ADITYA N BHATT(231057015)
8/21/23

PDV ASGINMENT 1

Q1. What is web scraping?

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

Example:

1. Comparison Shopping Sites

There are several websites and applications that can help you to easily compare pricing between several retailers for the same product.
One way that these websites work is by using web scrapers to scrape product data and pricing from each retailer daily. This way, they can provide their users with the price comparison data they need.

2. SEO (Search Engine Optimization)

Not many businesses will think to use web scraping for SEO. It can help you gather the right data that can help improve your online presence on search engines. You'll be able to find keywords and backlink opportunities.

Web scraping can be used for SEO in many ways! You can scrape SERPs, do some competitor research, find backlink opportunities and find influencers!

References:

1. https://www.zyte.com/learn/what-is-web-scraping/#What-is-web-scraping?

Q2. Web Scraping Challenges

We distinguish three categories of obstacles that we commonly encountered when scraping scientific Web repositories:
1. size-limitations of the result sets
2. dynamic contents and

3. access barriers.

In Size Limitations of the Result Sets, there are 2 types

1. Static Caps: The search functionality of scientific Web repositories often returns a static maximum number of search results per query.

2. Pagination: scientific Web repositories typically use pagination for listing the items in the result set. When employing classic pagination, the interface divides the list of items into multiple pages, each showing a fixed number of items.

Dynamic Contents: Scientific Web repositories, and Web pages in general, increasingly rely on content that is dynamically loaded using JavaScript. In this approach, JavaScript is employed to manipulate the currently loaded page at runtime instead of loading a different page.

In Access Barriers, there are 2 types:

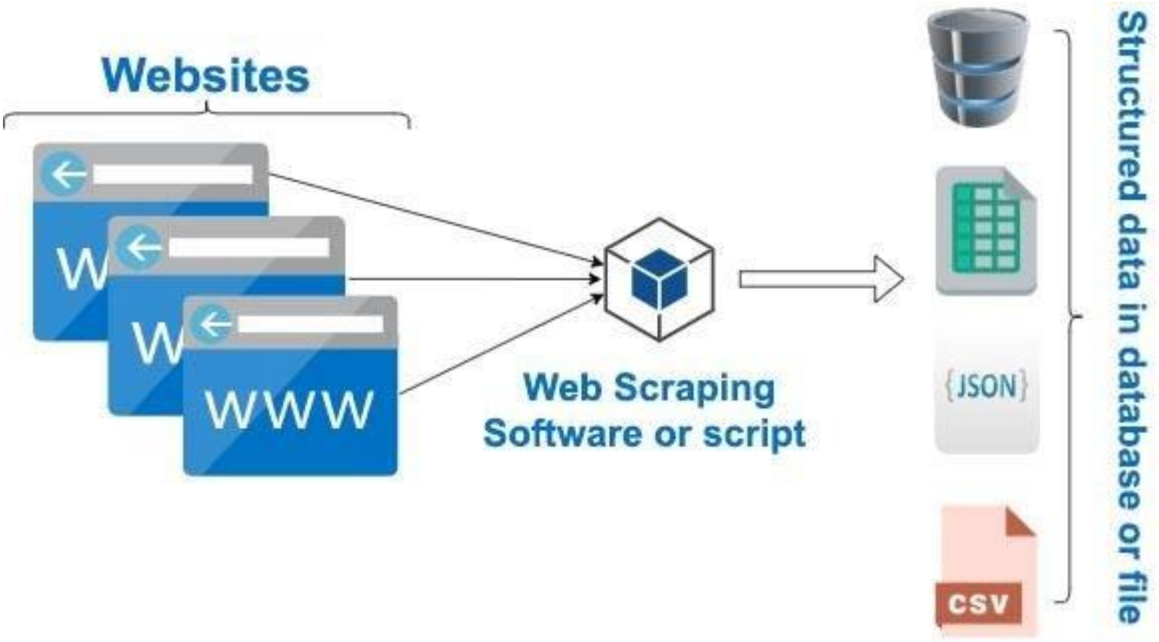
1. Obfuscated URL Parameters: The use of non-sequential URL parameters, e.g., for handling pagination, is a typical measure scientific Web repositories employ to impede content mining

2. Robot Detection and Reverse Turing Tests: Many scientific Web repositories implement methods to detect and stop robots.

Reference: https://www.dlib.org/dlib/september16/meschenmoser/09meschenmoser.html

Q3. Describe the basic architecture for performing web scraping.

1. First step, robots.txt: One of the most important and overlooked step is to check the robots.txt file to ensure that we have the permission to access the web page without violating any terms or conditions.
2. Secondly, the web scraper will be given one or more URLs to load before scraping. The scraper then loads the entire HTML code for the page in question. More advanced scrapers will render the entire website, including CSS and JavaScript elements.
3. Then the scraper will either extract all the data on the page or specific data selected by the user before the project is run.
4. Ideally, the user will go through the process of selecting the specific data they want from the page. For example, you might want to scrape an Amazon product page for prices and models but are not necessarily interested in product reviews.
5. Lastly, the web scraper will output all the data that has been collected into a format that is more useful to the user.



Reference: https://www.researchgate.net/publication/347999311_Importance_of_web_scraping_in_e-commerce_and_e-marketing

Q4. Describe various techniques of web scraping.

Different Web scraping methods have been developed in multiple types of research.

1. Traditional Copy and Paste: The copy-pasting method is simple: access the page using your browser, then manually copy and paste it onto other media.
2. HTML Parsing: Extensive collections of pages are produced programmatically from a fundamental organized source, such as a database, on many websites. A common script or template encodes data from the same category into similar pages.
3. DOM parsing: Programs can obtain dynamic material generated by client-side scripts by placing a developed web browser, such as Internet Explorer or the Mozilla browser control.
4. HTML DOM: The HTML DOM (Hyper Text Markup Language Document Object Model) is a yardstick for obtaining, altering, and editing HTML elements. By defining objects and properties for all HTML components, as well as ways to access them, DOM efficiency can be improved.
5. Regular Expression (Regex): Regex is a formula that explains a group of words that spans numerous alphabets and follows a precise pattern. It can be used to match specific character patterns across several strings.

6. XPath: XPath is a node selection language for XML documents that may also be used with HTML. The most useful XPath expression is the location path.
7. Vertical aggregation platform: With no manual intervention and effort tied to a single target site, these systems build and monitor a slew of bots for specific verticals.
8. Semantic annotation recognizing: Metadata, semantic markups, and annotations may be included on the scraped pages, which can discover data pieces.
9. Computer Vision Web Page Analyzer: Machine learning and computer vision are being used to recognize and extract information from web pages in a visual manner, analysing them as a human would. Based on the image of the rendered page, a computer vision-based system is used to analyse the semantic structure of web pages, and a rich representation of the page is produced as a tree of regions labelled according to their semantic role.
10. Comparison between web scraping methods: The comparison is conducted by putting each method to the test when extracting data from the required website, then computation and comparing the results

Reference: https://www.publications.scrs.in/chapter/pdf/download/38