Experiment No. 8

Title: Study Experiment on Web Scraping

Batch: B2 **Roll No.: 16010420117** **Experiment No:8**

Aim: Study Experiment on Web Scraping

Resources needed: Windows OS

Pre Lab/ Prior Concepts:
Students should have prior knowledge of PHP and Python

Theory:

What is Web Scraping?

Web Scraping is also termed as Screen Scraping, Web Data Extraction, Web Harvesting etc. Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. There are many different ways to perform web scraping to obtain data from websites. These include using online services, particular API's or even creating your code for web scraping from scratch. Many large websites, like Google, Twitter, Facebook, StackOverflow, etc. have API's that allow you to access their datain a structured format. This is the best option, but there are other sites that don't allow users to access large amounts of data in a structured form or they are simply not that technologically advanced. In that situation, it's best to use Web Scraping to scrape the website for data.

Web scraping requires two parts, namely the **crawler** and the **scraper**. The crawler is an artificial intelligence algorithm that browses the web to search for the particular data required by following the links across the internet. The scraper, on the other hand, is a specific tool created to extract data from the website. The design of the scraper can vary greatly according to the complexity and scope of the project so that it can quickly and accurately extract the data.

How web scraper works?

Web Scrapers can extract all the data on particular sites or the specific data that a user wants. Ideally, it's best if you specify the data you want so that the web scraper only extracts that data quickly. For example, you might want to scrape an Amazon page for the types of juicers available, but you might only want the data about the models of different juicers and not the customer reviews.

So, when a web scraper needs to scrape a site, first the URLs are provided. Then it loads all the HTML code for those sites and a more advanced scraper might even extract all the CSS and Javascript elements as well. Then the scraper obtains the required data from this HTML code and outputs this data in the format specified by the user. Mostly, this is in the form of an Excel spreadsheet or a CSV file, but the data can also be saved in other formats, such as a JSON file.

Activity:

1. **List and explain various applications of Web Scraping.**
2. **List different types of web scraper and explore any available web scraper and prepare a documentation by considering various aspect of working of selected web scraper**

Output (Detailed Documentation):

1. **List and explain various applications of Web Scraping.**

Market Research

Since manually finding such vast data can be a daunting task, implementing web scrapers can automate market research to extract accurate data in real time.
Finally, web scraping to conduct market research makes collection of data easy and cost-effective.

Lead Generation

No matter what industry you are in, customers are your lifeblood. Therefore, adding more potential customers should be the primary goal for the development of your business.
Web scraping for lead generation helps businesses find the best and most. Web scraping for lead generation is a need in almost every industry.

Competitive Intelligence

The easiest way for collecting and compiling such data can be done with the help of web scraping. But one should know how to use web scraping to gather competitive data that will help them in tracking information like real-time price, product updates, customer information, reviews, feedback, and many more. There are several web scraping tools that can provide businesses with data quickly and easily from multiple websites.

Product Pricing Comparison

Web scraping and data mining to find product information plays a key role in helping businesses and marketers identify and extract data points from multiple websites and eCommerce stores to make data-driven decisions.

Monitoring Consumer Sentiment

There are many software review aggregator websites containing reviews of almost every category and web scraping to collect sentiment analysis from marketplaces like Amazon helps businesses understand customer needs and preferences.

Brand Audits

Brands audits through web scraping social media and review sites reveal the branded products consumers prefer the most, trends in the market, and many more. The information gathered can lead to improvements in products and service, resulting in a stronger brand impression over time.

AI & Machine Learning Data

The quality of machine learning depends upon the accuracy, relevancy, and volume of data extracted to train the model. To make reliable data readily available AI web scraping solutions come into play. Web scraping for machine learning helps data scientists collect the required information to feed their datasets.

Creating a Job Board

A quality job board web scraping tool is required to create and run a successful job board. Due to the broad-based nature of the most popular job boards, it can be difficult to search for niche positions and new opportunities are posted daily. Focused job boards powered by crawling multiple websites, provide job seekers with relevant employment postings.

Monitoring Investment News & Stock Prices

Investment and stock firms rely on data scraping tools in order to collect investing news and information in real time. Web scraping stock market data helps monitor global economic trends, stock prices, and market moves to inform their investment strategies.

Affiliate Marketing

Creating high quality blog and social posts based on trending keywords are critical to an affiliate marketing website. Web scraping helps marketers compile the most relevant keywords and related information, then leverage the data to drive traffic and conversions.

Real Estate Data

Web scraping real estate data can even help finding historical sales as comps when pricing a property that’s about to hit the market. This type of automated data collection eliminates manual work saving time for busy agents who need to be meeting with clients in person and showing homes to prospective buyers.

Sports Betting

Casinos, gambling websites, and other betting organizations use web scraping to easily collect the sports statistics required to accurately create the official betting lines. The most sophisticated oddsmakers also offer new “prop bets” while competitions are happening live. It would be virtually impossible to capture and analyze this type of real-time data manually at the speed required to produce new betting opportunities and increase the number of bets placed.

2. **List different types of web scraper and explore any available web scraper and prepare a documentation by considering various aspect of working of selected web scraper**

You can have **Self-built Web Scrapers** but that requires advanced knowledge of programming. And if you want more features in your Web Scraper, then you need even more knowledge. On the other hand, pre-built **Web Scrapers** are previously created scrapers that you can download and run easily. These also have more advanced options that you can customize.

Browser extensions Web Scrapers are extensions that can be added to your browser. These are easy to run as they are integrated with your browser, but at the same time, they are also limited because of this. Any advanced features that are outside the scope of your browser are impossible to run on Browser extension Web Scrapers.

But **Software Web Scrapers** don’t have these limitations as they can be downloaded and installed on your computer. These are more complex than Browser web scrapers, but they also have advanced features that are not limited by the scope of your browser.

Cloud Web Scrapers run on the cloud, which is an off-site server mostly provided by the company that you buy the scraper from. These allow your computer to focus on other tasks asthe computer resources are not required to scrape data from websites.

Local Web Scrapers, on the other hand, run on your computer using local resources. So, if the Web scrapers require more CPU or RAM, then your computer will become slow and not be able to perform other tasks.

Browser extension web scrapper

Google Chrome web scrapper

Web Scraper utilizes a modular structure that is made of selectors, which instruct the scraper on how to traverse the target site and what data to extract. Thanks to this structure, data mining from modern and dynamic websites such as Amazon, Tripadvisor, eBay, as well as from lesser-known sites is effortless.

Data extraction is run on your browser and doesn’t require anything to be installed on your computer. You don't need Python, PHP, or JavaScript coding experience to start scraping. Additionally, it is possible to completely automate data extraction in Web Scraper Cloud.

Once the data is scraped, download it as a CSV or XLSX file that can be further imported into Excel, Google Sheets, etc.

Features
Web Scraper is a simple web scraping tool that allows you to use many advanced features to get the exact information you are looking for. It offers features like:
 * Data scraping from multiple pages;
 * Multiple data extraction types (text, images, URL’s, and more);
 * Scraping data from dynamic pages (JavaScript + AJAX, infinite scroll);
 * Browsing scraped data;
 * Exporting scraped data from a website to Excel;

It is dependent only on the web browser; therefore, no extra software needed for you to start scraping.

How to begin scraping?

There are only a couple of steps you will need to learn in order to master web scraping:

1. Install the extension and open the Web Scraper tab in developer tools (which has to be placed at the bottom of the screen);
2. Create a new sitemap;
3. Add data extraction selectors to the sitemap;
4. Lastly, launch the scraper and export scraped data.

It’s as easy as that!

Outcomes:

CO4: Demonstrate the use advanced features such as REST API, email handling, Localization and internalization in PHP.

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

From this experiment I got a through idea about web scrapping, its uses and its types

Signature of faculty in-charge with date

References:

1. Thomson PHP and MySQL Web Development Addison-Wesley Professional , 5th Edition2016.
2. www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it
3. towardsdatascience.com/web-scraping-basics-82f8b5acd45c