Suma-

Like any powerful tool, web scraping can be incredibly helpful, but it can be malicious when misused. Web scraping is essentially extracting data from websites, and although it is legal as long as it uses publicly available data, one should always be respectful and empathetic and create an “ethical” web scraper.

The rules to follow when creating a web scraper are to use publicly available data, avoid crashing servers, and share what we can for others to benefit from. Recursive scrapers should be tested on a smaller dataset first to ensure that they don’t send unnecessary queries and won’t slow down servers. Next, to follow ethical collection of data, the data must be publicly available and not copyrighted. To scrape data that requires purchase, or logging into an account, such as academic journals, agreements should be made with the publisher or website owners, and all original publishers or owners should be credited for their data. Robots.txt files can also be checked to learn how the website should be crawled and indexed through the format <domain\_url>/robots.txt. Apart from collecting data, how we use the data also matters. For example, scraped data can help hackers to perfect phishing attacks or improve password cracking algorithms. However, rather than use web scrapers for harm, if we get permission to share our data, it can be shared to benefit others who might want to use relevant data for other projects without re-scraping.

Ezra Cohen-

There are two major ethical concerns when it comes to web scraping, the first being that it repeatedly queries the server, and the second regarding copyright. The problem with repeatedly querying a server is that it overwhelms it, which makes it harder for the server to respond to actual humans trying to access it. This is an incredibly effective way of shutting down servers, to the point where some people intentionally do it, this being a DOS attack, as this is not our purpose we should try to avoid having our programs cause this effect. The next issue deals with copyright, sometimes websites have rules about what can and cannot be scraped, so if a website prohibits it we should not go ahead with scraping. Even if it is permitted, we should never try to pass any of the data as our own.

The best practice is to first try to see if the website has an API that allows for easy access of the data without scraping, if a website doesn’t we should check robots.txt and the terms and conditions to see what we are and aren’t allowed to do. Next we should either ask for permission from the owners of the website or include a user-agent string that tells the owner what we are doing and how to contact us. While scraping we should only take what we need from the websites, and be careful to not overwhelm the websites. And finally after scraping we should never try to pass any of it off as our own, credit the original websites that had the data so as to create more traffic towards those websites as thanks, and when publishing do it in a way that can be easy for others to use, like providing an API so that others don’t need to scrape our data.

Hawa Iman-

Ethical web scraping is critical for retaining access to data from specific websites. However, if a website disables you after scraping data, this might be a problem. One of the most crucial rules to follow when it comes to ethical web scraping is to only use it when you don't have any other options. Before you begin web scraping, you should check to see if the site in question has a public API and if it has the data you want. You may then avoid web scraping that website while still obtaining the information you want. You may utilize the data in the same way that you would if you scraped the web. However, you won't have to worry about being banned or having other problems.

The first and most crucial thing to keep in mind when creating a web scraper is that it often requires frequently querying a website and obtaining a potentially huge number of pages. A request will be made to the web server that is hosting the site for each of these pages, and the server will have to process the request and return a response back to the machine that is executing our code. Each of these requests will require server resources at which time the server will not be doing anything else, such as replying to another request to visit the same site. It is also vital to understand that online scraping might be unlawful in some instances. If the terms and conditions of the website we are scraping expressly prohibit downloading and duplicating its material, we may face legal consequences for scraping it. Keep in mind that copyright and data privacy laws vary from nation to country. Check the laws that apply in your situation. Scraping and storing personal information such as names, phone numbers, and email addresses, for example, might be prohibited in Australia, even though the information is publicly available.

Jeffrey Zhou-

Web scraping is the use of bots to extract data from a website. Web scraping can have legitimate uses, and many users and businesses rely on web scraping for data harvesting to gather prices, data from forums, or content from a site. However, there are some legal and ethical considerations to consider when web scraping. Firstly, one should be careful not to “break the web” with their web scraper which works by querying websites repeatedly to access data. These requests which must be handled by the web server must be processed and returned to the computer running the web scraper. This consuming of resources on the server takes away time allotted to other tasks such as someone else accessing the same site. Overloading a site can happen when too many requests of a site are sent over a short span of time, preventing other users from accessing the site and possibly causing the server to crash.

When web scraping, one must also consider the possibility that their actions may be illegal due to copyright or data privacy legislation. For example, if a website prohibits the downloading and copying of its content but you still scrape, you could be in trouble for violating its terms and conditions. Usually, web scraping is tolerated if regular use of the site is not disrupted and if the data being scraped is publicly available. One should refrain from sharing the data further which can be problematic. For example, some might try to pass the data off as their own or try to make money off it which is not “fair use.” Lastly, before web scraping for a project, one can try asking the owners of the data if they already have it available in the correct format which can save time and help clarify what can and cannot be done with the data. When publishing the data that you have collected, you can also consider sharing the data with others who might be interested in using it for their own project, saving them the trouble of scraping the data for themselves.

Diba Demir

Although web scraping is considered harmless, as users are accessing websites in the public domain, following guidelines is necessary. Since every website is owned by a certain person or organization, extracting information should be done with care. Web scraping requires multiple requests to be sent to web pages, and every one of these requests occupy the resources on the server. If a web scraper were to send too many requests in a short time period, other users would be unable to access the information on the page. When hackers utilize this weakness, it is called the Denial of Service attack. If web scraping isn’t done carefully, such accidents may occur.