M.Sc. In Data Science
Programming for Data Scientists
2021/2022 Academic Year

Assignment 2 – Web Scraping.

Motivation for Assignment

Although a lot of raw data is available in CSV, JSON, XML, and other "standard" formats, an *awful* lot of published data resides within webpages, embedded within HTML markup (usually within `<table>` tags, but not always).

Developing strategies and techniques for extracting usable data from raw HTML is an important skill, and this assignment is designed to expose you to the techniques which must be mastered and applied when working with "web data".

In addition, this assignments is designed to improve your SQL and database skills, as well as expose you to writing more complex Python code.

Details of Assignment

You are to study the following Wikipedia page:

https://en.wikipedia.org/wiki/List_of_prime_ministers_of_Canada

paying particular attention to any tables of data.

Your tasks are as follows (with individual mark allocations shown within square brackets):

1. Provide the commands required to create a new MariaDB database to store your scraped data. You are to further provide the commands which you used to define a new user-id and password to the database, as well the appropriate table schema needed to house your scraped web data. I only want the schema and user details, not the data. [6]

2. Write Python code which automatically scrapes the data you need from the above Wikipedia page and stores it into your database table(s). [9]

3. Use the data in your database table(s) to provide answers to the following questions/queries (Note: you are required to show the SQL statements used to determine your answers):

(a) Which political party has produced the most individual prime ministers? [2]

(b) Provide a list of prime ministers who served by province/territory. [3]

(c) Which political party held the office of prime minister for the longest amount of overall time, and for how long? [3]

(d) Which individual politician held the office of prime minister for the longest amount of uninterrupted time, and for how long? [2]

(e) Which individual politician has held the office of prime ministers for the longest amount of overall time? [1]

(f) Which individual politician held the office of prime minister for the shortest amount of overall time, and for how long? [1]

(g) [To answer this question and the next, you may need to adjust your Python scraping code to follow the link for each individual prime minister in order to automatically scrap their date of birth. Note, too, that a few prime ministers will feature more than once in in the answer to this question].

What age was each prime minister on the day they assumed office? [4]

(h) On the last day of their term of office, which politician was the oldest? [1]

(i) In your view, and based on statistics calculated from your scraped data, which political party is the most successful? Show and describe your reasoning. [3]

Marks Allocation, Submission, and Deadline

- This assignment is worth 35% of your total grade.
- You are to submit a Jupyter Notebook file (which has been through *Kernel… Restart & Clear Output* before saving), together with the commands needed to setup your database.
- The due date/time for this assignment is: **5:00pm on Friday November 19th 2021 –** your email to paul.barry@itcarlow.ie must arrive prior to this deadline.

This is an individual assignment: you are expected to work on your own, and that the work you submit is written by you. You must declare if this is not the case.

Note: you may use whichever tools you desire with this assignment, just so long as the scraping technology employed is Python-based (e.g., pandas) and you use SQL to answer the questions.