Detecting Mobility Patterns and Events in Switzerland using Twitter

Stylianos Agapiou, Athanasios Giannakopoulos, Dimitrios Sarigiannis

{firstname.lastname}@epfl.ch

École Polytechnique Fédérale de Lausanne — EPFL, Switzerland

Abstract

Our analysis is based on a dataset containing tweets in Switzerland starting from 2010. First, we analyse the data and reconstruct mobility flows of the users by getting insights into high-frequency migration patterns in the Swiss territory. We focus on people who live and work in Switzerland as well as on "frontaliers" who e.g. commute daily from France and Germany to Geneva and Zurich respectively. Secondly, we detect events using (i) DBSCAN and (ii) a heuristic approach by focusing on dates and number of users in the same location. Finally, we perform sentiment analysis on the tweets related to the detected events.

Data Wrangling and Cleaning

We use Spark to clean and process the raw twitter dataset and create files for each year separately, that can be processed on a single machine. The data processing pipeline is depicted in Fig. 1.

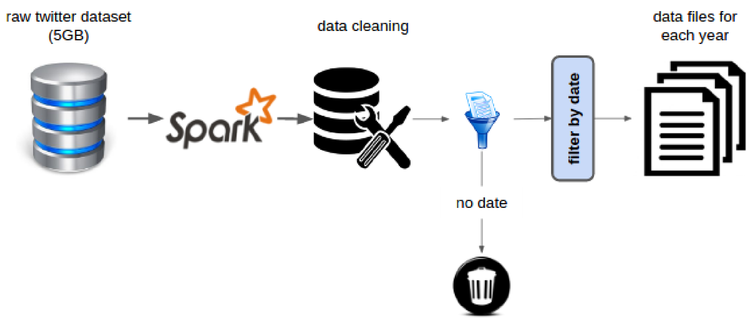


Figure 1: Pipeline for data processing and cleaning using Spark

In addition, the twitter dataset should be filtered based on the number of tweets of each user. This cleaning process removes both inactive users and users with hyperactivity who are able of biasing the results. According to [1], the twitter dataset has a *sweet spot* for analysis, depicted in Fig. 2.

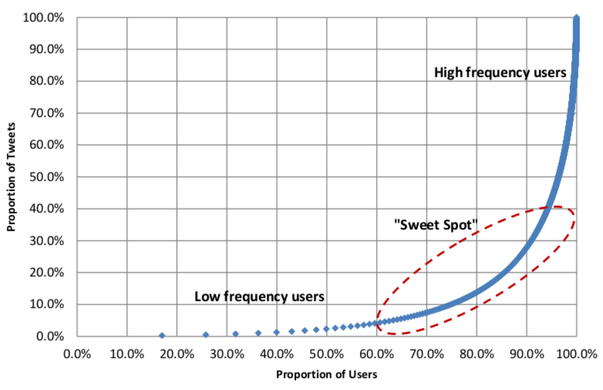


Figure 2: Sweet spot for twitter dataset

Detecting Mobility Patterns and Flows

Mobility patterns and flows should be studied while having a fixed point of reference for each user, i.e. the user's **home location**. Also, by determining the **workplace location**, we obtain useful information about the daily mobility patters of users. The pipeline for determining the aforementioned locations is depicted in Fig. 3.

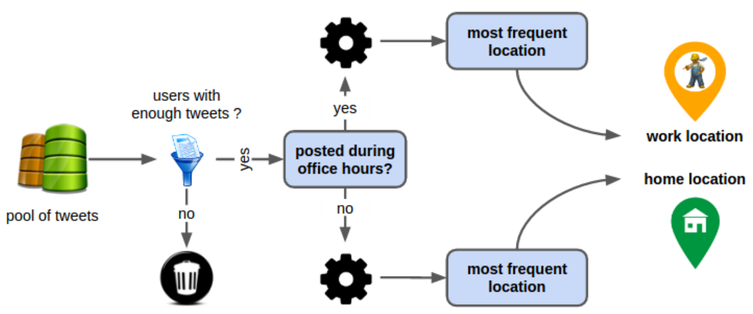


Figure 3: Pipeline for estimating home and workplace location

The GoogleMaps API is used to provide information about time and distance to work for each user, as well as canton of work and residence.

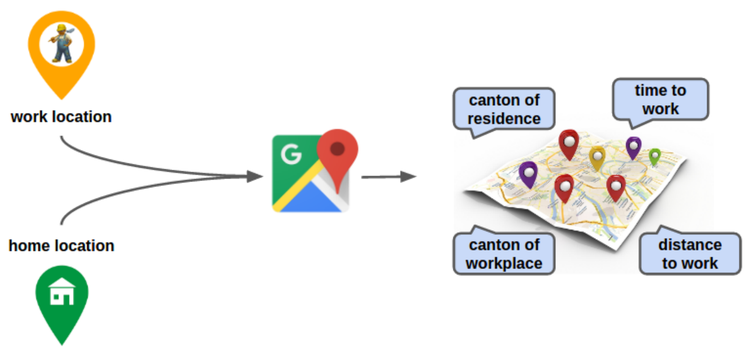


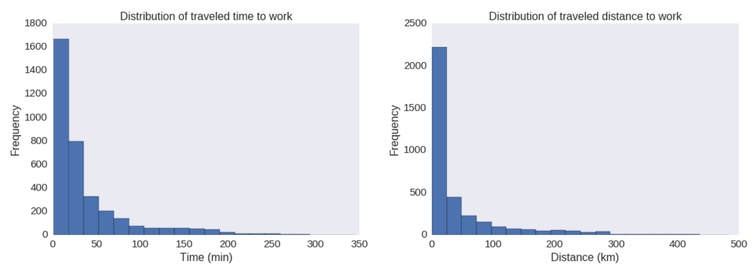
Figure 4: Extracting mobility information using the GoogleMaps API

Statistics for Mobility Patters and Flows

The gathered data are processed to extract information regarding:

- average time and distance to work
- location/canton of residence and workplace
- radius of gyration

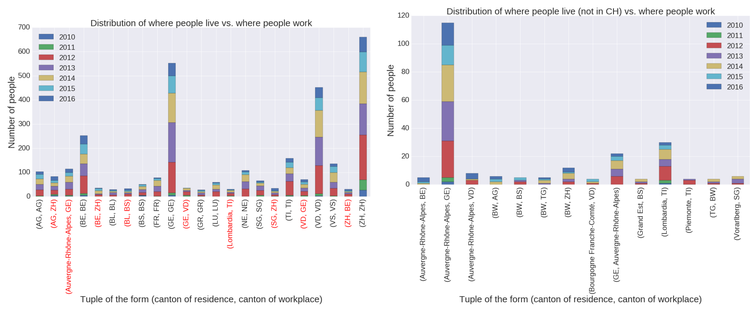
Average time and distance to work



(a) Distribution of average time to work **(b)** Distribution of average distance to work

Figure 5: Distributions of average time and distance to work

Where people live and where they work



(a) Where people live and where they work **(b)** People who work but do not live in Switzerland

Figure 6: Distribution of where people live and where people work

Radius of gyration

The radius of gyration is the standard deviation of distances between tweets and the user's likely home location and is estimated for each user [2]. It is given by

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_h)^2}$$

where

- n : the number of tweets
- r_i : the location of the tweet i
- r_h : the home location

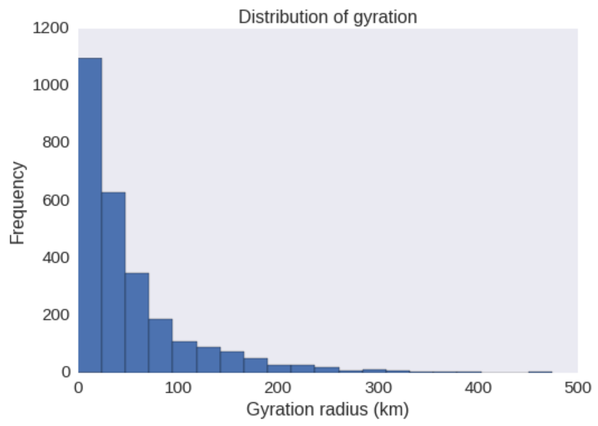


Figure 7: Distribution of the radius of gyration

Event Detection

We detect events using the tweet text and the geolocated information. The following assumptions are made:

- a tweet should contain at least one hashtag in order to possibly describe an event.
- events should happen during one day, i.e. events are detected as a unique pair of `(date, #hashtag)`.
- tweets about events should be posted in approximately the same location, assuming that events take place in a small area.

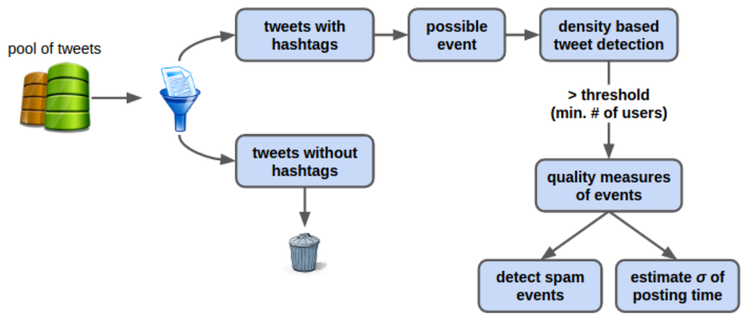


Figure 8: Event detection pipeline

In order to detect events, we use two algorithms:

- **DBSCAN** for density based clustering
- **heuristic algorithm** that forms square clusters. Here, the GPS accuracy is reduced to three decimal digits (accuracy of $80m$) [3].

Quality Measures of Detected Events

We provide two quality measures for the detected events:

- **users per event:** the higher the number of users per hashtag per date, the more likely a true event is detected.
- **standard deviation of posting time:** the smaller the standard deviation of the posting time of the tweets, the more likely a true event is detected.

Sentiment Analysis

Based on the detected events, a sentiment analysis is performed for each tweet. The results are aggregated based on (i) **area** of event and (ii) **event**, i.e. hashtag. The sentiment analysis pipeline is depicted in Fig. 9. The sampling is required in order to reduce the number of requests to the translation API.

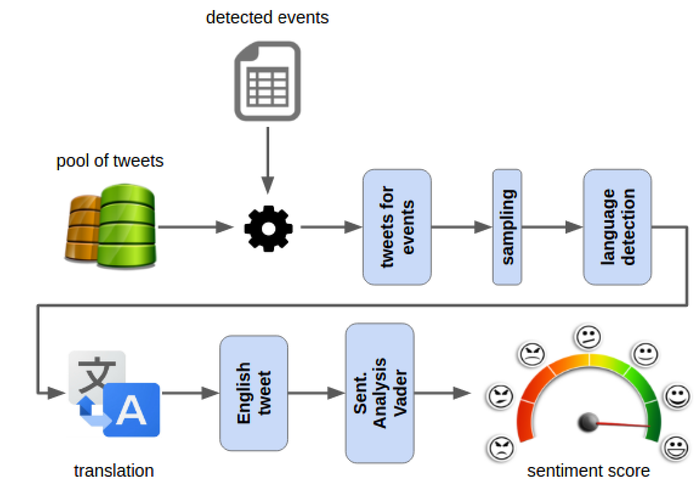


Figure 9: Sentiment analysis pipeline

Sentiment Analysis for Events

The sentiment analysis for the detected events in 2015 is depicted in Fig. 10.

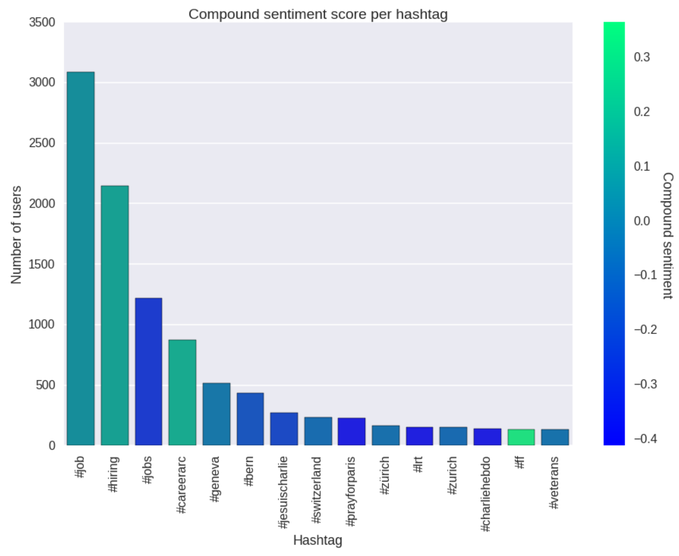


Figure 10: Sentiment analysis for detected events in 2015

Conclusion

We perform data analysis on the given twitter dataset. We conclude that such a dataset is useful in order to extract information regarding mobility patterns and statistics. Moreover, it is apt for studying event detection and sentiment analysis. However, a twitter dataset is also tricky, since most results are derived using heuristic approaches (e.g. a detected event may not be an actual event), and therefore should be cautiously interpreted.

Acknowledgements

We would like to thank M. Catasta for giving us the opportunity to work on this project in the context of the Applied Data Analysis course. Also, we would like to thank Swisscom AG for providing the dataset.

References

[1] Nigel Swier, Bence Komarniczky and Ben Clapperton. *Using geolocated Twitter traces to infer residence and mobility*. Office for National Statistics, GSS Methodology Series No 41, October 2015.

[2] Abdullah Kurkcu. *Evaluating the Usability of Geolocated Twitter as a Tool for Human Activity and Mobility Patterns: A Case Stude for NYC*. October 2015

[3] Wikipedia `https://en.wikipedia.org/wiki/Decimal_degrees`.