Web Scraping Write-Up

Web Scraping Overview: Web scraping is one of the most efficient ways of accessing information from a website and can help to sort through data to find whatever information wanted. Typically, the information gathered is put into a word or excel document, so it is easier to read and understand. While web scraping is not illegal, it must only be used to access public information. If it accesses any information intentionally hidden by the maker of the website, then it can be considered illegal. When using a scraper, you will give it a URL and then the scraper loads all the html code from the website. With some scrapers (like mine) it will change a certain part of the URL so if can access different pages of a website. Then the scrapers take the elements of the website and extract certain data chosen by the user. Web Scrapers can each be vastly different. Some may only work on one website, such as eBay or amazon, while others may work on many. They could extract only prices, reviews, sellers, job listings, or URLs listed on the websites. This means that there can be a variety of Web Scrapers and they can complete a variety of tasks designed by the maker.

Understanding Pip / libraries: In my Web Scraper, I had to use various libraries such as beautifulsoup4 and the requests library. To download these libraries, I used pip, or the Package Installer for Python. This allows me to import and download libraries to use in my code. One the libraries I previously mentioned, beautifulsoup4, helps isolate links from a web page, and extract text from an HTML. Another library that I used in my program was the requests library. This allows me to make requests to access a website, such as eBay, Amazon, or Wikipedia. However, this library cannot get into a website if it has a captcha, and it needs further code to help bypass it. One library that can be used to bypass the captcha is the selenium library. This library has numerous ways to get around a captcha, and access elements that can be used in a program. I did not use this library because i used a website without a captcha.

My Web Scraper: My Web Scraper takes a base eBay URL and changes a small part of it to go other eBay pages. The part that changes is a user entered value that looks for the item entered. Then it goes through all the prices, reviews, and sellers on the page. It then prints these out all prices, sellers, and reviews into neat groups. The program will also prompt the user to enter a maximum price that they are willing to pay for the item. It will then sort out any item that is priced over the amount you are willing to pay. This allows for a much more convenient way to find an item that is within your price range. When the program cannot access the website that was entered it will throw a status code of 0 showing that it has failed. The biggest problem that I was faced with when making this code was getting past the captcha. At first, I wanted to make the same program but for amazon instead. When I tried this however, the captcha blocked me from taking any information from the website. What i did to get around this problem was use a different website with no captcha, such as eBay.

Ethics: When making this project I had to ask myself a lot of ethical questions to make sure what I was doing was morally correct and within the confines of the law. To prevent being immoral on this project I decided to extract information that is easily avalible to any user of the website being accessed. Also, to avoid breaking any laws, I made sure not to access any information purposefully hidden in the source page.

Works Cited:

Perez, Martin. “What Is Web Scraping and What Is It Used For?” *ParseHub*, Web Scraping Blog (Tips, Guides + Tutorials) | ParseHub, 14 Apr. 2023, www.parsehub.com/blog/what-is-web-scraping/.

“Educative Answers - Trusted Answers to Developer Questions.” *Educative*, www.educative.io/answers/what-is-beautiful-soup. Accessed 26 Feb. 2024.

Miranda, Lucas. “Understanding Pip - The Package Installer for Python.” *Medium*, Medium, 31 July 2023, [medium.com/@pdx.lucasm/understanding-pip-the-package-installer-for-python-d3401de7072a](mailto:medium.com/@pdx.lucasm/understanding-pip-the-package-installer-for-python-d3401de7072a).

“Python’s Requests Library (Guide).” *Real Python*, Real Python, 17 Sept. 2022, realpython.com/python- requests/.

Zilberman, “Shades of Grey: The Good, the Bad, and the Ugly Side of Web Scraping – Radware Blog” [www.radware.com](http://www.radware.com) 30 Mar. 2016 [Shades of Grey: The Good, The Bad, and The Ugly](https://www.radware.com/blog/security/2016/03/good-bad-and-ugly-web-scraping/)  Side of Web Scraping – Radware Blog

GfG. “Selenium Python Tutorial.” *GeeksforGeeks*, GeeksforGeeks, 13 July 2023, www.geeksforgeeks.org/selenium-python-tutorial/.