

Few Things to Consider When Choosing a Scraping Tool

TechRounder PDF Edition

Live article: <https://www.techrounder.com/development/few-things-to-consider-when-choosing-a-scraping-tool/>

By Vipin PG | Published June 22, 2022 | Updated May 26, 2023 | Format: Article | 4 min read

In brief

Data provides companies with an all-encompassing view of the market conditions and consumer needs and preferences. It informs better decision-making.

Data provides companies with an all-encompassing view of the market conditions and consumer needs and preferences. It informs better decision-making. The immeasurable value attached to data has given rise to the web scraping industry, fraught with numerous scraping solutions that promise to deliver.

Against this background, choosing the right tool can be a daunting task. Should you use a proxy service, a CAPTCHA resolver, a headless browser, or an all-in-one web scraping API like ZenRows? If you are in this boat, this article is for you. Here, we'll discuss 5 factors to consider when choosing a scraping tool. But first, let's understand what web scraping is.

What is Web Scraping?

Web scraping mainly refers to the automated form of data collection. It utilizes bots known as web scrapers that work based on lines of code that guide them on how to extract data from websites. These scrapers usually extract publicly available data from websites. And depending on the type of bot or service provider, they typically come with additional features that boost the chances of successfully retrieving data without impediments. Successful web scraping offers the following benefits:

- It provides data that guides search engine optimization (SEO) strategies, enabling company's websites to rank high on search engines
- It enables businesses to undertake market research and identify their competitors
- Web scraping provides pricing data that guide pricing strategies
- The data collected can be used to optimize products
- Scraping solutions can extract contact information that is then used to generate leads
- It promotes reputation and review monitoring, allowing companies to protect their brand image

To learn more about web scraping, you should understand how to extract website data in detail, where the process and the associated challenges are explained in depth.

Anti-Scraping Techniques

Websites are a graphical window to the data stored in a server. To safeguard the data stored in their servers and prevent crashes that may occur when the server receives too many requests, web developers deploy anti-bot measures that negatively impact scraping efforts.

These measures include IP blocking, headers, user agents, CAPTCHA puzzles, honeypot traps, mechanisms to monitor the number of requests within a given period, sign-in and login requirements, dynamically changing site layout, complex web structures, and more.

Although anti-scraping techniques are common, that does not mean they cannot be bypassed, especially when this is done as part of ethical web scraping. And this is where scraping tools come in.

5 Things to Consider When Choosing a Web Scraper

Here are 5 considerations you should make when selecting a scraping tool:

Proxy Server

A proxy server or proxy is a computer through which requests are routed before being directed to the target website. It anonymizes the requests by assigning them a new, unique IP address, thereby masking their real identity. When scraping, this online anonymity prevents a web server from blocking your real IP address should it get to this point. Instead, it will block the newly assigned IP address. But you can still prevent this by following the second consideration below.

IP Rotation Capabilities

IP rotators or proxy rotators periodically change the assigned IP address. They achieve this by either assigning a different IP address to each outgoing request or changing the given identifier after a short period. This way, the rotators keep the number of requests originating from the same IP address to a minimum, thereby reducing the chances that it will be blocked.

Ability to Adapt to Layout Changes and Complex Web Structures/Formats

A good scraper should be capable of extracting data from websites that use JavaScript and AJAX, enabling sites to change their layouts or content dynamically. These bots achieve this through the use of headless browsers. At the same time, the scraping tool should be able to scrape data from tables, maps, and pages such as social media sites or review pages that can be scrolled infinitely.

CAPTCHA Solving Ability

Some scrapers use artificial intelligence and machine learning to solve CAPTCHA and reCAPTCHA puzzles. Therefore, it is important to choose a scraping tool with this capability.

Ability to Mimic Human Browsing Behavior

Typically, web servers check traffic volume from a single IP address. They block IPs that make numerous requests within a short time because only bots are capable of such a feat. Thus, your scraper should be capable of mimicking human browsing behavior by limiting the number of requests sent per hour or minute to a realistic figure.

Bonus: the scraper should also be able to surpass login screens, follow the directions stipulated in the robots.txt file, and handle scheduled data extraction projects.

Conclusion

Web scraping, carried out by tools programmed on how to extract data from websites, offers numerous benefits, but it is not without its impediments. To bypass these challenges, you must consider several aspects when choosing a web scraping tool.

These considerations include proxies, proxy rotation tools, CAPTCHA solving capabilities, the ability to adapt to layout changes and complex web structures, and the ability to limit the speed and intensity of scraping.

References

1. zenrows.com - <https://www.zenrows.com/>
2. oxylabs.io - blog / extract-data-from-website - <https://oxylabs.io/blog/extract-data-from-website>
3. whatismyip.com - what-is-a-proxy - <https://www.whatismyip.com/what-is-a-proxy/>