Web Scraping

What is Web Scraping?

- Web scraping is an automated method used to extract large amounts of data from websites.
- The data on the websites are unstructured.
  - Web scraping helps collect these unstructured data and store it in a structured form.

Image Source: https://www.edureka.co/

Why is Web Scraping Used?

- **Price Comparison:** Many services use web scraping to collect data from online shopping websites and use it to compare the prices of products.
- **Brand Monitoring and Competition Analysis:** Web Scraping is used to get customer feedback regarding a particular service or product so as to understand how a customer feels regarding that particular thing. It is also used to extract competitor data in a structural, usable format.
- **Email address gathering:** Many companies that use email as a medium for marketing, use web scraping to collect email ID and then send bulk emails.
- **Social Media Scraping:** Web scraping is used to collect data from Social Media websites such as Twitter to find out what's trending.
- **Research and Development:** Web scraping is used to collect a large set of data (Statistics, General Information, Temperature, etc.) from websites, which are analyzed and used to carry out Surveys or for R&D.

Techniques of Web Scraping

- **Manual Extraction Techniques:** Manually copy-pasting the site content comes under this technique. Though tedious, time taking and repetitive it is an effective way to scrap data from the sites having good anti-scraping measures like bot detection.
- **Automated Extraction Techniques:** Web scraping software is used to automatically extract data from sites based on user requirement.
  - **HTML Parsing:**
    - Parsing means to make something understandable to be analyzing it part by part.
    - HTML parsing means taking in the code and extracting relevant information from it based on the user requirement.
    - Mainly executed using JavaScript, the target as the name suggests are HTML pages.
  - **DOM Parsing:** The Document Object Model is the official recommendation of the World Wide Web Consortium. It defines an interface that enables a user to modify and update the style, structure, and content of the XML document.
  - **Web Scraping Software:** Nowadays, many web scraping tools are available or are custom build on users need to extract required desiring information from millions of websites.

Tools for Web Scraping

- Web Scraping tools are specifically developed for extracting data from the internet.
- Some of the most popular Web Scraping tools are:
  - WebScraper.io
  - Import.io
  - Webhose.io
  - Dexi.io
  - Scrapinghub
  - Parsehub

Python Libraries used for Web Scraping

- Python has various applications and there are different libraries for different purposes

  - **BeautifulSoup**: Beautiful Soup is a Python package for parsing HTML and XML documents. It creates parse trees that is helpful to extract the data easily.
  - **Selenium**: Selenium is a web testing library. It is used to automate browser activities.
  - **Pandas**: Pandas is a library used for data manipulation and analysis. It is used to extract the data and store it in the desired format.

Web Scrapping – Legal or Illegal?

- The legalization of web scraping is a sensitive topic, depending on how it is used
- On one hand, web scraping with good bot enables search engines to index web content, price comparison services to save customer money and value
- But web scraping can be re-targeted to meet more malicious and abusive ends. Web scraping can be aligned with other forms of malicious automation, named *"bad bots"*, which enable other harmful activities like *denial of service attacks*, *competitive data mining*, *account hijacking*, *data theft* etc.
- Legality of Web Scraping is a grey area that tends to develop as time goes on

Image Source: https://www.geeksforgeeks.org/

Challenges to Web Scraping

- **Data Warehousing:** Data extraction at a scale will generate a large amount of information to be stored. If the data warehousing infrastructure is not properly built then the searching, storing and exporting of this data will become a cumbersome task. Hence, for large-scale data extraction, there needs to be a perfect data warehousing system without any flaws and faults.
- **Website Structure Changes:** Every website periodically updates its user interface to improve its attractiveness and experience. This requires various structural changes too. Since the web scrapers are set up according to the code elements of the website at that time, they require changes too.
- **Anti-Scraping Technologies:** Some websites use anti-scraping technologies that thwart away any scraping attempt. They apply a dynamic coding algorithm to prevent any bot intervention and use the IP blocking mechanism. It requires a lot of time and money to work around such anti-scraping technologies.
- **Quality of Data Extracted:** Records that do not meet the quality of information required will affect the overall integrity of the data. Making sure that the Data Scraped meets the quality guidelines is a difficult task as it needs to be done in real-time.

Hand-on Session

- Using Python's BeautifulSoup
- Using 'Web Scraper' tool

Using Python: BeautifulSoup package

- Step#1:
  - From command prompt:
    - pip install requests
    - pip install html5lib
    - pip install bs4

- Step#2: Import libraries
- Step#3: Extraction HTML file using request method
- Step#4: Prettify using Beautiful Soup
- Step#5: Extract relevant data from the HTML
- Step#6: Save the data using Pandas library

Example: Amazon's Product review

Using ‘Web Scraper’

Source: https://webscraper.io/

- Step#1: Add WebScraper extension to browser
- Step#2: Create sitemap
- Step#3: Check selector graph
- Step#4: Scrape the data

Example:

1. Scraping data from webscraper test website
2. Scraping data from Amozon.in
   - Phone name, price and images
   - Product review

Refrences

- https://www.edureka.co/blog
- https://www.geeksforgeeks.org/
- https://webscraper.io/