Lab 7 Specification – Exploring Web Scraping
Due (via your git repo) no later than 2 p.m., Monday, 9th Nov 2020.
50 points

Lab Goals

- Implement automation procedures using Web Scraping.

Summary

We will do a few hands-on exercises to automate web scraping procedures. We had started discussing the techniques, by which we can collect data from the web. It is the right time to understand how the web scraping works practically, and doing the basic Jsoup coding to achieve this goal. Additionally, we will watch a video clip, which is a segment of a video series in ParseHub.com.

Learning Assignment

If not done so already, please read all of the relevant ”GitHub Guides”, available at https://guides.github.com/, which explains how to use many of the features that GitHub provides. In particular, please make sure that you have read guides such as ”Mastering Markdown” and ”Documenting Your Projects on GitHub”; each of them will help you to understand how to use both GitHub and GitHub Classroom. To do well on this assignment, you should also read

- **HTML and Jsoup reading material in the class (lesson-5) repository.**

Assignment Details

It is required for all students to follow the honor code. Some important points from the class honor code are outlined below for your reference:

1. Students are not allowed to share code files and/or other implementation details. It is acceptable to have a healthy discussion with your peers. However, this discussion should be limited to sharing ideas only.
2. Submitting a copy of the other’s program(s) is strictly not allowed. Please note that all work done during laboratory sessions will be an opportunity for students to learn, practice, and master the materials taught in this course. By doing the work individually, students maximize the learning and increase the chances to do well in other assessments such as lab assignments, skill tests, projects, etc.

At any duration during and/or after the lab session, students are recommended to team up with the Professor and/or the Technical Leader(s) to clarify if there is any confusion related to the items in the lab sheet and/or class materials.

Section 1: Web Scraping

This section is worth 15 points. The points breakdown is provided below:

- Task 1 = 15 points, a maximum of 3 points awarded for each question.

In this section, we will watch a video and reflect on the points discussed in the clip. This reflection is instrumental to further advance our understanding of web scraping in general to foster the learning from our recent discussions on dataset collection. To complete this part, it is required to do the following:

- **Task 1:** Watch these two short video clip(s), by using the link below:

  `https://www.youtube.com/watch?v=Ct8Gxo8StBU`

  `https://www.youtube.com/watch?v=2XfA0e4Bzkk&t=45s`

  After watching the videos, create a markdown file and name it as `video-reflection`. In this file, provide detailed answers to the questions provided below:

  1. What is web scraping?
  2. What are the practical uses of web scrapers?
  3. How is web scraping different from API based data collection?
  4. How is web scraping used in a Real Estate business?
  5. What is the connection between web scraping and sentiment analysis?

Section 2: Lyrics Data Scraper

This section is worth 35 points. The points breakdown is provided below:

- Task 2 = 15 points

In this section, we will modify the starter code to scrape the data from the lyrics website. The underlying principle that tried out here are the generating our own dataset by web scraping. To complete this part, it is required to complete the tasks listed below:

1. **Task 2:**
   - Review the code provided in the starter code repository. The LyricsDriver program is an extension of what we tried out in DogTime program.
   - The getLyrics method accepts a Song URL as input. This method is currently incomplete and requires some implementation to be done. Look for a place holder comment in the method, that indicates to add your logic here. This place holder is placed after parsing through the HTML file and a document object is created using the Jsoup library.
   - Create an object called resultLinks

     Elements resultLinks = doc.select((”div[class=container main-page]”));
   - Open the URL, used in the main method, using a browser. Click on view - page source. Identify the DOM model for the HTML document and logically reason why we are using the div class provided in the previous bullet point?
   - Once an object is created, setup a for loop to iterate through the resultLinks.
   - Inside the for loop, create a string variable called lyrics and assign the value from link.text() to it. The next two steps should be implemented within the for loop.
   - The text for lyrics may include some unrelated data. This may start from ”Submit Corrections”. By using substring and lastIndexOf functions, remove this unrelated data from the string variable lyrics.
   - Call the method writeToFile by passing the arguments outputFile and lyrics.
   - At this point, compiling and executing the program should output the lyrics for the given URL in the upload directory.
   - You may choose to try out a different URL in the main method. Please make sure to wait a few seconds before executing the program. In this way, the scraping is done as a soft pull and the website doesn’t block your ipaddress.

Submission Details

For this assignment, please submit the following to your GitHub lab repository.

1. **video-reflection** markdown file.
2. An updated version of the Lyrics project..
3. It is recommended to upload a readme file, with the details that you would like the Professor to know while grading the work. For example, it may be reflection of your experience in the lab by highlighting some of the challenges faced and a brief mention of how you had addressed those challenges while implementing this lab. The readme file may also include a brief mention of any details that one should know about executing your program and what to expect during the execution.
4. It is highly important, for you to meet the honor code standards provided by the college and to ensure that the submission is made before the deadline. The honor code policy can be accessed through the course syllabus. Make sure to add the statement ”This work is mine unless otherwise cited.” in all your deliverables such as source code and PDF files.

Grading Rubric

1. Details including the points breakdown are provided in the individual sections above.
2. If a student needs any clarification on their lab credits, it is strongly recommended to talk to the Professor. The lab credits may be changed if deemed appropriate.