Rovira i Virgili University

Master in Artificial Intelligence

Big Data Analytics

Twitter Lab

Authors: Rafael Bianchi

November 11, 2019

UNIVERSITAT ROVIRA i VIRGILI

Contents

1 Introduction **2**

2 Analysis **3**

2.1 Languages 3

2.2 Content Type 3

2.3 Other Hashtags 4

2.4 Users' Countries 5

2.5 Sentiment Analysis 7

3 Conclusion **9**

References **10**

1 Introduction

The idea of this lab is to stream data from twitter, given a hashtag, save the data on `MongoDB` and create insight analysis from it. In order to setup the environment, it is necessary to create a developer app on Twitter, and also download the necessary python packages in order to connect and download the tweets' information.
The selected hashtag for this assignment was `#brexit`, as the deadline for a deal no-deal approaches, this subjects is quite trending. The number of tweets saved for this analysis is close to 11k tweets, on the same day in a time window of around three hours.

2 Analysis

For each tweet, Twitter also has the metadata related to the tweet (date, language, location, retweeted, favorited, media and etc.) and also about the user that posted that tweet (location, verified, timezone, friends count and etc.). Even though the information is not always complete and, sometimes some properties of this data is not reliable, it is possible to make some analysis on the data gotten from Twitter API's.

2.1 Languages

One of the information available on the tweet metadata, is the language. On figure 1 it is shown the distribution of the languages among the tweets collected. It is clear that the English language is far more used than the others. Also, there is some missing data as a und(undefined) language.

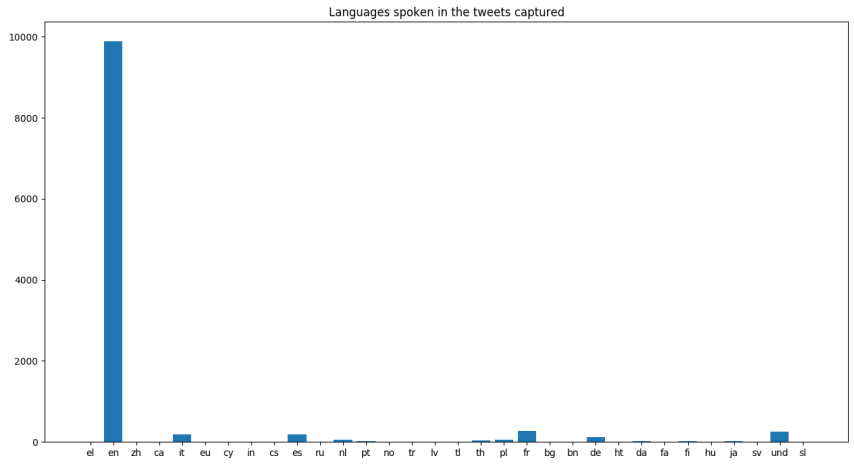

Figure 1: Languages used on tweets for `#brexit`

2.2 Content Type

In figure 2 it shows the percentage of each content type. The vast majority of the tweets are retweets and original content and quotations have similar amount. There are only few replies, so there are few people actually

discussing this topic with each other.

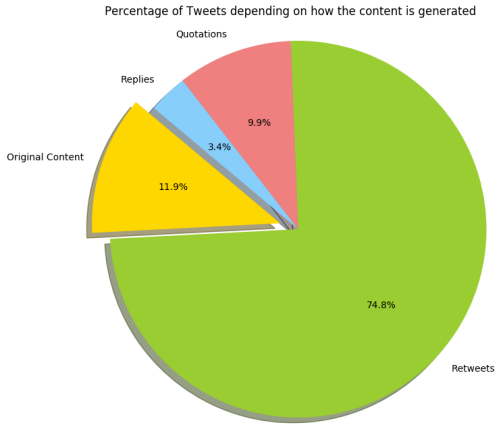

Figure 2: Type of the tweets for `#brexit`

2.3 Other Hashtags

For the same tweet, the user might add another hashtags that might be related to the chosen hashtag. Figure 3 shows the top fifteen hashtags found with the hashtag brexit. It is easy to see most of them are tight related to the brexit process.

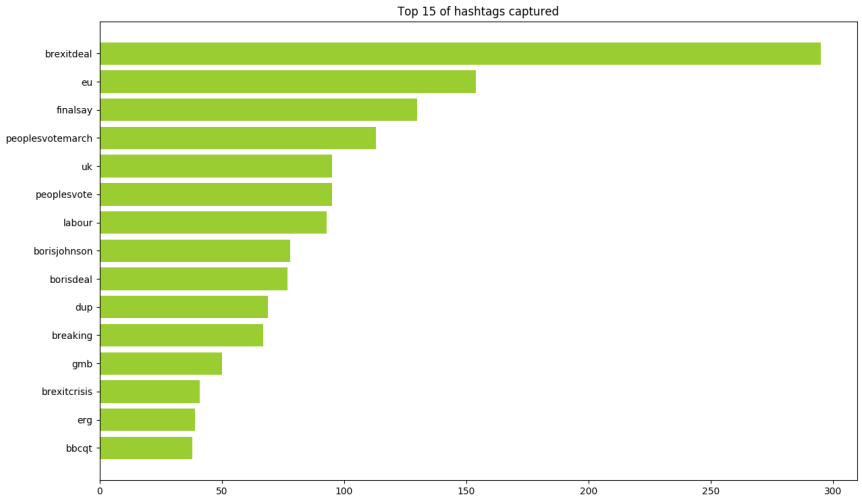

Figure 3: Top 15 hashtags found together with `#brexit`

2.4 Users' Countries

Not always the geographical information is available for the tweets. In this test, only 13 out of the 11203 tweets had the information about the location. On the other hand, a considerable number of the users (7815 out of 11203) had some information of the location on their profiles. The problem is that the information is not standardized and the user can fill anything on the user location field. To make it more reliable, it was used the Geocode[1] library in order to make calls to the OpenStreetMap[2] so that the user location information text can be geocoded and then, the country information is available for analysis. Figure 4 shows the top fifteen users' countries that tweeted `#brexit`. The relation on the language found in figure 1 is clear, most of the users are from UK and the second place is also a English speaker country: USA. It is curious to find out that countries that are not directly impacted by the brexit are interested in the subject (Canada, Japan and etc.).

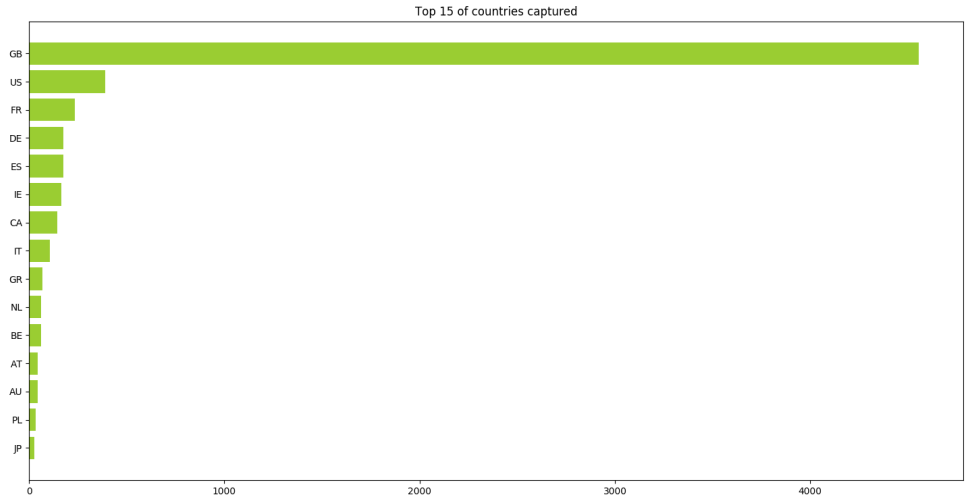

Figure 4: Top 15 users' countries that tweeted `#brexit`

Another interesting way to see the data is plotting it as a choropleth map. In image 5, it gives a big view of the twitter distribution around the world for the `#brexit` hashtag. On thing to notice is that for obvious reasons UK has more twitters than any other country, but also, that the twitters fetching app was running while was afternoon in Catalonia, so others countries might not have a good representation at that time period.

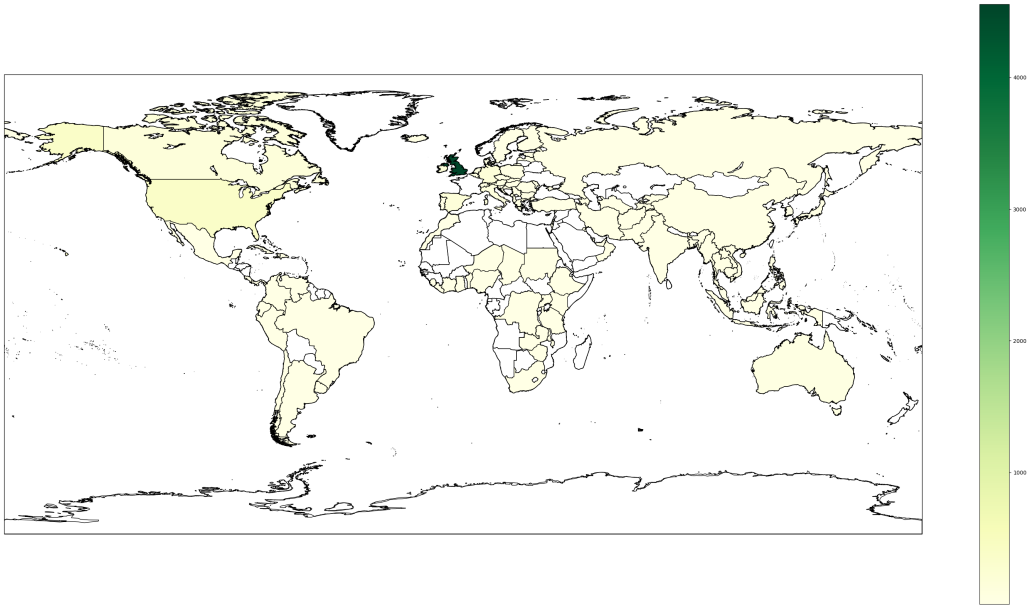

Figure 5: Choropleth map: Twitters per country. `#brexit`

2.5 Sentiment Analysis

In image 6, the map show another choropleth map but this time instead of the number of tweets, it shows the average of the polarities of the twitters for each country.

Figure 6: Polarities map: Averaged polarities scores for each country. `#brexit`

The polarity score takes into consideration all elements from the tweet string. It is curious to see countries like Uruguay and South Korea so with negative polarities and others with so positive polarity like Ethiopia and Bangladesh.

3 Conclusion

Getting twitter information is easy ans straightforward. The difficulty appears when we try to explain the data using plots that transmits the analysis to the reader. Also, other types of information could be released (sex, age, real location and etc.) that could help to build a more robust and solid analysis. The sentiment analysis is fun, but it could be more elaborated to separate hashtags supporting and opposing the brexit so the polarity could be more readable (people are supporting/opposing to staying/leaving UE).

References

[1] D. Carriere, "Python geocoder," https://github.com/DenisCarriere/geocoder, 2014.

[2] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org, 2017.