

# 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

```
rithvik@Rithviks-MacBook-Pro Downloads % ./crawler.sh --subreddits ucr --limit 100 --depth 5
INFO:root:Fetchd and processed post: 638mh5 from ucr
INFO:root:Completed processing post: 638mh5 in ucr
INFO:root:Fetchd and processed post: 1c7wqtw from ucr
INFO:root:Completed processing post: 1c7wqtw in ucr
INFO:root:Fetchd and processed post: 1clqt5v from ucr
INFO:root:Completed processing post: 1clqt5v in ucr
INFO:root:Fetchd and processed post: 1clqlh1 from ucr
INFO:root:Completed processing post: 1clqlh1 in ucr
INFO:root:Fetchd and processed post: 1cl9ztz from ucr
INFO:root:Completed processing post: 1cl9ztz in ucr
INFO:root:Fetchd and processed post: 1clu8dd from ucr
INFO:root:Completed processing post: 1clu8dd in ucr
INFO:root:Fetchd and processed post: 1clu6m7 from ucr
INFO:root:Completed processing post: 1clu6m7 in ucr
INFO:root:Fetchd and processed post: 1ckzofh from ucr
INFO:root:Completed processing post: 1ckzofh in ucr
INFO:root:Fetchd and processed post: 1ckzn7s from ucr
INFO:root:Completed processing post: 1ckzn7s in ucr
INFO:root:Fetchd and processed post: 1clt8mn from ucr
INFO:root:Completed processing post: 1clt8mn in ucr
INFO:root:Fetchd and processed post: 1cluntl from ucr
INFO:root:Completed processing post: 1cluntl in ucr
INFO:root:Fetchd and processed post: 1clvtvt from ucr
INFO:root:Completed processing post: 1clvtvt in ucr
INFO:root:Fetchd and processed post: 1cltg5c from ucr
INFO:root:Completed processing post: 1cltg5c in ucr
INFO:root:Fetchd and processed post: 1cl4de2 from ucr
INFO:root:Completed processing post: 1cl4de2 in ucr
INFO:root:Fetchd and processed post: 1cloakf from ucr
INFO:root:Completed processing post: 1cloakf in ucr
INFO:root:Fetchd and processed post: 1clrj6h from ucr
INFO:root:Completed processing post: 1clrj6h in ucr
INFO:root:Fetchd and processed post: 1clvn0y from ucr
INFO:root:Completed processing post: 1clvn0y in ucr
INFO:root:Fetchd and processed post: 1clvjxd from ucr
INFO:root:Completed processing post: 1clvjxd in ucr
```

```
INFO:root:Fetchd and processed post: 1cja20z from ucr
INFO:root:Completed processing post: 1cja20z in ucr
INFO:root:Fetchd and processed post: 1cj6b1o from ucr
INFO:root:Completed processing post: 1cj6b1o in ucr
INFO:root:Fetchd and processed post: 1ciztcf from ucr
INFO:root:Completed processing post: 1ciztcf in ucr
INFO:root:Fetchd and processed post: 1cj0ovb from ucr
INFO:root:Completed processing post: 1cj0ovb in ucr
INFO:root:Fetchd and processed post: 1cj0b8x from ucr
INFO:root:Completed processing post: 1cj0b8x in ucr
INFO:root:Fetchd and processed post: 1cixoi from ucr
INFO:root:Completed processing post: 1cixoi in ucr
INFO:root:Fetchd and processed post: 1cj0e3a from ucr
INFO:root:Completed processing post: 1cj0e3a in ucr
INFO:root:Fetchd and processed post: 1ciy8jp from ucr
INFO:root:Completed processing post: 1ciy8jp in ucr
INFO:root:Fetchd and processed post: 1cj0fxz from ucr
INFO:root:Completed processing post: 1cj0fxz in ucr
INFO:root:Fetchd and processed post: 1cj2e2o from ucr
INFO:root:Completed processing post: 1cj2e2o in ucr
INFO:root:Fetchd and processed post: 1cj2b74 from ucr
INFO:root:Completed processing post: 1cj2b74 in ucr
INFO:root:Fetchd and processed post: 1cj1tm3 from ucr
INFO:root:Completed processing post: 1cj1tm3 in ucr
INFO:root:Fetchd and processed post: 1ciuxrl from ucr
INFO:root:Completed processing post: 1ciuxrl in ucr
INFO:root:Fetchd and processed post: 1cj0v3v from ucr
INFO:root:Completed processing post: 1cj0v3v in ucr
INFO:root:Fetchd and processed post: 1ci5d9f from ucr
INFO:root:Completed processing post: 1ci5d9f in ucr
INFO:root:Fetchd and processed post: 1ci7ddy from ucr
INFO:root:Completed processing post: 1ci7ddy in ucr
INFO:root:Fetchd and processed post: 1cilg2k from ucr
INFO:root:Completed processing post: 1cilg2k in ucr
INFO:root:Completed crawling subreddit: ucr
finished crawling!!
```

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

```
rithvik@Rithviks-MacBook-Pro files_cs172 % ./crawler.sh --subreddits ucr books --limit 15 --depth 3
INFO:root:Fetchd and processed post: 638mh5 from ucr
INFO:root:Completed processing post: 638mh5 in ucr
INFO:root:Fetchd and processed post: 1c7wqtw from ucr
INFO:root:Completed processing post: 1c7wqtw in ucr
INFO:root:Fetchd and processed post: 1clwkke from ucr
INFO:root:Completed processing post: 1clwkke in ucr
INFO:root:Fetchd and processed post: 1clqt5v from ucr
INFO:root:Completed processing post: 1clqt5v in ucr
INFO:root:Fetchd and processed post: 1clu8dd from ucr
INFO:root:Completed processing post: 1clu8dd in ucr
INFO:root:Fetchd and processed post: 1clqlh1 from ucr
INFO:root:Completed processing post: 1clqlh1 in ucr
INFO:root:Fetchd and processed post: 1clu6m7 from ucr
INFO:root:Completed processing post: 1clu6m7 in ucr
INFO:root:Fetchd and processed post: 1clzcg4 from ucr
INFO:root:Completed processing post: 1clzcg4 in ucr
INFO:root:Fetchd and processed post: 1cl9ztz from ucr
INFO:root:Completed processing post: 1cl9ztz in ucr
INFO:root:Fetchd and processed post: 1cluntl from ucr
INFO:root:Completed processing post: 1cluntl in ucr
INFO:root:Fetchd and processed post: 1cm0Sk4 from ucr
INFO:root:Completed processing post: 1cm0Sk4 in ucr
INFO:root:Fetchd and processed post: 1cm0gej from ucr
INFO:root:Completed processing post: 1cm0gej in ucr
INFO:root:Fetchd and processed post: 1cm00lp from ucr
INFO:root:Completed processing post: 1cm00lp in ucr
INFO:root:Fetchd and processed post: 1clzyep from ucr
INFO:root:Completed processing post: 1clzyep in ucr
INFO:root:Fetchd and processed post: 1clt8mn from ucr
INFO:root:Completed processing post: 1clt8mn in ucr
INFO:root:Completed crawling subreddit: ucr
INFO:root:Fetchd and processed post: 1clfah9 from books
INFO:root:Completed processing post: 1clfah9 in books
INFO:root:Fetchd and processed post: 1cj4v1h from books
INFO:root:Completed processing post: 1cj4v1h in books
INFO:root:Fetchd and processed post: 1clmia2 from books
```

```
INFO:root:Completed processing post: 1clzyep in ucr
INFO:root:Fetchd and processed post: 1clt8mn from ucr
INFO:root:Completed processing post: 1clt8mn in ucr
INFO:root:Completed crawling subreddit: ucr
INFO:root:Fetchd and processed post: 1clfah9 from books
INFO:root:Completed processing post: 1clfah9 in books
INFO:root:Fetchd and processed post: 1cj4v1h from books
INFO:root:Completed processing post: 1cj4v1h in books
INFO:root:Fetchd and processed post: 1clmia2 from books
INFO:root:Completed processing post: 1clmia2 in books
INFO:root:Fetchd and processed post: 1clm19h from books
INFO:root:Completed processing post: 1clm19h in books
INFO:root:Fetchd and processed post: 1clm70t from books
INFO:root:Completed processing post: 1clm70t in books
INFO:root:Fetchd and processed post: 1clsm12 from books
INFO:root:Completed processing post: 1clsm12 in books
INFO:root:Fetchd and processed post: 1clg522 from books
INFO:root:Completed processing post: 1clg522 in books
INFO:root:Fetchd and processed post: 1ckyj10 from books
INFO:root:Completed processing post: 1ckyj10 in books
INFO:root:Fetchd and processed post: 1cllu1q from books
INFO:root:Completed processing post: 1cllu1q in books
INFO:root:Fetchd and processed post: 1clxdpq from books
INFO:root:Completed processing post: 1clxdpq in books
INFO:root:Fetchd and processed post: 1clm1hf1 from books
INFO:root:Completed processing post: 1clm1hf1 in books
INFO:root:Fetchd and processed post: 1cl047 from books
INFO:root:Completed processing post: 1cl047 in books
INFO:root:Fetchd and processed post: 1clqanq from books
INFO:root:Completed processing post: 1clqanq in books
INFO:root:Fetchd and processed post: 1clm1men from books
INFO:root:Completed processing post: 1clm1men in books
INFO:root:Fetchd and processed post: 1cl190x from books
INFO:root:Completed processing post: 1cl190x in books
INFO:root:Completed crawling subreddit: books
finished crawling!!
rithvik@Rithviks-MacBook-Pro files_cs172 %
```