Assignment 1

Task 1: Static Website Extraction – Dawn News Headlines

In this task, the Dawn news website was scraped to extract the latest 30 headlines. The requests library was used to fetch the HTML content, and BeautifulSoup was employed to parse and locate the news section. Each headline and its corresponding URL were stored in a pandas DataFrame. Headlines without valid URLs were marked as 'No URL available'.

Task 2: Website Data Extraction – Pakistan Stock Exchange

The PSX website was scraped to extract market summary data. The requests library was used to retrieve the page content, and BeautifulSoup was applied to locate the index table. Data such as Index Name, LDCP, Open, High, Low, and Change were extracted. The results were structured into a pandas DataFrame.

Task 3: University Ranking Scraper – QS World Rankings

QS World University Rankings were scraped using Selenium due to dynamic content loading. The top 50 universities were extracted, including details like University Name, Country, Overall Score, and Subject-Specific Ranking. The extracted data was stored in a DataFrame for further trend analysis.

Task 4: Product Data Extraction – Daraz.pk

Daraz.pk was scraped to collect product information for a given search query (e.g., 'iPhone 15'). Selenium was used to load the page and scroll dynamically. Data for 20 products including Title, Price, Seller, Rating, and Delivery Info were stored in a DataFrame. Price and rating values were cleaned for analysis, and the best deal (lowest price with highest rating) was identified.

Task 5: Book Recommendation Data Extraction – Goodreads

Goodreads genres (e.g., Fiction and Science) were scraped using Selenium. For each genre, 10 books were extracted with details such as Title, Author, Rating, Number of Reviews, and Publication Date. Missing values were marked as 'Not Available'. Ratings were cleaned into numerical values, and average ratings across genres were compared to identify the highest-rated genre.