Event Detection on a Twitter Dataset

Tanguy Albrici
tanguy.albrici@epfl.ch

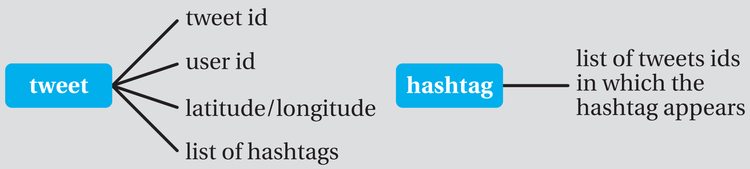
Davide Di Dio
davide.didio@epfl.ch

Bruno Wicht
bruno.wicht@epfl.ch

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

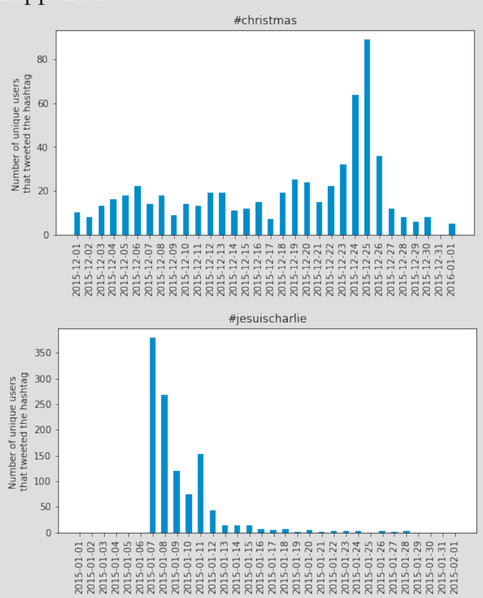
Data Cleaning

Our dataset consisted of more than 20M tweets localized in Switzerland between 2010-2016, but only we decided to keep only the 3.4M of them that contained hashtags, as our idea was to detect events based on hashtags. We further cleaned our dataset by keeping only the tweet metadata that will be used later on, which is shown below. Then we group our data by hashtag and remove those that either don't appear frequently enough or don't have enough unique authors. Our cleaned data structure is thus as follows:



Data Analysis

Before performing event detection, it is important to get a feel of how the number of hashtags tweeted typically varies for some given events. We can observe that they are two main types of events that have very different characteristics: predictable and unpredictable events.
To show this, we can take as example Christmas and the Charlie Hebdo attacks. Indeed, Christmas produce a huge interest for the whole month of december with a peak on the 25th, while Charlie Hebdo attacks were not expected at all and nobody was talking about it before it happened.

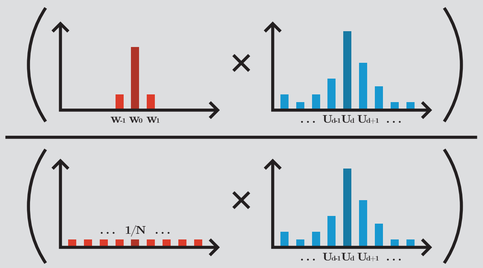


Event Detection

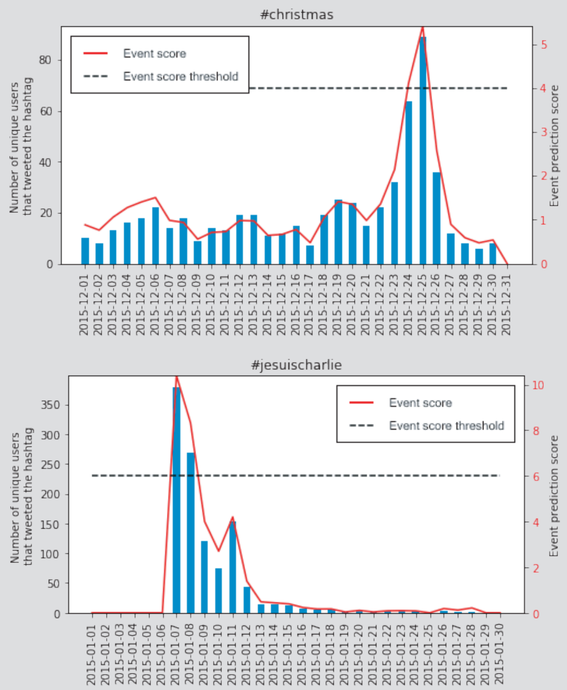
We used an algorithm that considers the number of unique users that used a certain hashtag per day. We computed an "event score" for each day, which is given a local Gaussian-like average of the number of unique users, divided by a more global average around this day. Once we have an event score for each day, we used a threshold to determine which days should be considered as events linked to the hashtag. The exact formula used and a vizualisation of it are shown below.

$$\text{event\_score}(d) = \frac{\sum_{i=-1}^{1} w_i * U_{d+i}}{\frac{1}{N}\sum_{j=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N}{2} \rfloor} U_{d+j}}, \qquad [w_{-1}, w_0, w_1] = [0.2, 0.8, 0.2]$$

U_d : unique users on day d



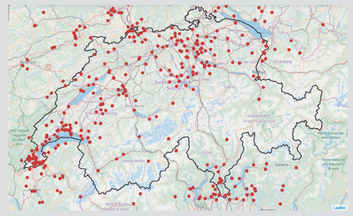
The event score for the two examples used previously are shown here:



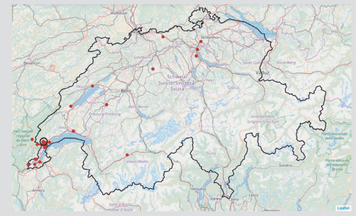
Event Localization

Now that we have a list of events, we want to determine the location of an event based on the location of the tweets that caused this event to be detected. To do so, we compute the median of the latitude and longitude of the tweets.

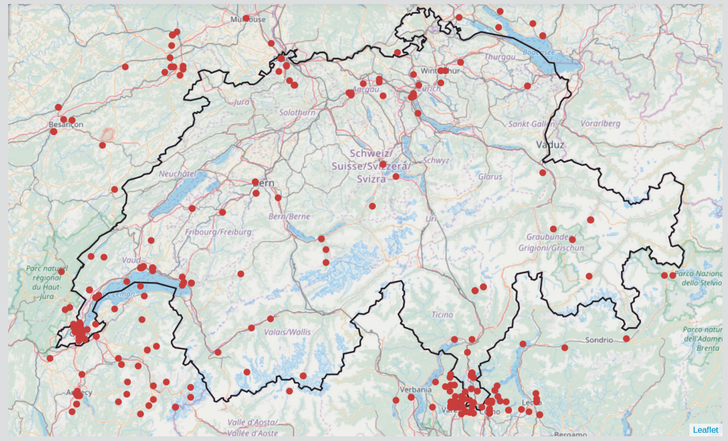
But for events that are not physically happening at some place, the location might not be relevent, as we can see in the following examples. Thus we compute the mean standard deviation to determine if the location we compute for an event is relevent. To illustrate this, two maps of tweets for a local and a global event are shown below.



Map of tweets for Eurovision:
We see that this is a global event.



Map of tweets for Paléo Festival:
We see that this is a local event.



Map of all detectd event locations

Results

Here are some statistics to quantify our results:

- Out of 6197 different hashtags, we found events for 2108 of them.
- By grouping by similar dates, we then found 7028 different events.
- We found a relevant location for 440 events out of 7028 .