PROJECT REPORT
WEB SCRAPPER

Omar ALLAL
Hamza BOUKHATEM
Skander RADHOUANE
Mohamed Ali BEN GHARBIA

Contents

What is Web Scraping

Web Scraping is an automatic way to retrieve unstructured data from a website and store them in a structured format.
Website scrapers are commonly used for data mining, web research, and competitive analysis. However, they can also be used for malicious purposes, such as stealing copyrighted content or personal information. Therefore, it is important to use website scrapers ethically and legally, and to respect the website's terms of service and privacy policy.

How does web scraping work?

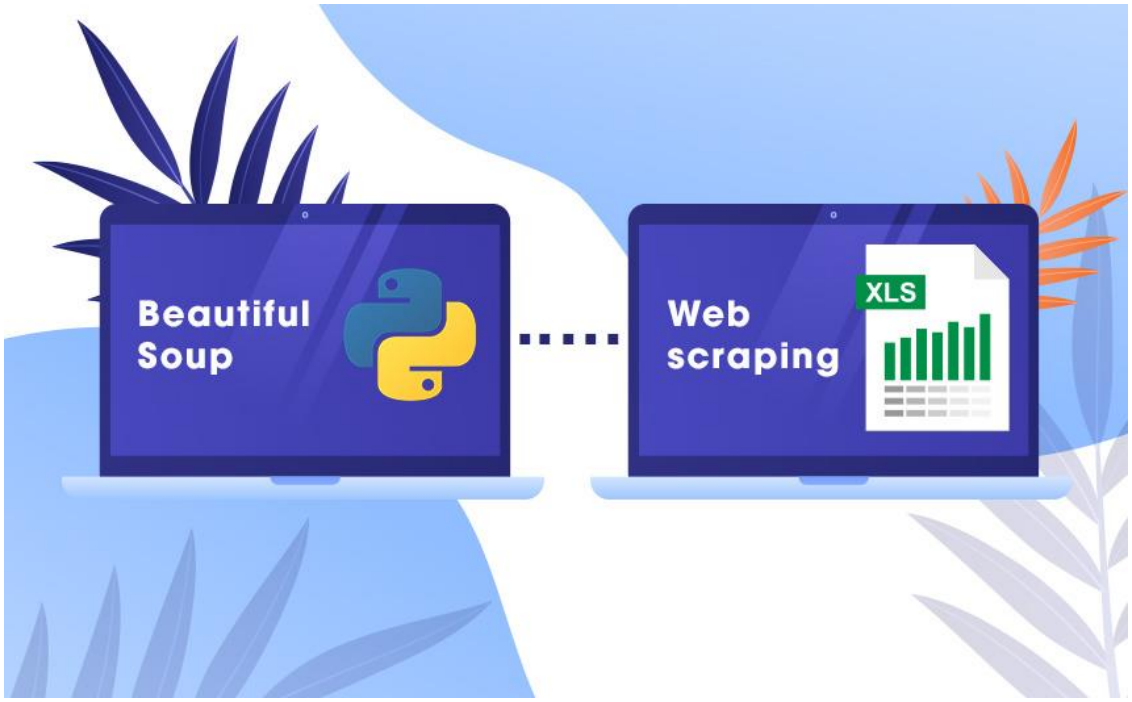
Web scraping just works like a bot person browsing different websites and copy paste down all the contents. When you run the code, it will send a request to the server and the data is contained in the response you get. What you then do is parse the response data and extract out the parts you want.
There are various tools and programming languages that can be used to build a website scraper, such as Python, Scrapy, BeautifulSoup, and Selenium. However, it is important to note that some websites may have anti-scraping measures in place,such as CAPTCHAs or rate limiting, which can prevent or hinder scraping efforts.

Different tools and library for web scraping

Beautiful Soup: It helps you parse the HTML or XML documents into a readable format. It allows you to search different elements within the documents and help you retrieve required information faster.

Requests: It is a Python module in which you can send HTTP requests to retrieve contents. It helps you to access website HTML contents or API by sending Get or Post requests.

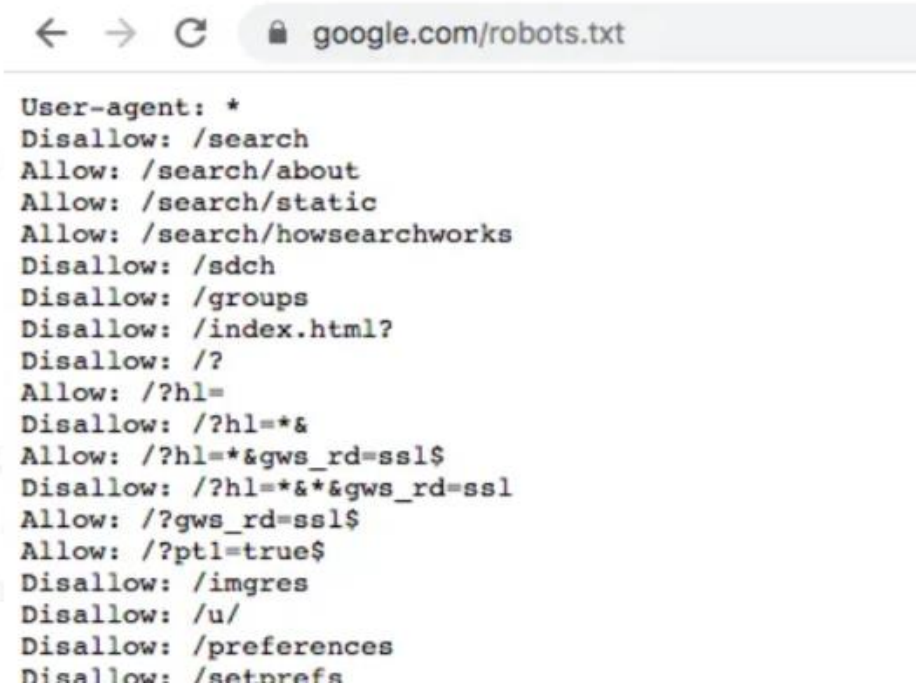
Selenium: It is widely used for website testing and it allows you to automate different events(clicking, scrolling, etc) on the website to get the results you want.



Can you scrape from all the websites?

Scraping makes the website traffic spike and may cause the breakdown of the website server. Thus, not all websites allow people to scrape. How do you knowwhich websites are allowed or not? You can look at the ‘robots.txt’ file of the

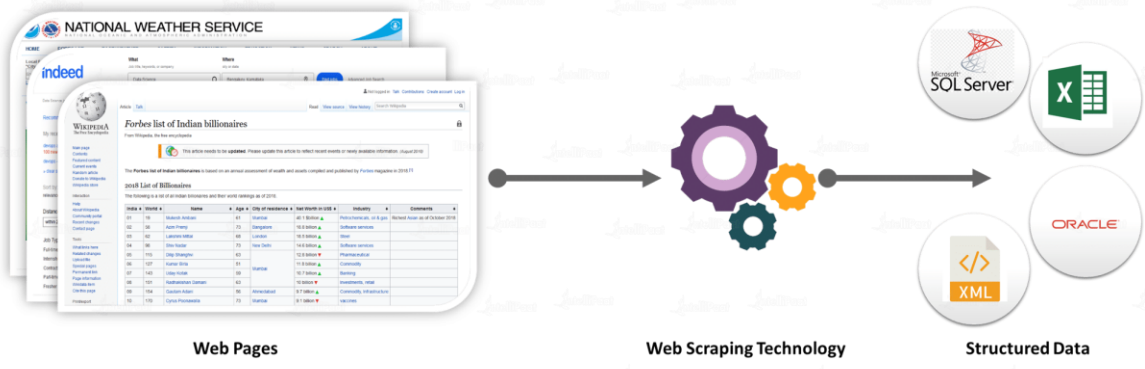
website. You just simply put robots.txt after the URL that you want to scrape and you will see information on whether the website host allows you to scrape the website.



Web scraping workflow

A web scraping project workflow is commonly categorized into three steps:

1. First, fetch web pages that we want to retrieve data from;
2. Second, apply web scraping technologies,
3. And finally, store the data in a structured form.



What are the elements of a web scraping project?

the building blocks you need to take care of, in order to develop a healthy web data pipeline:

- Web spiders
- Spider management
- Javascript rendering
- Data QA
- Proxy management

Web spiders :

Web spiders, also known as web crawlers, are automated software programs that browse the World Wide Web in a systematic, automated manner. They start at a given website and follow links to other pages, and then follow links on those pages to other pages, and so on, until they have visited as many pages as possible within a given time frame or until they have reached a specific depth of the website's hierarchy.

Spider management:

Spider management refers to the process of controlling and managing the behavior of web spiders or crawlers that visit a website. It involves various techniques and tools to ensure that the spiders are behaving ethically and not causing harm to the website or its users; and that is by:

1. Analyzing spider traffic
2. Implementing robots.txt
3. Using rate-limiting techniques
4. Blocking malicious spiders
5. Monitoring spider behavior

Javascript rendering :

JavaScript can be used to dynamically generate content on a website, which can make it more difficult to scrape that content using traditional scraping methods.it gives more complexity to your web scraping project if the data you want to extract is rendered using JS.

To scrape dynamically generated content that is loaded via JavaScript, you can use a headless browser like Puppeteer or Selenium. These tools allow you to programmatically control a browser, including executing JavaScript on the page and waiting for content to load before scraping it.

Data quality:

The next building block is data quality assurance. All our scraping efforts are worth it only if the output data is the right data in the correct format. To make sure this is the case, we can do several things:

- validate the output data against a predefined JSON schema
- check the coverage of the extracted fields
- check duplicates
- compare two scraping jobs

Proxy management:

Proxy servers can be used in website scraping to help manage IP address restrictions, bypass geographical restrictions, and prevent the target website from blocking or detecting your scraping activity.

Web scraping ethics

Before finishing up this article, it’s important to talk about web scraping ethics. When you scrape a website, you have to make sure you also respect it. Here are some best practices you can follow to scrape respectfully.

Don't be a burden

The most important rule when you scrape a website is not to harm it. Do not make too many requests. Making requests too frequently could make it hard for the website server to serve other visitors. Limit the number of requests in accordance with the target website.

Robots.txt

Before scraping, always inspect the robots.txt file first. This will give you a good idea of what parts of the website you are free to visit and what pages you should not.

User-agent

Define a user-agent that clearly describes you or your company. Also, it’s best to include contact information in your user-agent as well, so they can let you know if they have any issues with what you’re doing.

Pages behind a login wall

There are cases when you can only access certain pages if you are logged in. If you want to scrape those pages, you need to be very careful. By logging in and/or explicitly agreeing to the website's terms and conditions that state you cannot scrape, then you CAN NOT scrape. You should always honor the terms of any contract you enter into, including website terms and conditions and privacy policies.

Responsible Web Scraping:

gathering web data without becoming Dr. Evil

SOSHACE