DSLab: Topic detection in social media

Report I.

Bálint Hantos, supervisor: István Márkusz

The original plan was to gather historical data from Twitter from a given time interval mentioning a certain word. I had to register a developer account at Twitter and authorize my profile via phone. Then created a project and waited for approval from Twitter devs. At the same time I installed tweepy, which I used to make queries to the Twitter API using python.

Later on I had to face the fact, that simply I'm unable to gather tweets from the past in any meaningful quantity and quality. The standard plan is restricted to only 5k tweets/month with a limitation of 128 characters per tweet [1]. We agreed with my supervisor that it was not enough to tackle with the task at hand, as a substantial amount of this 5k tweets are retweets.

So, we had to come up with a new idea for the subject of the project. We decided to choose the primary elections in Hungary and the surrounding discussion about it on the social media platform Reddit. In order to do that I needed to gather data from the r/hungary subreddit. As in the case of Twitter Reddit also has an official API, which can be accessed through Python with PRAW (Python Reddit API Wrapper). PRAW also needs some authentication on the part of Reddit, so I needed to create an app at the Reddit dev page and generate a token.

I tried to write queries using this PRAW package, which became quite easy to use after a while. But sadly, I came to the conclusion that using this package I can only go back as far as 1000 posts (which is the average number of posts in a week on r/hungary). Moreover, time horizon for the posts couldn't be specified neither.

So, instead I looked for another method to go even further into the post history and found PSAW (Python Pushshift.io API Wrapper). PSAW comes handy when one wants to gather a massive amount of data from Reddit, as it is much more powerful than PRAW. Using PSAW I could finally gather posts from even the beginning of September 2021, and then I had to make queries for the comments using the post IDs.

The current state of the project is the following. We had to find a new subject for the project, I walked around a bunch of caveats of API limitations and gathered the necessary data for further analysis.

[1] - https://developer.twitter.com/en/docs/twitter-api/premium/search-api/overview