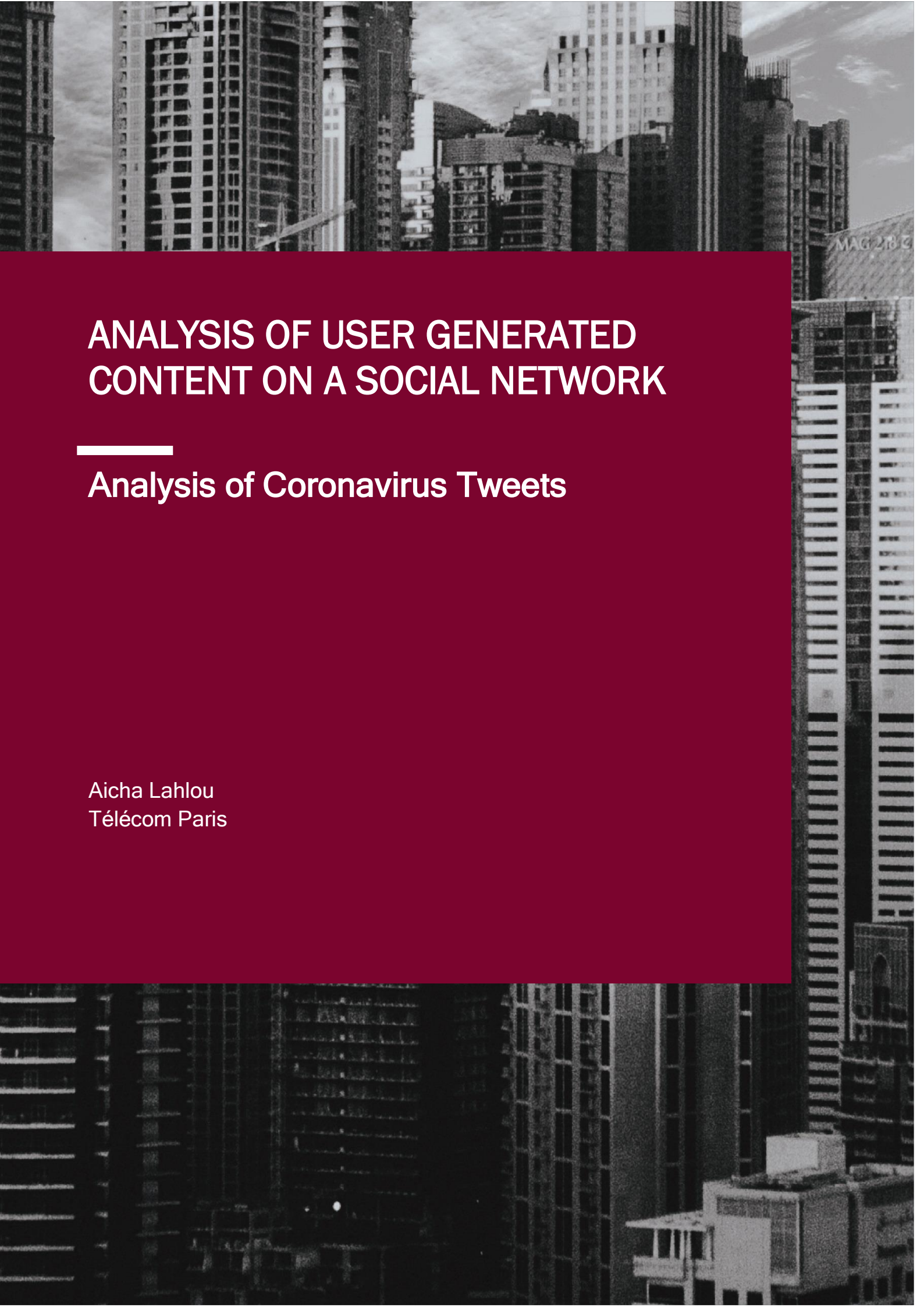


ANALYSIS OF USER GENERATED CONTENT ON A SOCIAL NETWORK

Analysis of Coronavirus Tweets

Aicha Lahlou
Télécom Paris

Table of Content

I. Presentation of the study

Since last year, the world has been completely turned upside down by the coronavirus. Up to today, more than 68 million of people have been affected by the virus, and more than 1.55 million of people died because of the latter. This pandemic has also drastically impacted the social, political, and economical systems at a worldwide level and has consequently transformed our societies, and more specifically the way humans interact with each other.

Last year, more than 50% of the world's population experienced a lockdown, and most countries have applied social restrictions in order to regulate the evolution of the epidemic. These restrictions have led to the restructuring of our social, economic, and work organizations. They also have a significant impact not only on our mental and physical health, but also on the way we interact with each other. Indeed, studies have shown that since last year, the commitment to social media is strongly increasing, and has increased globally by 61% compared to the usage rates usually observed. Thus, the social link among people is now mainly provided by social networks, and that makes people express and share more content on these new virtual platforms.

The purpose of the study is to analyze the content related to the covid, and this unprecedented situation on one of the most used social network platforms: Twitter. Other aspects of the study will be related to the study of the different groups of similar users that share content about covid on this platform, and to analyze the general sentiment of text posted by users.

II. Collection of Data

Data has been collected with twitter using Twitter API. Given the limitations of Twitter API, only recent tweets have been gathered, which have been acquired during multiple sessions of collection. Therefore, the final data that will be used for the analysis is the concatenation of all data collected on Twitter. In total, more than 2.4 million tweets have been collected.
The key words used for this data collection are: "corona", "coronavirus", "covid", and "covid19". I chose these words, because they are directly pointing to the actual virus, and the situation. I chose not to include words such as pandemic, disease, confinement and others because they could be vague for describing this specific situation. And they are not necessarily pointing to the virus, but also to regulations or mental health. However, the previous fact does not exclude the presence of these words in the tweets collected. The tweets gathered are in English, and from all over the world.
In order to nurture the study 20 other characteristics of tweets have also been collected, related for example to the characteristics of the twitter users, the time of posts, geographic information, etc.

III. Data preparation and structuration

A. *Initial dataset*

The final dataset contains 2429302 lines and 21 columns. The 21 columns represent features associated with the tweet: the user characteristics, dates of posts, the text of the tweet, some features related to the popularity of the user and others. For more information, please refer to the data dictionary in the appendix of the report.

B. *Data Cleaning*

Many transformations have been performed to clean the data and make them consistent with their definition. In fact, Twitter API provides a data definition on their website.
Features related to dates have been normalized in order to make them recognizable by the system as dates. Some features related to the verified profile of users or retweeted have also been transformed to make them consistent with their definition as Boolean type features, I have also decided to drop null values from the dataset, because dropping them from the database does not seem to impact the purpose of the project, since the volume of data remaining after removing them is still sufficient. At the end, the final dataset approximatively contains 1.8 million of tweets, and more than 21 features.
NB: The text preprocessing part is explained and performed in the part related to natural language processing.

C. *Feature Engineering*

The objective here was to add more information using the features provided in the dataset.
A first line of work consists in transforming locations label provided by non-structured and nonvalid string into valid geographic format that is analyzable. In fact, the location feature is a string and is thus not recognized by the platform as a location. In the dataset, we could find some substantially creative examples such as: "my mind", "my world"… In

order to work on consistent locations, the work has been performed on location data that occur more than a certain number of times in the dataset collected. And geolocator library has been used to retrieve a latitude and a longitude matching from the text informed. In order to accelerate the processing on the main notebook, this part has been performed separately, and the location information is saved in a separate file, easily downloadable and mergeable with the general data frame. The second line of word consisted in transforming dates features into more easily analyzable features such as date, weekday or month.

IV. Data visualization and understanding

The objective of this part was to perform analysis on the different variables of the dataset in order to better understand the information provided by the features in the general dataset.

A. Analysis of categorical features

First, the aim was to analyze categorical features. Several functions were defined in order to do so: plots of repartition of occurrence, plots of repartition of percentage of a value… The main conclusions of this analysis are listed here.
We can see that some users are highly active on the platform. In fact, some of them have generated more than 5000 tweets (e.g Jeremy Hume) on it. This does not reflect the relevance of the information provided, but just show that some users are highly interested in the subject and publish a lot in a short period of time. We could also notice that more than 13% of tweets collected are published by verified users, whereas they represent only 0.065% of Twitter accounts. Besides that, we have information about the utilization of the platform. In fact, we could notice that the main sources of utilization are Twitter Web APP, and mobile applications such as Twitter IOS, or Twitter for Android. These three represent more than 60% of the sources of utilization in terms of number of publications per application.

B. Analysis of numerical features

The analysis of numerical features was mainly based on distributions, boxplots, and density plots. The main conclusions are presented below.
We can see that the average number of followers by users is approximatively 2500. However, the standard deviation is substantially high, that means there are huge differences in terms of number of followers. Indeed, the minimum number of followers is 0, and the maximum in the data collected seems to be more than 1.8 million.
We can also notice that some users are highly interactive with the platform and others not so much. As a matter of fact, some of them have liked more than 1 million posts since the creation of their accounts, and that reflects an overwhelming use of the platform by these individuals. We can also notice that among all posts only a few are re-shared (retweeted) on the platform by other users. These posts represent a modest contribution of the shared content on the platform, but once they are retweeted, they tend to be broadcasted on a substantially large scale.

C. Analysis of the tweets collected in time

The graph on the right represents the evolution of number of tweets collected per day. We can see that the tweets collected were posted between the 17th of November and the 9th of December. This evolution does not show the number of tweets published per day, but the number of tweets collected.
However, this graph does not show the real evolution of number of tweets published on the platform, since the data used in this graph is only related to the data collected for this study. Even, if the number of tweets seems to increase, we cannot confirm the evolution of the interest in the subject, since we do not have access to all the data related to the subject.

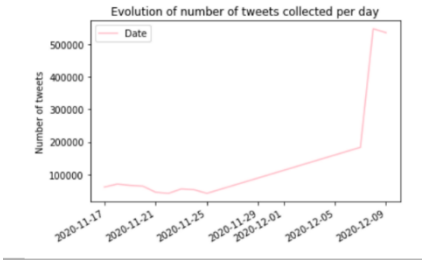


Figure 1: Evolution of number of tweets collected per day

D. Analysis of the popularity of the platform

A study about the popularity of the platform has been performed. In order to assess the popularity and the evolution of the utilization of the platform, I chose to work on the evolution of the number of accounts created per day. This evolution is represented in this opposite chart. We can see that the number of accounts created exploded during 2010, when the platform's audience exploded. We could possibly explain this evolution by the utilization of the platform by new celebrities (e.g., Barrack Obama), or magazines and other medias (e.g., CNN). We could also notice that since the beginning of the coronavirus, the platform seems to attract more people, in fact on the chart since 2020 the number of users has increased in comparison with other previous years, when the number of accounts created per day seem to be stable. This could be explained that the social media where increasingly used during the lockdown.

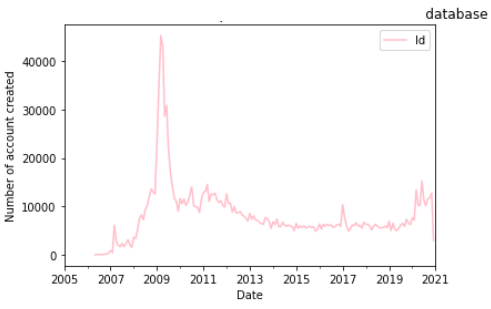


Figure 2: Number of creation of account per month

E. Analysis of the popularity of the users

I was also interested in understanding more the different types of popularities among users. There are two types of features representing the popularity on the platform: the number of followers and the number of friends.

The users that have the highest number of followers are usually users that are well known for their political engagement / position, or for being a popular media. We can notice among the most followed accounts: Barack Obama, Donald Trump, Joe Biden, CNN, BBC, The New York Times… The popularity of these accounts also shows that this platform is particularly adapted to American context.

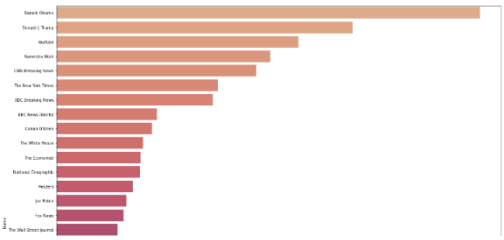


Figure 3: Number of followers

When taking an interest in the users who have the highest number of friends, they do not seem to be the ones who either have high political positions or belong to an information media platform. After further analyses, we can say that these figures are generally influencers, that are the new modern influencing figures among the young generation.

F. World maps

When looking at the map, we can see realize that most tweets collected come mainly from the US, the UK and Europe, India, South Africa, and Australia. This could be explained by the fact that the tweets collected are in English, and these countries are English speaking countries. Moreover, these regions of the world represent the places where the platform is most used.

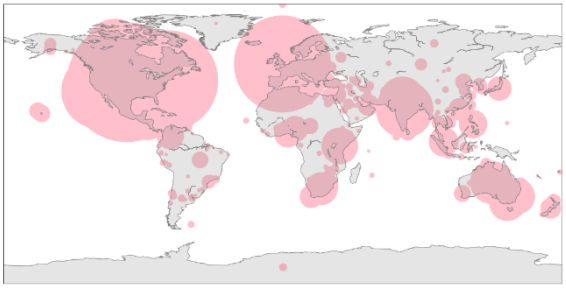


Figure 4: Maps of number of tweets in the world

V. Determination of groups of similar users

The objective of this part is to determine what are the different types of users who are on the platform. In order to do so, I have used a clustering approach.

A. Explanation of the chosen method

In order to determine the groups of similar users, I wanted to focus on three main axes:

- The popularity of the user
- The amount of interaction of the user
- The location of the user

For the popularity, I have selected two variables: the number of followers and the number of friends. I think these are useful features, and good indicators to assess the size of the social network of the user, and consequently describe the impact and the popularity of the user on the platform.

For the matter of the interaction of the user, I chose to consider the time spent on the platform by the user based on the creation date of the account, and the number of favorites count, which is the number of posts liked by the user, because they seem to be good indicators for assessing the time spent on the platform and the interactions of posts of other users.

Finally, in order to consider, the potential cultural and social aspects of the users, I also considered the latitude and longitude of the user in order to give information about its location. I chose not to consider the country, but the latitude and longitude because I do not want to lose the information of proximity among users, that will be lost if replacing the latter by the country, which is a categorical feature.

For all the criteria specified above, I applied a classic dimensionality reduction algorithm (PCA) applied to normalized features, in order to represent each axis of analysis by only one feature. So, I ended up with three axes/features representing respectively: the popularity, the interaction of the user with the platform, and the geographic repartition.

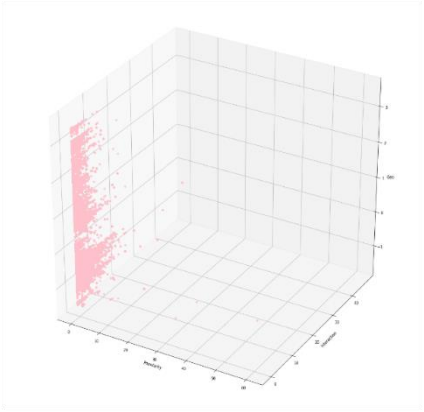
Then I applied a clustering to these three features in order to determine groups of users with similar behaviors.

I also could have let the different features as it was, but I wanted to simplify the study of the different clusters and aggregate the information by axes of analyses. Moreover, my experiments show that reducing these groups of features in one axis each is not a bad representation, because the variance explained by the PCA obtained for each group of variables is significantly high (more than 99% for popularity and interactions groups of features, and approximatively 95% for the geographic features). Another interesting thing when training the algorithm is that using dimension reduction algorithm prior to clustering algorithms such as K-means (the one I chose for this part) speeds up the computation of the algorithm and it is really useful in my case because I have a large dataset.

B. The choice of the clustering algorithm

There exist multiple types of clustering techniques. However, for a problem of clustering applied to determining users with similar behaviors I think choosing a technique that let the clusters determined be convex in space distribution

seem to be more consistent with the definition of users having the same behavior. For this reason, I chose to consider clustering method that guarantees this type of results, such as K-means.
Moreover, to decrease the problem of convergence to the local minima for the Kmeans algorithm, I chose to consider the Kmeans ++ initialization parameter. It helped to avoid running multiple times my algorithm with different starting points because I have a limited computing capacity, a large dataset, and a limited time. So, to keep things simple, I chose to consider the default distance which is the Euclidian distance. Of course, other distances could have been tested.



C. Plots of clustering data before applying algorithms

We can see in this plot that the geographic feature is the one that presents the biggest variation. However, it seems to me that one of the most important features in determining clusters of similar users are popularity and interaction. So, I decided to weight the different axes in function of their importance for determining clusters of users with similar behaviors before applying a clustering algorithm.

D. Application of clustering on normalized dataset – no weights applied

Determination of best k
In order to determine the best number of clusters for the clustering algorithm (k in kMeans). I have used the elbow method by plotting the evolution of the distortions for multiple numbers of clusters: 1 to 14. It appears that a good number of clusters is 4.

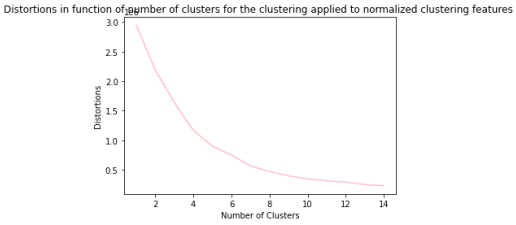


Figure 5 : Evolution of distortions in function of K - no weights applied

Analysis of cluster centers
After determining the cluster of each user, many plots have been performed in order to better understand the characteristics of each cluster.

Characteristics of users from different clusters:

- Cluster 0: users that are the most followed, with an average number of friends, are the less like to put posts as their favorites, most of them are verified, many people like their posts, most ancient users of the platform
- Cluster 1: Low number of friends, a small number, have an important number of followers but a small number in comparison with the fist cluster, but they have a low number of friends, do not interact a lot with the platform, just a few of them have a verified profile, and they do not like other posts a lot. Their profiles are in average the most recent ones.
- Cluster 2: The lowest number of followers, a really low number of friends, they like many other users' posts. Most of them are not verified profiles, and other users do not interact a lot with their posts, and their profile are quite recent.
- Cluster 3: a small number of followers but a high number of friends, they interact a lot with other users' posts and other users also interact with their posts, some of them are verified users, and their profiles are in average older than users from cluster 1 and 2.

NB: this analysis comes from the graph in the appendix.
For next parts, the label of these clusters will be kept in the users' dataset.

E. Application of clustering on normalized dataset – weights applied

Application of K-means clustering
The same procedures of clustering have been applied to clustering data modified with the weighting coefficients.
For the weighted clustering data, we have obtained 3 clusters as the best parameter.
Let's look at the characteristics of each cluster:

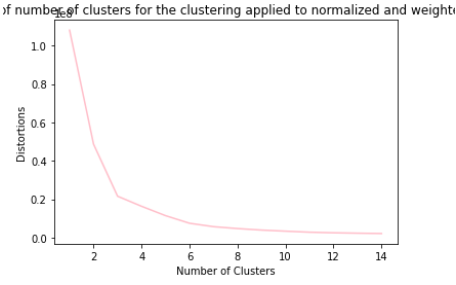


Figure 6: Evolution of distortions in function of K -weights applied

- Cluster 0: A very small number of followers, a small number of friends, interact a lot with other's people content, none of them is verified, a low number of likes from other users, pretty recent account.
- Cluster 1: An average number of followers, an average number of friends, almost all of them are verified, an average number of likes from other users, average age of account
- Cluster 2: The highest number of followers, a really large number of friends, does not interact a lot with other people's content, almost all of them are verified. The highest number of likes from other users, significantly old accounts

Analysis of cluster centers

It seems like the clustering with weights result seem be satisficing since they are understandable and generalizable in terms of explanation in the real word. In fact, from the characteristics of the user, we could match a profile type:

- Cluster 0 -> average user
- Cluster 1 -> influencer
- Cluster 2 -> public figure or media

VI. Analysis of user generated content

The objective of this part is to analyze contents generated by the users on twitter platform.

A. Preprocessing text data

Usually, tweets are not the easiest text to analyze. Indeed, the spelling is not always correct, the text sometimes contains links, emojis, or repetition, that have to be checked and cleaned before analysis.
In order to clean these texts generated by users before analyzing them, some transformations have been performed: correction of spelling, removing links, removing emojis, removing hashtags, putting to lower cases… In order to keep information in the dataset, other features have been added in separate columns (links, hashtags, emojis…). We have also enriched the dataset with other features (e.g. Number of words). For future analyses performed for this part, more adapted text cleaning and structuration have been performed, such as stemming, tokenization, and vectorization (with Bag of Words or TfIFD).

B. What are the most used words?

With the goal of determining the most used words, we have paid attention to removing stop words from the text. In addition to the stop words, words linked to COVID-19 ('corona', coronavirus…) have been also removed from the text in order to have more significant results. In fact, without removing them from the text, the results obtained were not informative enough.
Among the most used words are vaccine, new, positive, mask case, people, cases, and trump. This could be explained by the fact that the most important concern recently is the vaccine and the political regulations given by the US government in the American-speaking world. Given the distribution of occurrence of words in the different tweets, we could also notice that the content is remarkably similar, and that people usually use the same vocabulary, and probably express the same main ideas.

Figure 7: Word Cloud

C. What are the most used combinations of words?

After studying the most common words in the tweets, I was interested in studying the most popular sequence of words when talking about covid. (NB: For this part, the words in tweets have been steemed).
The analysis of bigrams shows that the most frequent sequence of two words that appear together are: 'new' and 'cases. That reflects the major interest in the evolution of the epidemic. Given the most common bigram, we can conclude, that the main concerns are in general related to the evolution of the epidemic, with the most frequent occurrences of: "tested positive", "highest daily", "cases deaths", "new deaths, and others. The other aspects expressed on the platform, seem to be related to the social and economic impact of the pandemic. We can see that with the high occurrence of "massive unemployment peak", "americans going hungry", "unemployment claims", and "growing number Americans going hungry".

D. What is the approximative length of the tweets?

I have chosen to study the complexity of tweets by using the number of words by tweet. We can notice that the tweets related to covid contain approximatively 15 words. However, we can also notice that the distribution of number of words in tweets is squeezed to the right. Given these results we can confirm that the tweets are generally simple, short sentences, and points to only one main idea.

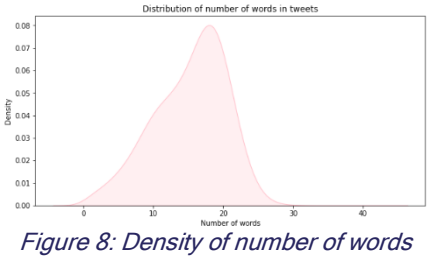


Figure 8: Density of number of words per tweet

E. What are people talking about?

In order to understand more, the content shared on twitter about covid, I chose to perform a topic extraction. For this purpose, I used three different approaches: K-means clustering, Hierarchical clustering, and Latent Dirichlet Allocation (LDA).

In order to do so, I have first defined a corpus using a part of tweets in the general dataframe. (NB: I was limited by the resources available on my computer, and the complexity of the algorithms I wanted to use). Then I applied a vectorization with Tfldf, and K-means clustering method. In order to choose the optimal parameter k, I have used the elbow method. I have determined 4 number of clusters. In order to understand the characteristics of these clusters I have Identified the top words per cluster. We can conclude from the result that the first cluster is manly related to health, and the evolution of the pandemic, the second one is related the vaccine and the pharmaceutical sector, the third one is related to Donald Trump and him being tested positive, and the fourth cluster is mainly related to the reports, and records of the evolution of the epidemic. We could thus respectively label them as: Health, Medicine, Politics, and Reports. I have then used TSNE for visualization of results. I preferred this technique rather than applying PCA, because I have a high-dimensional dataset. The results in figure 9 do not seem to

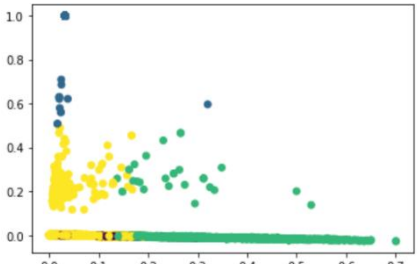


Figure 9: Visualization of Kmeans clustering results with T-SNE

The second approach of clustering was based on hierarchical clustering. Given the highly dimensional dataset that I must deal with, and in order to avoid having a non-uniform hierarchy I have used a cosine similarity, that is particularly suited for text mining applications. In fact, the Euclidian distance tends to group tweets that have the same number of words in common, and I wanted to avoid this case. I have then plotted the dendrogram to determine the optimal number of clusters, which in this case was 2. Similarly, I have plotted the results using TSNE, and I have studied the results of the algorithm, based on the most used words per cluster defined.

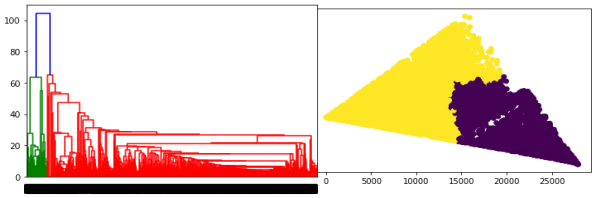


Figure 10: Dendrogram, and visualization of repartition of clusters with TSNE

The last approach was based on Latent Dirichlet Allocation (LDA), a popular algorithm for topic modeling. This model considers the description of each document with a distribution of topic, and each topic is described by a distribution of words. I have applied the algorithm to two different types of vectorizations of tweets: with bag of words, and Tfldf. Also, in order to consider the limited space, I have on my memory and the complexity of the algorithm used, I chose to limit the bag of words to be above a certain number of repetitions rather than reducing again the number of tweets used in the algorithm.
After studying the repartition of these topics in space, it seems that the best results Is obtained using Tfldf vectorization. Moreover, after trying multiple numbers of clusters, I have defined 4 as a substantially good parameter. The results are shown in the plot in the right.

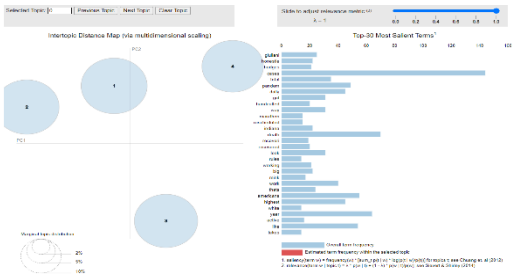


Figure 11: Intertopic distance map and Top 30 most salient terms of LDA applied to TFIDF

When analyzing the characteristics words of the four clusters identified. We can see that the first one is mainly related to tests and people, the second is related to vaccine, celebration and politics. The third clusters are more related to the evolution of the number of deaths and the social life. And the last cluster focus on the reported cases, and the evolution of the disease.

After combining the results of these different analyses, we can conclude that even if the approaches were different, we can still see that the general topics detected are quite similar. Even if the number of clusters found from an algorithm to another is different, we still find a topic related to the evolution of the pandemic, the politics, and health, even if in some results of algorithms, the topics found are combined differently.
In this case, it is hard to identify what is the best metric to evaluate the clustering results, but I would say the most coherent I obtained when analyzing the clusters determined were the results of the LDA method.

F. What is the general feeling expressed on the platform?

After cleaning and structuring text data, I was interested in studying the sentiment of the content of the text shared by the users. For that I have used VADER, which is a pre-trained model available on NLTK package that is useful because it could be directly applied to unlabeled text data. In fact, in my case, I do not have labels that could help me label a tweet as positive or negative? Moreover, VADER is also sensitive to the intensity of an emotion.

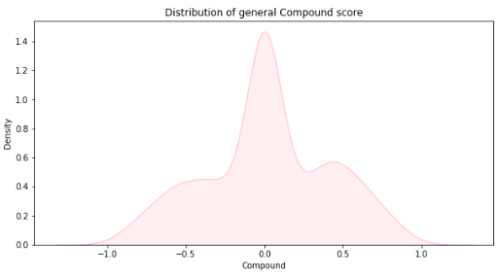
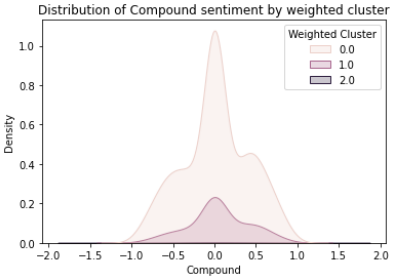


Figure 12: Distribution of general compound score

After analyzing the repartition of positive and negative comments it seems like more than approximatively 70% of the content shared on this platform related to covid seems to be positive. However, we should deepen the analysis, and see whether some combinations of words are interpreted correctly (e.g., tested positive). Otherwise, we can see that the distribution of the compound score given by the model shows that the comments which are negative, but most positive comments have an average positive score. That means the negative comments are usually more intense than the positive comments. But it seems most comments approximatively have a neutral comment. That could be explained by the fact that most of the comments were related about facts. Indeed, we have shown before that most words are linked to daily reports, politics, health, vaccines…



G. Is there a relation between the type of user and the sentiment expressed in a tweet?

When plotting the distribution of sentiments of users, we can notice that for different groups of users, the shape of the curves is approximatively the same. That means the group of users does not seem to impact the feeling expressed on the platform.

VII. Conclusions

To conclude, this study was the opportunity to explore the last trends in a worldwide used social media applied to a recent and impactful subject. The data collection part was time consuming, but luckily the data collected has a lot to offer and provides numerous axes of analyses. Furthermore, we have shown with the data collected that the utilization of the platform has significantly increased since the beginning of the covid, and this confirms that people are more available and interested in interacting on social media than they were before. Moreover, we could notice that the topic of the epidemy interests a significant number of users on the platform. We have also determined that we could identify three main different types of users with a clustering approach, that represent the reality classes because they are easily identifiable as logical types of users (public figure, influencer, average). During the analyses of tweets content we have also identified that the main topics discussed are linked to the evolution of the epidemic, health, vaccine, politics, and the economic impacts on society. After studying the sentiment of each tweet generated by the user, we can affirm that most of the content has a neutral tone, however there are approximatively 70% of tweets that are perceived as positive. So, this could show that users are attracted by the platform in order to share positive comments and to encourage each other in this unprecedented time. When we combine these axes of analyses at the end, we can see that there are no big gaps between the sentiments expressed and the type of words used by the user given the group of users that they belong to. That means the type of user does not necessarily determine the tone, the general perceived sentiment, or the shared content on the platform. Finally, further axes could be analyzed, such as the impact of other variables on the positivity of the comments, or an in-depth analysis of the themes addressed. This will hopefully be the subject of a future work.

Appendix

RESOURCES

Data collection

- https://developer.twitter.com/en/docs/labs/covid19-stream/quick-start
- https://github.com/gabrielpreda/covid-19-tweets/blob/master/covid-19-tweets.ipynb

Clustering of users

- TP2 MDI343

Analysis of user generated content

- Courser NLP Specialization
- https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24#:~:text=Topic%20modeling%20is%20a%20type,document%20to%20a%20particular%20topic.
- https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0
- http://ericmittelhammer.com/clustering-with-tf-idf.html
- Sklearn library

STRUCTURE OF THE FOLDER

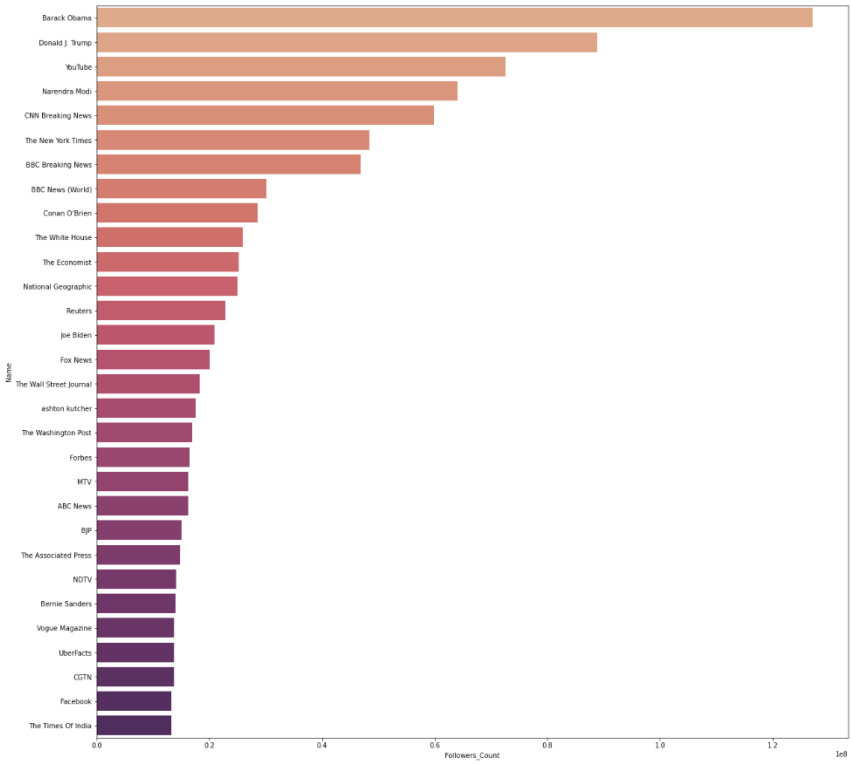
Report: report of the project
Notebooks: directory that contains all notebooks and codes developed for the project
Data: contains all data collected or generated in csv format

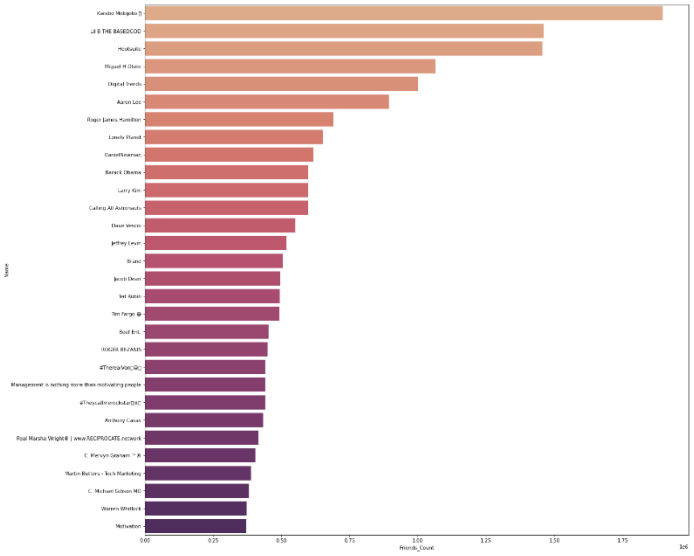
DATA DICTIONNARY

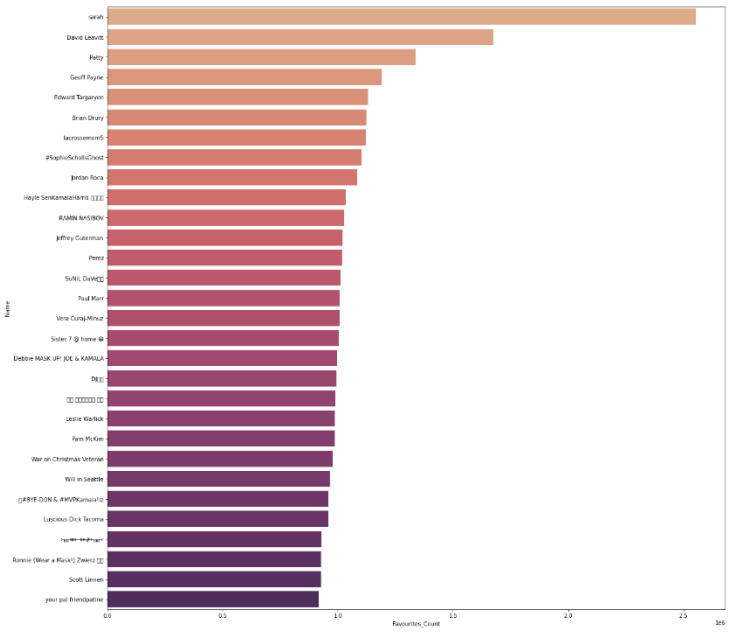
Attribute	Type	Description
id	Int64	The integer representation of the unique identifier for this User. This number is greater than 53 bits and some programming languages may have difficulty/silent defects in interpreting it. Using a signed 64 bit integer for storing this identifier is safe. Use id_str to fetch the identifier to be safe. See Twitter IDs for more information. Example: "id": 6253282
name	String	The name of the user, as they've defined it. Not necessarily a person's name. Typically capped at 50 characters, but subject to change. Example: "name": "Twitter API"
screen_name	String	The screen name, handle, or alias that this user identifies themselves with. screen_names are unique but subject to change. Use id_str as a user identifier whenever possible. Typically a maximum of 15 characters long, but some historical accounts may exist with longer names. Example: "screen_name": "twitterapi"
location	String	Nullable . The user-defined location for this account's profile. Not necessarily a location, nor machine-parseable. This field will occasionally be fuzzily interpreted by the Search service. Example: "location": "San Francisco, CA"

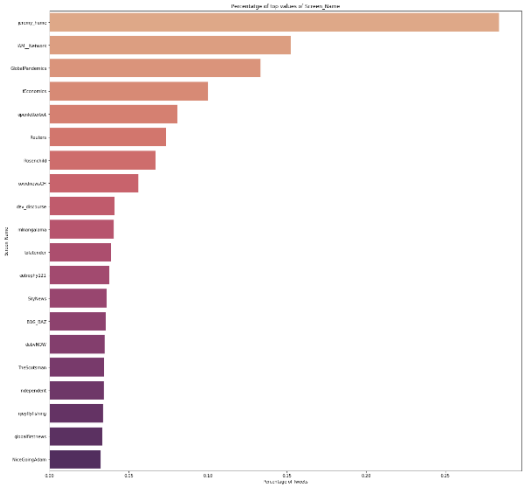
description	String	Nullable . The user-defined UTF-8 string describing their account. Example: "description": "The Real Twitter API."
verified	Boolean	When true, indicates that the user has a verified account. See Verified Accounts . Example: "verified": false
followers_count	Int	The number of followers this account currently has. Under certain conditions of duress, this field will temporarily indicate “0”. Example: "followers_count": 21
friends_count	Int	The number of users this account is following (AKA their “followings”). Under certain conditions of duress, this field will temporarily indicate “0”. Example: "friends_count": 32
favourites_count	Int	The number of Tweets this user has liked in the account’s lifetime. British spelling used in the field name for historical reasons. Example: "favourites_count": 13
created_at	String	The UTC datetime that the user account was created on Twitter. Example: "created_at": "Mon Nov 29 21:18:15 +0000 2010"

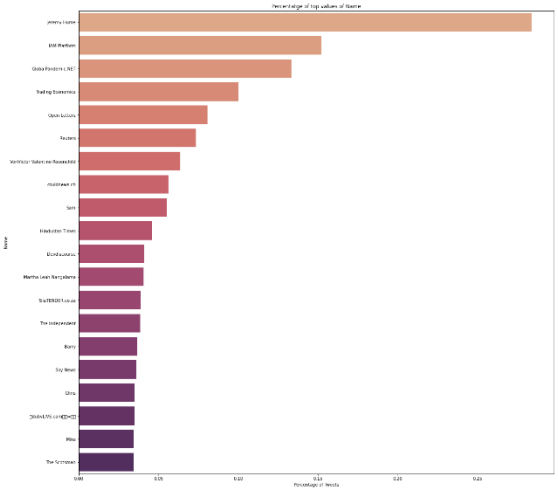
Occurrence of categorical features

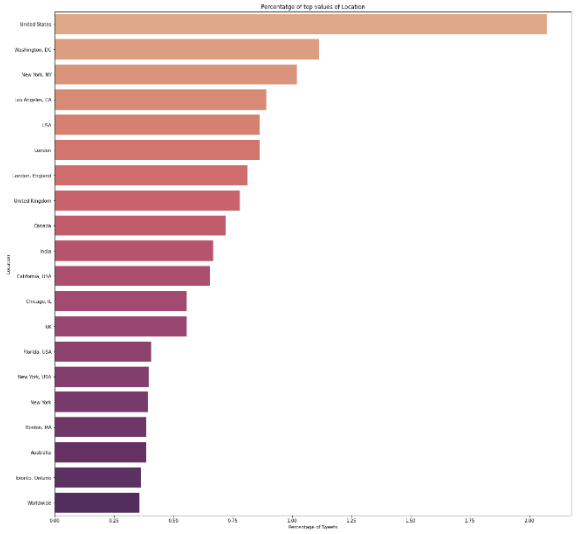


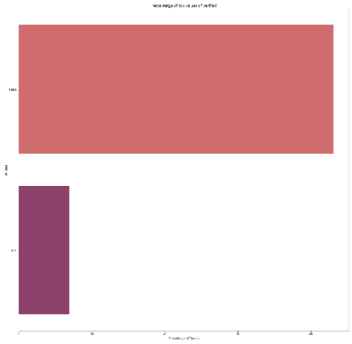


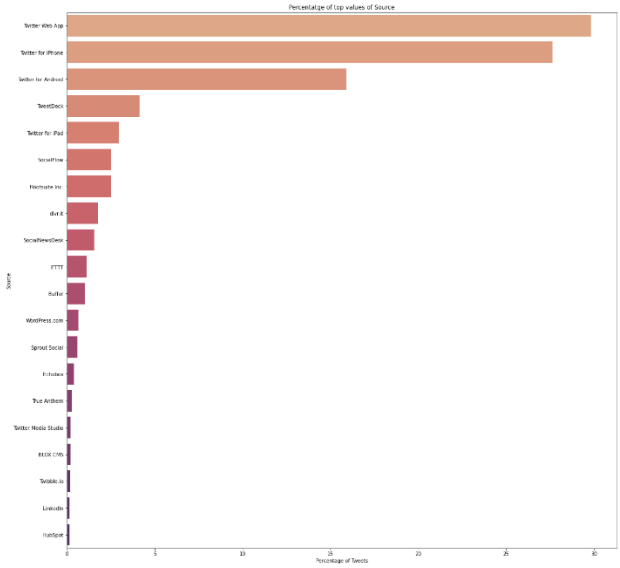

Name
Favourites_Count
0.0
0.5
1.0
1.5
2.0
2.5
1e6
Sarah
David Leavitt
Patty
Geoff Payne
Edward Targaryen
Brian Drury
lacrossemom5
#SophieSchollsGhost
Jordan Roca
Hayle SenKamalaHarris
RAMIN NASIBOV
Jeffrey Cuterman
Perez
SuNiL DaVe
Paul Marr
Vera Cunaj-Mrkaj
Sister 7 @ home
Debbie MASK UP! JOE & KAMALA
DJ
Leslie Warlick
Pam McKim
War on Christmas Veteran
Will in Seattle
#BYE-DON & #MVPKamala
Luscious Dick Tacoma
Ronnie (Wear a Mask!) Zwierz
Scott Linnes
your gal friendpaltine

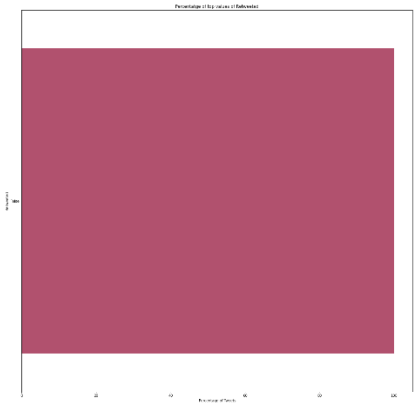

Percentage of top values of Screen_Name
Screen_Name
Percentage of Tweets

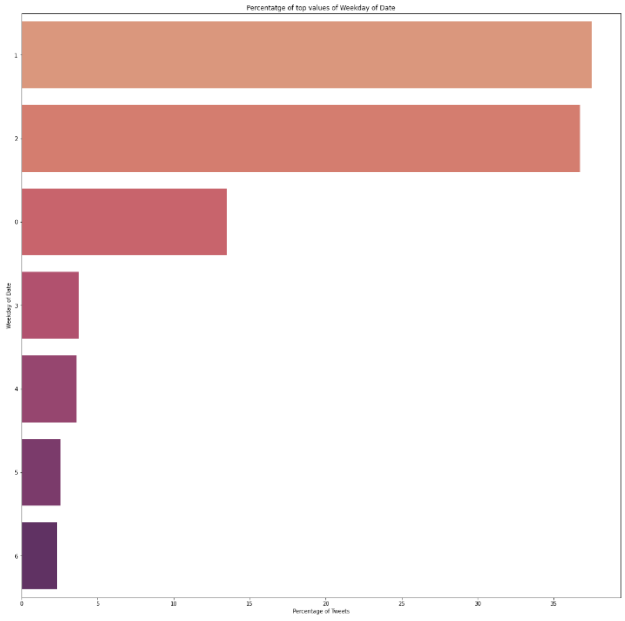

Percentage of top values of Name
Name
Percentage of Tweets

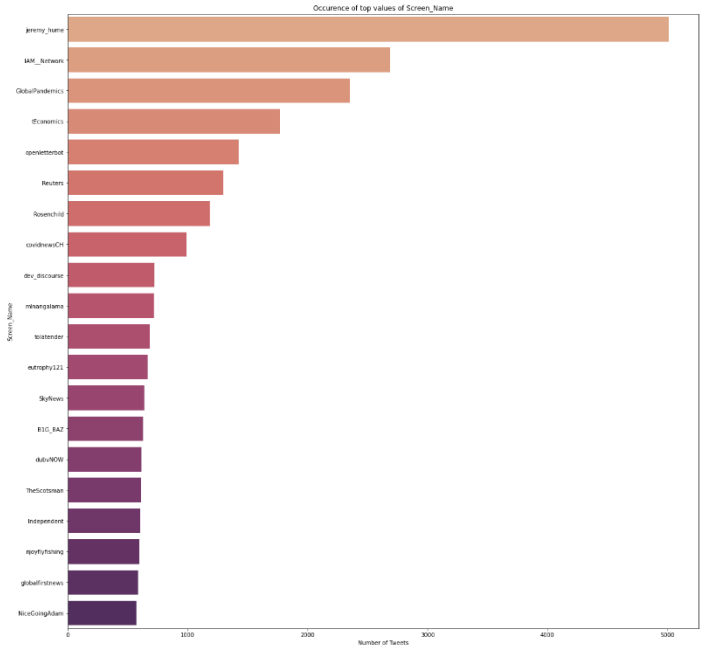

Percentatge of top values of Location
United States
Washington, DC
New York, NY
Los Angeles, CA
USA
London
London, England
United Kingdom
Canada
India
California, USA
Chicago, IL
UK
Florida, USA
New York, USA
New York
Boston, MA
Australia
Toronto, Ontario
Worldwide
Location
Percentage of Tweets

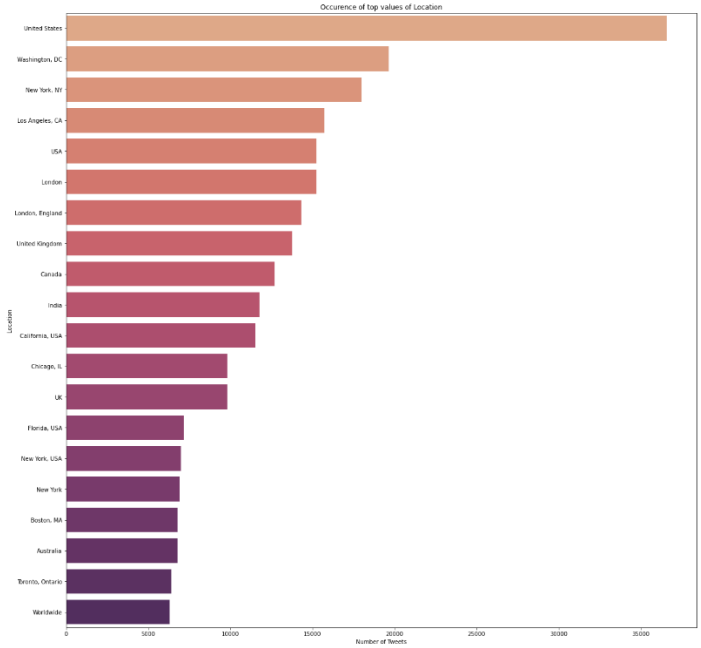


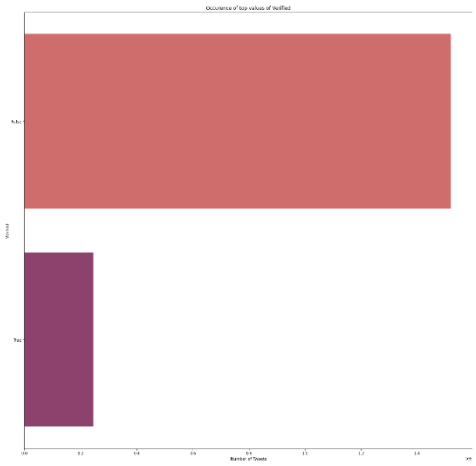

Percentatge of top values of Source
Twitter Web App
Twitter for iPhone
Twitter for Android
TweetDeck
Twitter for iPad
TweetFlow
Hootsuite Inc.
dlvr.it
SocialNewsDesk
IFTTT
Buffer
WordPress.com
Sprout Social
Echobox
True Anthem
Twitter Media Studio
IBLOX CMS
Twibble.io
LinkedIn
HubSpot
Source
Percentage of Tweets

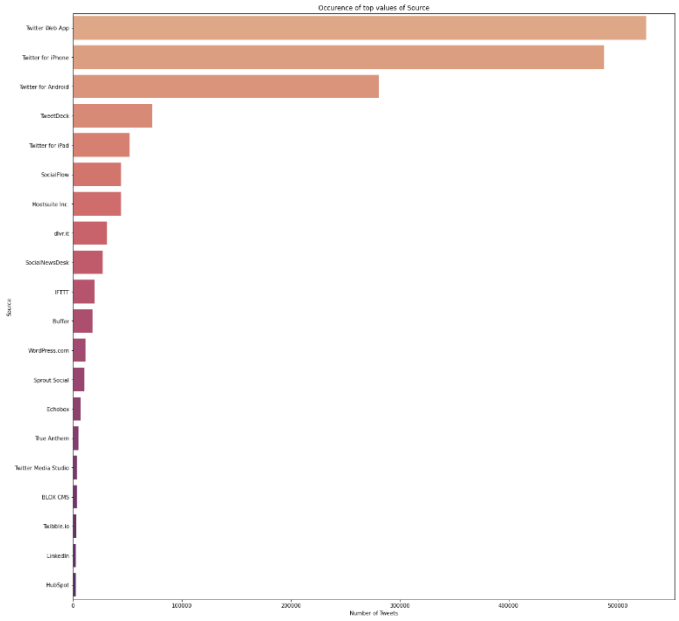



Percentatge of top values of Weekday of Date
Weekday of Date
1
2
0
3
4
5
6
0
5
10
15
20
25
30
35
Percentage of Tweets

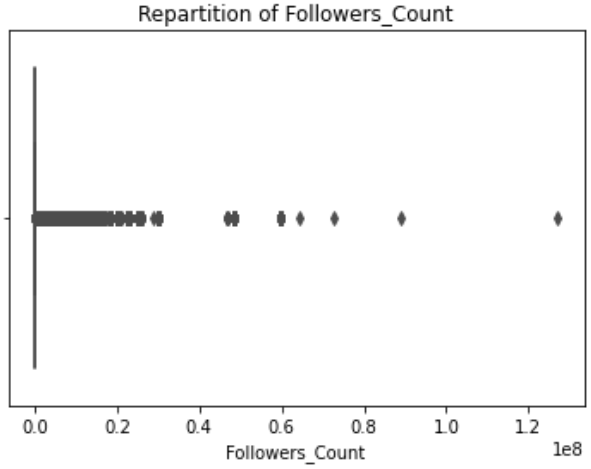

Occurence of top values of Screen_Name
Screen_Name
jeremy_hunt
IAM__Network
GlobalPandemics
tfEconomics
openletterbot
Reuters
Rosenchild
covidnewsCH
dev_discourse
mlnangalama
tolatender
eutrophy121
SkyNews
BIG_BAZ
dubvNOW
TheScotsman
Independent
ejoyfiyfishing
globalfirstnews
NiceGoingAdam
0
1000
2000
3000
4000
5000
Number of Tweets


Occurence of top values of Location
Location
United States
Washington, DC
New York, NY
Los Angeles, CA
USA
London
London, England
United Kingdom
Canada
India
California, USA
Chicago, IL
UK
Florida, USA
New York, USA
New York
Boston, MA
Australia
Toronto, Ontario
Worldwide
0
5000
10000
15000
20000
25000
30000
35000
Number of Tweets



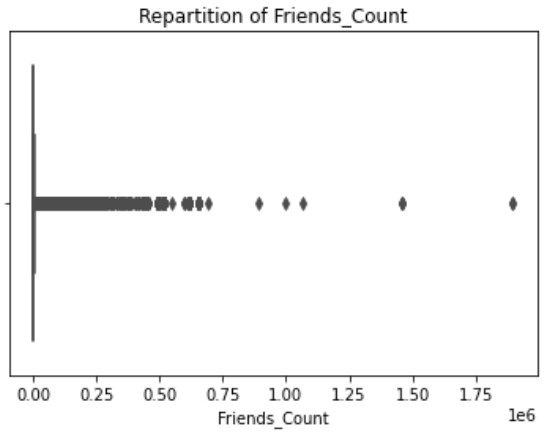


Boxplots of repartition of numeric features

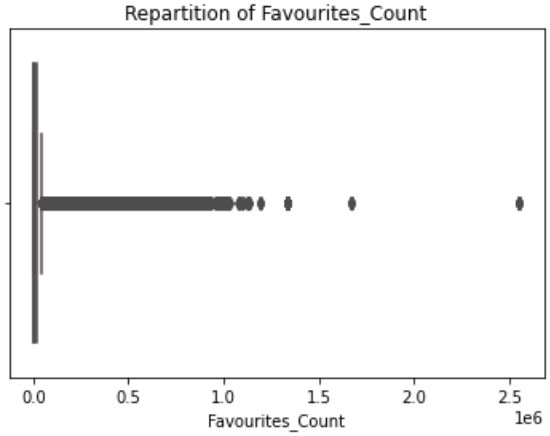



```
The detailed description is the following
count    1.764917e+06
mean     1.027216e+05
std      1.153894e+06
min      0.000000e+00
25%      2.220000e+02
50%      1.050000e+03
75%      5.343000e+03
max      1.270600e+08
```

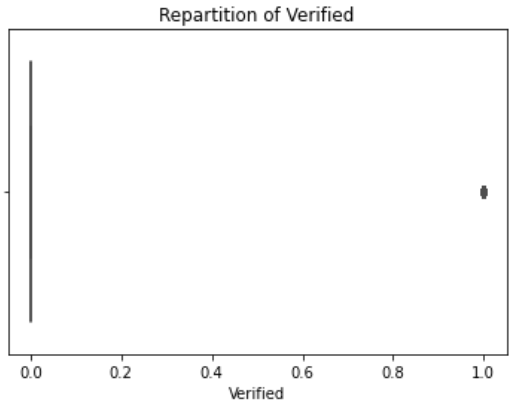
```
Name: Followers_Count, dtype: float64
```



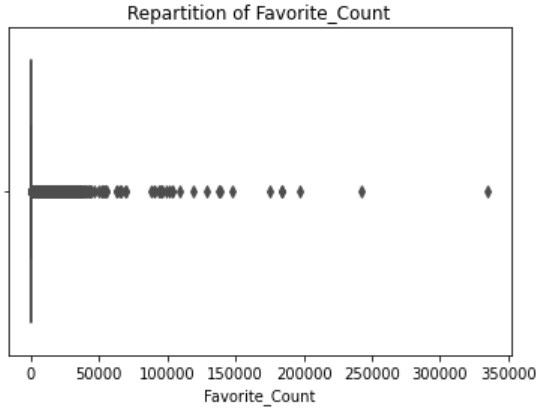

```
The detailed description is the following
count    1.764917e+06
mean     2.537532e+03
std      1.284156e+04
min      0.000000e+00
25%      2.530000e+02
50%      7.060000e+02
75%      1.908000e+03
max      1.896632e+06
Name: Friends_Count, dtype: float64
```



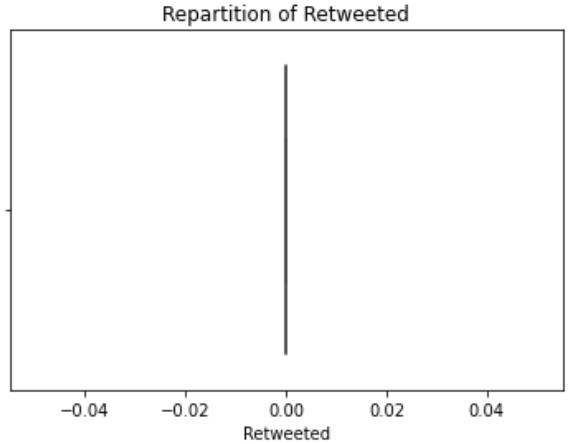

```
The detailed description is the following
count    1.764917e+06
mean     2.137906e+04
std      5.427805e+04
min      0.000000e+00
25%      6.160000e+02
50%      4.005000e+03
75%      1.807600e+04
max      2.554744e+06
Name: Favourites_Count, dtype: float64
```




```
The detailed description is the following
count     1764917
unique          2
top         False
freq      1520227
Name: Verified, dtype: object
```

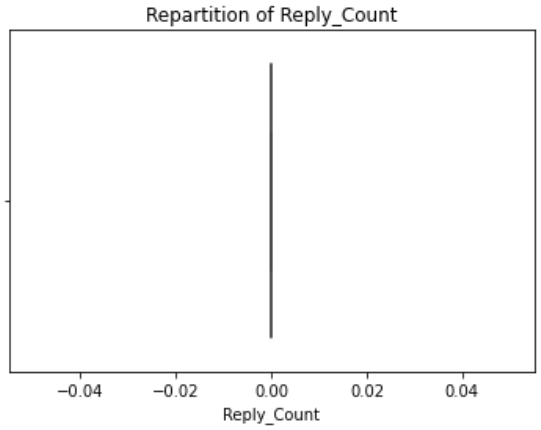



```
The detailed description is the following
count    1.764917e+06
mean     1.354005e+01
std      6.275416e+02
min      0.000000e+00
25%      0.000000e+00
50%      0.000000e+00
75%      2.000000e+00
max      3.350690e+05
Name: Favorite_Count, dtype: float64
```

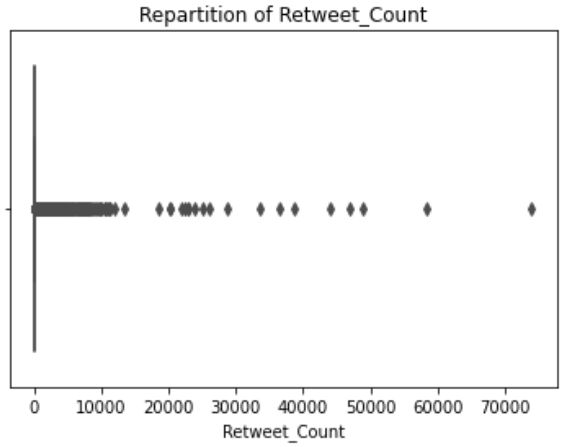



```
The detailed description is the following
count     1764917
unique          1
top         False
```

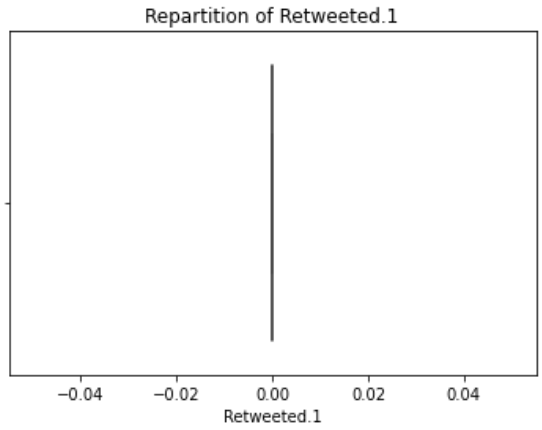
```
freq      1764917
Name: Retweeted, dtype: object
```




```
The detailed description is the following
count    1764917.0
mean           0.0
std            0.0
min            0.0
25%            0.0
50%            0.0
75%            0.0
max            0.0
Name: Reply_Count, dtype: float64
```

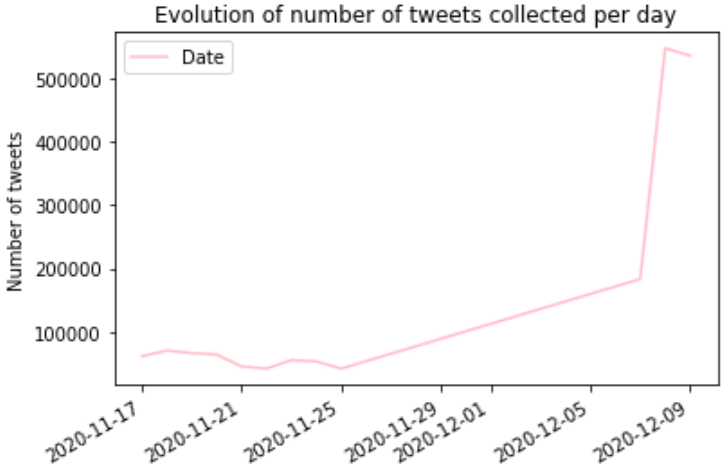



```
The detailed description is the following
count    1.764917e+06
mean     2.793678e+00
std      1.329558e+02
min      0.000000e+00
25%      0.000000e+00
50%      0.000000e+00
75%      0.000000e+00
max      7.399300e+04
Name: Retweet_Count, dtype: float64
```

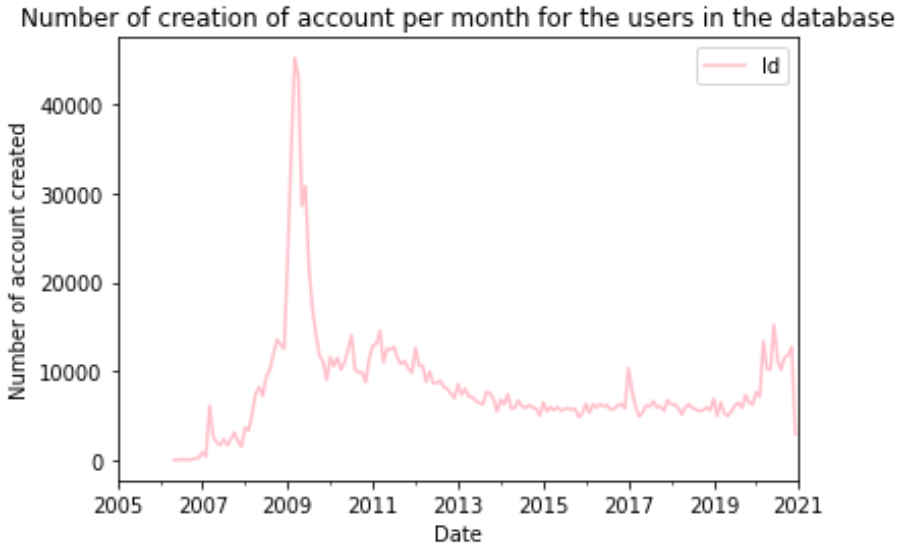


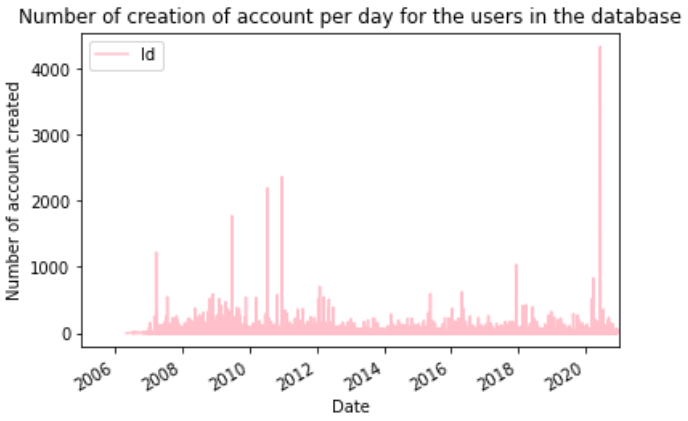

```
The detailed description is the following
count      1764917
unique           1
top          False


freq       1764917
Name: Retweeted.1, dtype: object
```

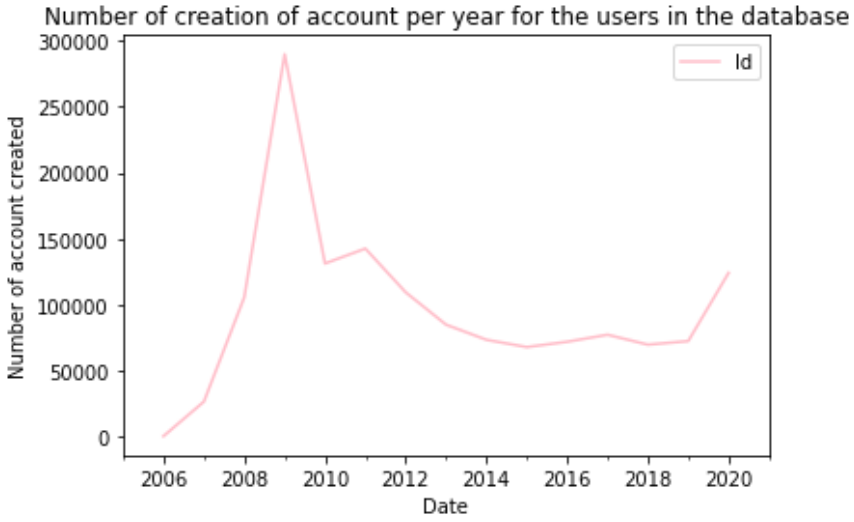
Number of tweets collected per day



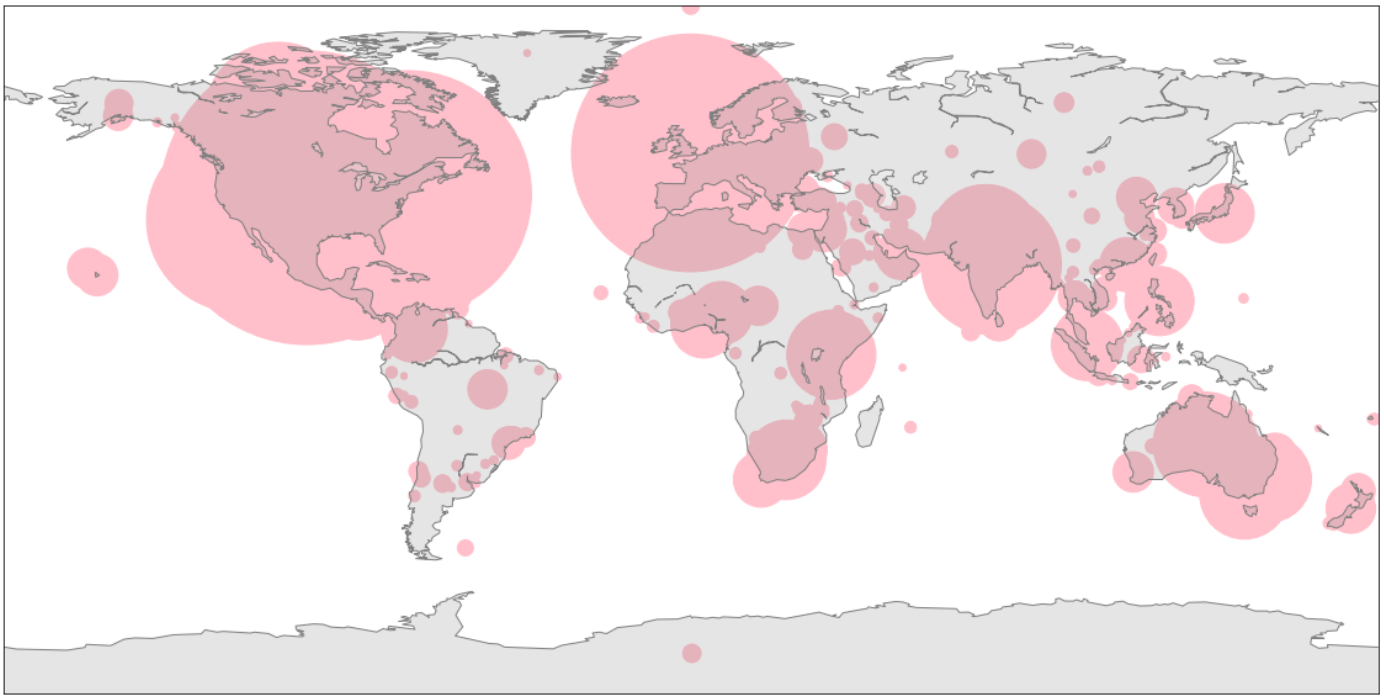
Number of creation of accounts



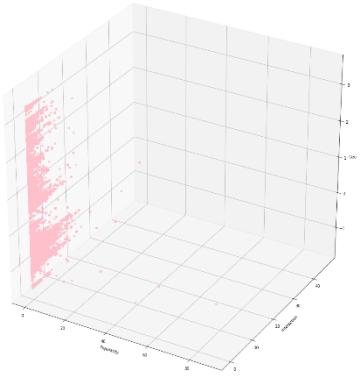




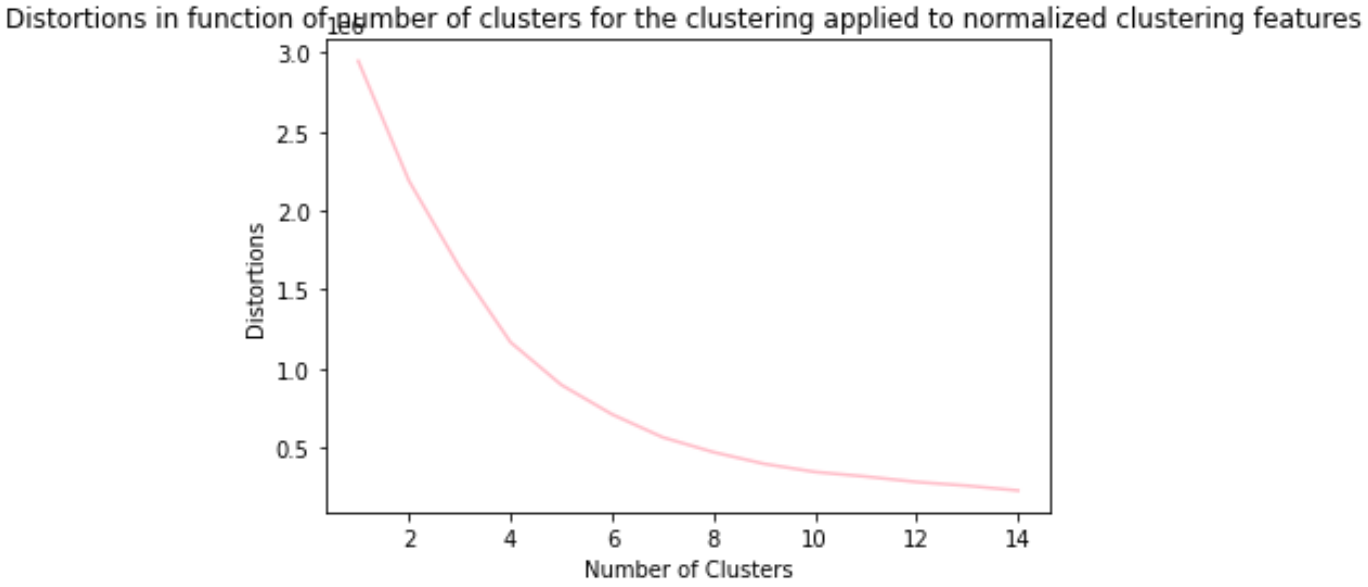
Maps: Number of Covid Tweets in the world



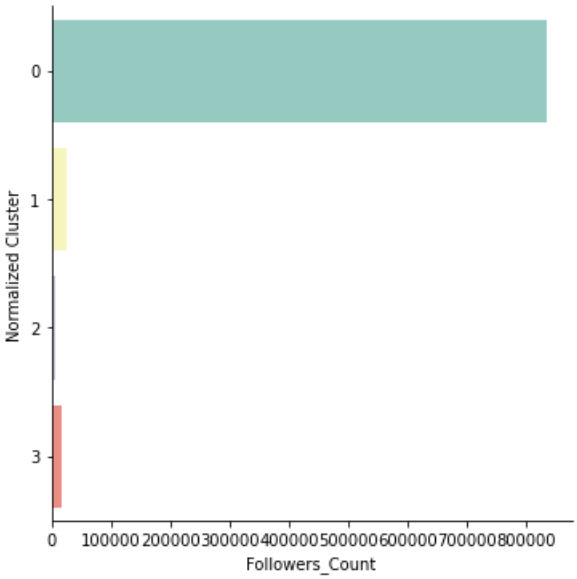
Plots of repartition in 3D Space of clustering users data

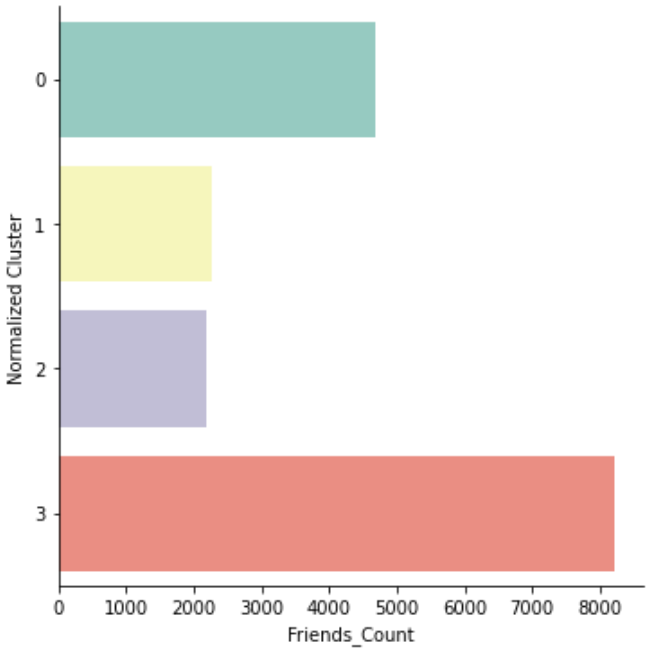


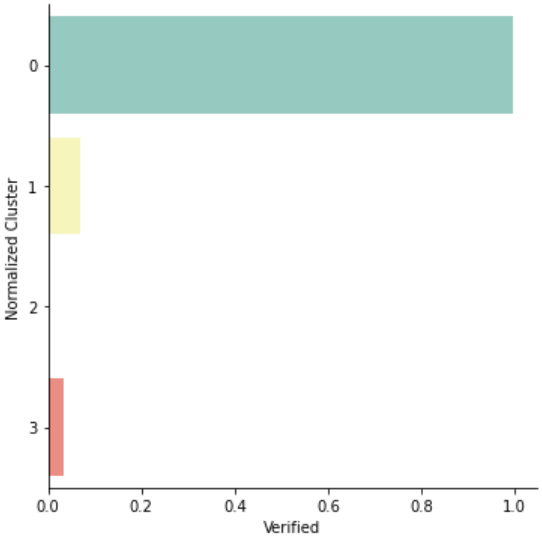
Elbow method clustering of users

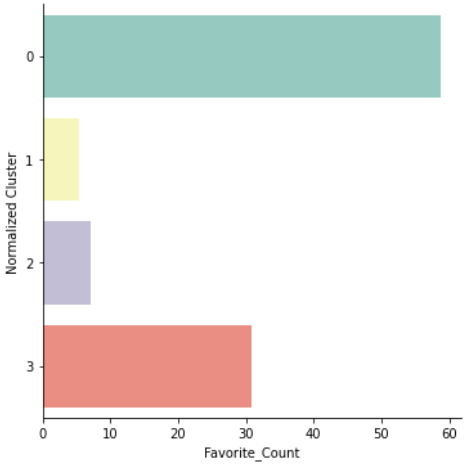


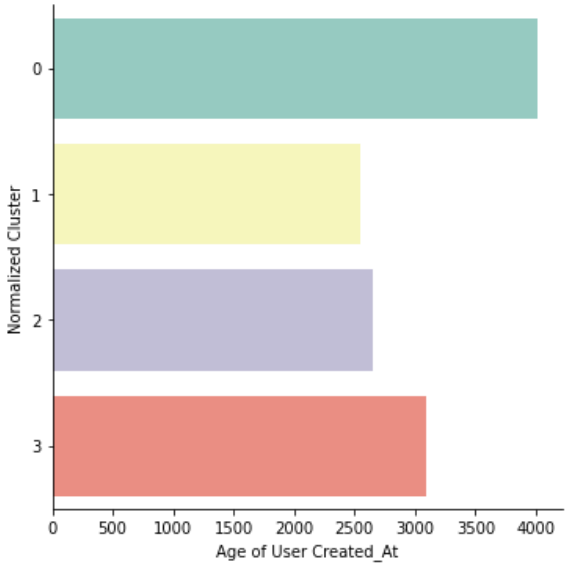
ANALYSIS OF CLUSTERS FOR CLUSTERING APPLIED TO NORMALIZED FEATURES


Normalized Cluster
0
1
2
3
0
100000
200000
300000
400000
500000
600000
700000
800000
Followers_Count

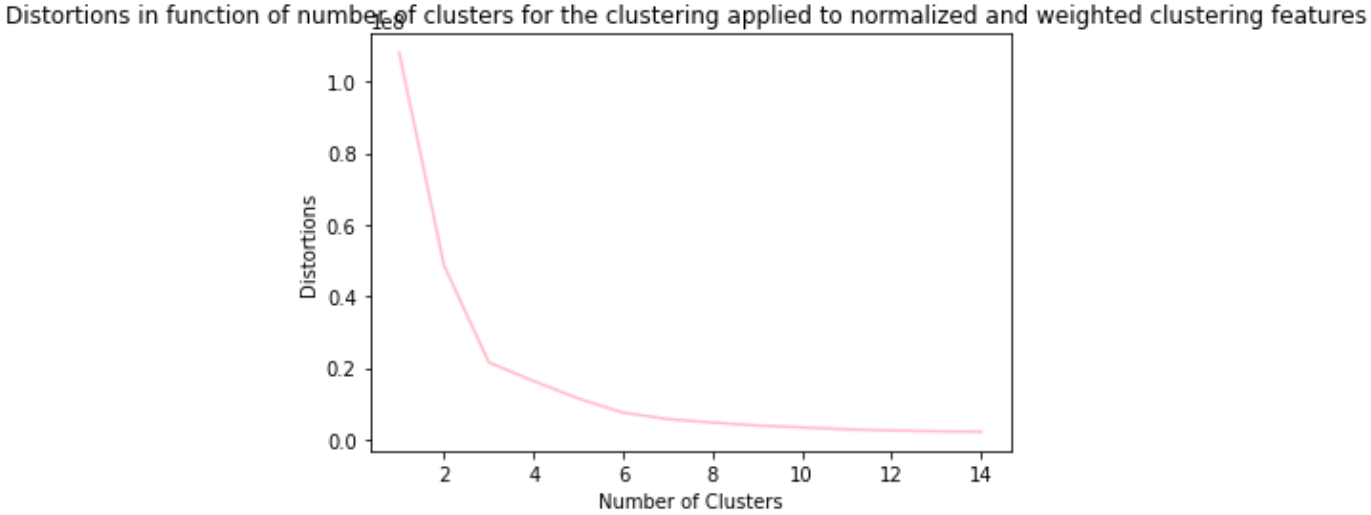

Normalized Cluster
0
1
2
3
0
1000
2000
3000
4000
5000
6000
7000
8000
Friends_Count


Normalized Cluster
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
Verified

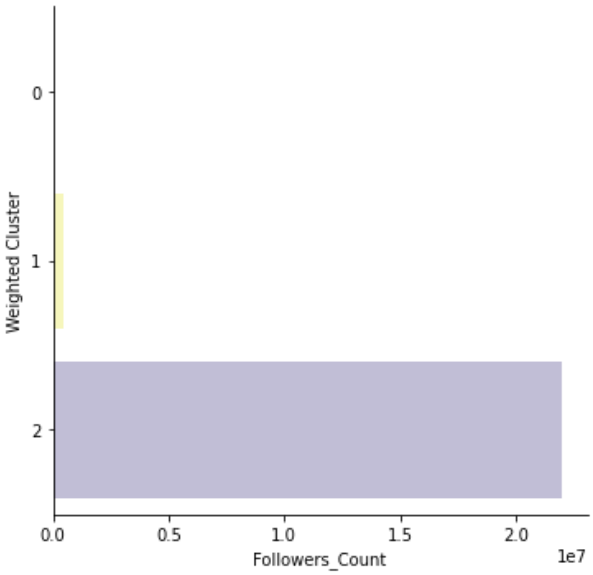


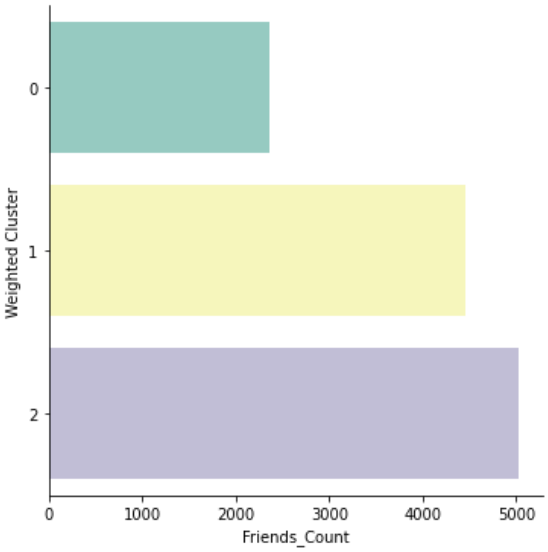


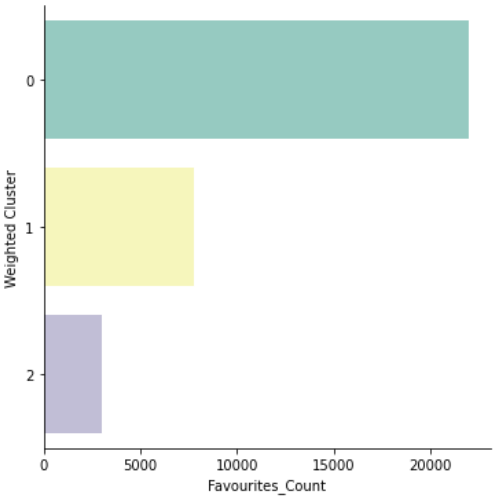
Elbow method clustering of users with weights appplied

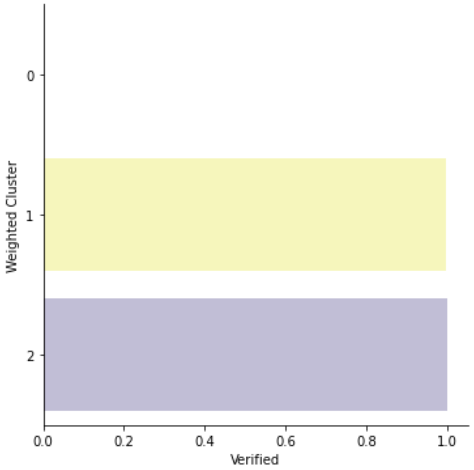


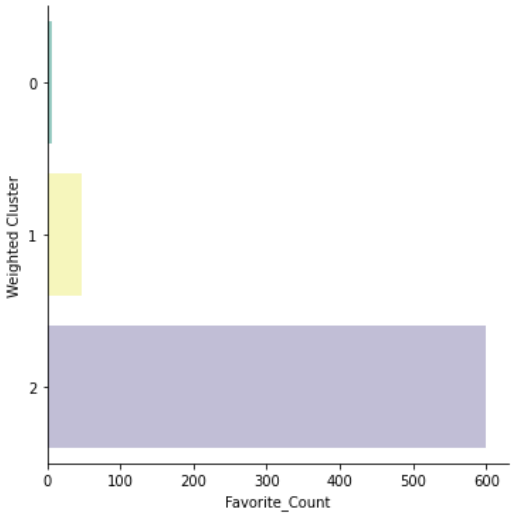
ANALYSIS OF CLUSTERS FOR CLUSTERING APPLIED TO NORMALIZED AND WEIGHTED FEATURES

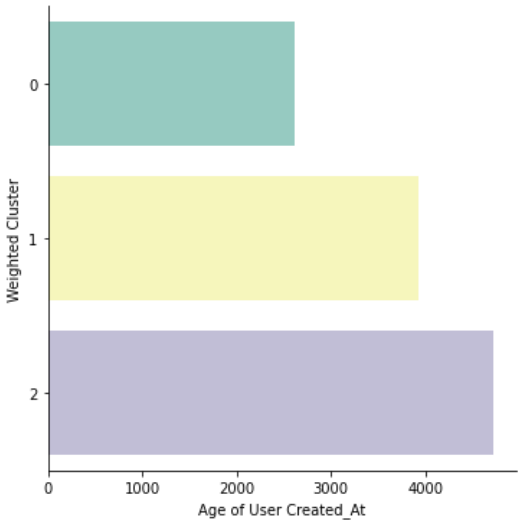

Weighted Cluster
0
1
2
0.0
0.5
1.0
1.5
2.0
Followers_Count
1e7


Weighted Cluster
0
1
2
0
1000
2000
3000
4000
5000
Friends_Count

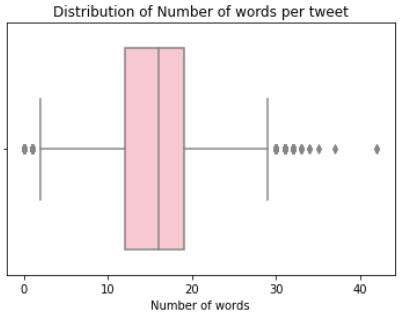

Weighted Cluster
0
1
2
0
5000
10000
15000
20000
Favourites_Count


Weighted Cluster
0
1
2
0.0
0.2
0.4
0.6
0.8
1.0
Verified

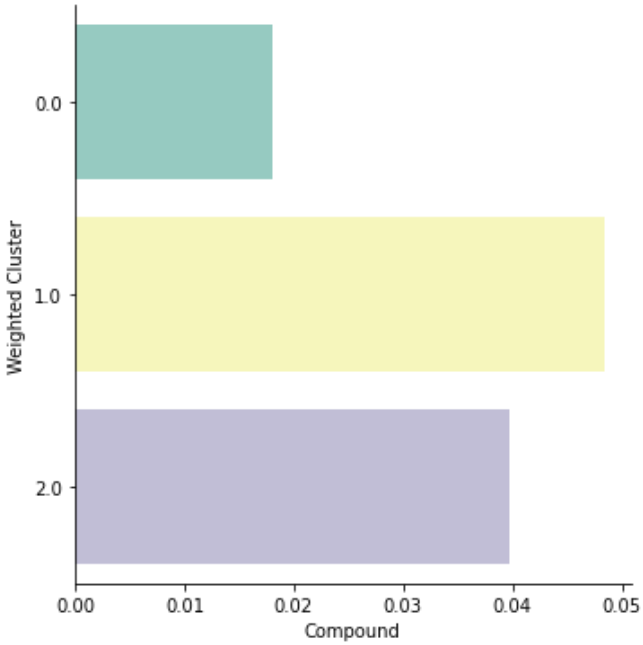

Weighted Cluster
0
1
2
0
100
200
300
400
500
600
Favorite_Count

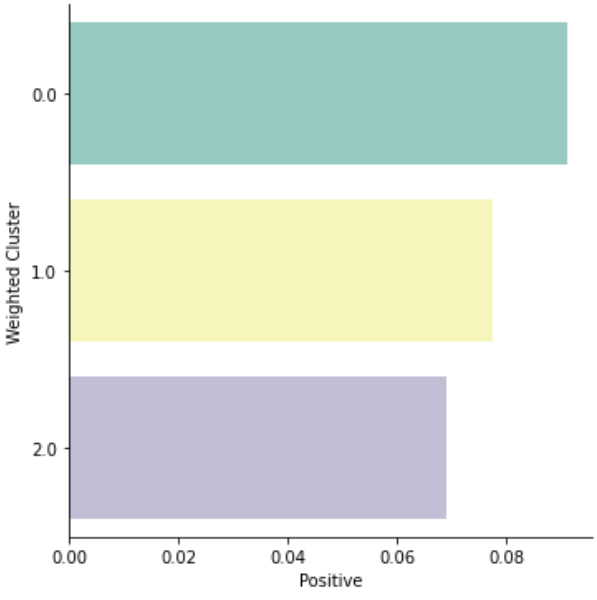

Weighted Cluster
0
1
2
0
1000
2000
3000
4000
Age of User Created_At

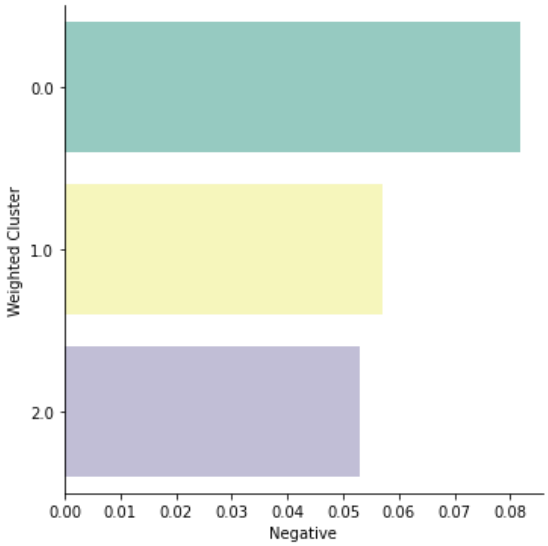
DISTRIBUTION OF NUMBER OF WORDS BY TWEET

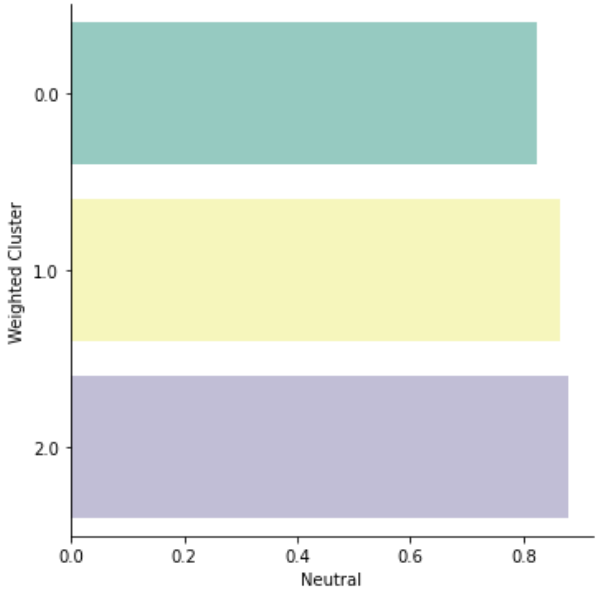


AVERAGE VALUES OF NUMERIC FEATURES BY WEIGHTED CLUSTER



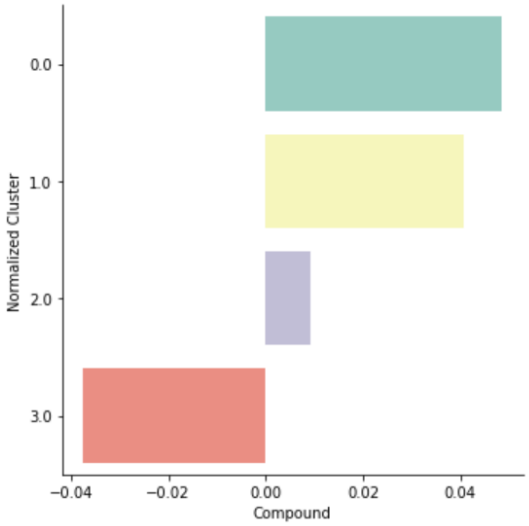


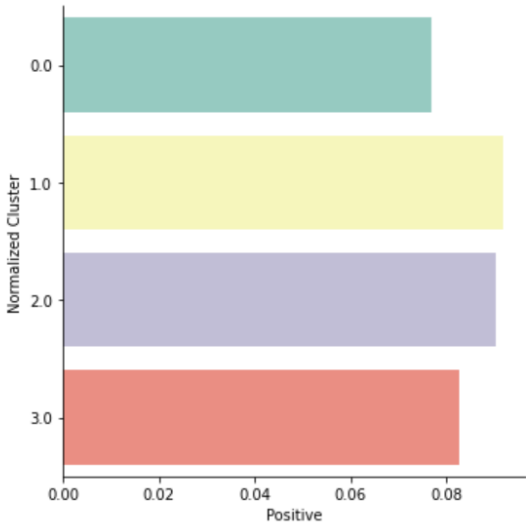


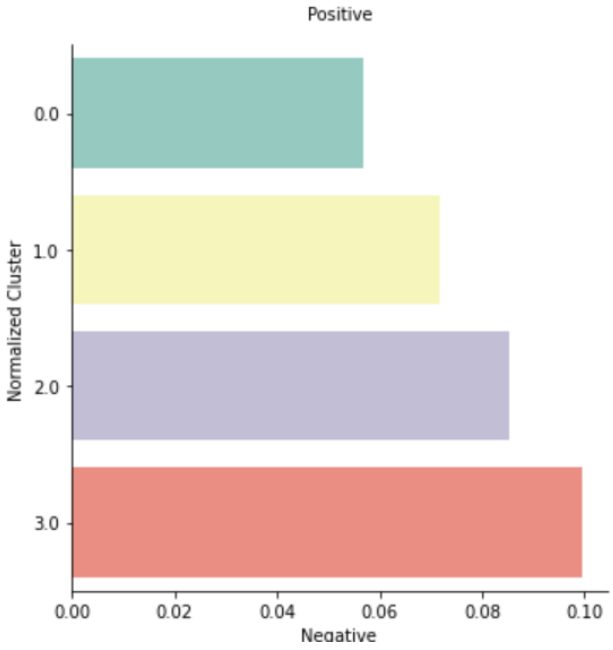


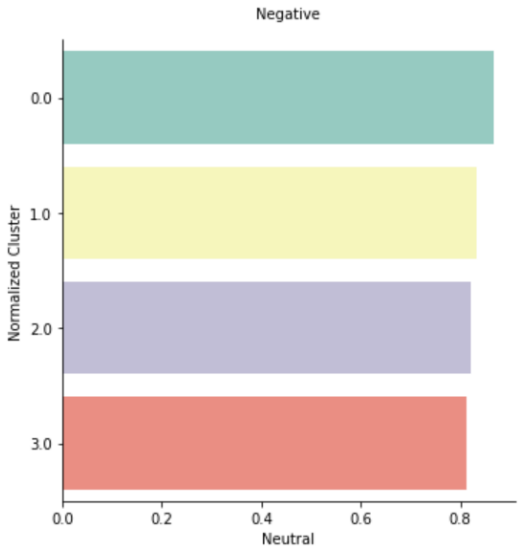
	Weighted Cluster	Compound	Positive	Negative	Neutral
0	0.0	0.018027	0.090916	0.081657	0.823563
1	1.0	0.048418	0.077547	0.056945	0.865296
2	2.0	0.039676	0.068982	0.053080	0.877941

AVERAGE VALUES BY NORMALIZED CLUSTER



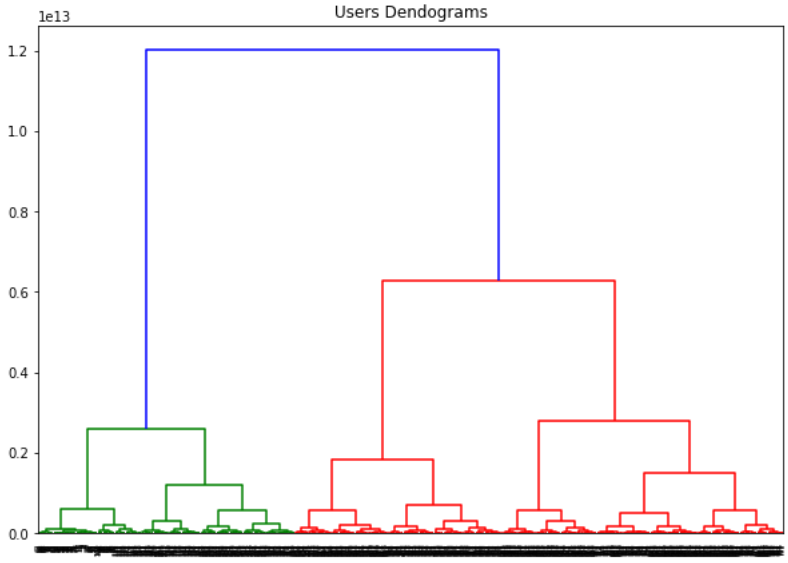




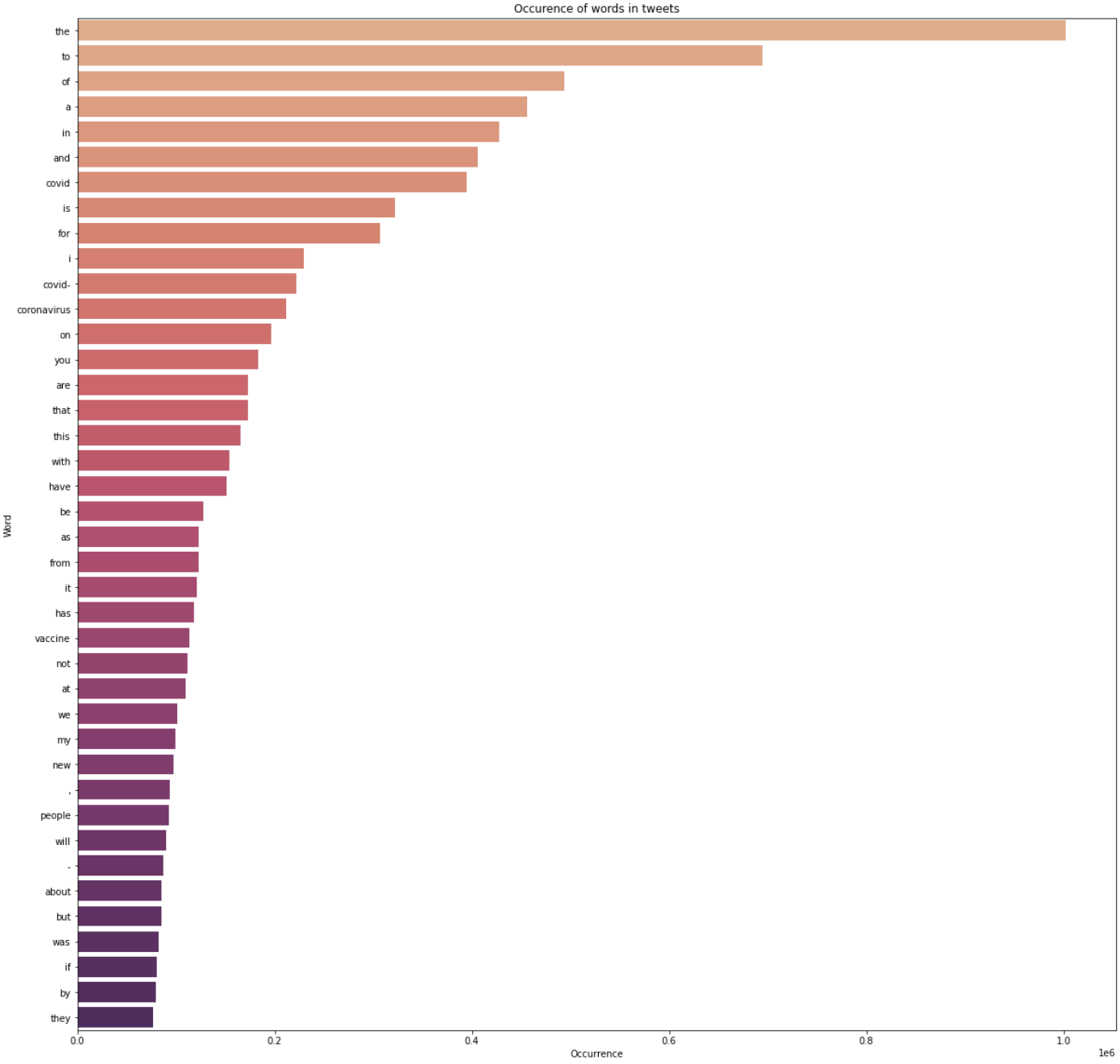


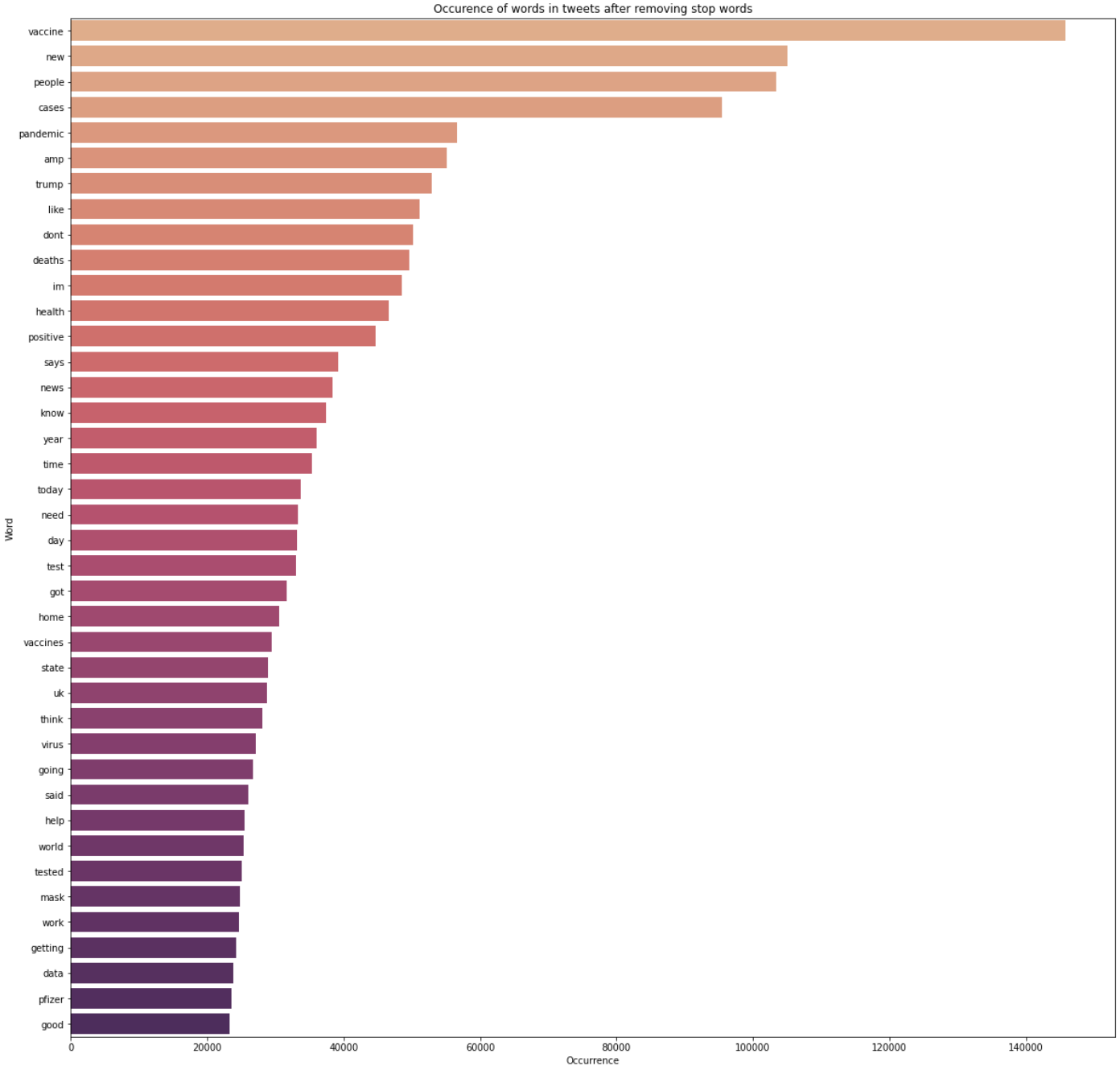
	Normalized Cluster	Compound	Positive	Negative	Neutral	Weighted Cluster
0	0.0	0.048546	0.077059	0.056748	0.866005	1.017697
1	1.0	0.040622	0.091785	0.071861	0.831362	0.068282
2	2.0	0.009273	0.090376	0.085281	0.821320	0.000126
3	3.0	-0.037426	0.082809	0.099367	0.811686	0.034713

DENDOGRAM OF USERS CLUSTERING



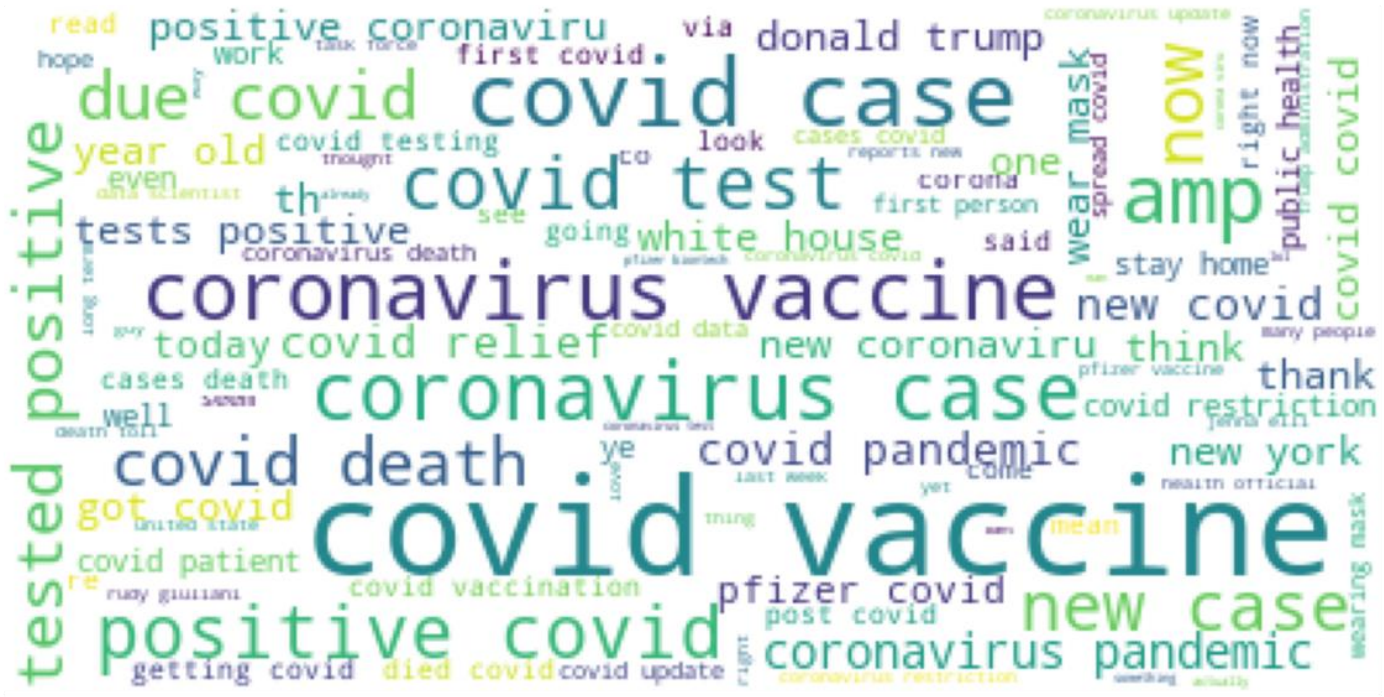
MOST COMMON WORDS


Occurence of words in tweets
the
to
of
a
in
and
covid
is
for
i
covid-
coronavirus
on
you
are
that
this
with
have
be
as
from
it
has
vaccine
not
at
we
my
new
,
people
will
-
about
but
was
if
by
they
Word
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Occurrence

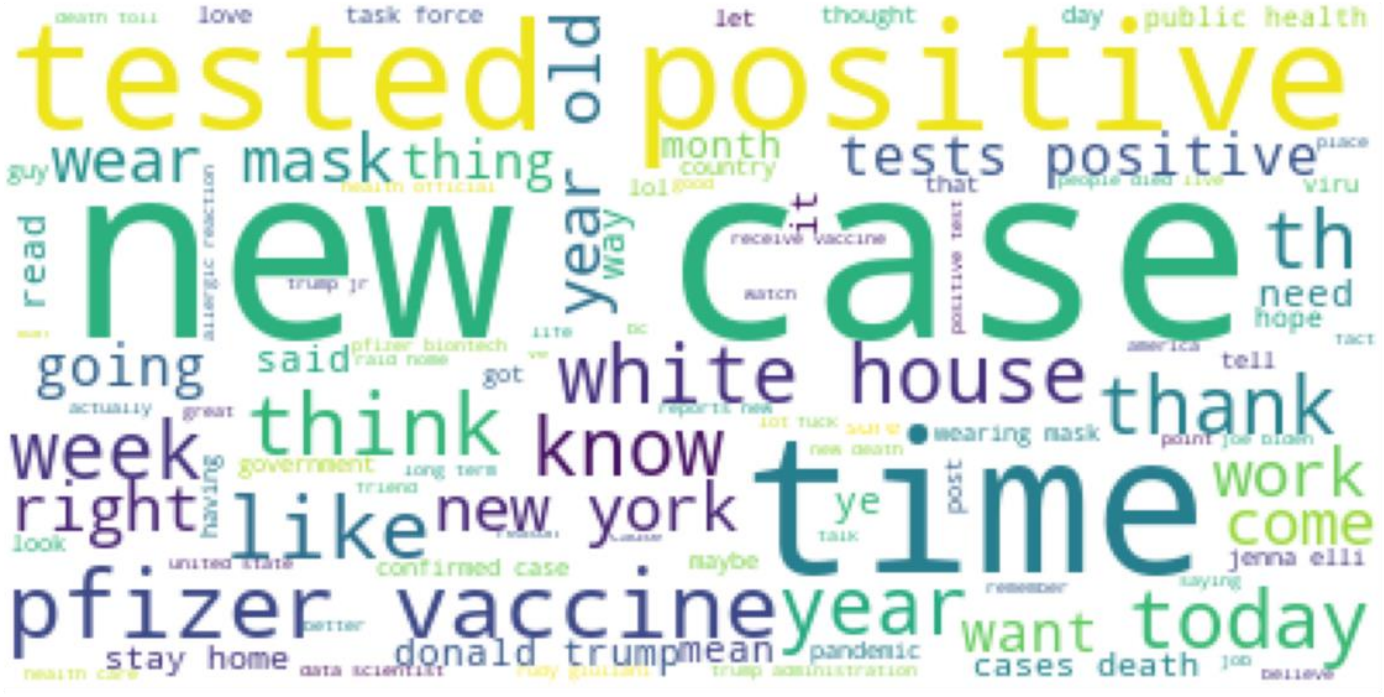

Occurence of words in tweets after removing stop words
vaccine
new
people
cases
pandemic
amp
trump
like
dont
deaths
im
health
positive
says
news
know
year
time
today
need
day
test
got
home
vaccines
state
uk
think
virus
going
said
help
world
tested
mask
work
getting
data
pfizer
good
Word
0
20000
40000
60000
80000
100000
120000
140000
Occurrence

Word Clouds

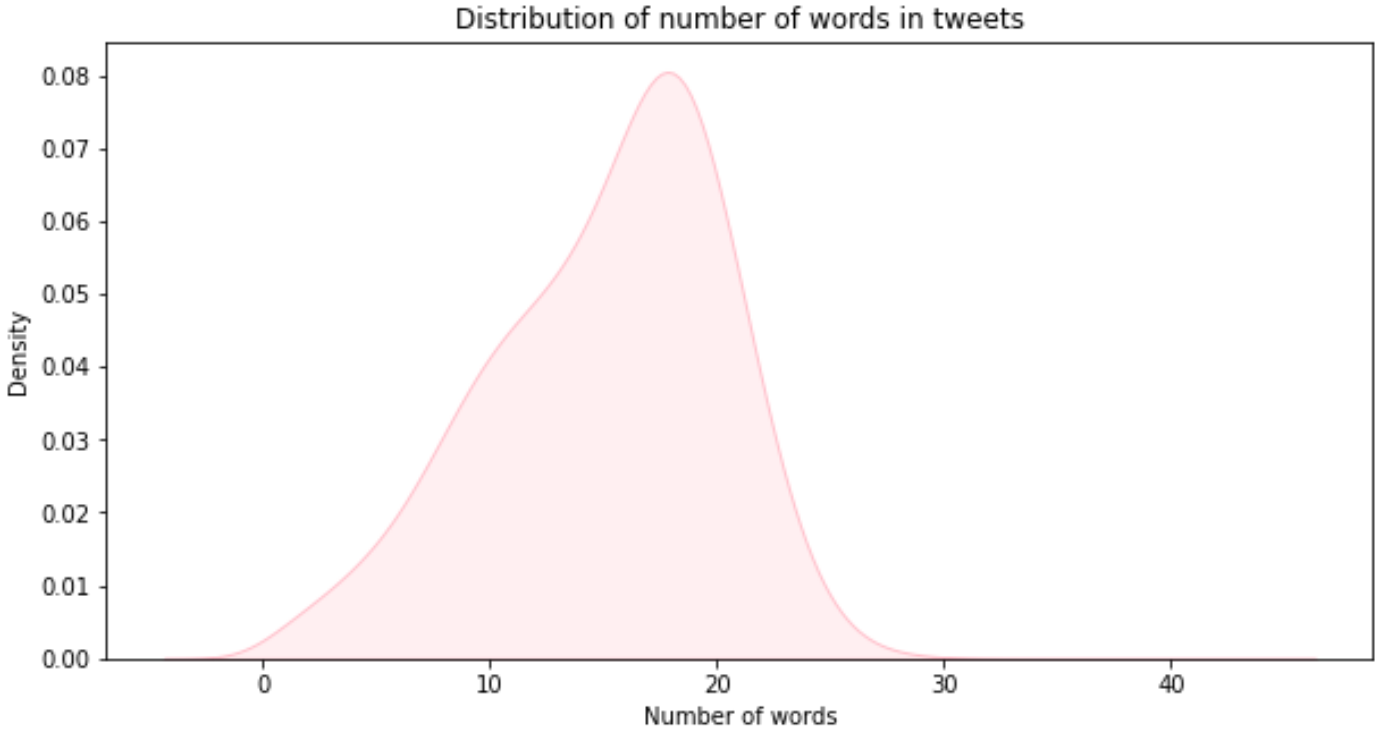
Word cloud of covid tweets

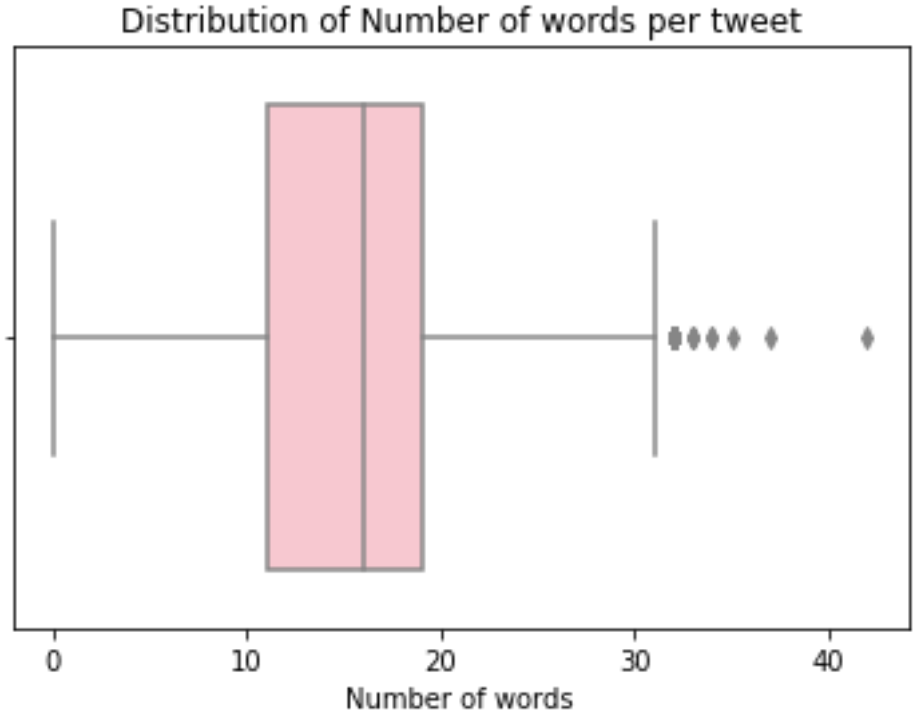


Word cloud of covid tweets without covid words.



Distribution of number of words in tweets





N-Grams

```
[(('new', 'cases'), 323),
 (('death', 'toll'), 131),
 (('tested', 'positive'), 98),
 (('public', 'health'), 87),
 (('cases', 'deaths'), 77),
 (('highest', 'daily'), 76),
 (('new', 'deaths'), 71),
 (('new', 'york'), 71),
 (('reports', 'new'), 67),
 (('tests', 'positive'), 67),
 (('coronavirusrelated', 'deaths'), 66),
 (('stay', 'home'), 62),
 (('honestie', 'hodges'), 60),
 (('health', 'officials'), 59),
 (('unemployment', 'claims'), 55),
 (('uk', 'records'), 54),
 (('cases', 'new'), 52),
 (('denver', 'mayor'), 52),
 (('holiday', 'season'), 52),
 (('americans', 'going'), 51),
 (('strict', 'spain'), 50),
 (('watch', 'live'), 49),
 (('deaths', 'highest'), 49),
 (('daily', 'total'), 48),
 (('avoid', 'travel'), 47),
 (('active', 'cases'), 47),
 (('deaths', 'new'), 45),
 (('task', 'force'), 45),
 (('confirmed', 'cases'), 44),
 (('going', 'hungry'), 44)]
```

```
[(('reports', 'new', 'cases'), 58),
 (('highest', 'daily', 'total'), 47),
 (('deaths', 'highest', 'daily'), 44),
 (('highest', 'death', 'toll'), 44),
 (('coronavirusrelated', 'deaths', 'highest'), 43),
 (('uk', 'records', 'coronavirusrelated'), 42),
 (('records', 'coronavirusrelated', 'deaths'), 42),
 (('daily', 'total', 'start'), 42),
 (('americans', 'going', 'hungry'), 41),
 (('thanksgiving', 'night', 'game'), 40),
 (('recovered', 'persons', 'recovered'), 40),
 (('persons', 'recovered', 'persons'), 40),
 (('strict', 'spain', 'masks'), 39),
 (('need', 'people', 'recovery'), 39),
 (('people', 'work', 'september'), 39),
 (('work', 'september', 'far'), 39),
 (('september', 'far', 'massive'), 39),
 (('far', 'massive', 'unemployment'), 39),
 (('massive', 'unemployment', 'peak'), 39),
 (('unemployment', 'peak', 'o'), 39),
 (('deaths', 'new', 'cases'), 37),
 (('honestie', 'hodges', 'handcuffed'), 36),
 (('cancels', 'cheering', 'crowds'), 36),
 (('cheering', 'crowds', 'wild'), 36),
 (('crowds', 'wild', 'celebrations'), 36),
 (('wild', 'celebrations', 'coaches'), 36),
 (('celebrations', 'coaches', 'players'), 36),
 (('coaches', 'players', 'look'), 36),
 (('players', 'look', 'hopeful'), 36),
 (('new', 'cases', 'new'), 35)]
```

```
[(('deaths', 'highest', 'daily', 'total'), 44),
 (('records', 'coronavirusrelated', 'deaths', 'highest'), 42
 (('coronavirusrelated', 'deaths', 'highest', 'daily'), 42),
 (('highest', 'daily', 'total', 'start'), 42),
 (('uk', 'records', 'coronavirusrelated', 'deaths'), 41),
 (('recovered', 'persons', 'recovered', 'persons'), 40),
 (('people', 'work', 'september', 'far'), 39),
 (('work', 'september', 'far', 'massive'), 39),
 (('september', 'far', 'massive', 'unemployment'), 39),
 (('far', 'massive', 'unemployment', 'peak'), 39),
 (('massive', 'unemployment', 'peak', 'o'), 39),
 (('persons', 'recovered', 'persons', 'recovered'), 39),
 (('cancels', 'cheering', 'crowds', 'wild'), 36),
 (('cheering', 'crowds', 'wild', 'celebrations'), 36),
 (('crowds', 'wild', 'celebrations', 'coaches'), 36),
 (('wild', 'celebrations', 'coaches', 'players'), 36),
 (('celebrations', 'coaches', 'players', 'look'), 36),
 (('coaches', 'players', 'look', 'hopeful'), 36),
 (('thanksgiving', 'night', 'game', 'ravens'), 34),
 (('night', 'game', 'ravens', 'steelers'), 34),
 (('game', 'ravens', 'steelers', 'switched'), 34),
 (('ravens', 'steelers', 'switched', 'sunday'), 34),
 (('steelers', 'switched', 'sunday', 'issu'), 32),
 (('growing', 'number', 'americans', 'going'), 31),
 (('hits', 'highest', 'death', 'toll'), 30),
 (('number', 'americans', 'going', 'hungry'), 29),
 (('strict', 'spain', 'masks', 'worn'), 27),
 (('highest', 'death', 'toll', 'hospitals'), 27),
 (('new', 'cases', 'new', 'deaths'), 26),
 (('nyc', 'checkpoints', 'sheriff', 'warns'), 26)]
```

total

```
[(('coronavirusrelated', 'deaths', 'highest', 'daily', 'total'), 42)
 (('deaths', 'highest', 'daily', 'total', 'start'), 42),
 (('uk', 'records', 'coronavirusrelated', 'deaths', 'highest'), 41),
 (('records', 'coronavirusrelated', 'deaths', 'highest', 'daily'), 4
 (('people', 'work', 'september', 'far', 'massive'), 39),
 (('work', 'september', 'far', 'massive', 'unemployment'), 39),
 (('september', 'far', 'massive', 'unemployment', 'peak'), 39),
 (('far', 'massive', 'unemployment', 'peak', 'o'), 39),
 (('recovered', 'persons', 'recovered', 'persons', 'recovered'), 39)
 (('persons', 'recovered', 'persons', 'recovered', 'persons'), 39),
 (('cancels', 'cheering', 'crowds', 'wild', 'celebrations'), 36),
 (('cheering', 'crowds', 'wild', 'celebrations', 'coaches'), 36),
 (('crowds', 'wild', 'celebrations', 'coaches', 'players'), 36),
 (('wild', 'celebrations', 'coaches', 'players', 'look'), 36),
 (('celebrations', 'coaches', 'players', 'look', 'hopeful'), 36),
 (('thanksgiving', 'night', 'game', 'ravens', 'steelers'), 34),
 (('night', 'game', 'ravens', 'steelers', 'switched'), 34),
 (('game', 'ravens', 'steelers', 'switched', 'sunday'), 34),
 (('ravens', 'steelers', 'switched', 'sunday', 'issu'), 32),
 (('growing', 'number', 'americans', 'going', 'hungry'), 29),
 (('hits', 'highest', 'death', 'toll', 'hospitals'), 27),
 (('heres', 'look', 'different', 'vaccines', 'work'), 26),
 (('nyc', 'checkpoints', 'sheriff', 'warns', 'consequences'), 25),
 (('warnings', 'avoid', 'travel', 'denver', 'mayor'), 25),
 (('avoid', 'travel', 'denver', 'mayor', 'hancock'), 24),
 (('travel', 'denver', 'mayor', 'hancock', 'flies'), 24),
 (('checkpoints', 'sheriff', 'warns', 'consequences', 'violating'),
 (('mayor', 'hancock', 'flies', 'visit', 'family'), 23),
 (('sheriff', 'warns', 'consequences', 'violating', 'quarantine'), 2
 (('denver', 'mayor', 'hancock', 'flies', 'visit'), 22)]
```

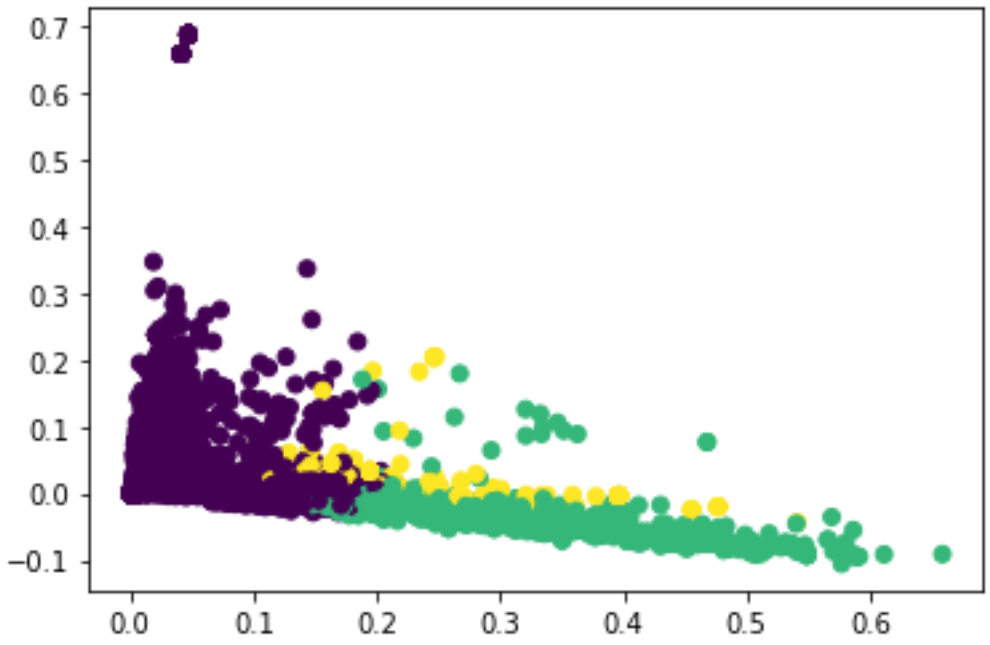
Topic extraction

Kmeans describing words per cluster

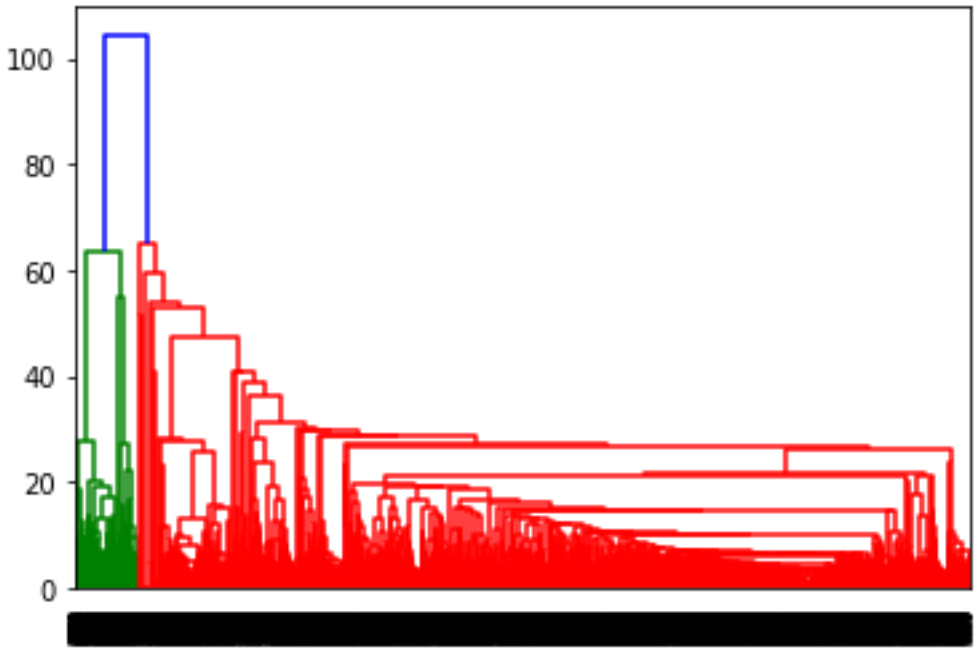
```
For cluster 0 the most describing words are:
people
thanksgiving
pandemic
new
health
americans
vaccine
need
year
week
```

```
For cluster 1 the most describing words are:
recovered
person
honestie
hodges
police
handcuffed
dead
flies
warnings
hancock

For cluster 2 the most describing words are:
cases
new
deaths
reports
reported
wednesday
confirmed
death
total
today

For cluster 3 the most describing words are:
highest
toll
death
daily
records
uk
deaths
hits
coronavirusrelated
```

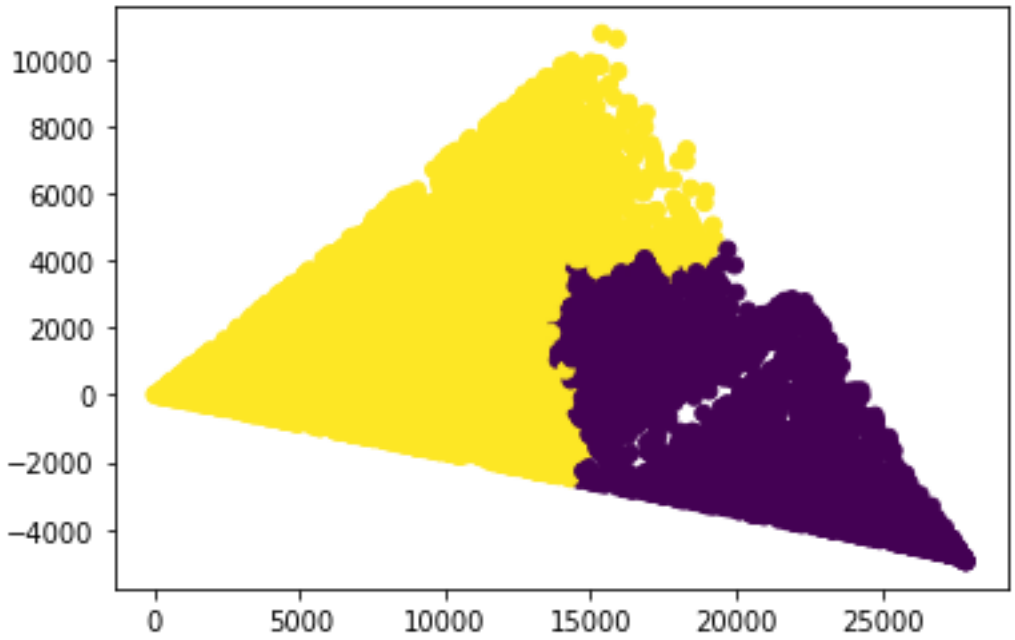
Kmeans clustering visualization with TSNE



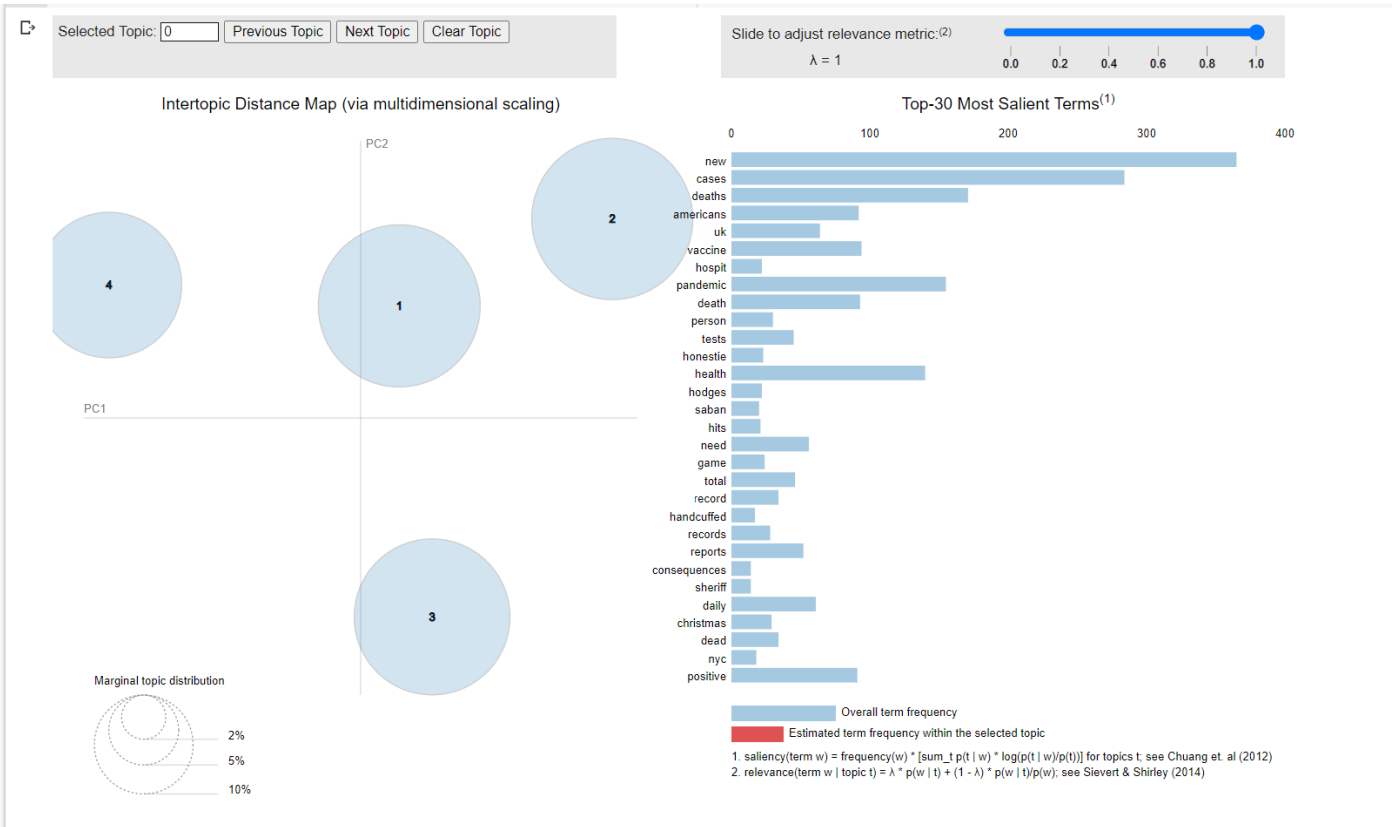
Dendrogram of clustering of tweets



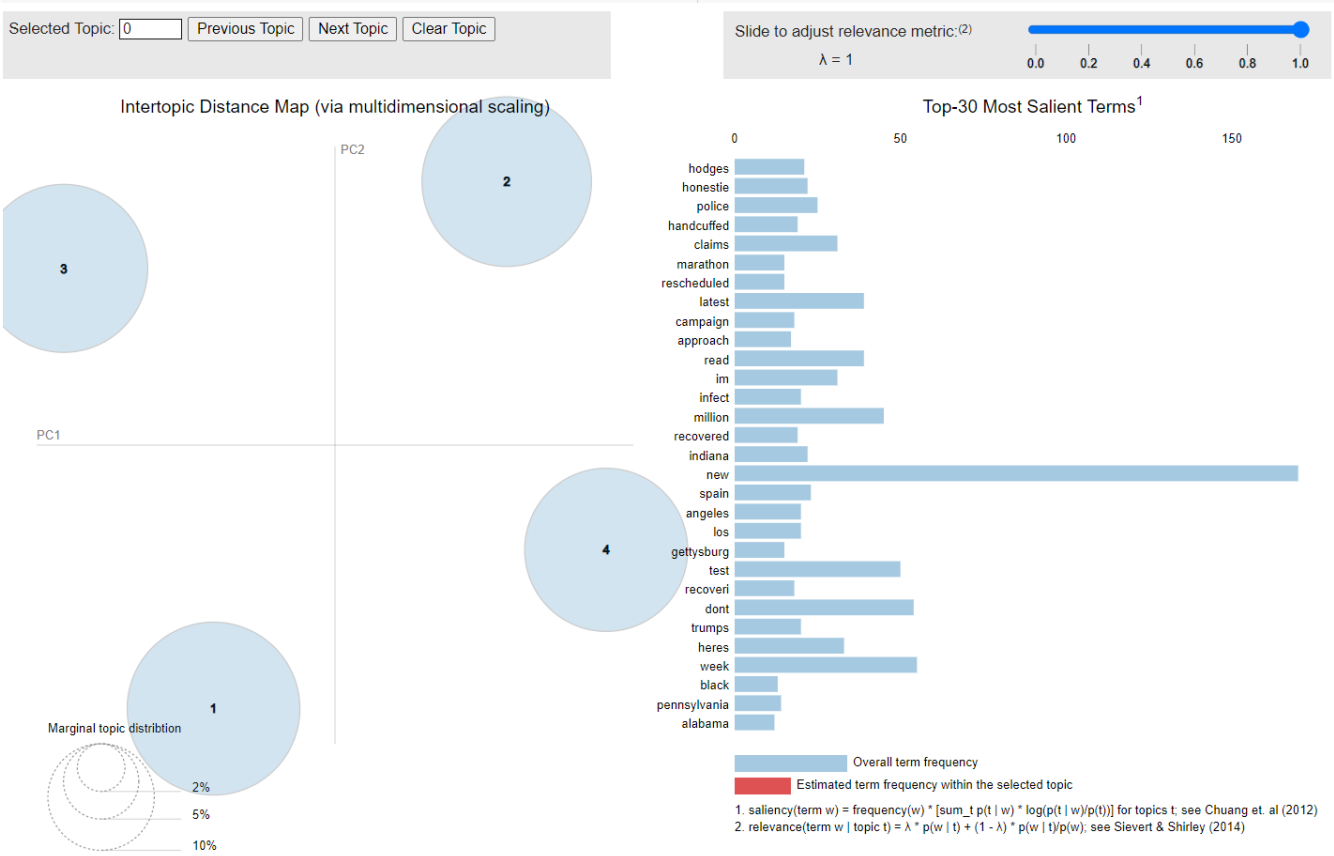
Visualization of agglomerative clustering of content with TSNE



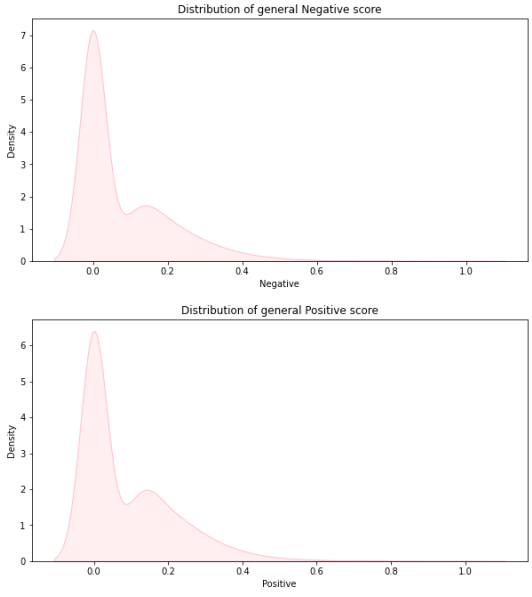
LDA Results on Bag Of Words

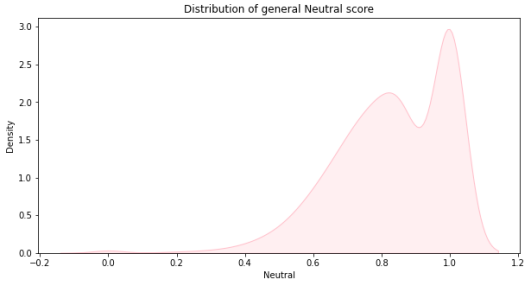


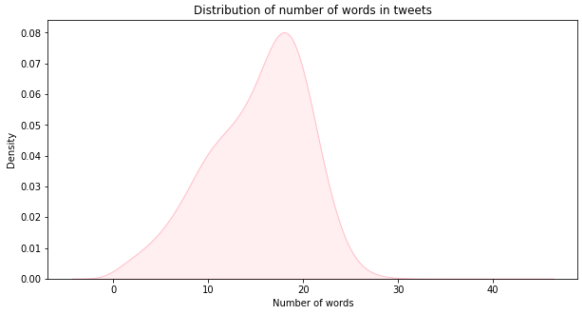
LDA Results on Tfldf



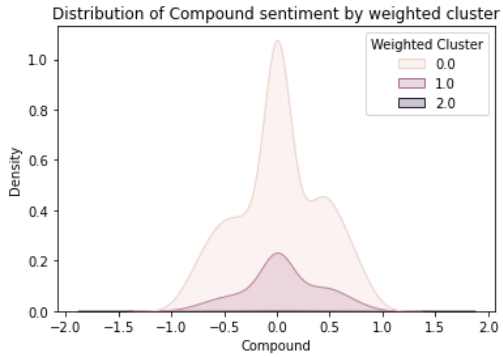
General density of sentiment analysis

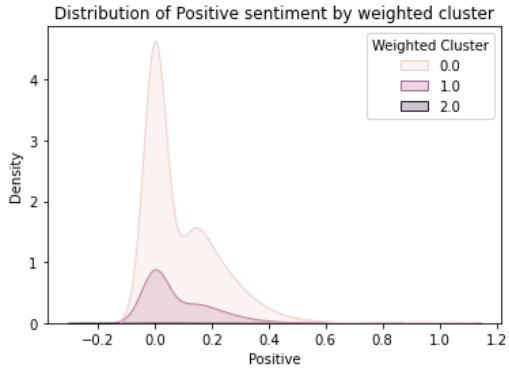


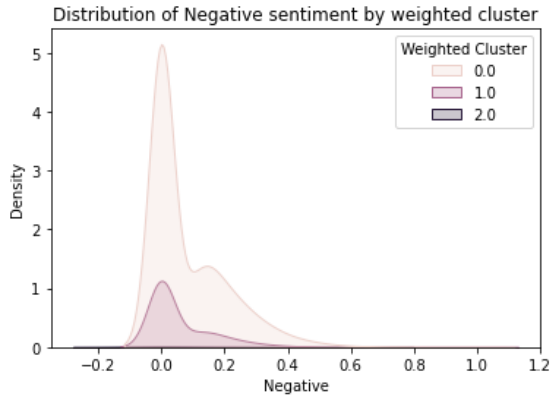


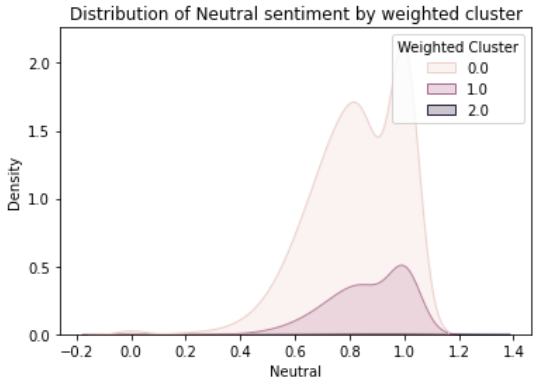


SENTIMENT ANALYSIS BY WEIGHTED CLUSTER

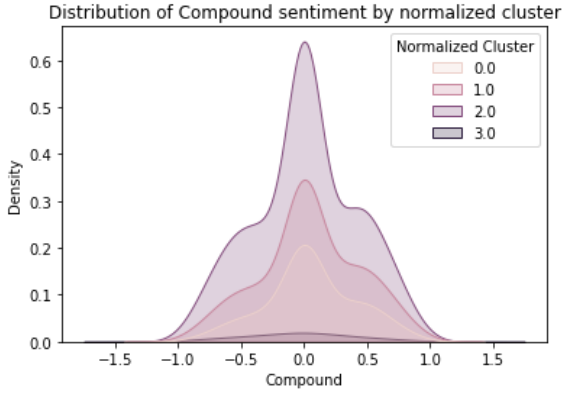


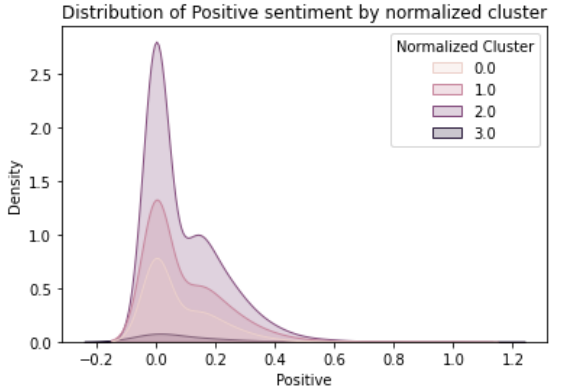


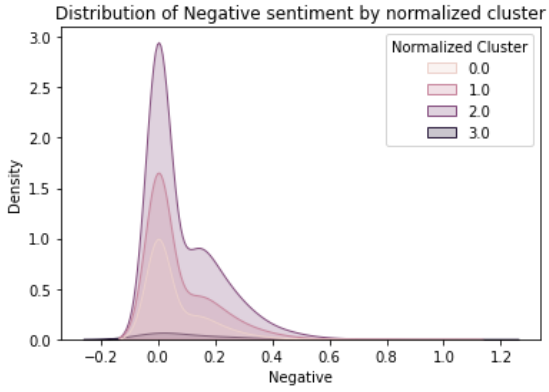


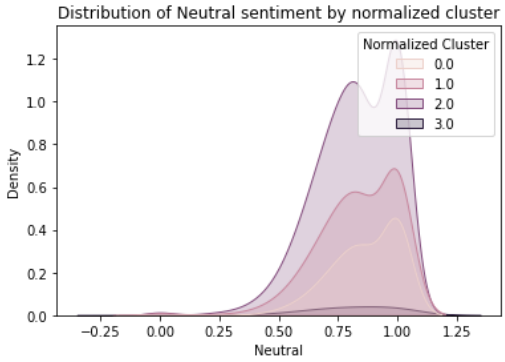


SENTIMENT ANALYSIS BY NORMALIZED CLUSTER









MOST FREQUENT WORDS PER CLUSTER

For cluster 0 the most describing words are:
people
new
pandemic
test
health
trump
amp
thanksgiving
dont
says

For cluster 1 the most describing words are:
vaccine
pfizer
effective
spain
oxford
astrazeneca
says
strict
moderna
effect

For cluster 2 the most describing words are:
jr
donald
trump
tests
posit
positive
son
tested
eldest
spokesman

For cluster 3 the most describing words are:
cases
new
deaths

reported
death
reports
day
total
daily
record