Event detection on a Twitter dataset

Albrici Tanguy
tanguy.albrici@epfl.ch

Di Dio Davide
davide.didio@epfl.ch

Bruno Wicht
bruno.wicht@epfl.ch

Abstract

Social networks now have a huge importance in our lives and many people use them to comment about events that are happening around the globe. With this project, we would like to see how the Swiss Twitter community reacts to important events happening in Switzerland or around the world. Our main goal is to determine to what extent and how well we can learn about what is happening in the world or in our country based on the Swiss tweets. The story we want to tell is the evolution of tweets during important events between 2010 and 2016 and discover what kind of events Swiss people are tweeting about the most. We are motivated to do this project and tell this story because none of us are active on Twitter and we're interested in understanding better how twitter is used in Switzerland.

1 Dataset Description and Pre-processing

At first, we extract the hashtags from each tweets. Each tweet is stored in a dataframe that only contains the informations we require, namely:

- The tweet id
- The user id
- The longitude and latitude
- The hashtags extracted from the text
- The day, month and year of the tweet creation

After the hashtag extraction, we only keep the tweets with at least one hashtag. This dataframe is called `df_tag` in our notebook

2 Data Manipulation

2.1 Grouping by hashtag

In order to implement event detection efficiently, we need to compute a dictionary containing for each hashtags the ids of each post that contains it. This is done in the function `group_by_hashtag(...)`. This function iterates over every post in the database and adds the id of the post in the dictionary entry of the hashtag. In this method, we also compute the number of unique authors for this hashtag. With this value, we can easily filter out hashtags that were tweeted by only a few users.

3 Data analysis and visualization

This section covers how we visualize our data and allow us to get a feel on how to detect events.

3.1 Visualizing hashtag frequency

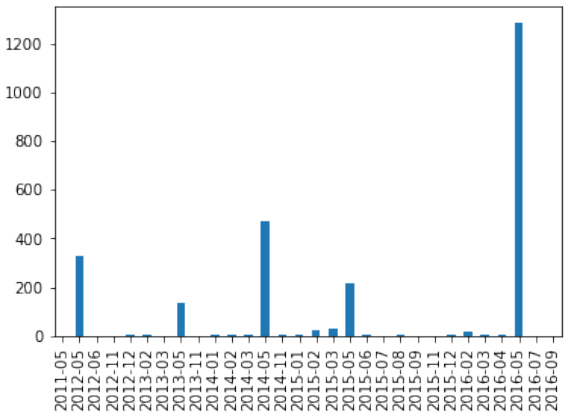

Figure 1: For the top months, plot the number of tweets with `#eurovision`

With frequency visualisation, we can visualize spikes in the number of tweets. The method `plot_frequency_tags(...)` computes for a given hashtag the numbers of tweets per frequency that contain this hashtag. Here we can

chose frequency as either day, month or year. Then it takes the `n` most tweeted dates and displays them in chronological order with a bar plot. We only take the `n` most tweeted dates because it provides the most compact visualisation. Note however that these plot are not homogenous in time.

Now, let us look a the frequency plot 1 for `#eurovision` . We can very clearly see an increase in the number of tweet during the month of may of each years. Since eurovision takes place in may, it is obvious that people will more likely talk about it during it's happening. We will be using this knowledge to develop an algorithm to detect those sudden spikes.

3.2 Visualizing event localisation

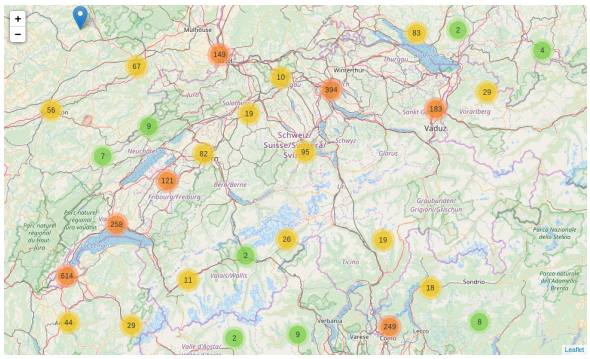

Figure 2: Map with every tweet for `#eurovision`

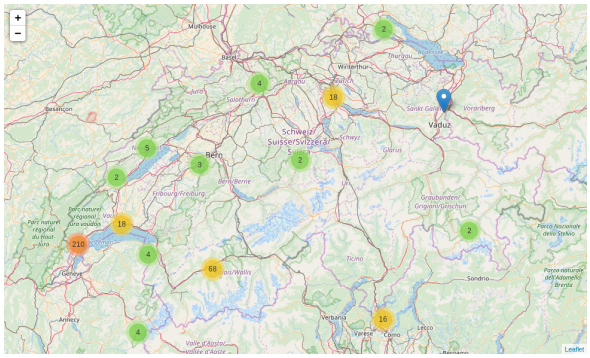

Figure 3: Map with every tweet for `#paleo`

The map are displayed in order to get an intuition on the geographic repartitions of tweets. The map are created using folium. We use a MarkerCluster to group every tweet close in space. Other than that, it is a pretty straightforward folium map. Now, look at the first map**??**. We see that the distribution is regrouped in the cities but overall it is quite uniform if we consider the population density. On the other side, in the second map 3, we see that most of the tweets are located in Nyon, which is indeed where this music festival takes place. The logic here people close to the actual event are much more likely to talk about it. Whereas international event are known in switzerland as a whole. Hence a local events lead to tweet localisation around the position of the event and international events have a spread tweet geographic distribution. We will use this knowledge to find the type of an event.

4 Event detection

In this part of the project we focus on automatic event detection.

4.1 Filtering out irrelevant hashtags

In the data manipulation, we constructed a dictionary containing each hashtag and their respective number of post and unique authors. Since hashtags that do not have enough posts and that are posted by only a few users are not likely to be considered events by our detection algorithm. We will be filtering out every hashtag that have less than `100` overal posts and less than `50` total authors.

5 Event Localization

6 Results

7 Conclusion

References

[Gusfield1997] Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.