Switzerland’s Tweets sentiment analysis

C. Han & D. Rodrigues
ADA - CS401

Goals

- Analyze the sentiments expressed by each tweet
- Visualize its relation with respect to time and locations

Challenges

- Multiple languages
- Twitter specific vocabulary
- Assign a location to each tweet

Sentiment Analysis Difficulties

- Tweets contain unusual Unicode/ASCii characters
- Tweets contain Emojis and Emoticons
- Tweets contain Urls/hashtags and @user mentions
- Tweets are written in multiple languages.
- One tweet can be written multiple languages.

Sentiment Analysis Process

- Clean each Tweet, remove Url/hashtags and @username mentions
- Predefine Sad and Happy set of Emoji/Emoticons
- Classify Tweets based on Happy and Sad Emoji and Emoticon usage
- Tokenize the extracted tweets and remove and strip French/English and German stopwords and punctuation and single character words.
- Frequency analysis on two sets of data and subtract the frequency of the most popular words to remove conflicts.
- Build Liu_Hu lexicon with appropriate size.
- Count the words with sentiment score in the tweets in big dataset.
- Produce a total sentiment score of each tweet.

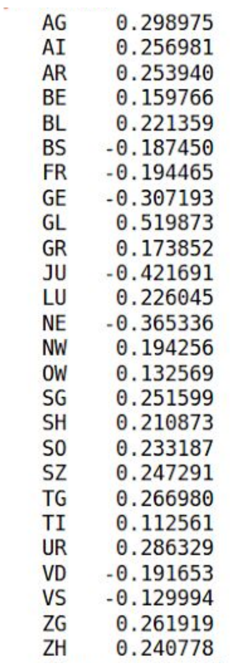
Visualization

- Each tweets has a pair of longitude/latitude
- Compute the closest district capital
- Faster than library, e.g *geopy.*

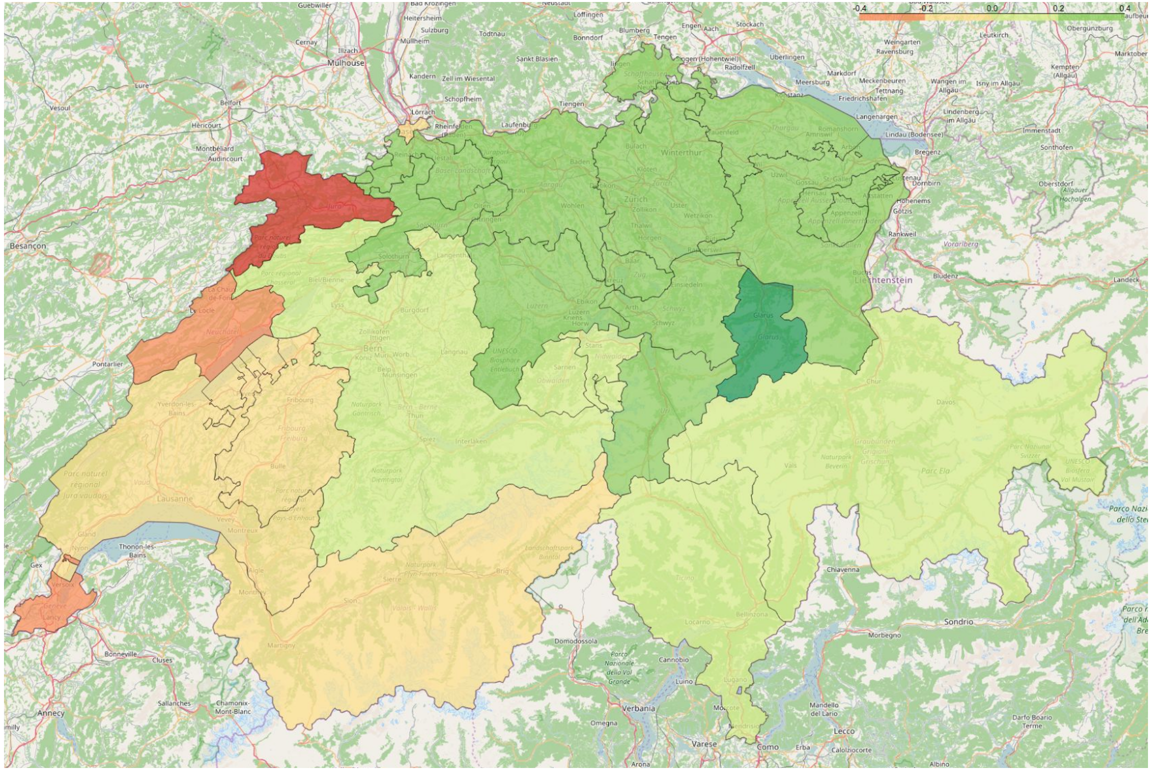
This method is not perfect: capitals are not in the middle of districts and districts aren't perfect cycles.

Tweets from other countries are removed based on the distance between a tweet location and its nearest district.

Results by Canton



```
AG     0.298975
AI     0.256981
AR     0.253940
BE     0.159766
BL     0.221359
BS    -0.187450
FR    -0.194465
GE    -0.307193
GL     0.519873
GR     0.173852
JU    -0.421691
LU     0.226045
NE    -0.365336
NW     0.194256
OW     0.132569
SG     0.251599
SH     0.210873
SO     0.233187
SZ     0.247291
TG     0.266980
TI     0.112561
UR     0.286329
VD    -0.191653
VS    -0.129994
ZG     0.261919
ZH     0.240778
```



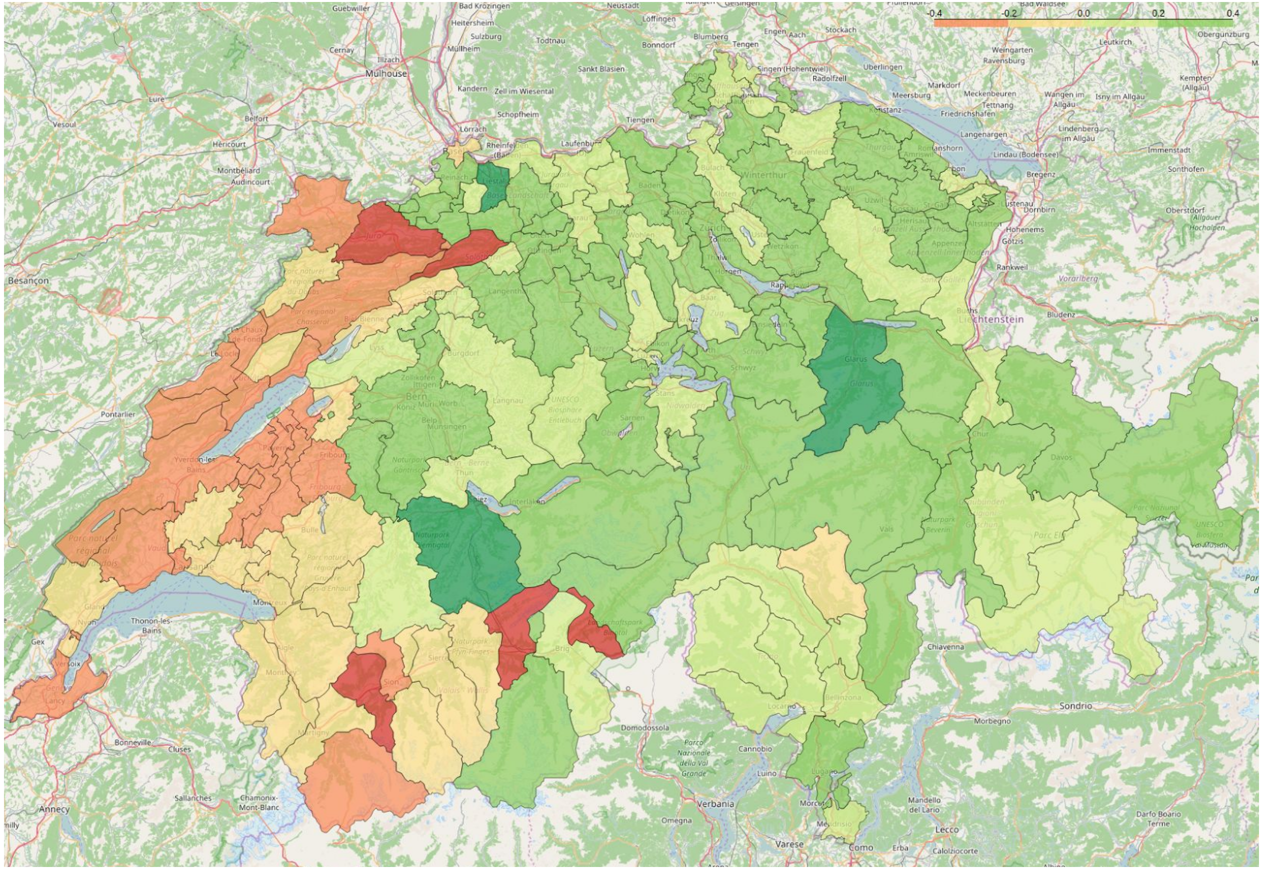
Röstigraben

- As we can see, it appears that the *French* part of Switzerland has more negative feelings that the *German* part.

- The *Italian* part is kind of in-between

- No canton express *extreme* feelings

Districts


-0.4
-0.2
0.0
0.2
0.4
Mulhouse
Besançon
Montbéliard
Belfort
Vesoul
Héricourt
Pontarlier
Lörrach
Tiengen
Winterthur
Konstanz
Friedrichshafen
Lindau (Bodensee)
Bregenz
Dornbirn
Liechtenstein
Bern
Burgdorf
Langnau
Thun
Interlaken
Glarus
Davos
Chur
Sion
Brig
Annecy
Chamonix-Mont-Blanc
Thonon-les-Bains
Domodossola
Verbania
Varese
Como
Lecco
Sondrio
Chiavenna
Morbegno
Locarno
Lugano
Bellinzona

Valais - Wallis

- Almost all district follow the same score as their cantons

- Valais is a bilingual canton, and we can see that the French-German canton have the same score difference as for the entire country.

- Fribourg, another bilingual canton, also has this separation

Analysis

- Do we score *french* tweets more negatively than *german* ones ?

- A more proper cleaning of the data would have been useful. Remove *robot* accounts for example.

- No differences between urban and rural areas.