Data process

Before working on this dataset, I picked the attributes that used in tasks from JSON and saved them in a new csv file. The final size of file is about 3GB.

Basic Stats

In this dataset, there are about 14.36 million tweets in the dataset. Twitter allocates a unique id to each tweet so that we can use it to detect duplicates and to make sure we get the correct data to analysis.

I plot the number of tweets by days in June with bar plot (Figure 1), we can find that the lowest number of tweets appeared in 10th, June, and the numbers were larger than 400000 in all days. After that, I split days into weekend and weekday, and plot box and whisker diagrams (Figure 2) for them, the result shows that weekday has larger means in this number.

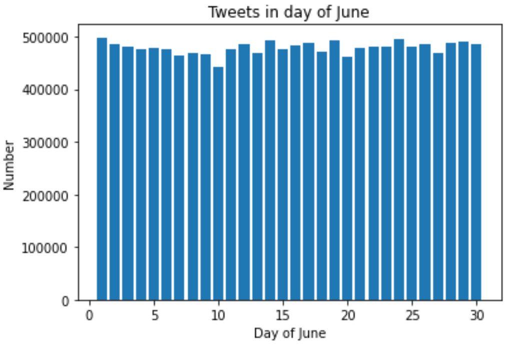


Figure 1

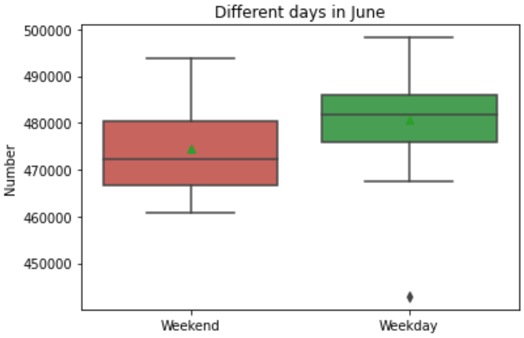


Figure 2

I used bar plot to research the time-series by hour in all days and weekdays (Figure 3-4), the result show that in the mid-night (12PM-5AM), the number of tweets is the lowest in a day. In the morning, the number grows at 6AM, and be stable at 9AM. There are most tweets in the evening (6PM-9PM), then the number drops after 9AM.

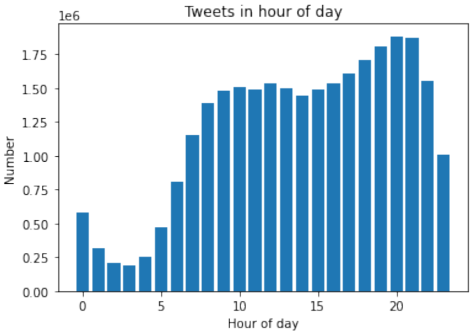


Figure 3

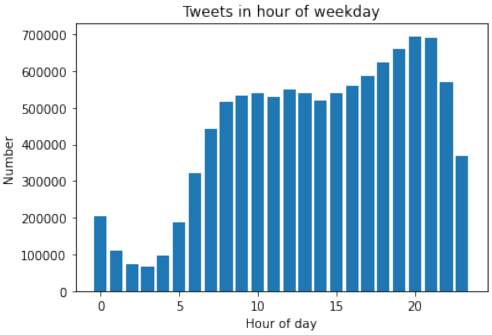


Figure 4

Users

Twitter gives a unique ID to each user, and each user has a screen name so that I constructed a dictionary to store their relations.

To visualize the number of tweets for users, I generated histogram for all data collected first (Figure 5). Unfortunately, it was difficult to read. I used box plot to learn data distribution (Figure 6) and it clearly showed that there are a lot of outliers in the data. I picked users whose number was smaller than 100 and 12 (75% data) to draw histogram and get the result (Figure 7-8). The results illustrate that most of users did not send many tweets, and there are most users who only sent one tweets in June.

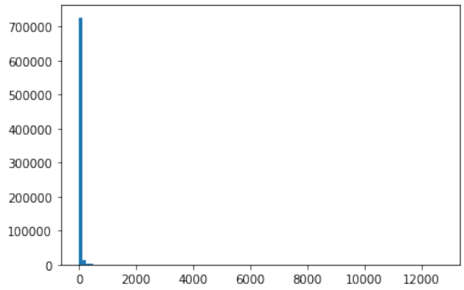


Figure 5

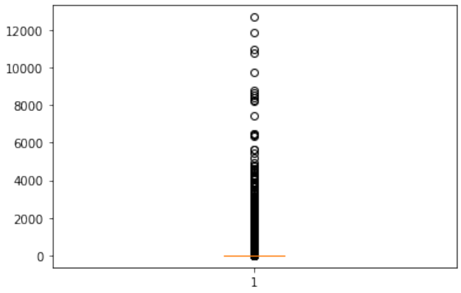


Figure 6

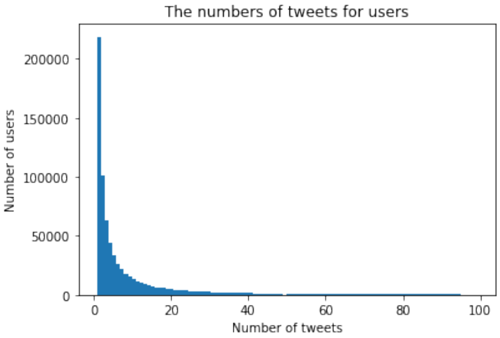


Figure 7

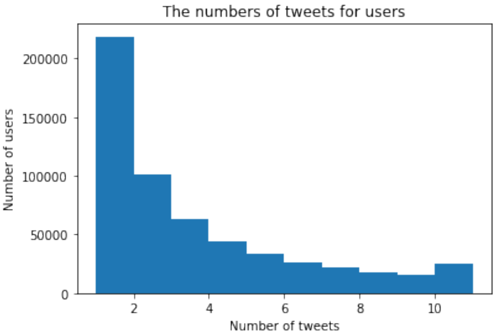


Figure 8

The top-5 users by total number of tweets are:

Screen name	Number
Kardeimcin1	12693
DailyNews79	11859
c_antolic	10980
HoraCatalana	10775
minijobanzeigen	9752

If we visit their twitter pages, we can know that all of them are automated accounts (it seems that the second user is frozen).

Top-5 users get most mentions:

User: YouTube	Number: 19822
User: RTErdogan	Number: 17162
User: BorisJohnson	Number: 15456

User: elonmusk Number: 10075
User: GBNEWS Number: 7547
Top-5 users get most comments:
UserID: 2866804900 Number: 5167
UserID: 1503799593405800448 Number: 5101
UserID: 44196397 Number: 4121
UserID: 1339166129110065152 Number: 3547
UserID: 3131144855 Number: 3356

I did not get the screen name for the first two users; they may be frozen. The last three users are: @elonmusk, @GBNEWS and @BorisJohnson

I selected Italia, France, Spain, and UK to compute how often they mention each other, the result is visualized by heatmap (Figure 9). We find that in these four countries, except Spain, other countries mentioned themselves mostly. In addition, UK mentions other countries most among them.

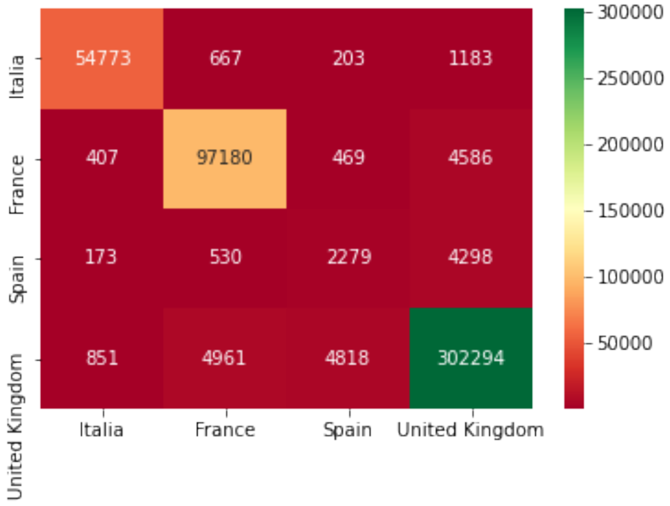


Figure 9

Mapping

There are about 600000 GPS coordinates in the dataset. As we use them to draw map with dataset, the points are transformed into EPSG3857 format. I used matplotlib to generate scatter plot for these point coordinates (Figure 10) and I went to the Eurostat website to download geospatial vector data and generated map with country boarder (Figure 11).
Dataset: https://gisco-services.ec.europa.eu/distribution/v2/nuts/download/ref-nuts-2021-01m.json.zip
Using scatter plot without any other information, we can also see the outline of Europe, and some outliers can also be observed.

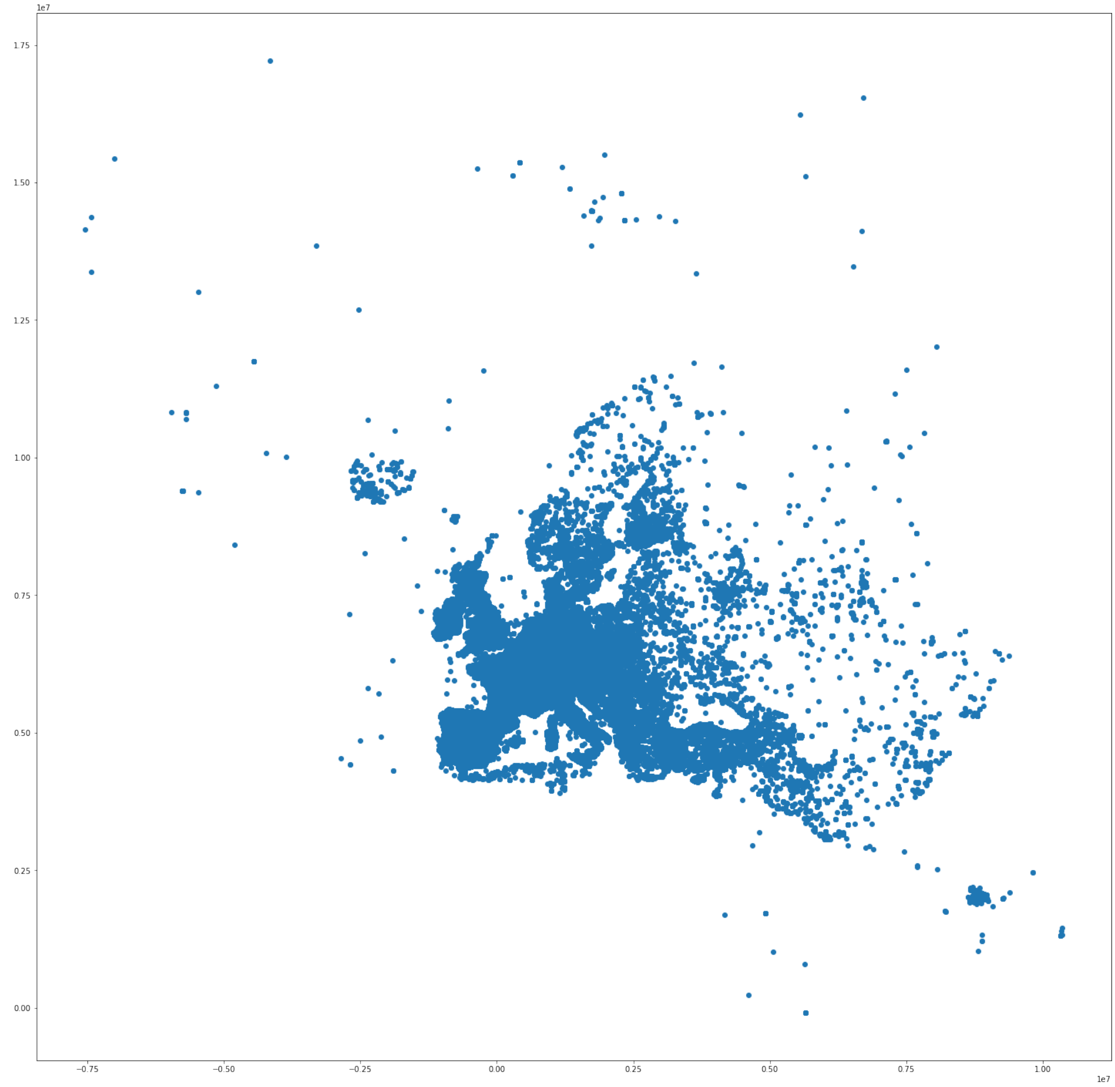


Figure 10

And after adding country outline in the map, the map illustrates that some countries that had more Twitter activities such as UK, France, German and other countries in Europe. The data used in this part doesn’t contains some countries such as Ukraine. Iceland, Norway, Sweden, and Finland had less uses of Twitter than countries in southern Europe.

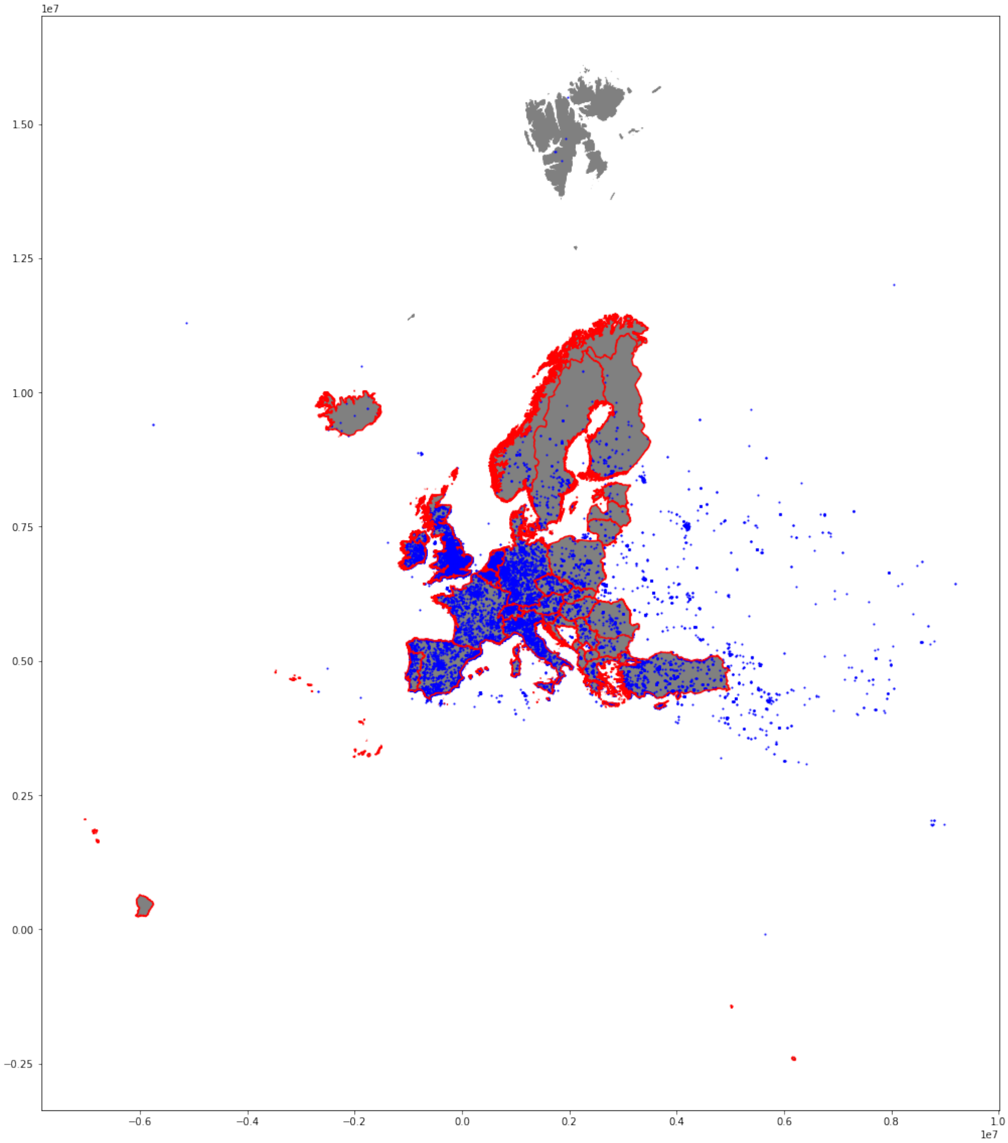
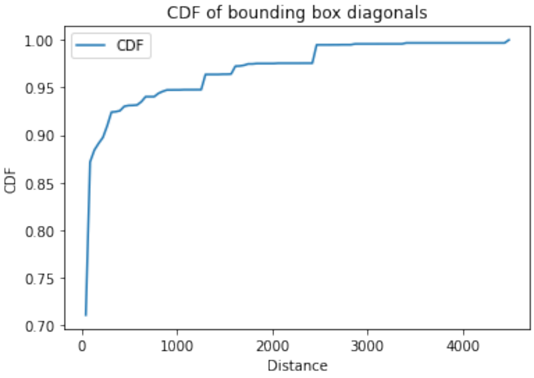


Figure 11

There are more than 10 million record that has bounding box information, and the range of diagonals is from 0km to 5000km. I generated two CDF plot (Figure 12-13), one for the whole dataset and another is for values smaller than 500. We can find that most of diagonals is smaller than 1000km (95%), and for values smaller than 500, most of them are less than 100km.



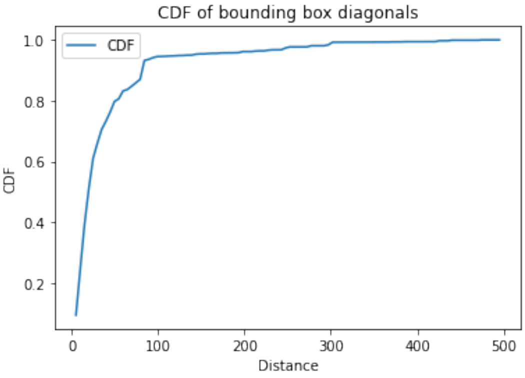


Figure 12

Figure 13

I filter the records in UK and created another map in United Kingdom (Figure 14). Southern UK has more tweets, especially in London and Birmingham

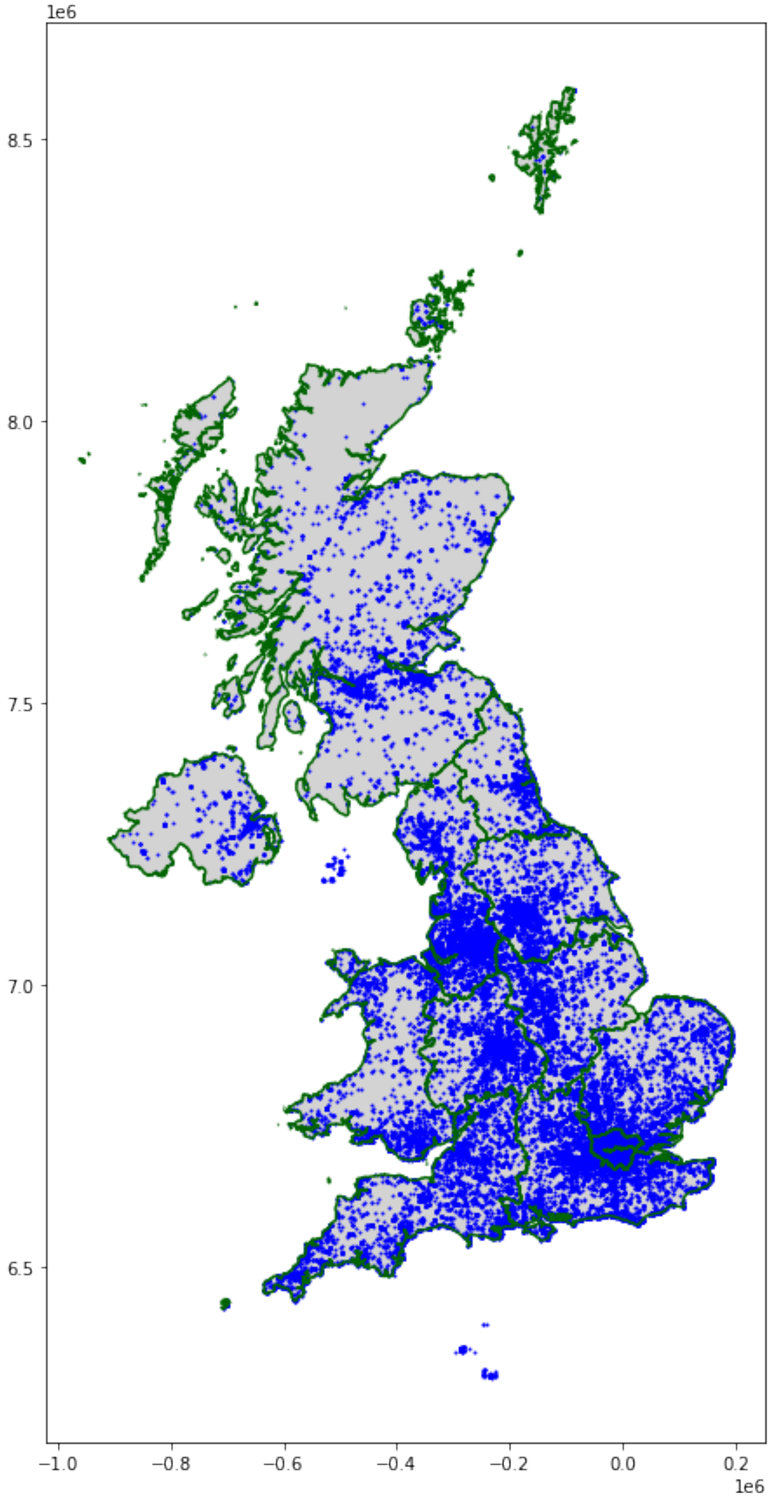


Figure 14

Events

I picked UK, France, and Turkey to research this topic. I used bar plot to show time series at first (Figure 15, 17, 19) and box plot in numbers of tweets (Figure 16, 18, 20) to find unusually high activity. The June 23 in UK, June 19 in France, and June 29 in Turkey is unusual.

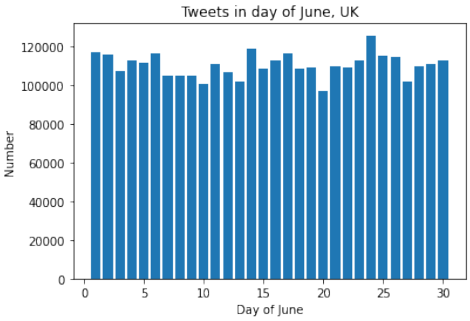


Figure 15

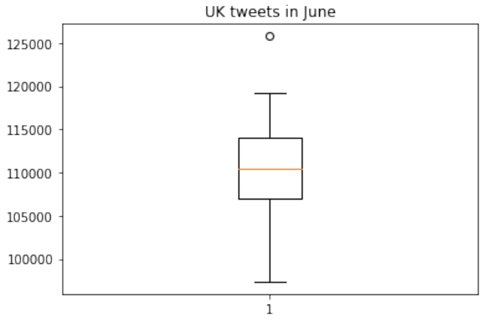


Figure 16

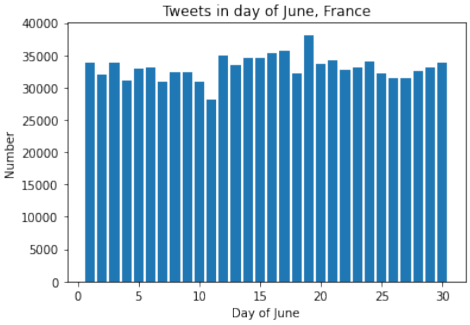


Figure 17

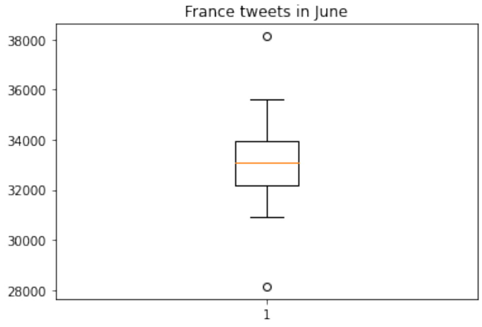


Figure 18

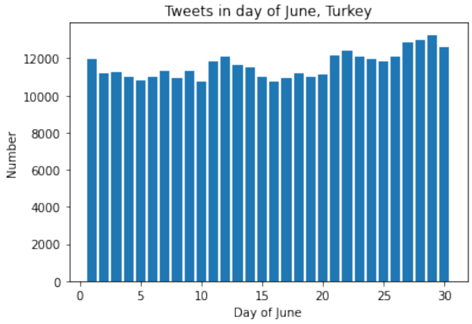


Figure 19

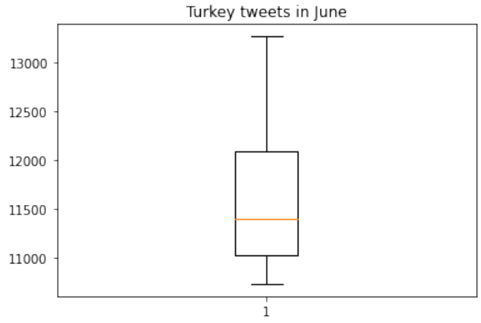
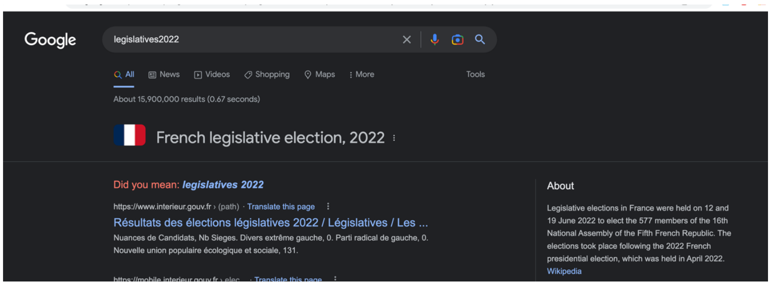


Figure 20

Before generating word cloud, stop words and other frequent words such as twitter, t.co, and https are removed from text, and all words are in low case. I noticed that in word cloud for France, legislatives2022 looks like an activity and I search it in Google:



About 15,900,000 results (0.67 seconds)

French legislative election, 2022

Did you mean: ***legislatives 2022***

https://www.interieur.gouv.fr › (path) · Translate this page

Résultats des élections législatives 2022 / Législatives / Les ...

Nuances de Candidats, Nb Sieges. Divers extrême gauche, 0. Parti radical de gauche, 0. Nouvelle union populaire écologique et sociale, 131.

https://mobile.interieur.gouv.fr › elec... · Translate this page

About

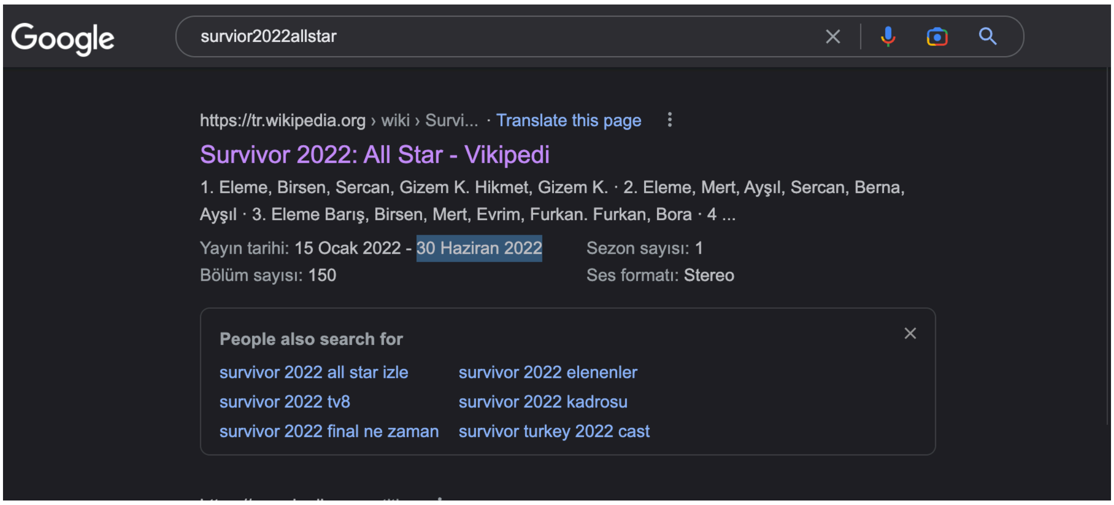
Legislative elections in France were held on 12 and 19 June 2022 to elect the 577 members of the 16th National Assembly of the Fifth French Republic. The elections took place following the 2022 French presidential election, which was held in April 2022.

Wikipedia

And it was National Assembly of the Fifth French Republic so that it can explain the unusually high activity in France on that day.

23 June - International Olympic Day, the frequent words in UK’ s word cloud such as thank, great, think, and love is matched with Olympic spirit. Hence, we can assume that there are lots of tweets about Olympic on that day so that unusually high activity is reasonable.

As for Turkey, unfortunately most of the words were not in English but I still noticed surviorr2022allstar and I searched it. This is a TV program in Turkey, and it was ended in June 30 so that we can assume that most of Twitter users discussed about ending and story of this TV program



Google

survior2022allstar

https://tr.wikipedia.org › wiki › Survi... · Translate this page

Survivor 2022: All Star - Vikipedi

1. Eleme, Birsen, Sercan, Gizem K. Hikmet, Gizem K. · 2. Eleme, Mert, Ayşıl, Sercan, Berna, Ayşıl · 3. Eleme Barış, Birsen, Mert, Evrim, Furkan. Furkan, Bora · 4 ...

Yayın tarihi: 15 Ocak 2022 - 30 Haziran 2022

Sezon sayısı: 1

Bölüm sayısı: 150

Ses formatı: Stereo

People also search for

survivor 2022 all star izle

survivor 2022 elenenler

survivor 2022 tv8

survivor 2022 kadrosu

survivor 2022 final ne zaman

survivor turkey 2022 cast

Word cloud:

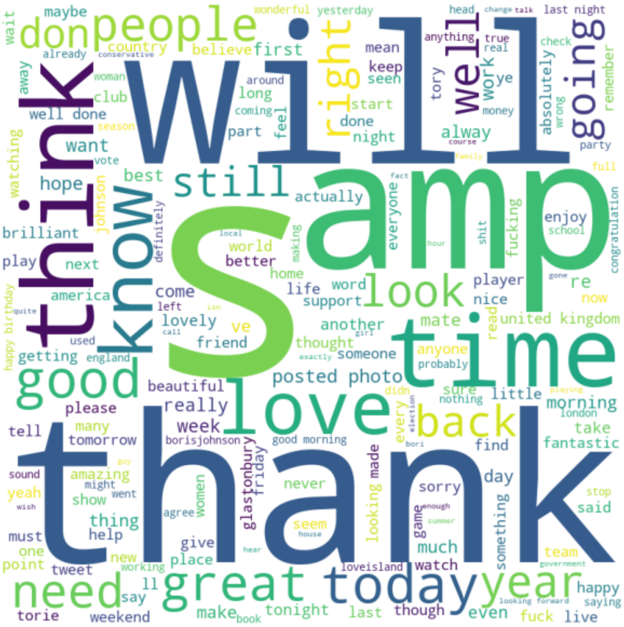


Figure 21 Word cloud for UK

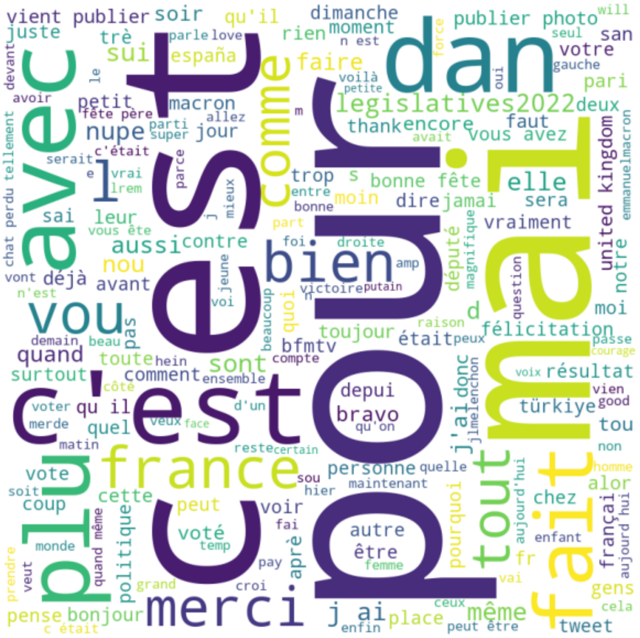


Figure 22 Word cloud for France

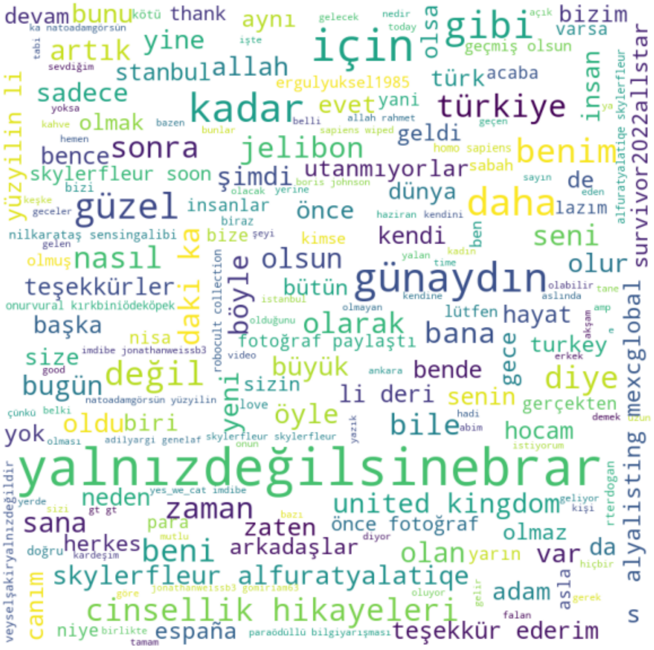


Figure 23 Word cloud for Turkey

Reflection

Twitter data is huge so that it can be difficult for data scientist to process and analysis, and it is more expensive in computing. As there are lots of automated robot in Twitter, it increases the cost in storing data and cleaning data because some useless tweets are difficult to detect. We can also observe some strange expression such as abbreviate words and Emoji, which is difficult for computer to learning. In addition, there are also lots of comment data in Twitter, most of them are short and if we need to research them, we need consider the tweet connected to it, which can be very complex and difficult.

Twitter publishes its data without informing their users (most of users do not know it) and it is almost impossible to inform all Twitter user that how their data had been used. User ID and tweets for a certain can be obtained easily for developers, other researchers, and companies. Twitter data include some sensitive privacy data such as accurate location coordinates, personal information, favorites. With specific techniques, we can get activity tracks, personal hobbies and lifestyle habits for a user, this information can be used by sharks to implement fraud or by advertisers to construct accurate user portrait, which is unfair and horrible to users. In addition, processed data such as data with cleaning process or data concentrates on specific group of people are more related to personal privacy, it should be used carefully and secrecies for this kind of data deserve more attentions.