Assignment 1 – Reddit Crawler

Description

Reddit (<https://www.reddit.com/>) is an online social media site that allows users to discuss any topic. Some of these discussions are available and valuable to researchers and developers. In this assignment, you will learn to build a crawler that collects posts information from Reddit given query (e.g., coronavirus, mask, WHO, etc.). Some of the posts have no contents or only an image. You only need to extract the plain text from posts and comments. It is possible to get empty results.

You can use any keywords as search query. Return the top 3 posts and 3 comments for each post as the results and write out to a xml file, name it ***result.xml***. The format should be the same as the sample file (***sample.xml***).

There are many useful python libraries for web crawling. ***Selenium*** is a powerful tool to automate browsers and easy to debug. We recommend it to you.

Requirement

* Python programming language only
* Allow to use any libraries or tools

Grading

* Completed source code (50%)
* Result file ***result.xml*** (40%)
* Report (10%)

Submission Rule

Please pack up your **source code**, **result file** and **report** into a

.zip file named as *studentid\_hw1.zip*, and upload it to LMS. The following files must be included in your submission:

* Python source code (one or more).
* Result file ***result.xml***
* requirements.txt: installed libraries in your Python environment. To create a requirements file, use:

*pip freeze > requirements.txt*

* report.pdf: assignment report. You can write anything you want.

Deadline: 2020/04/03