

7 important Things You Should Never Do When Scraping a Website

TechRounder PDF Edition

Live article: <https://www.techrounder.com/tools/crucial-things-you-should-never-do-when-scraping-a-website/>

By Vipin PG | Published November 14, 2022 | Updated March 10, 2023 | Format: Article | 6 min read

In brief

There are many ways to get data from websites, some of which are more ethical than others, like scraping a website. Scraping a website is one way to get data, but you should do it in a way that doesn't violate the site's terms of service or break the law.

There are many ways to get data from websites, some of which are more ethical than others, like scraping a website. Scraping a website is one way to get data, but you should do it in a way that doesn't violate the site's terms of service or break the law. And to ensure you are following the right track, there are some things you should never do when scraping a website. This article discusses some important points regarding what to do and what not to do in website scraping. So, without further ado, let's dive into it.

1. Avoid Sticking to One Web Scraping Tool

Individuals usually adapt to one of these methods to parse web pages: (1) using an HTML parser and (2) using a web scraping tool. Specifically, when using a web scraping tool, many individuals tend to stick to only one tool they have used for a long time. But did you know if you use the same technique, it could block you from several other potentials of web scraping? It means you'll miss out on the benefits of using different tools. Different tools have different strengths and weaknesses. Using multiple tools, you can cover more ground and get better results. For example, if you're looking to scrape data from a website that uses HTTPS encryption, you'll need a tool that supports that type of encryption. Let's say if you are going to scrap one of the biggest public financial resource API websites like Crunchbase, security should be followed at most. In short, for scrapping Crunchbase API, great encryption is needed. Additionally, some online scraping technologies complete particular tasks better than others. For instance, if you're trying to parse web pages that use JavaScript code for their layout and design, you may find that one tool is better at parsing through that code and extracting the data you need. Using different web scraping tools makes you more valuable as a web scraper. And if you ever find yourself in a situation where your preferred tool doesn't work for some reason, being able to fall back on another tool will save you a lot of headaches. So don't be hesitant to try out different web scraping tools. The more you use them, the better you'll get at scraping.

2. Avoid Sending Several Requests

One of the important things you should never do when scraping a website is to send too many requests within a short time frame. If you do, it can trigger a website's security measures and make getting the information you need more challenging. It can also increase your chances of being flagged as a bot or spider by the website's owner, which eventually runs the risk of getting your IP address banned. So, how do you avoid this? One way is to space out your requests so you don't make too many of them quickly. Another way is to use a proxy server, making it look like your requests come from different IP addresses. Either way, you should be careful when web scraping so that you don't get banned and can continue to get the data you need.

3. Don't Overuse Synchronous Requests

Synchronous requests are outstanding for small amounts of data, but when web scraping on a large scale, they can become a bottleneck. Each request must be finished before moving on to the next. Therefore, as the number of requests increases, so does the time needed to complete them. Synchronous requests mean you make a single request to the website that you are trying to scrape and then wait for a response before continuing. So, when you overuse synchronous requests, it can strain the website, resulting in slower performance or downtime. Furthermore, using synchronous queries excessively can make writing and managing your code more challenging. It's generally advisable to use asynchronous requests instead. It will make your scrapes faster and more reliable. However, if you are to use any synchronous requests, it is essential to consider the potential drawbacks of this approach before moving forward.

4. Avoid Scraping Website Data Behind a Login

Don't acquire data from a website that requires logging in! Scraping data from behind a login is both difficult and dangerous. If caught scraping data from a site that requires a login, you could get penalized by the site or even sued. So save yourself the trouble and steer clear of sites that require a login. If you must scrape data from a website with a login, it is best to ask for permission first. Once you have the go-ahead, you can log in and access the data you're interested in scraping. Again, you can use APIs (application programming interfaces) or advanced web scraping tools to do so legitimately. Whichever route you choose, be sure to tread carefully. Scraping data from behind a login can be tricky, so you should do it right.

5. Avoid Website Scraping Without an API

One of the other things you should never do when scraping a website is not to use an API. Usually, an API on the website will make the process much easier. Compared to manually scraping a page, using an API has several advantages.

1. APIs are more efficient. You can access data in real-time and in a more automated fashion.
2. You can submit requests and receive responses in a more user-friendly way. This saves time and effort, especially if you need to make a lot of requests.
3. APIs are more reliable. Websites constantly change, which can break your scraper. An API is less likely to change, so it's less likely to damage your code.
4. You also don't have to worry about keeping up with website layout or design changes. An API will usually remain consistent even as a website changes over time.

6. Never Forget to Save the Raw Data

It's happened to all of us at one point or another - we're scraping a website for data, and we get so caught up in the process that we forget to save the raw data. Then, when we analyze the data, we realize that all we have is a bunch of processed data that will not be useful. This is easily avoidable by simply remembering to save the raw data before processing it. Then, if you want to change how you process the data, you'll have the raw data to fall back on. It's also helpful to have a backup of the raw data if the website changes and you need to re-scrap it. The raw data may take up more space, but it's worth it for the peace of mind and flexibility it gives you.

7. Don't Use Unreliable Scraping Pipelines

If you're scraping a website for data, you must ensure your scraping pipeline is reliable. Without it, you might obtain data that is erroneous or incomplete. There are a lot of scrapers out there that promise to be the best, but not all of them are the best. So, do your research. Look for scrapers that provide high-quality data analysis outcomes, better business insights, or even a more accurate machine learning algorithm. Remember, a reliable scraping pipeline will result in accurate data collection, protect your account from suspension or banning, and be cost-effective. This is why you should only use reliable sources when scraping a website, or all your efforts will be wasted with no results to show for it.

Conclusion

Scraping a website isn't so hard. With the right scraping tools and techniques, you can quickly get the best insights into different web pages for your specific needs. However, as mentioned, you must know what you should never do when scraping a website. Throughout the blog post, we highlighted seven important points everyone should bear while scraping a website. It will ace your performance and ensure driving out high-quality data from the other websites.

FAQs

What Is Web Scraping?

Web scraping involves collecting data such as contact information and product prices or creating a content database. Although it can be carried out manually, automated tools are more frequently used. These tools can extract data from a single website or multiple websites.

Can Websites Detect Scraping?

Yes, there are a few ways that a website can detect scraping. One method is by examining the user agent string included in each request. Another way is to look at the IP address the requests are coming from.

Is Web Scraping Legal?

Web scraping can be used for both legal and illicit activities. There is no clear line between legal and illicit web scraping, and the legality of web scraping depends on the specific circumstances. For instance, it's considered legal if you scrap a website that allows scraping or has publicly accessible content.

References

1. scrapingant.com - blog / scrape-dynamic-website-with-python - <https://scrapingant.com/blog/scrape-dynamic-website-with-python>

2. scrapfly.io - blog / how-to-scrape-crunchbase - <https://scrapfly.io/blog/how-to-scrape-crunchbase/>