Assignment

Q1. What is Web Scraping? Why is it Used? Give three areas where Web Scraping is used to get data.

Answer.

Web scraping is the process of collecting structured web data in an automated manner. It's also widely known as web data extraction or web data scraping.

Web scraping can help companies gather the correct contact information from their target market—including names, job titles, email addresses, and telephone numbers. Then, they can reach out to these contacts and generate more leads and sales for their business.

Some of the main use cases of web scraping include price monitoring, price intelligence, news monitoring, lead generation, and market research among many others.

Q2. What are the different methods used for Web Scraping?

Answer:

The various methods used for the Web scraping are mentioned below,

- Human copy-and-paste. The simplest form of web scraping is manually copying and pasting data from a web page into a text file or spreadsheet.
- Text pattern matching.
- HTTP programming.
- HTML parsing.
- DOM parsing.
- Vertical aggregation.
- Semantic annotation recognizing.
- Computer vision web-page analysis.

Q3. What is Beautiful Soup? Why is it used?

Answer: Beautiful Soup is a Python library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser and provides Pythonic idioms for iterating, searching, and modifying the parse tree.

Q4. Why is flask used in this Web Scraping project?

Answer: Flask is a lightweight framework to build websites. It is used to parse the collected data and display it as HTML in a new HTML file. The requests module allows us to send http requests to the website we want to scrape. The first line imports the Flask class and the render template method from the flask library.

Q5. Write the names of AWS services used in this project. Also, explain the use of each service.

Answer:

***Pipeline: AWS Data Pipeline is a web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premises data sources, at specified intervals.**

***Beanstalk: Elastic Beanstalk is a service for deploying and scaling web applications and services. Upload your code and Elastic Beanstalk automatically handles the deployment—from capacity provisioning, load balancing, and auto scaling to application health monitoring.**