What gives your company a sustainable competitive advantage in the era of increasing digitalization, where several new technologies arise, where the walls that kept new entrants away are collapsing, where transaction as well as communication costs are decreasing, where computational power is almost available to everyone, and where more and more powerful algorithms are forged? The answer is straightforward: data is the main factor that will determine whether you will be able to keep up with your competitors. The more data you have that your competitors cannot gain access to, the stronger the competitive advantage.

You wonder why? It might be true that there are more powerful algorithms emerging and that the computational power available becomes more accessible to everyone, but the important point is, who owns the data that we will use to make decisions and to improve the company? If everyone has access to the algorithms and computational power, then the only thing the competitors won't have access to will be your data. And one way to stay ahead, is the understand, how you can harvest additional data with web scraping. This article will show you have you can use web scraping and crawling to gather further data for your company.

Web scraping is the process of automating the data extraction from the World Wide Web in an efficient and fast way. This is at the heart of market research and business strategy, for instance when you want to compare the prices of your online-store to the prices of the competitors regularly.

In this article we will go through the advantages of web scraping, the applications of web scraping and finally all possible forms of web scraping for your company. Depending on the strategy of your company, the goal of the web scraping and the complexity of the website to be scraped, different forms of web scraping might be preferable. At the same time, if you are just an individual data scientiest looking for a good introduction into the web scraping world, this article will also give you first good insight on how to precede.

Table of Contents

- Business Area & Impact: Create a constant stream of incoming data
- Approaches: From simple static scraping to automated browsing
    - Complexity of Scraping
    - Web Scraping Approaches Explained
        - Static Webscraping
        - Automated Browsing (Selenium)
        - Application Programming Interface (API)

        - Intercepting AJAX Calls
        - Web Scraping Tools
        - Web Scraping Services
    - Comparison of the Approaches
- Advantages: Fast, efficient and reliant
- Disadvantages: Complexity determines costs of scraping
    - Share this:

Business Area & Impact: Create a constant stream of incoming data

There is hardly no area, where web scraping does not have a profound influence. Where data is increasingly becoming a main resource to compete, acquiring the data has also become especially important.

- **Marketing & Sales:** Web scraping can help you with gathering additional leads, analyzing people's interests, and monitoring consumer sentiment by regularly extracting customer ratings from different platforms
- **Competitor Analysis & Pricing:** If your business is working on a pricing strategy, web scraping could help you extract the pricings of your competitors. Furthermore, you could track all moves of your competitors on the news, the development of the competitors as well as their discounts and pricings on aregular basis.
- **Strategy Development:** For developing a strategy, you often need hard facts. For this, scraping could be useful to conduct a one-time extraction for an initial analysis and to monitor the strategy later. Furthermore, you might want to regularly capture the latest trends in the industry, so you could develop a web crawler that checks the news in the area relevant for your company.
- **Product Development:** If you need the customer ratings on platforms like Amazon or the product descriptions, then web scraping is also a valid option.
- **PR & Brand Management:** Web scraping could help you extract information about how often your company was mentioned on the World Wide Web and what the associated sentiment is. That way your business could identify any negative development early on and prevent that the brand is being damaged.
- **Risk Management & Compliance:** Web crawlers can also be used to conduct automated background checks to ensure, that everything is running smoothly for

your company. Furthermore, it could help you crawl legal databases. Another interesting development is that web crawling is increasingly used to detect fraudulent reviews where fraudsters are writing fake reviews for your products.
- **Business Intelligence:** You can use web scraping to enrich your machine learning data and to improve your machine learning model. Additionally, you can enrich different reports with additional data that is only available on the internet.

Approaches: From simple static scraping to automated browsing

Complexity of Scraping

In order to show you the advantages and disadvantages of each method, we will have a look at the following categories mentioned below. For each category, we will assign a score ranging from 1 (poor performance) to 5 (very good performance).

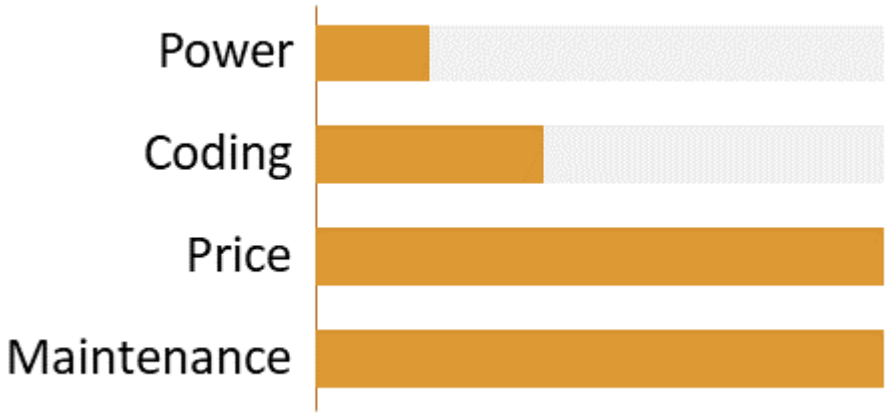
1. **Power:** This category tells you how well this approach can deal with homepages with a complex structure. A low flexibility (1) indicates that I can only scrape simple static homepages, whereas a high flexibility (5) indicates that this approach can also master complex web pages with several exceptions and that require interaction with the homepage.
2. **Coding:** This category indicates how coding intensive the approach is. If the application involves lots of coding with complex algorithms, then it will score low on this category (1). If approach can be realized without any coding, it will receive a high score (5).
3. **Price:** This category indicates how costly this approach is compared to the others. A very costly approach will score low (1), while a less costly approach will score high (5).
4. **Maintenance:** This category will rate the associated maintenance effort with each approach. High maintenance efforts will result in a low score (1), little maintenance efforts will result in a high score (5)

Web Scraping Approaches Explained

Static Webscraping

Static webscraping only extracts the html code without any interaction

Almost every programming language you will use will have a library that will let you scrape dynamic pages, or at least, that will let you send GET-request through the internet. For Python it would be for instance Scrapy, and for R it would Rvest. This is simplest coding-approach, that can let you extract a high amount of data in a short time. However, it is also the least powerful coding based approach. You will be able to scrape only static homepages. As soon as the structure of the homepages becomes more complex or interaction with the homepage is required, the approach fails.



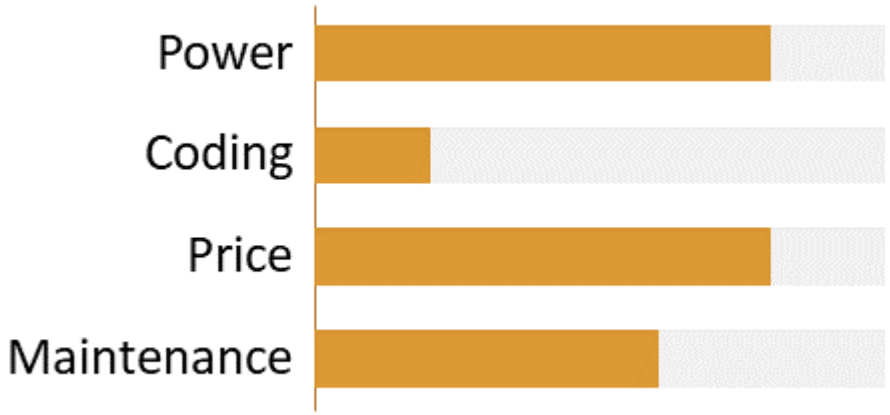
Scores for static webscraping

Automated Browsing (Selenium)

With selenium you automate everything you would do on a simple browser

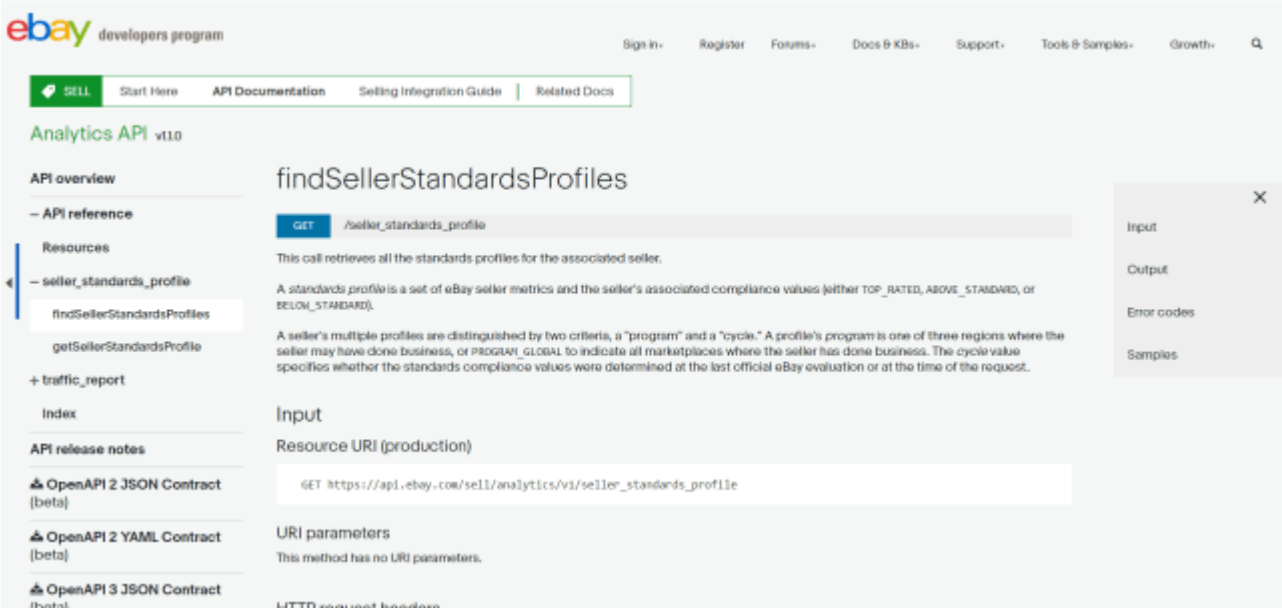
Automated browsing is also based on a programming language. A programmer basically writes down in a programming language that supports Selenium (Python, R, Java and more)

the instructions, what should be done in a Browser. In the backend, you automate all the steps that you would usually do manually on your browser (for example type in the URL and then press enter, click on the first link in the navigation, copy the values from a certain area and paste them into an local excel sheet). The written script will then execute all your instructions by opening a browser and simulating each step as if a human was behind the steps. This is a rather more complex approach compared to simple static webscraping, but at the same time a much more powerful approach, because you can scrape AJAX based homepages, interact with homepages to retrieve certain information that would not be accessible otherwise. At the same time you can undergo several security measures because from the other side it will look like a normal human is accessing the homepage.



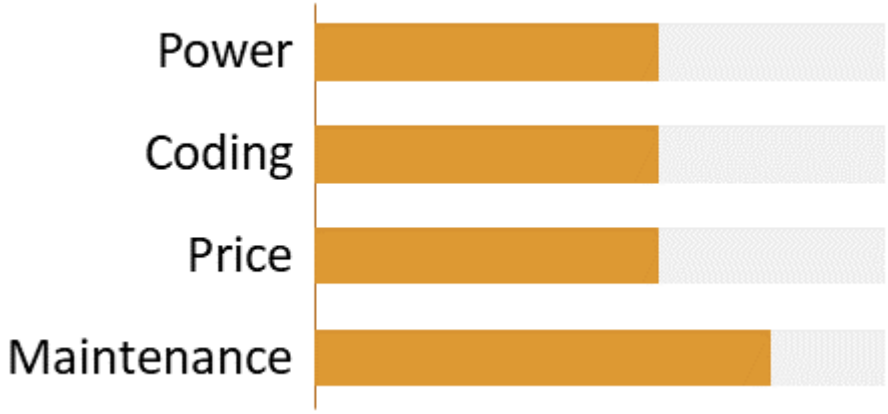
Scores for automated browsing

Application Programming Interface (API)



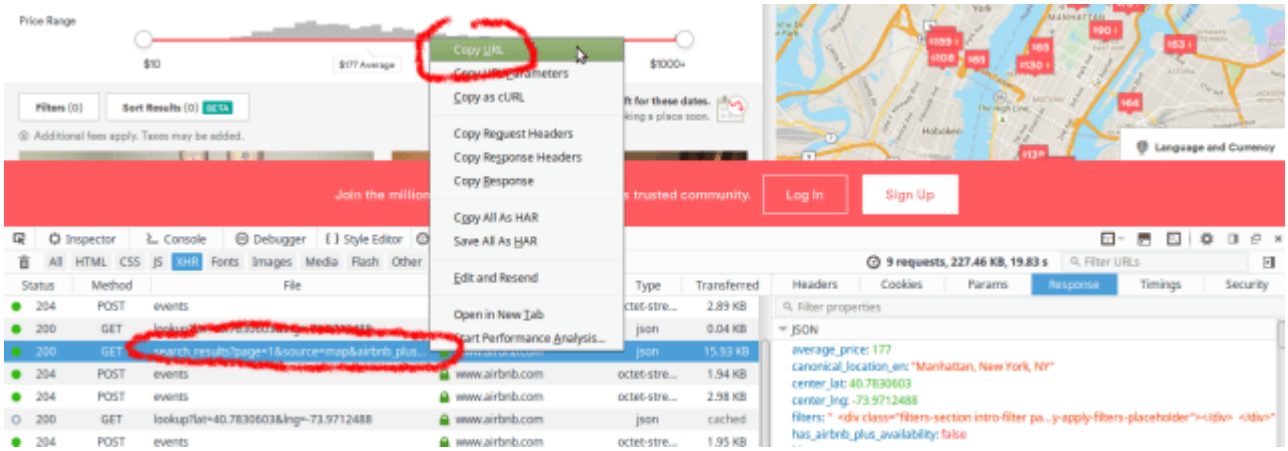
Ebay for instance has an extensive and wide API library that let's you access the data directly

Many homepages and internet-based companies provide own APIs in order to let you access their data. This makes the scraping process much easier and faster, as the data can be scraped with little amount of coding and will be provided in a format that is ready for use. However, the disadvantage of the official APIs is that there usually not for free and cost depending one the amount of the data that you want to scrape. Additionally, APIs are less flexible, because you will be only able to scrape data, that the homepage owner lets you scrape.



Scores for APIs

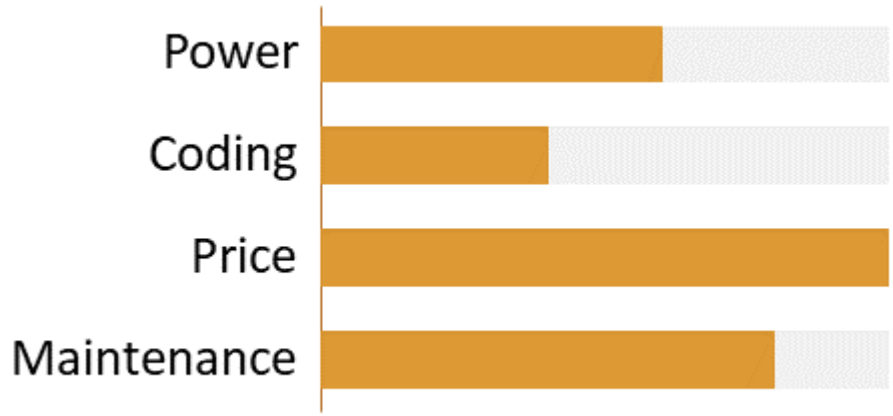
Intercepting AJAX Calls



A hidden API can be spotted by watching the traffic when you access the homepage

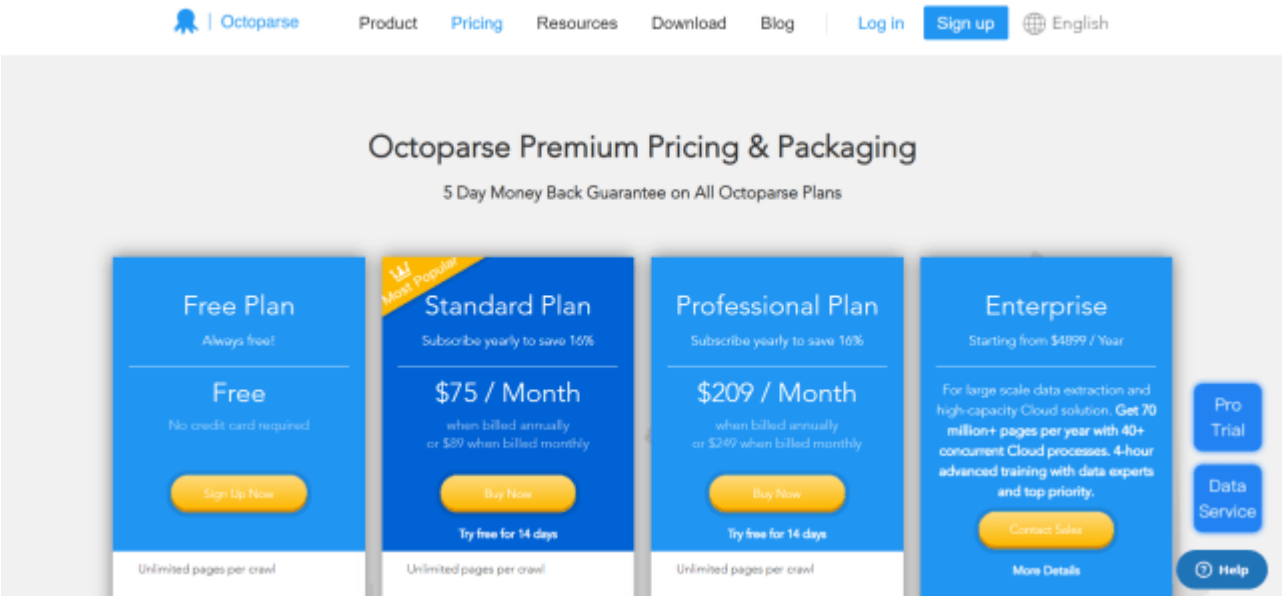
Even if the homepage you want to scrape does not provide an official API, there are chances that there is a "hidden API", escpecially if the hompage works with AJAX-calls. A proficient programmer could easily access the AJAX-interface, send requests with little code and extract all information necessary in an easy interpretable format like JSON. While this approach can give you access to large amounts of data, it is generally less flexible and requires advanced knowledge of how homepages are developed. If you want to know more about hidden APIs and how to implement them, then I would suggest you consult the following two homepages:

- https://www.codementor.io/codementorteam/how-to-scrape-an-ajax-website-using-python-qw8fuitvi
- https://ianlondon.github.io/blog/web-scraping-discovering-hidden-apis/



Scores for intercepting AJAX calls

Web Scraping Tools

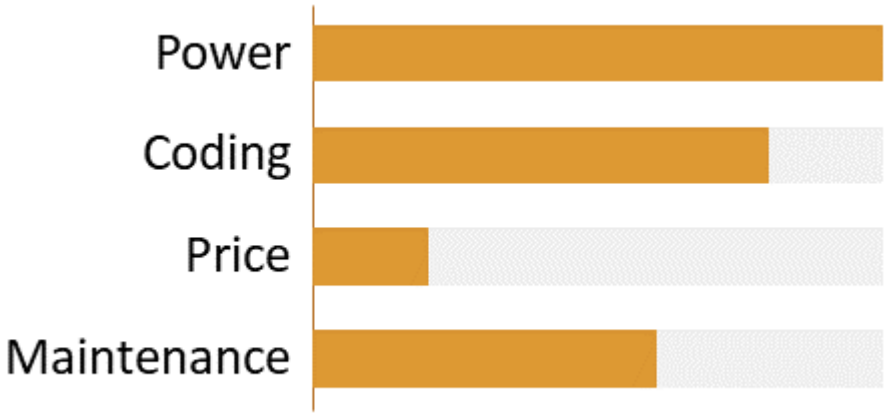


Octoparse is a popular low-code tool for web scraping with a free tier

There is a vast variety of different web scraping tools that will suit your need and help you implement your web scraper with little coding. There are different tools ranging from very powerful ones that regularly change the IP-address and can overcome even captcha, to simple ones that can simply scrape only static homepages. There are tools that can help you scrape data regularly on a continuous basis or that can help you conduct a one-time scraping. Many tools also offer additionally customer support. The only advantage of this approach is that it is very costly depending on the capabilities of the tool. Some tools like Octoparse, let you scrape data for free up to a certain limit. Here is a description of the abilities of Octoparse:

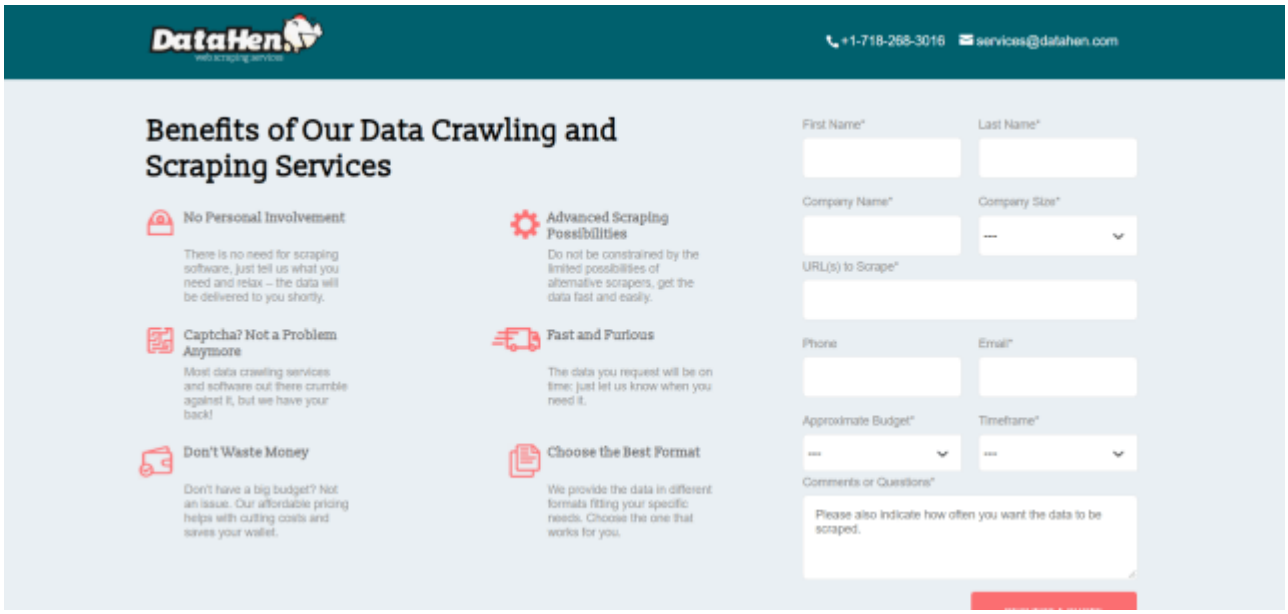
"Octoparse is a fantastic tool for people who want to extract data from websites without having to code. It includes a point and click interface, allowing users to scrape behind login forms, fill in forms, input search terms, scroll through infinite scroll, render javascript, and more. It also includes a hosted solution for users who want to run their scrapers in the cloud. Best of all, it comes with a generous free tier allowing users to build up to 10 crawlers for free."

In case you want to dive further into this approach, here is a homepage that compares 10 web scraping tools.



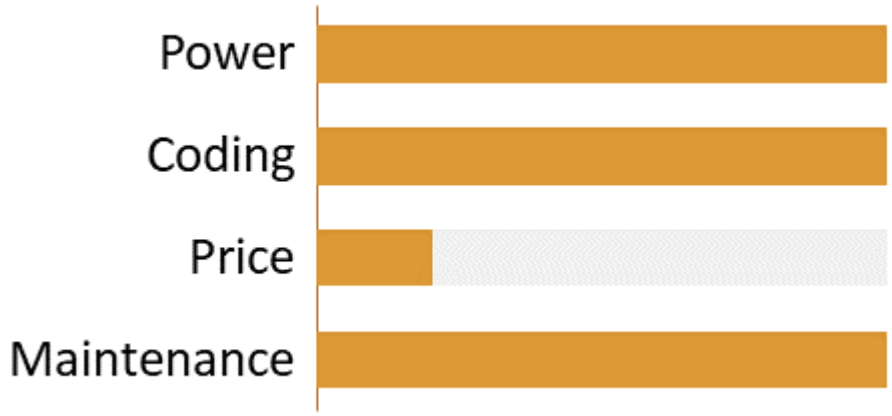
Scores for web scraping tools

Web Scraping Services



DataHen is fast provider and ideal for getting started or if you need to get it done quickly

This is the approach to go if you plan to outsource the scraping completely. From your side, all that is required is to hire a web scraping service and to explain exactly what information you need and the rest will be taken care of by the service. This approach is especially useful for one-time scraping. However, this approach can also be quite costly. A popular web scraping service is DataHen, that is regularly recommended. To get more information on the pricings for services, I would recommend you consult the following quora post, which explains the pricing for different services.



Using services removes heavy work form you to focus on the analysis part

Comparison of the Approaches

Approach	Power	Coding	Price	Main.
Static Web Scraping	2	2	5	5
Automated Browsing (Selenium)	4	1	2	3
Application Programming Interface (API)	3	3	3	5
Intercepting AJAX Calls	2	4	5	4
Web Scraping Tools	5	3	1	3
Web Scraping Services	5	5	1	5

Comparison of all web scraping approaches

When choosing the right approach, you should consider whether you want to outsource the web scraping process or develop it internally. For your web scraping project, try to keep it a simple as possible. That implies that you should only use powerful tools, if they are really necessary. If you settle for a complex approach that is not required, you will overspend on maintenance and features that are not required.

Advantages: Fast, efficient and reliant

Web scraping offers several advantages including the following ones:

1. **Faster:** What would take days or weeks to be extracted by manual work, scraping can reduce effort substantially and increase the decision speed.
2. **Reliable & consisted:** Manually scraping data will easily lead to errors, e.g. typos, forgotten information or information put in the wrong columns. Automating the

scraping process ensures data consistency and quality. Furthermore, you can instruct the scraper directly to sort, organize and put the data in the format you desire without any extra manual effort.

3. **Less costly:** Once implemented, the overall cost of extracting data is reduced significantly, especially if you compare it to the manual work that would be necessary to scrape the data.
4. **Organized:** The scraper can be scheduled to scrape data on a regular basis or at the occurrence of certain events (e.g. when new data is available) at any time. That way you can rely upon the fact that you will always have the most recent data.
5. **Low maintenance:** Web scrapers usually do not require a lot of maintenance over longer periods of time.

Disadvantages: Complexity determines costs of scraping

While web scraping can provide the company with tremendous benefits, there are also a few downsides and assumptions it rests on:

1. **Less complex pages:** the more complex the homepage is you want to scrape, the more difficult scraping will become. The reasons are two. First, setting up the scraper becomes more difficult, and second, maintenance costs can increase, because your scraper is more likely to run into errors.
2. **Stable homepage:** Automated web scraping makes only sense if the target homepage does not change its structure frequently. Each structure change implies additional costs, because the web scraper also needs to be adjusted.
3. **Structured data:** Web scraping will not work, if you want to scrape data from 1000 different homepages and each homepage has an entirely different structure. There will need to be some basic structure that differs only in certain situations.
4. **Low protection:** If the data on the homepage is protected, then web scraping can also become a challenge and drive up the costs. A simple form of protection is for instance captcha, when the homepage requires you to log in or when the data is only accessible through API's that cost.

- Click to share on LinkedIn (Opens in new window)
- Click to share on Twitter (Opens in new window)
- Click to share on WhatsApp (Opens in new window)