Build a Web or Social Network Search Engine

Part A: A Reddit Post Collector

Team 18 Members:
Vishal Gondi
Venkata Sai Vineeth Gudela
Alan Ngo
Rithvik Vukka
Rajeswari Pedaballi

1. Collaboration Details

Each team member played an equal role in writing the Python code and in the data collection process. The team consists of five members, and each member was responsible for fetching 100MB of data, to get a total of 500MB.

2. Overview of system

a. Architecture:

The system's architecture is designed around a Python script (*reddit_crawler.py*) which is controlled through a Bash shell script (*crawler.sh*). The Python script is powered by the PRAW library (Python Reddit API Wrapper), which interacts with Reddit's API in a structured manner. This allows for easy adjustments and scalability. The script supports multithreading, which enhances the efficiency of data fetching by processing multiple threads in parallel. Rate-limiting is implemented using semaphores to manage API calls and prevent exceeding Reddit's API usage limits, ensuring the crawler operates within the terms of service.

b. Crawling or Data Collection Strategy:

The data collection process begins by parsing command-line arguments to determine the target subreddits, the number of posts to fetch per subreddit, and the depth of comment threads to explore. The script fetches top posts from the specified subreddits using the Reddit API and then iteratively explores the comments of each post up to the specified depth. Each post and its comments are processed in real-time, and the data is structured to preserve the context and threading of discussions.

c. Methodology:

The subreddits we want to scrape, the post and comment limits, are passed as parameters to shell script. We created two functions, *crawl_multiple_subreddits* to get the data from multiple subreddits simultaneously and *crawl_subreddit* function to get the data from a single subreddit. Firstly we get the ids of all the posts of that subreddit and have stored it in *submission_ids*. We loop through each of the post ids and get the post

details using the *process_submission* function. In this function *process_submission*, we get the *'selftext'*, *'title'*, *'id'*, *'score'*, *'permalink'* and *'comments'* till the comment depth which was passed as a parameter to the shell script for each post. Once the collection is done for all the post ids in the *submission_ids* list, we dump the data into a json file.

d. Data Structures Employed:

The script uses data structures for efficient data management. A set is used to keep track of already processed IDs, ensuring no duplicate data processing. A Semaphore controls the concurrency level of network requests, to maintain a balance between speed and compliance with API rate limits. Data is stored in dictionaries, which allows for dynamic data access and manipulation before being saved to JSON files to efficiently handle large data while preventing data loss during the collection process.

3. Limitations

a. As each member independently collects a portion of the data, there is no built-in mechanism to ensure the uniqueness of the data across different machines.

b. Our data collection process is vulnerable to disruptions caused by network issues or unexpected changes to API limitations which made us re-run the script multiple times.

4. Instructions on Deploying the System

a. Preparation:

- Install python3, dependent packages and get the Reddit API, secret for the PRAW module from here
- Place both the *reddit_crawler.py* and *crawler.sh* scripts in the same directory on your machine.

b. Configuration:

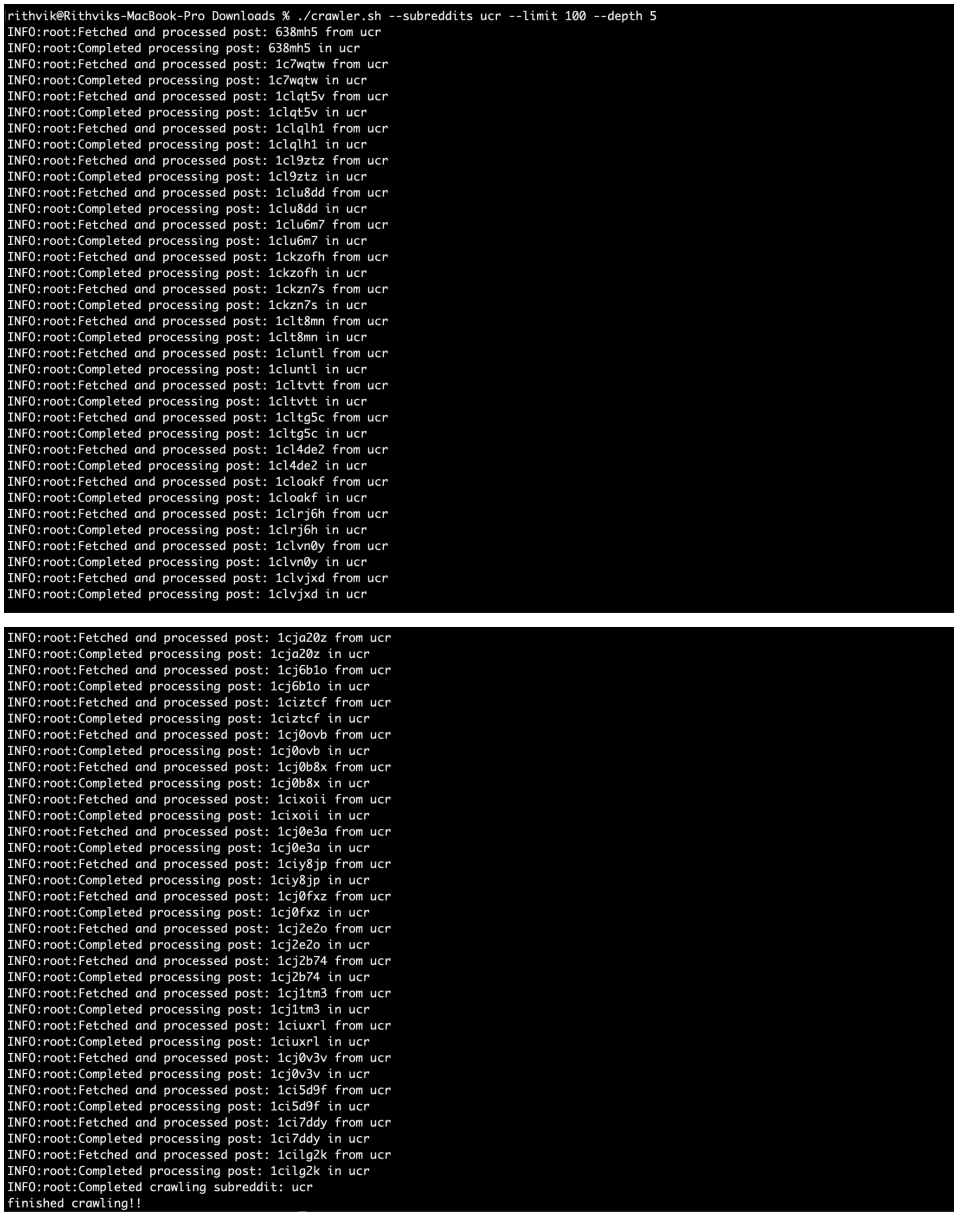
- Add the Reddit API configuration details at the specified part in the *reddit_crawler.py* script.
- Modify file permissions to make the bash script executable. This can be done by: *chmod +x crawler.sh*.

c. Execution:

- To run the crawler, *./crawler.sh --subreddits <subreddit1> <subreddit2> --limit <number> --depth <depth>*, substituting *<subreddit1>, <subreddit2>, <number>,* and *<depth>* with appropriate values where the number is number of posts to fetch and depth is level of comments need to be fetched in a single post.

The script will start fetching data according to the specified parameters and will save it in JSON format in the same directory.

5. **Screenshots**
    a. When one argument is used for subreddits



b. When more than one argument is used for subreddits. Here the execution happens in parallel for both the subreddits /ucr and /books.