Data Cleaning :

We had a dataset made of more than 20M of tweets, but only 3.4M of them were containing hashtags. As our idea was to detect events using the hashtags, we decided to keep only tweets with hashtags. So, a tweet was made of: a unique id, its user id, its longitude and latitude, and a non-empty list of hashtags. Not all the hashtags are interesting, and we can already filter them and remove those that either doesn’t appear enough frequently or doesn’t have enough unique author.

Data Analysis:

There are some obvious events we can use to see what our data looks like: Our example events are Christmas, Charlie Hebdo attacks and eurovision contest because these are different types of events: Christmas produce a huge interest for the whole december month with a peak on the D-Day, Charlie Hedbo attacks were not expected at all, and eurovision happens every year around the same date.

Event Detection:

We used an algorithm that considers the number of user that used a certain hashtag per day. We computed an “event score” which is the ratio of a hashtag’s use on a day compared to a period around this day. When we have a score for each day, we used a threshold to keep only days with high enough score that can be considered as an event linked to the hashtag.

Results:

Out of the 6197 hashtags we had, we found events for 2108 of them. Some hashtags were linked to multiple events, so we got a total of 7028 events and could locate 440 of them in Switzerland.