QS World University Rankings

Summary:

This document involved scraping data from the QS World University Rankings website, which comprised extensive information about universities. Due to the large volume of data, the scraping process was divided into smaller tasks, with each task focused on retrieving specific information and storing it in CSV files segregated into various folders. This approach was necessitated by system limitations, as the entire dataset couldn't be processed in a single iteration.

The code was structured to handle each scraping task separately, enabling the extraction of data points without overwhelming the system's capacity. By breaking down the scraping process into manageable segments, it facilitated efficient data retrieval while circumventing technical constraints.

Each CSV file generated from the scraping process contained distinct subsets of university data, organized into folders based on their respective categories. This systematic approach allowed for better management and analysis of the scraped data, ensuring its usability despite the challenges posed by its size.

In summary, the assignment's implementation involved a structured approach to scrape and store university data in CSV files distributed across various folders, addressing the limitations posed by the system's capacity constraints.

Folders:

- University Urls
- Admissions
- Students Staffs
- Locations
- Programs Available
- Rankings
- Details

Detailed Description of each folder and CSV files:

1. University Urls

- The code for extracting university names and URLs from a website is contained within a Python notebook named "UniversityUrls.ipynb".
- The data extracted from the website is distributed across seven separate CSV files. Each CSV file contains data from 100 pages, with the first file named "universities_with_no_100.csv" and the last one named "universities_with_no_700.csv".

- To consolidate all the extracted data into a single file, a separate Python notebook titled "MergedCsvs.ipynb" was used.

- The consolidation process involved merging the content of all seven CSV files into one comprehensive file named "UniversityUrls_Filesmerged_file.csv".

  Note: I had extracted all the information in the below folders using the university urls extracted from this code “UniversityUrls.ipynb”

2. Admissions

- The code for fetching admission scores required for various courses from a website is contained within a Python notebook named "admissions.ipynb".
- The admission data extracted from the website is distributed across seven separate CSV files. Each CSV file contains admission scores for a set of universities and courses, with the first file named "university_admission_data1.csv" and the last one named "university_admission_data7.csv".
- To consolidate all the admission data into a single file, a separate Python notebook titled "MergedCsvs.ipynb" was used.
- The consolidation process involved merging the content of all seven CSV files into one comprehensive file named "Admissions_Filesmerged_file.csv".

3. Students Staffs

- The code for gathering student and staff count data from a website is contained within a Python notebook named "StudentsStaffs.ipynb".
- The student and staff count data extracted from the website is distributed across seven separate CSV files. Each CSV file contains data for a specific timeframe, with the first file named "studentstaff_count_1.csv" and the last one named "studentstaff_count_7.csv".
- To consolidate all the student and staff count data into a single file, a separate Python notebook titled "StudentsStaffs_Merge.ipynb" was utilized.
- The merging process involved combining the content of all seven CSV files into one comprehensive file named "StudentsStaffs_Filesmerged_file.csv".

4. Locations

- The code for retrieving university locations from a website is contained within a Python notebook named "Locations.ipynb".
- The university location data extracted from the website is spread across seven separate CSV files. Each CSV file contains location information for a specific set of universities, with the first file named "university_locations_1.csv" and the last one named "university_locations_7.csv".
- To consolidate all the university location data into a single file, a separate Python notebook titled "Locations_Merge.ipynb" was employed.
- The merging process involved aggregating the content of all seven CSV files into one comprehensive file named "Locations_Filesmerged_file.csv".

5. Programs Available

- The code for extracting available programs from a website is contained within a Python notebook named "ProgramsAvailable.ipynb".
- The data regarding available programs extracted from the website is distributed across seven separate CSV files. Each CSV file contains program information for a specific set of universities, with the first file named "programsavailable1.csv" and the last one named "programsavailable7.csv".
- To consolidate all the program data into a single file, a separate Python notebook titled "ProgramsAvailable_Merge.ipynb" was utilized.
- The merging process involved combining the content of all seven CSV files into one comprehensive file named "ProgramsAvailable_Filesmerged_file.csv".

6. Rankings

- The code for extracting final rankings from a website is contained within a Python notebook named "FinalRanks.ipynb".
- The final ranking data extracted from the website is spread across seven separate CSV files. Each CSV file contains ranking information for a specific set of entities, with the first file named "FinalRankings1.csv" and the last one named "FinalRankings7.csv".
- To consolidate all the final ranking data into a single file, a separate Python notebook titled "Rankings_Merge.ipynb" was used.
- The merging process involved aggregating the content of all seven CSV files into one comprehensive file named "Rankings_Filesmerged_file.csv".

7. Details

- The code for extracting details about universities from a website is contained within a Python notebook named "DetailsUniversity.ipynb".
- The details include University Name, Country, City, Region.
- The data regarding university details extracted from the website is stored in a CSV file named "details_universities.csv".
- Note: I tried to extract these details for all the colleges, but it is taking long time for running and the kernel fails. So, I had extracted for 1422 universities under 2023 rankings.