|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Technical  Documentation | |  | | |
|  |  |
|  |  |
| Inside this Report | |
| PG. 3  Why is Web Scrapping Used for?  PG. 5  Libraries used for Web Scraping  PG. 7  Available Web Scraping Tool | | |  | Web Scrapping ToolsWhat is Web Scrapping? **Web scraping** may be accomplished in a variety of methods, including manual data collection (simple copy/paste), bespoke scripts, and web scraping platforms.  Author: - Pandula Pallewatta  Data Analyst Consultant (Part – Time)  Table Contents  [Inside this Report 1](#_Toc109232587)  [Web Scrapping Tools 1](#_Toc109232588)  [What is Web Scrapping? 1](#_Toc109232589)  [Why We Use 3](#_Toc109232590)  [Brief Go through of Web Scrapping 3](#_Toc109232591)  [Libraries used for Web Scraping: 4](#_Toc109232592)  [Identification of Web Tools 4](#_Toc109232593)  [Demonstration Site used for scrapping 4](#_Toc109232594)  [Improving Scrapping Tool 7](#_Toc109232595)  [Factors to Consider when Choosing Web Scraping Tools 7](#_Toc109232596)  [Top Web Scraping Tools 7](#_Toc109232597)  [Target Audience 8](#_Toc109232598) |

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | Why We UseBrief Go through of Web Scrapping There are many things you can do with the extracted data from web scraping like:   * **Competitor research:**   Learn how your rivals are pricing their items or discover the phrases they're targeting.   * **Financial data:**   Used to scrape financial information such as stocks, income statements, balance sheets, and stock market news.   * **Gather data for research:**   Some large data websites and libraries have the information you need for your research; you may scrape and export the data from these websites.   * **Industry insights:**   You may scrape articles, stocks, prices to determine how well a given industry is going.   * **Generate leads:**   Numerous web scrapers will trawl internet directories to identify firms in their target market and compile contact lists.  ***Web scraping*** enables data scientists to obtain web data more effectively and is an essential ability for data scientists. Since data science entails the collection of **internet data, many data scientists employ web scrapers** to assist them. Manual and automatic web scraping are both possible, however automated web scrapers are faster and more efficient.  Why Web Scrapping is needed to be followed  **Web scraping** is essential to the field of data science. It was one of the numerous technologies required to capture web data efficiently and successfully. Since one of the first steps in data analysis is data collection, web scraping can facilitate the initial step. |
|  |  | |  |
|  |  | |
|  | |  |
|  | |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |
| Libraries used for Web Scraping:Identification of Web Tools As we know, Python is having various applications and there are different libraries for different purposes. In our further demonstration, we will be using the following libraries:  **1) Selenium:**  Selenium is a web testing library. It is used to automate browser activities.  **2) BeautifulSoup:**  Beautiful Soup is a Python package for parsing HTML and XML documents. It creates parse trees that is helpful to extract the data easily.  **3) Pandas:**  Pandas is a library used for data manipulation and analysis. It is used to extract the data and store it in the desired format. Demonstration Site used for scrapping So, with this, we will get started with writing our own code for **Laptops for gaming under 80000/-** on Flipkart website "<https://www.flipkart.com/search?q=best%20laptops%20under%2080000&otracker=search&otracker1=search&marketplace=FLIPKART&as-show=on&as=off>"  I will add some of the Jupyter notebook code segments utilized for the example.  **GitHub Repo: -** [**Web Scrap Demo**](https://github.com/Pandula1234/Data-Analyst-Consultancy/tree/master/Web%20Scrapping)    *Figure 0‑1 Installing Bs4*    Figure ‑ URL Scrapping    Figure ‑ Beautiful Soup is a Python package for parsing HTML    Figure ‑ HTML container analyzation  Figure 0‑5 Data Retireval from conatiners in the div body    Figure ‑ Data retrival from the URL | | |

Figure 0‑7 ParseHUb Table

|  |  |  |
| --- | --- | --- |
| Improving Scrapping ToolFactors to Consider when Choosing Web Scraping Tools |  |  |
|  |  |
|  |  |
| Most of the data on the Internet is not set up in a certain way. Because of this, we need to set up systems to get useful information from it. Web scraping is one of the most important things you have to do if you want to play around with data and get some useful information from it. Top Web Scraping Tools [ParseHub](https://www.parsehub.com/)  **ParseHub** is an incredibly powerful and elegant tool that allows you to build web scrapers without having to write a single line of code. It is thus as simple as selecting the data you require. ParseHub is intended for anyone who wants to experiment with data. Analysts, data scientists, and journalists could all fall within this category.  Key Features of **ParseHub**   * Clean Text and HTML before downloading data. * Simple to use graphical interface. * **ParseHub** allows you to collect and store data on servers automatically.  **ParseHub Pricing****Parse Hub’s** pricing structure looks like this:  * **Everyone**: It is made available to the users free of cost. Allows 200 pages per run in 40 minutes. It supports up to 5 public projects with very limited support and data retention for 14 days.   [Scrapy](https://scrapy.org/) Target Audience **Scrapy** is a Web Scraping library used by python developers to build scalable web crawlers. It is a complete web crawling framework that handles all the functionalities that make building web crawlers difficult such as proxy middleware, querying requests among many others.  Key Features of Scrapy   * Open-Source Tool. * Extremely well documented. * Easily Extensible.  **Scrapy Pricing** It is an open-source tool that is free of cost and managed by Scraping hub and other contributors | | |
| Figure 0‑1 Scrapy Tool **Shortcomings** In terms of JavaScript support, inspecting and developing the crawler to emulate **AJAX/PJAX** requests takes time.  [OctoParse](https://www.octoparse.com/)  **OctoParse**, like ParseHub, caters to users who wish to scrape data without having to write a single line of code, while still having complete control over the entire process thanks to their incredibly intuitive user interface. **Key Features of OctoParse** For users that want to run scrapers on the cloud, there is a site parser and a hosted solution.  Screen scraper that allows you to scrape behind login forms, fill in forms, render **JavaScript**, navigate through the infinite scroll, and many other things. **OctoParse Pricing**  * **Free**: This plan offers unlimited pages per crawl, unlimited computers,10,00 records per export, and 2 concurrent local runs allowing you to build up to 10 crawlers for free with community support.     Figure 0‑2 Octoparse **Shortcomings** If you run the crawler using local extraction rather than cloud extraction, it will automatically terminate after 4 hours, making the process of recovering, saving, and starting over with the next set of data very time-consuming.  [Scraper API](https://www.scraperapi.com/)  **Scraper API** is intended for web scraper designers. It supports browsers, proxies, and CAPTCHAs, so raw HTML from any website may be accessed with a simple API call.  Figure 0‑3 Scraperapi | | |
| **Scraper API Pricing** Scraper API offers 1000 free API calls to start. Scraper API thereafter offers several lucrative price plans to pick from.   * **Hobby($29/month)**: This plan offers 10 Concurrent requests, 250,000 API Calls, no Geotargeting, no JS Rendering, Standard Proxies, and reliable Email Support.   [Mozenda](https://www.mozenda.com/)    Figure 0‑4 Mozenda  **Target Audience**  Mozenda is a cloud-based self-service Web Scraping software for businesses. Mozenda has scraped over 7 billion pages and has enterprise customers all around the world.  **Key Features of Mozenda**   * Offers point and click interface to create Web Scraping events in no time. * Request blocking features and job sequencer to harvest web data in real-time. * Best customer support and in-class account management.   **It then highlighted** the critical factors to consider when making an informed decision about purchasing a **Web Scraping Tool**, followed by a sneak peek at the best Web Scraping Tools on the market based on a variety of characteristics. A user should select the Web Scraping Tools that are best suited to their needs. Extracting complicated data from a broad mix of data sources might be difficult.  **References**   1. Anon (n.d.) *What is the best way to scrape data from a website?* Quora. <https://www.quora.com/What-is-the-best-way-to-scrape-data-from-a-website> [Accessed: 12 July 2022]. 2. Wu, S. (2022) *Web Scraping Basics. How to scrape data from a website in… | by Songhao Wu | Towards Data Science*. 3 May 2022. Medium. <https://towardsdatascience.com/web-scraping-basics-82f8b5acd45c> [Accessed: 14 July 2022]. 3. Phaujdar, A. & Agrawal, V. (n.d.) *8 Best Web Scraping Tools - Learn | Hevo*. Learn | Hevo. https://hevodata.com/learn/8-best-web-scraping-tools/ [Accessed: 16 July 2022]. 4. Anon (2019) *14 Best Web Scraping Tools For Developers - ScraperAPI*. 31 December 2019. ScraperAPI. https://www.scraperapi.com/blog/the-10-best-web-scraping-tools/ [Accessed: 15 July 2022]. | | |
|  | | |