WEB SCRAPPING PROJECT DESCRIPTIONS

Name & ID: Yamini Kuntal 455723

Description: I am doing a project on Web scrapping where I am scraping some details from webpage: Richest 250 People In the World (therichest.com). This page has the details of Top richest people in the world; however, I am fetching top 100 richest people from this overall list. I will be accomplishing this activity using Beautiful Soup, Selenium and Scrapy.

Scrapy Steps:

- Uses Scrapy to build a web scraping spider called "ScrapySpider"
- Begins from a URL with the top 250 richest people's list, extracting links to individual profiles.
- For each profile link:
  - Gathers data like title, source, URL, and net worth.
  - Handles pagination to move through multiple list pages.
  - Stores collected data in a CSV file for analysis and reference.

Output: The output from Scrapy is a CSV file named "scrapy_richest.csv" containing details of the top 250 richest people in the world. Each row in the CSV file represents a person's profile and includes their title, source, URL, and net worth.

Total Time: 10 seconds

Beautiful Soup (bs_richest.py):

- Utilizes CSV, requests, and BeautifulSoup libraries.
- Initiates a CSV file and writes header for 'title', 'source', 'url', and 'worth'.
- Sets a base URL for the richest people list and initializes a page number.
- In a loop:
  - Constructs the URL using the base URL and page number.
  - Makes a request to the URL and checks for a successful response.
  - Parses the page content using BeautifulSoup.
  - Extracts individual profile information like title, net worth, and source.
  - Writes the data into the CSV file.
  - Checks for a "next" link to continue to the next page or breaks the loop.
- Completes the data scraping process and outputs a completion message.

Output: The output from BeautifulSoup is a CSV file named "bs_richest.csv" containing details of the top 250 richest people in the world. Each row in the CSV file represents a person's profile and includes their title, source, URL, and net worth.

Total Time: 98.75 seconds

Selenium (selenium_richest.py):

- Imports necessary modules, including csv for CSV operations and webdriver from Selenium for browser automation.
- Sets up the Chrome WebDriver using the provided ChromeDriver executable path.
- Opens a CSV file and writes the header for 'title', 'source', 'url', and 'worth'.

- Constructs a base URL for the richest people list and initializes page number and a list to store profile links.
- In a loop:
    - Visits the URL using the WebDriver.
    - Extracts links to individual profiles using Selenium's element location methods.
    - Handles pagination and updates the page number accordingly.
- Visits each profile link:
    - Extracts the title, source, and net worth using WebDriver and CSS selectors.
    - Writes the data into the CSV file.
- Closes the WebDriver and prints a completion message.

Output: The output from Selenium is a CSV file named "selenium_richest.csv" containing details of the top 250 richest people in the world. Each row in the CSV file represents a person's profile and includes their title, source, URL, and net worth.

Total Time: 309 seconds