**Event detection on Swiss tweets**

This model detects past events based on a database of 28 million of tweets. This data set was only composed of Swiss tweets.

Twitter is an online news and social networking service. Users post and interact with messages, "tweets," restricted to 140 characters. They are using Twitter to report real-life events. It constructs a big base of data on events sometime located.

|  |  |
| --- | --- |
| The available data set was consisting of:   * 27.632.392 tweets * Each tweets have 20 features   A tweet is mostly composed of:   * a username * the text * the created time * location information   We worked with another team which implemented all the visualisation of the events while we were in charge of the event detection. | **What is an event?**  In this project context, we defined an event as an unusual outcome that occurs in a certain place during a particular interval of time.  We based our event detection on the assumption that if an event occurs, an unusual high number of users will tweet about this, i.e. post information on the context of the event and maybe its location.  Based on this hypothesis, we looked for peaks of tweet related to a certain field and state that it is an event.  We used the constraint of event over their duration to identify them. Time series were pretty useful to visualized peaks and then process them.  Figure 1 Election of François Hollande, the French president, the 6th of May |

|  |  |  |
| --- | --- | --- |
| **Our strategy**  We needed to infer three different information from out tweets.   1. The context: processing the tweet in order to extract the different topics 2. The location: where does the tweet occurs, where does the event it may represents happens. 3. If it is an event: if this tweet is part of an event or just a regular tweet     **Topic’s extraction**  What defines Twitter is the limited number of character and the wide use of hashtag. A hashtag is a type of label or metadata tag used on social network and microblogging services which makes it easier for users to find messages with a specific theme or content.  According to Manish and al.[[1]](#endnote-1) using only hashtag to deduce the subject of a tweet is a fair approximation. Indeed, applying natural language processing on tweet is a challenge as users often misspell their messages and the limited number of characters forces them to massively use abbreviation.    **Location**  Determining the location of an event asked the location of the tweets it was generated from.  We compute the median of the latitude and longitude of all tweet composing the event.  Gupta Manish, Gao Jing, Zhai ChengXiang and Han Jiawei,,”Predicting future popularity trend of events in microblogging platforms”, Proceedings of the American Society for Information Science and Technology, 2012 |  | **Event or not Event**  We studied the time series of each hashtag. Basically, we defined an event as an unusual peak in this time series.  Peaks were detected by comparing every day number of tweet to a threshold. If the number of tweet containing a specific hashtag for a specific day is greater than the threshold, this day is stamp as an event for this hashtag.  The threshold was compute as follow:  We determined during the exploration phase. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Duration of event**  Some event can last for several days. In this case, we grouped detected events for the same hashtag closed in time. It avoids to falsely duplicate events and inflate the number of detected events.  We stated that two detected events close to each other represent the same event and should be merged.  To automatized this decision, we implemented an algorithm using a time window.  If the time separating two events was smaller than this window, we merge the two events.  For these events grouped because they are close, we computed the weighted mean over their date to determine the representative day of these aggregated events. We merged them into only one event with the previous computed date.  Boumelala Sabrine  Guggenheim Daniel  Shynkaruk Sergii | **Spamming user**    The analysis of the detected events highlighted events created by only one user. On twitter some users try to create trend by overusing a hashtags related to their activities. This behaviour implies the detection of false event. To overpass this issue, event needed to be composed of tweet from at least two different users.  **Recurrent event**  Some peaks appear periodically. We arbitrarily decided that peaks appearing every months was not events whereas peaks appearing every year were events. #salary, appearing every month, is not event while other like #christmas is. So, we stated as false positive the detected events if they are recurrent with a frequency shorter than a parameter .  **Parameters**   |  |  | | --- | --- | | MIN\_TOT\_NB\_TWEETS = 20  MIN\_NB\_DAYS\_WITH\_HASHTAGS = 3  MIN\_NB\_TWEETS\_DURING\_EVENT = 5 | THRESHOLD PARAMETER = 2.5  TIME WINDOW = 10 | |  |

|  |  |  |
| --- | --- | --- |
| **Results**  More than 5000 events were detected.  We observed that with our model, the French tweets are correctly located in the Swiss Romandie and the Swiss German tweet in the German side of Switzerland, and the Italian tweets in the Italian part.  For example, we have in the center of Lugano.   * Name: #natale2014 * Date: 25 Dec 2014 * Number of tweets: 9     We can detect real events happening here in Switzerland or near Switzerland, like the white dinner in Basel. |  |  |

1. [↑](#endnote-ref-1)