|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **National University of Computer and Emerging Sciences, Lahore Campus** | | | | |
| C:\Users\saif\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.Word\final design.jpg | **Course:** | **Data Science** | **Course Code:** | **CS-4048** |
| **Program:** | **BS(Computer Science)** | **Semester:** | **Spring 2023** |
| **Duration:** | **-** | **Total Marks:** | **10** |
| **Due Date:** | **10-Sept-23** | **Weight** | **3.33 %** |
| **Section:** | **A** | **Page(s):** | **2** |
| **Exam:** | **Assignment 1** | **Roll No.** |  |
| **Instruction/Notes:**   * Read the assignment carefully. Make sure you have understood the requirements and expectations of the assignment. * Ensure that you have all the necessary files and documents ready for submission in the CORRECT format. * This is an individual assignment. Any form of plagiarism will result in an award of ZERO marks. * The assignment must be submitted before the announced DEADLINE. One mark will be deducted for each day of late submission. | | | | |

**Web scraping**

Web scraping is a technique for extracting information from websites. It involves making HTTP requests to a website's server, downloading the HTML content of the web page, and then parsing that HTML data to extract the information you're interested in. The extracted data can then be stored in a local file or database for later analysis or use. Web scraping can be useful for a variety of tasks, such as collecting data for data analysis or research, tracking changes on websites, or even for automating certain tasks on the web. However, it is important to follow the terms of use and ethical guidelines of the websites you're scraping, as some websites may have restrictions on automated data extraction.

**Website**

For this assignment you have to collect the data of all faculty members from the following website.

<http://lhr.nu.edu.pk/faculty/>

<http://isb.nu.edu.pk/Faculty/allfaculty/>

<https://khi.nu.edu.pk/>

**Step 1:**

There are different departments and each department has a different number of faculty members. Faculty page for all campuses may not have the same structure. You may find faculty details inside the department's page for some campuses. Step one is to analyze the structure of the faculty pages for all campuses. You have to collect the following details of the faculty members for each department and for each campus.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Name | Designation | HEC Approved PHD Supervisor | Highest Education | Email | Department | Extension | ImageURL |
| Integer | String | String | Bool | String | EMAIL | String | Integer | URL |

Collect the data of the Lahore campus faculty members in the above format and transform it into a data frame. Save the data frame as a .csv file named **lhr.csv**.

Do the same for Islamabad, Karachi and save the files as **isb.csv** and  **khi.csv** respectively.

**Step 2:**

Concatenate all data frames to create a single data frame having the information of all faculty members from all campuses. Save this data frame as a .csv file named **fast\_faculty.csv**.

**Step 3:**

Load **fast\_faculty.csv** into a data frame named **faculty**. Divide last digit of your Student ID by 10 and pass it into sample method as frac value. If the last digit of your Student ID is 0, keep the below code as -it-is. Save the **sample\_** data frame as **sample.csv** file.

sample\_ = faculty.sample(frac=0.1)

**Deliverables:**

* A jupyter notebook with properly formatted and documented code.
* A zip file having all **.csv** files
* Rename both files to your student ID before submission