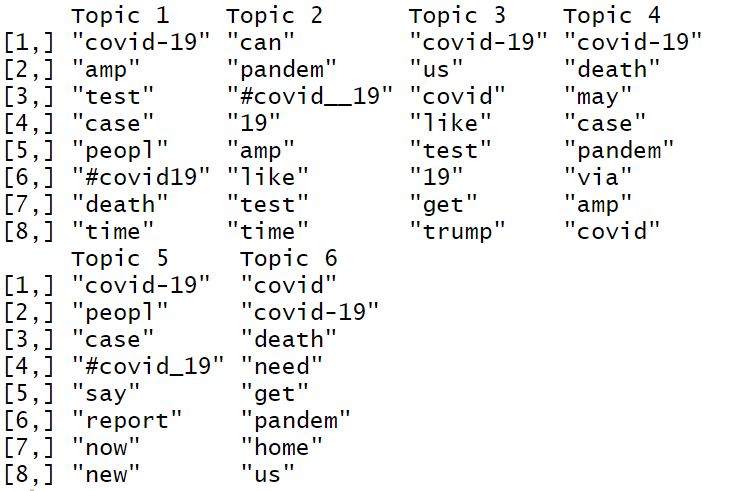
As we can see , the majority of the accounts are very old ( 10 years old ) , almost 2000 accounts are created in 2010 and before. We have less than 10 000 created in 2020.



**Topic modelling :**

**K=6 , n=8**

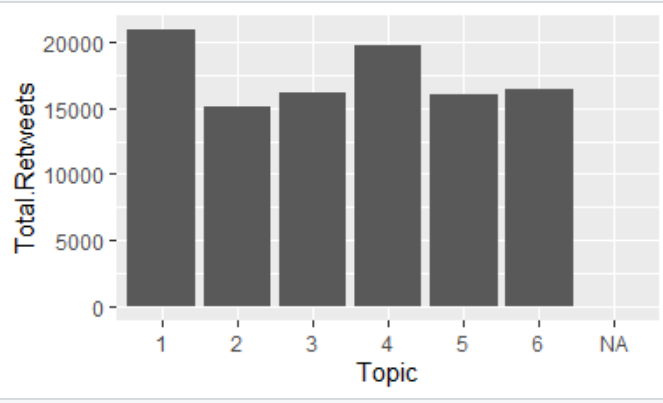
In the 6 topics , Covid-19 is present the most , and as we can notice there are topics more negatively oriented than others ( topic 1 , 2 , 4 )



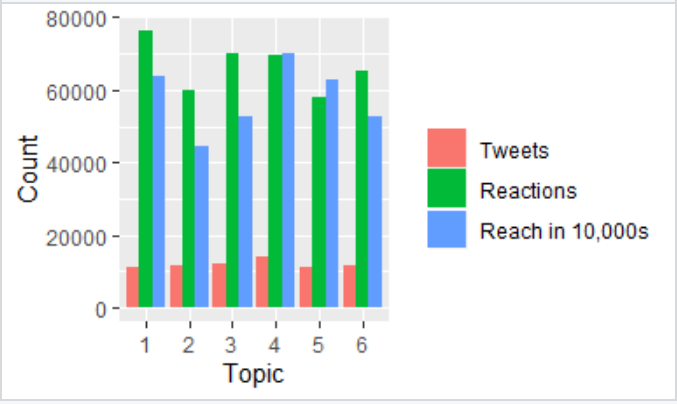
Topic = 1 DIM 11074

Here , I plot the total retweets per topic ( As we can see , we have a total of 20 000 tweets for the topic 1 and almost 20 000 for topic 4 ) ==) People are responding more on these two topics which have more negative insights ( covid , death , pandem ..)

The lowest one is topic 2 ( less than 15 000 retweets )

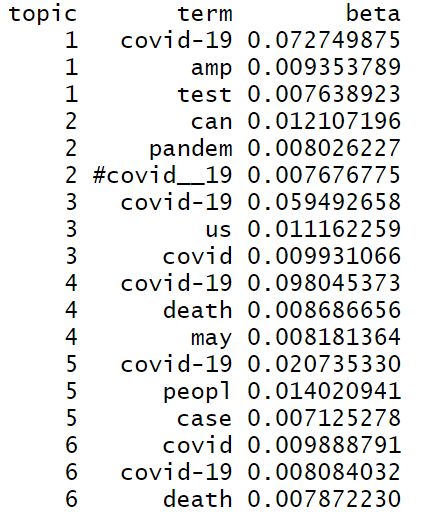


As we can see here , the most topic that are reacted on are topic 1 and 4 , and the lowest one is topic 2 ( topic 1 reachs almost 80 000 reactions for 10 000 tweets reached on 10 000 seconds , and topic 4 has 70 000 reactions reached on 70 000 seconds for less then 10 000 tweets )

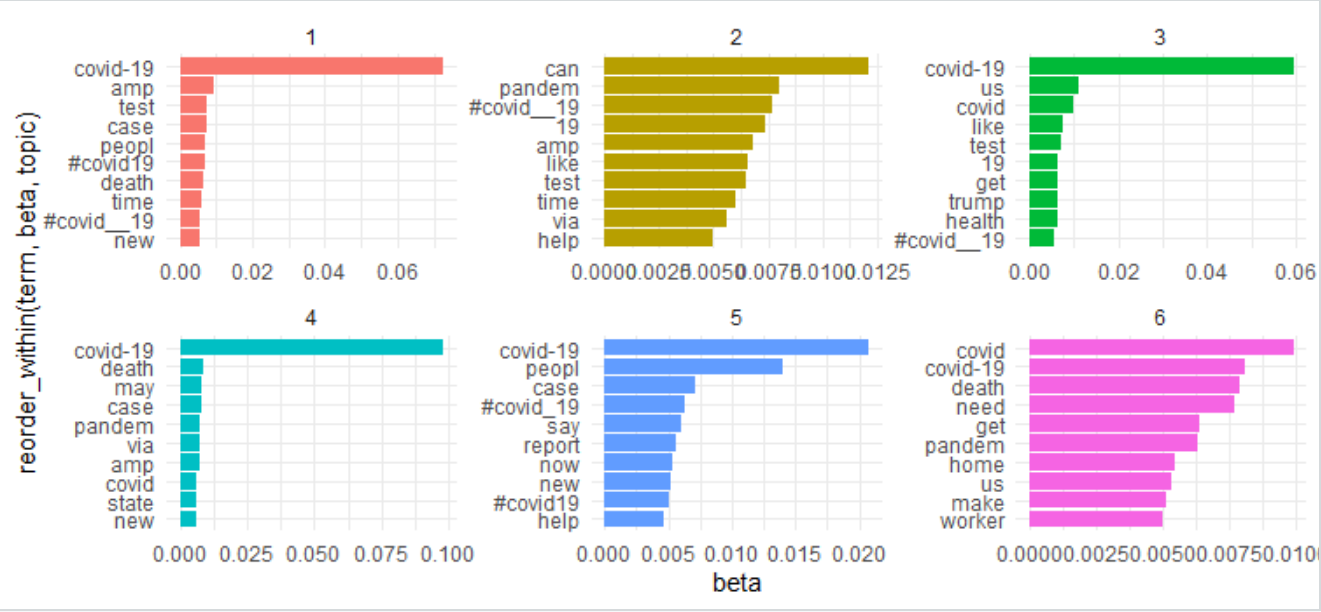


As we can see here , we have beta coefficients for each term per topic , and as I mentioned before , Topic 1 has the word covid-19 most used ( beta 0.072 very high comparing to other beta )

For topic 2 , can and pandem have the highest beta , covid-19 for topic 3 , 4 ,5 and covid for topic6



Here ,I plot the reorder for each term per topic and its beta ( same last analysis )

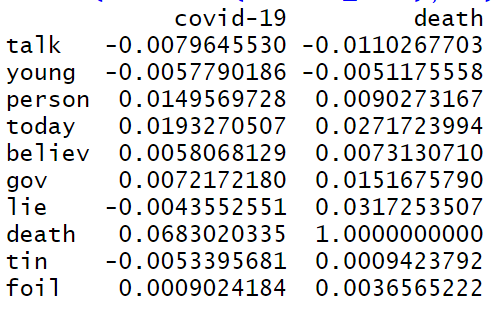


**Hierarical clustering**

Here I wanted to see the relation or association between the two most presented word with other features from the tweets ( covid-19 and death )

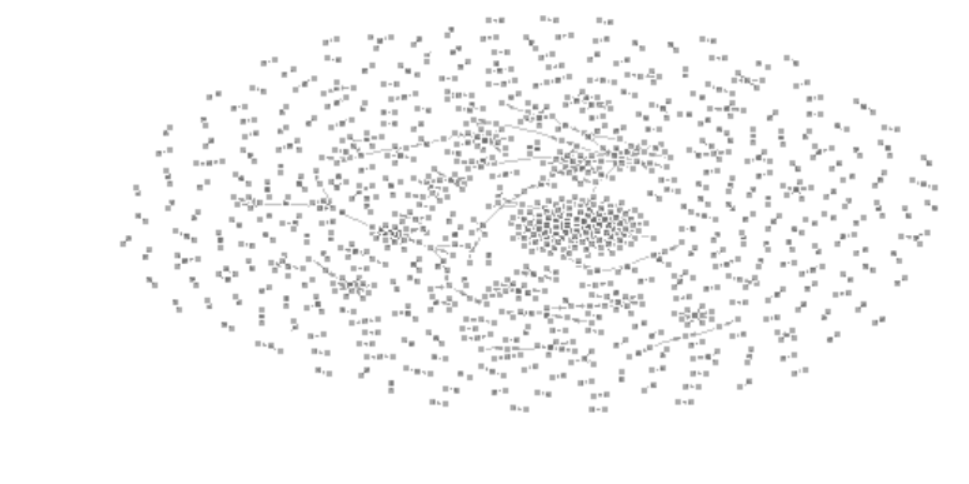
I Found that covid-19 is more associated with death ( 0.068 beta ) , today ( 0.019 ) ==) peple relate death to covid-19 , and many talks about the present moment ( situation today)

For death , it’s more associated with lie ( people maybe don’t give trust to geverment announcement about death numbers )



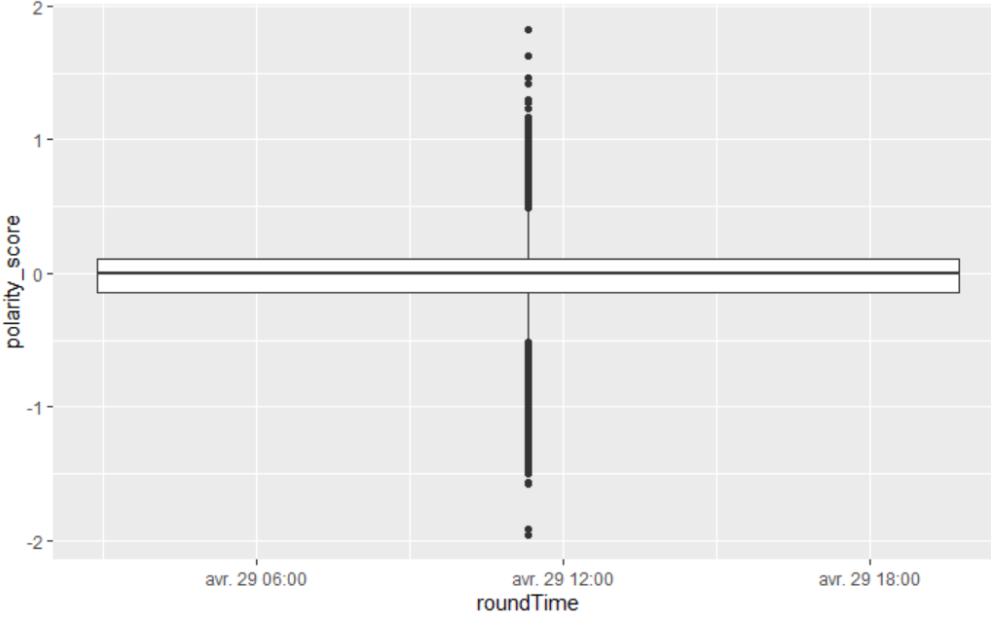
**Network**

This is the shape of network distribution of our tweets

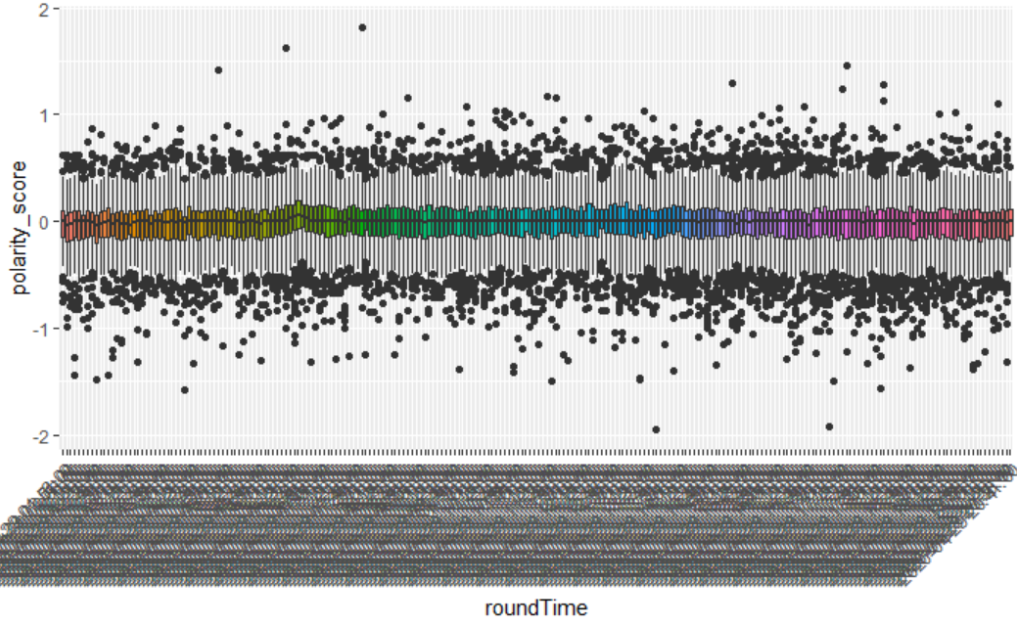


**Sentimental analysis :**

**As we can see the polarity score that shows the score or degree of sentiment within our tweets is centered around 0 .**

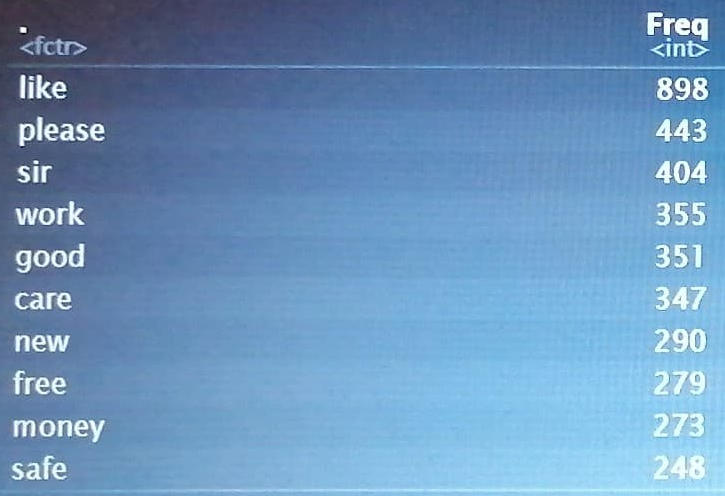


As we can see , we have almost equal degrees of sentiment in the second half of time , But in the last half of time , we have more negative sentiments ( points are more in negative side )



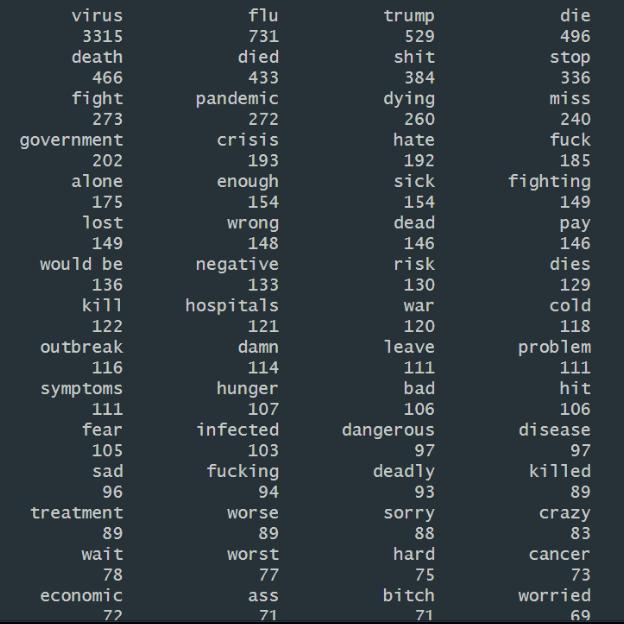
**Positive words :**

**We don’t have a lot of positive words comparing to negative ones ( like , please , sir are the most presented for + words )**

****

**Negative words :**

**We have many more negative words than positive ( as we can see , virus 3315 , trump also 529 ==) people are seeing Trump as negative sentiment for them , die , death , sh\*t ..)**

****

**Conclusion**

**As expected , virus and covid are the most frequent words that we found in our analysis , also it was not expected to find this degree of negative insights , negative topics and negative sentiments that are more presented in the 70 000 tweets than positive ones despite of the deconfinement which is soon in many countries .**

**Thank you**

**Ala Bouslama**