Data Collection System Proposal

Omer Feyzullahoglu
Binghamton University
Binghamton, New York, USA
ofeyzul1@binghamton.edu

ABSTRACT
In this paper, I describe the road map that I will follow while developing my data collection pipelining system.

Keywords
Proposal, data collection, Twitter API, Reddit API, python, requests.

1. DESCRIPTION OF SOURCES

First data source will be Twitter. Twitter is a microblog and social network platform that allows users post 240 characters-limited tweets. [3] In the first years of the platform, users could only tweet, but now Twitter has many features. Users can now also share photos and videos. Even after the application called Clubhouse was released, Twitter took the voice room feature of Clubhouse and added as a new feature, called Twitter Spaces to the platform. Thanks to the new feature, users can create rooms and interact with each other and followers by voice. Since Twitter has a lot of users, millions of data are produced every day, so Twitter's public data is one of the most used novel data in data science projects. On the other hand, it is much easier to follow trends with the Twitter's hashtag feature.

Second source will be Reddit. Reddit is a social news community which includes content rating, discussions etc. It is like a huge forum; everything has their own fan page. Ideal for finding own tribe. Users can post their contents as a picture, gif, links, and text. There is a strict organization mechanism which is user created boards called subreddits. Reddit has millions of users from many countries, mainly in the USA, which means that it is a huge data pool.

2. INTENDED COLLECTION METHODS

To collect public data from Twitter, Twitter provides Twitter API to developers so I intent to collect data with the help of Twitter API v2. Since I need to collect continuously, I will use sampled-stream method of the API that delivers roughly %1 random sample in real time and allows 50 connection per 15 minutes window. Since the streaming endpoint delivers Tweet objects through HTTP GET connection, I will send GET requests with my OAuth 2.0 Bearer Token using python requests library then I will store delivered Tweet objects in MongoDB for preliminary examination.

My intent to collect data from Reddit is using Reddit API. Reddit provides many endpoints to their users. Since I need a streaming service not a single snapshot at a time. I will collect all new posts. Unlike twitter, Reddit does not provide streaming, so I must develop my own stream-like method. According to API to documentation, limit is the 100 posts at time and if you manipulate the query little bit, API works up to 1000 posts. After 1000 posts it starts to show posts that you have already seen. Every json response has before and after fields that keeps track of post id's so If I arrange the queries sequentially according to the ids, I can pull all the new posts without missing any posts.[1] Then, I will store it MongoDB for preliminary examination

3. MEASUREMENT AND ANALYSIS IDEAS

I am planning to pull just "subreddit, title, self-text, permalink, URL" field from Reddit and "tweet, media, user" fields from the Twitter, since pulling everything can cause storage problems. I think I can analyze most frequently used emojis according to text's context annotations, create word cloud or emoji cloud according to couple of key terms that is from the post's text. Or a sentiment analysis of posts and it's like count.

4. ESTIMATION OF HOW MUCH DATA WILL BE COLLECTED

According to Reddit 2020 data, 303.4 million posts were created in 2020. [2] This makes 830,000 posts on average. So, I can say that each week I will collect around 5 million posts. On the other hand, as of May 2020 users have shared 6000 tweets on average every second.[4] Since, sampled stream of Twitter provides %1 of all twitter. This makes 60 tweets per second. So, I can say that each week I will collect around 36 million tweets.

5. REFERENCES

[1] Lewis Van Winkle, 2018. How F5Bot Slurps All of Reddit.(July 2018).Retrieved October 3, 2021, from https://intoli.com/blog/f5bot/

[2] Reddit, 2020, Reddit :Dive into Anything. Retrieved from https://www.reddit.com/r/blog/comments/k967mm/reddit_in_2020/

[3] [3] Wikipedia. 2017. WikipediA: the Free Encyclopedia. Retrieved from https://www.wikipedia.org/.

[4] David Sayce,2020. The Number of tweets per day in 2020. Retrieved October 3, 2021, from https://www.dsayce.com/social-media/tweets-day/