Scraping Reddit data

I used Pushshift API in combination with pmaw to extract the raw dataset from Reddit. Pushshift API is a RESTful API which provides features for data scraping according to specific criteria, i.e. it allows to specify the subreddit and a period of time. Pmaw is a wrapper for Pushshift API and provides a convenient way to retrieve all submissions which correspond to the defined criteria. Pmaw is a wrapper which makes multiple requests to the Pushshift server until all of the , respecting a rate limit of 60 requests per minute.

rate limits the requests to the Pushshift server.

can be used to search reddit data according to specific criteria and create data aggregations. Pmaw is a wrapper for Pushshift API which provides a convenient way to retrieve large datasets by using multithreading and intelligent rate limiting.

<https://medium.com/swlh/how-to-scrape-large-amounts-of-reddit-data-using-pushshift-1d33bde9286#:~:text=Under%20the%20hood%2C%20pmaw%20makes,of%2060%20requests%20per%20minute>.

<https://www.reddit.com/dev/api/>

Data Scraping

I used Pushshift API in combination with a wrapper called PMAW to retrieve the raw dataset. Pushshift API is a RESTful API and a powerful tool in providing search capabilities for reddit submissions and comments, i.e., it allows the specification of a subreddit and a period of time. Pmaw makes multiple requests to the Pushshift server until all data corresponding to the defined criteria is retrieved. It uses intelligent rate limiting in order to decrease the amount of rejected requests. It is recommended to not make more than 60 requests per minute. I scraped both the r/conspiracy and r/ukraine subreddit for all submissions made between the 23rd of February and the 3rd of March 2022. This resulted in 7806 submissions in r/conspiracy and 36737 submissions in r/ukraine. However, some of the data in the Pushshift API is not up to date. Submissions and comments are scraped once shortly after they are posted. There has been no attempt to re-ingest data since August 2021. Therefore, information about submissions such as the number of comments or the score are not accurate. Aiming to retrieve more precise information, I extracted the unique ids of the submissions and used them as search parameters for PRAW. PRAW is a wrapper for the official Reddit API. I retrieved recent information about score, comments and deletion of submissions scraping Reddit API on the 28th of May 2022.