Name : Aditi Dhepe

Roll no : 2213774

Topic: Introduction to Web Scraper

Experiment No:6

Develop a web scraper to mine structured data from any website according to a given application

Objective:

- To understand and implement a web scraper in Python using popular libraries.
- To extract specific structured data from a website.
- To handle common issues like pagination, login requirements, and data storage.

Theory:

Understanding the Web:

Choosing the right website is crucial for successful scraping. Consider factors such as the website's structure, data availability, and relevance to your project.

Ensure compliance with legal and ethical guidelines. Review the website's terms of service and robots.txt file to understand scraping policies and restrictions.

Python and Libraries for Scraping:

requests: This library allows you to send HTTP requests easily. It is commonly used for fetching web pages during the web scraping process.

Beautiful Soup (beautifulsoup4): Beautiful Soup is a powerful Python library for parsing HTML and XML documents. It provides functions and methods to navigate and manipulate the parse tree, making it ideal for extracting data from web pages.

pandas: Pandas is a versatile data manipulation library in Python. While it's not directly related to web scraping, it's commonly used for data analysis and manipulation tasks, including storing and processing scraped data in tabular format (DataFrames)

Initial Setup:

Develop a Python script to implement the scraping logic.

Start by defining the URL(s) of the webpage(s) to be scraped.

Use the requests library to send HTTP GET requests to fetch webpage content.

Utilize Beautiful Soup to parse the HTML content and extract desired data elements.

Implement logic to handle different data extraction scenarios, such as retrieving text, links, images, or structured data.

Building the Web Scraper:

After extracting and transforming the data, you typically organize it into a tabular format using a DataFrame, often provided by libraries like pandas.

Each row of the DataFrame represents a data record, and each column represents a different attribute or field of the data.

You can create a DataFrame and populate it with the extracted data, using pandas' functionality to manipulate and structure the data as needed.

Exporting to CSV:

Once the data is organized in a DataFrame, you can easily export it to a CSV file using pandas' to_csv() function.

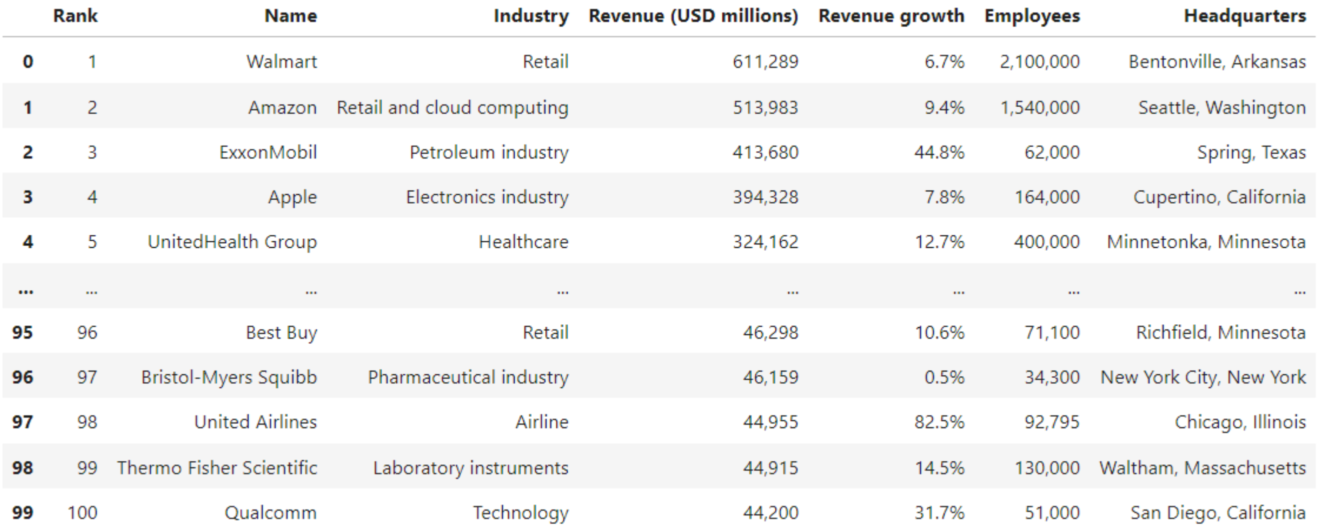
This function allows you to specify the file path and name for the CSV file, as well as various options for formatting the CSV output (e.g., delimiter, quoting, encoding).

The DataFrame's data will be written to the CSV file, with each row representing a record and each column representing a field, separated by commas.

Conclusion:

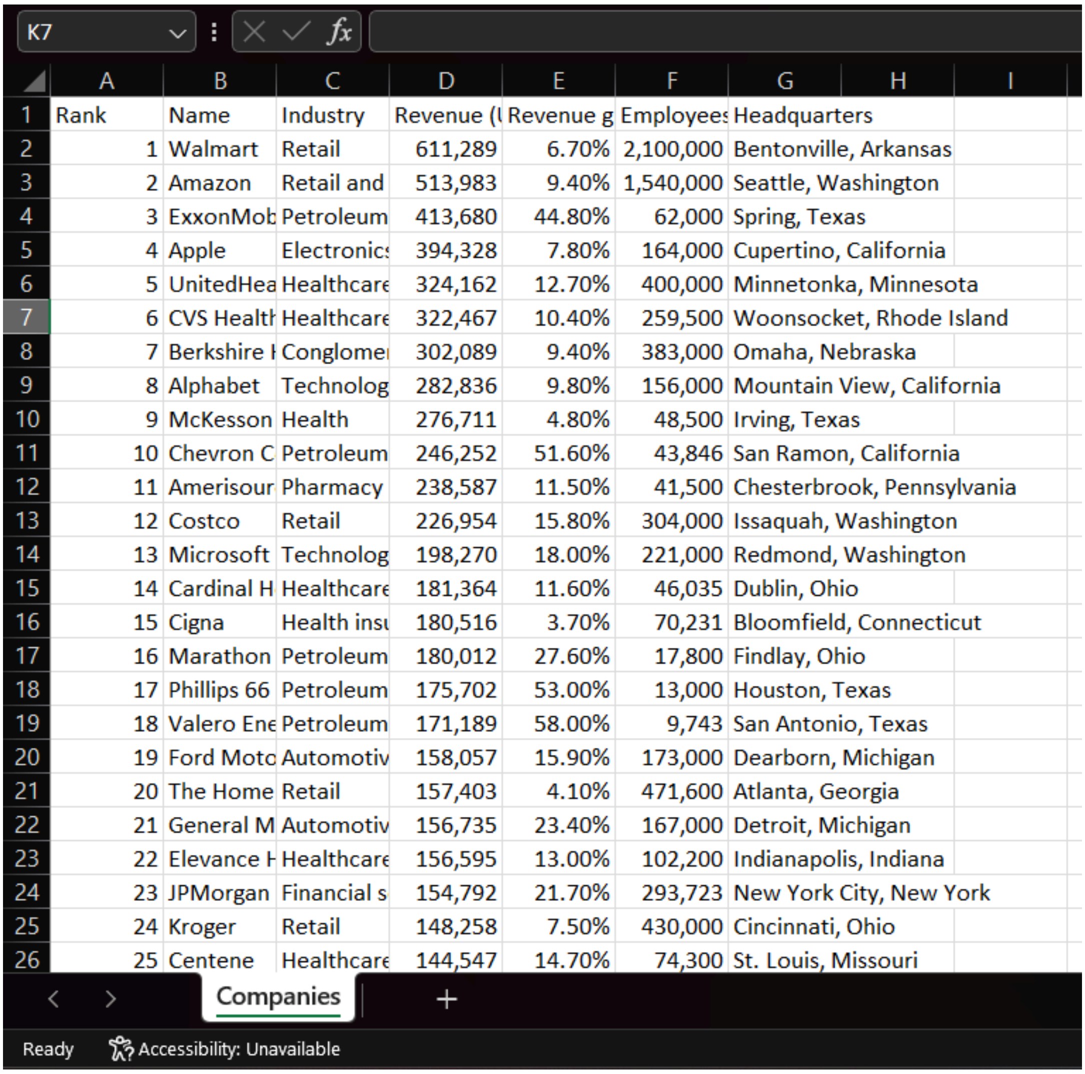
In conclusion, converting web data to a CSV file in a web scraper involves extracting, cleaning, and transforming data from web pages into a tabular format. This process typically includes parsing HTML content, organizing data into a DataFrame, and exporting it to a CSV file using libraries like pandas. Error handling, testing, and validation are essential steps to ensure the reliability and accuracy of the extracted data. By following these steps, web scrapers can efficiently generate CSV files suitable for further analysis and processing.

Output (Screenshots):



	Rank	Name	Industry	Revenue (USD millions)	Revenue growth	Employees	Headquarters
0	1	Walmart	Retail	611,289	6.7%	2,100,000	Bentonville, Arkansas
1	2	Amazon	Retail and cloud computing	513,983	9.4%	1,540,000	Seattle, Washington
2	3	ExxonMobil	Petroleum industry	413,680	44.8%	62,000	Spring, Texas
3	4	Apple	Electronics industry	394,328	7.8%	164,000	Cupertino, California
4	5	UnitedHealth Group	Healthcare	324,162	12.7%	400,000	Minnetonka, Minnesota
...	...	...	...	...	...	...	...
95	96	Best Buy	Retail	46,298	10.6%	71,100	Richfield, Minnesota
96	97	Bristol-Myers Squibb	Pharmaceutical industry	46,159	0.5%	34,300	New York City, New York
97	98	United Airlines	Airline	44,955	82.5%	92,795	Chicago, Illinois
98	99	Thermo Fisher Scientific	Laboratory instruments	44,915	14.5%	130,000	Waltham, Massachusetts
99	100	Qualcomm	Technology	44,200	31.7%	51,000	San Diego, California

CSV FILE GENERATED :



	A	B	C	D	E	F	G	H	I
1	Rank	Name	Industry	Revenue (	Revenue g	Employees	Headquarters		
2	1	Walmart	Retail	611,289	6.70%	2,100,000	Bentonville, Arkansas		
3	2	Amazon	Retail and	513,983	9.40%	1,540,000	Seattle, Washington		
4	3	ExxonMob	Petroleum	413,680	44.80%	62,000	Spring, Texas		
5	4	Apple	Electronics	394,328	7.80%	164,000	Cupertino, California		
6	5	UnitedHea	Healthcare	324,162	12.70%	400,000	Minnetonka, Minnesota		
7	6	CVS Health	Healthcare	322,467	10.40%	259,500	Woonsocket, Rhode Island		
8	7	Berkshire	Conglomer	302,089	9.40%	383,000	Omaha, Nebraska		
9	8	Alphabet	Technolog	282,836	9.80%	156,000	Mountain View, California		
10	9	McKesson	Health	276,711	4.80%	48,500	Irving, Texas		
11	10	Chevron C	Petroleum	246,252	51.60%	43,846	San Ramon, California		
12	11	Amerisour	Pharmacy	238,587	11.50%	41,500	Chesterbrook, Pennsylvania		
13	12	Costco	Retail	226,954	15.80%	304,000	Issaquah, Washington		
14	13	Microsoft	Technolog	198,270	18.00%	221,000	Redmond, Washington		
15	14	Cardinal H	Healthcare	181,364	11.60%	46,035	Dublin, Ohio		
16	15	Cigna	Health insu	180,516	3.70%	70,231	Bloomfield, Connecticut		
17	16	Marathon	Petroleum	180,012	27.60%	17,800	Findlay, Ohio		
18	17	Phillips 66	Petroleum	175,702	53.00%	13,000	Houston, Texas		
19	18	Valero Ene	Petroleum	171,189	58.00%	9,743	San Antonio, Texas		
20	19	Ford Moto	Automotiv	158,057	15.90%	173,000	Dearborn, Michigan		
21	20	The Home	Retail	157,403	4.10%	471,600	Atlanta, Georgia		
22	21	General M	Automotiv	156,735	23.40%	167,000	Detroit, Michigan		
23	22	Elevance H	Healthcare	156,595	13.00%	102,200	Indianapolis, Indiana		
24	23	JPMorgan	Financial s	154,792	21.70%	293,723	New York City, New York		
25	24	Kroger	Retail	148,258	7.50%	430,000	Cincinnati, Ohio		
26	25	Centene	Healthcare	144,547	14.70%	74,300	St. Louis, Missouri		