Reddit Scraper, V 1.0

March 25th, 2019

Moutasem Zakkar

About Reddit Scraper

I have created this application to scrape the content of specific health-related subreddits, which I have used in the research of my Ph.D dissertation in Health Informatics at the University of Waterloo, Ontario, Canada.

The program saves the data on two media: a Cassandra Database, and the file system (in the form of Json files, where each file corresponds to a specific sub-reddit).

Licensing Prerequisites

This program uses the "https://pushshift.io/" REST API to capture the Reddit Data. Therefore, it is your responsibility as the user of this program to obtain any necessary licenses from "pushshift" and from Reddit.

JAVA Prerequisites

1. Java SDK or JDK (1.8 or later)
2. Apache HttpClient 4.5.6
3. GSon 2.8
4. Cassandra Driver from DataStax
5. Apache Commons-text 1.6

Preparing the Program Execution Environment

1. Setup your Cassandra DB and create the required table(s) as described in the DB_Schema.txt
2. Create A folder for the Json files. For example, let's assume that you name this folder as "RedditData_Store".
3. Inside the RedditData_Store, create a sub-folder called "Metadata"
4. Assuming that your RedditData_Store is on D: drive, you now have
    a. D:\RedditData_Store
    b. D:\RedditData_Store\Metadata

5. Inside the subfolder D:\RedditData_Store\Metadata, create a text file called "List_subreddits.txt"

The format of "List_subreddits.txt"

In this file, you will put the list of subreddits that you want to download.

The first line of the file is the header.

Starting from the second line, each line contains the name of the subreddit (without the r/), followed by "0,0" , as illustrated in the following table:

```
#Name of the subreedit, Latest reading date (TimeStamp), Total posts saved so far, Latest reading date (Normal Format)
environment,0,0
news,0,0
politics,0,0
todayilearned,0,0
```

How the program works

1. The Application controller class is "AppController.java"
2. In the "main" method of "AppController.java" , you need to provide the full path to your RedditData_Store: "D:\\RedditData_Store\\"
3. Prepare your "List_subreddits.txt" file. You can always add subreddits to it.
4. When you run the program, it will fetch a fixed amount of submissions in each subreddit. Therefore, you need to run the program several times. The rationale for this is that submissions are endless, and they are added daily, and, therefore, you will never be able to download all the submissions. What the program does, is that, in each run, it try to download all submissions submitted from the beginning of the platform (I think it started in 2008) till the end of the previous day of the day you are executing the program in. For each subreddit, the program will read 1000 posts on each run. You can change this by adjusting the variables _LoadingSize and AttemptsPerSubereddit.