Ben Carroll

Jordan Dahmen

05 / 04 / 15

CS 475

A Concurrent Scraper

for Reddit

User and Comment Statistics

Table of Contents

2 : Table of contents

3 : Scraper description

3 - 4: Program design and structure

4 : Possible future additions

5 : Test design

5 : Analysis of results

6 – n: Code listings

Scraper Description

Our program concurrently scrapes and concurrently analyzes Reddit user and user comments to generate interesting metrics. This is accomplished by multi-threaded scraping of all posts made by a specified user or list of users, which are then returned to a series of multi-threaded user metrics. Currently, the program implements post frequency and word frequency analysis, but could be easily extended to implement other user metrics. In writing this program, we are trying to provide information that cannot be gleaned simply by viewing the user or users’ Reddit homepage.

Program Design and Structure

Our program was designed to be extendable, and currently consists of a **MainMenu**, where a user can enter Reddit user names or a file containing a list of Reddit usernames. This then invokes the **ScraperManager,** which handles much of the concurrency, through the use of the Java ExecutorService framework and the Java Future interface. The ExecutorService handles thread creation, while the Future interface allows for the asynchronous return of objects. The **ScraperManager** creates a fixed number of Scraper threads, which process up to that number of Reddit users concurrently, re-using threads if necessary for additional users. Each **Scraper** loads and parses the user’s comments page by page through the jsoup (HTML parsing) library and inserts them into a list of **Comments**, which is then returned to the Future in the **ScraperManager**.

Our program uses **Comment, Post, Sub, User, and Word** objects to store data extracted from Reddit pages, while it is in different stages of processing. A **Comment** objectstores a single comment by a single user, while a **Post** object is a comment that begins a discussion thread. The **Sub** and **Word** objects contain subreddits and individual words, respectively. All of these are encapsulated by the **User** object, which contains a user’s Reddit content and later user / comment metrics.

From here, the **ScraperManager** creates a dynamically allocated thread pool to analyze user comments. A second such pool is then created to analyze the user.

* Expand on this point
  + AnalyzeComment
  + AnalyzeUser
* Any other design decisions?

Possible Future Additions

Our program could be extended to check for “gilded” Reddit comments. Originally, we had planned to, but this would require an additional set of scrapers for each user. It would also be reasonable to make a GUI interface to replace the current CLI interface. It could also be extended to calculate readability metrics for a user.

In terms of additional metrics, there are many possibilities. Our program could be extended to calculate readability metrics for a user. Readability metrics rank text based on how difficult it is to read, and provide a guess as to the user’s writing or education level. Finally, our program could be extended to calculate these same metrics on a subreddit basis in addition to the current user basis.

Test Design

* Junit test cases?

Our program has been tested extensively on real Reddit users (mostly picked at random). As the program is now, we have not seen any errors from our code, although it is possible that there are edge cases that might occur when certain input is parsed or analyzed, that we do not know about.

Server-side errors are somewhat common, especially when run on the GMU network. HTTP 429 (Too Many Requests) errors occur frequently when Reddit is rate-limiting requests. Reddit appears to rate-limit based on IP address; therefore, all people accessing Reddit from the GMU network appear as one IP address or a pool of IP addresses and are treated as a single “entity.” This error is mostly gone if run from a private network, or better still, a VPS with a dedicated IP address. When Reddit is under peak load, it is still possible to see HTTP 503 (Service unavailable) errors occur. Reddit does not appear to have the server-capacity to meet peak demand.

Analysis of Results

* Results of running it on actual user’s Reddit homepages
* What was implemented successfully and what was not?
* Anything else?

Code Listing

The code for ScraperManager.java has been included below. It contains much of the concurrent code within our program. The code for the other java files that makeup our program can be viewed in the included zip file, or in our GitHub repository at:

<https://github.com/jdahmen/slash-u-scraper>

* Make this public
* Insert cleaned up ScraperManager.java code