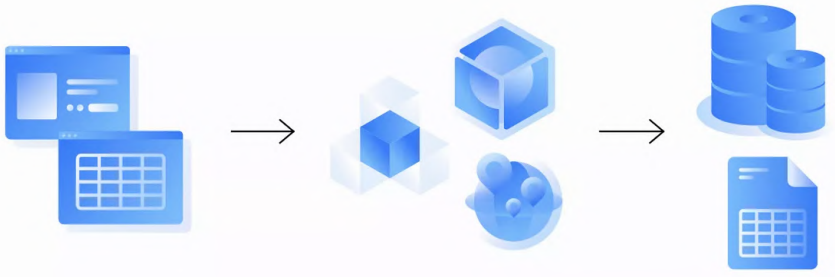
APIFY

The Beginner's Guide To Web Scraping

What is web scraping?

Web scraping is the process of automatically extracting data from websites.

Any publicly accessible web page can be analyzed and processed to extract information – or data. These data can then be downloaded or stored so that they can be used for any purpose outside the original website.



Contents

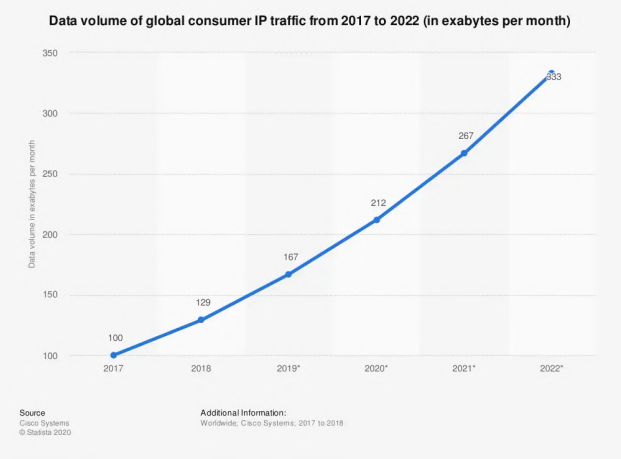
What is the point of web scraping?

The web is the greatest repository of knowledge and data in the history of humanity.

But that information was designed to be read by human beings, not machines. Web scraping enables you to create rules for computers to access those data in an efficient and machine-readable way.

It is already impossible for humans to process even a fraction of the data on the web. That's why web scraping is becoming essential. We need machines to read that data for us so that we can use it in business, conservation, protecting human rights, fighting crime, and any number of projects that can benefit from the kind of data that the Internet is so good at accumulating.

To ignore the potential of web scraping is to ignore the potential of the web.



Did you know?

According to World Bank/ITU, the number of worldwide Internet users increased from 3.5 billion people in 2017 to 4.2 billion in 2019, growing 8% annually (CAGR).

What is web scraping used for?

Web scraping allows you to collect structured data. Structured data is just a way to say that the information is easy for computers to read or add to a database.

Instead of relying on humans to read or process web pages, computers can rapidly use that data in lots of unexpected and useful ways.

To illustrate the difference, imagine how long it might take you to manually copy and paste text from 100 web pages.

A machine could do it in less than a second if you give it the correct instructions. It can also do it repeatedly, tirelessly, and at any scale. Forget about 100 pages. A computer could deal with 1,000,000 pages in the time it would take you to open just the first few.

```
2021-03-15T16:03:02.156Z ACTOR: Pulling Docker image from repository.
2021-03-15T16:03:02.240Z ACTOR: Creating Docker container.
2021-03-15T16:03:03.055Z ACTOR: Starting Docker container.
2021-03-15T16:03:11.563Z INFO: System info {"apifyVersion":"0.16.1","apifyClientVersion":"0.5.26","osType":"Linux","nodeVersion":"v12.18.4"}
2021-03-15T16:03:11.564Z WARNING: You are using an outdated version (0.16.1) of Apify SDK. We recommend you to update to the latest version (0.2
2021-03-15T16:03:11.564Z       Read more about Apify SDK versioning at: https://help.apify.com/en/articles/3184510-updates-and-versioning-of-a
2021-03-15T16:03:11.571Z INFO  System info {"apifyVersion":"1.0.1","apifyClientVersion":"1.0.5","osType":"Linux","nodeVersion":"v12.18.4"}
2021-03-15T16:03:11.576Z INFO  PHASE -- STARTING ACTOR.
2021-03-15T16:03:11.882Z Added finish webhook http://static.22.149.201.195.clients.your-server.de/api/v2/check_new_url/true
2021-03-15T16:03:11.883Z INFO  ACTOR OPTIONS: -- {"siteCrawlType":"newUrls","finishWebhookUrl":"http://static.22.149.201.195.clients.your-server.
2021-03-15T16:03:13.897Z INFO  Starting the crawl.
2021-03-15T16:03:14.281Z INFO  CheerioCrawler:AutoscaledPool: state {"currentConcurrency":0,"desiredConcurrency":2,"systemStatus":{"isSystemIdle'
2021-03-15T16:03:18.266Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid=&rh_cb_advert_type
2021-03-15T16:03:18.545Z INFO  Enqueued 4 URLs.
2021-03-15T16:03:21.324Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=3&ps=50&ob=
2021-03-15T16:03:21.673Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=1&ps=50&ob=
2021-03-15T16:03:21.713Z INFO  Enqueued 4 URLs.
2021-03-15T16:03:22.004Z INFO  Enqueued 7 URLs.
2021-03-15T16:03:23.846Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=2&ps=50&ob=
2021-03-15T16:03:23.884Z INFO  Enqueued 6 URLs.
2021-03-15T16:03:25.316Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=5&ps=50&ob=
2021-03-15T16:03:25.642Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=4&ps=50&ob=
2021-03-15T16:03:25.738Z INFO  Enqueued 8 URLs.
2021-03-15T16:03:25.978Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=7&ps=50&ob=
2021-03-15T16:03:26.021Z INFO  Enqueued 7 URLs.
2021-03-15T16:03:27.050Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=6&ps=50&ob=
2021-03-15T16:03:27.082Z INFO  Enqueued 8 URLs.
2021-03-15T16:03:27.880Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=9&ps=50&ob=
2021-03-15T16:03:27.922Z INFO  Enqueued 5 URLs.
2021-03-15T16:03:28.432Z INFO  Enqueued 9 URLs.
2021-03-15T16:03:29.361Z INFO  Page opened. {"url":"http://www.realhit.cz/vyhledavani/nemovitosti?seid=rh_cb_advert_functionid%3D&pi=8&ps=50&ob=
2021-03-15T16:03:29.408Z INFO  Enqueued 6 URLs.
2021-03-15T16:03:32.260Z INFO  CheerioCrawler: All the requests from request list and/or request queue have been processed, the crawler will shut
2021-03-15T16:03:32.798Z INFO  CheerioCrawler: Final request statistics: {"requestsFinished":10,"requestsFailed":0,"retryHistogram":[10],"request
2021-03-15T16:03:32.800Z INFO  Crawl finished.
```


APIFY

Did you know?

The majority of Internet traffic is generated by bots. 61.5% of all website traffic is automated.

Ways web scraping can benefit business

Web scraping gives you access to a lot of data.

Those data can be:

- loaded into databases
- added to spreadsheets
- used in apps
- repurposed in surprising and unexpected ways

Here are just some of the ways web scraping can help your business be more efficient and profitable:

Price tracking

Be more competitive by tracking the prices of your competitors in real time and with the ability to adjust your own prices on the fly. You can even tell your own customers what your competitors are up to so that they see the advantages of buying from you instead.

Lead generation

Generate smart leads by scraping publicly available contact information and social media platform profiles to find new customers and potential business leads.

Content aggregation

Aggregate content to create new uses for data, make data easier to read or add value by notifying users when prices or content changes.

Market analysis

Gain market insights by scraping data about your business, customer demand, feedback in the wild, or even identify opportunities in the real world by analyzing demographic changes and trends.

SEO

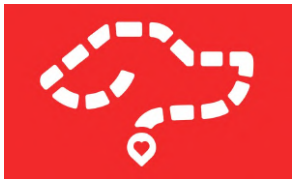
Improve your SEO by monitoring keywords, popularity, and trends across the web.

If you would like to read more about other businesses and industries that use web scraping, check out our use cases and success stories. You'll find examples of how retailer price monitoring, machine learning, copyright protection, and even moms returning to work can benefit from web scraping.

Web scraping can also benefit humanity

Web scraping isn't only used for financial gain. Organizations around the world are using web scraping to **help**.

Find missing animals



Combat human trafficking networks

Encourage forest restoration

Track COVID-19

Advantages of web scraping

Saves time

When you use web scraping, you don't have to manually collect data from websites and you can rapidly scrape many websites at the same time.

Data at scale

Web scraping gives you data at much greater volume than you could ever collect manually.

Cost-effective

A simple scraper can often do the job, so you don't need to invest in complex systems or extra staff.

Modifiable

Create a scraper for one task and you can often retrofit it for a different task by making only small changes.

Accurate and robust

Set up your scraper correctly and it will accurately collect data directly from websites, with a very low chance of errors being introduced.

Maintainable

Changes to websites can usually be accommodated by slightly tweaking the scraper.

Structured data

Scraped data arrive in a machine-readable format by default, so simple values can often immediately be used in other databases and programs.

Disadvantages of web scraping

Learning curve

The initial creation of a scraper can be time-consuming, especially if you start from scratch.

Data processing

Although the data will arrive in a structured format, more complex data will need to be processed so that they can be used in other programs.

Continued maintenance

Because your scraper depends on an external website, you have no control over when that website changes its structure or content, so you need to react if the scraper becomes outdated.

Blockable

Websites can use several different methods such as IP blocking to stop you from scraping their content.

Is web scraping legal?

Web scraping is just a way to get information from websites.

That information is already publicly available, but it is delivered in a way that is optimized for humans.

Web scraping simply optimizes it for machines. Web scraping is not hacking, and it is not intended to cause problems for the websites that are scraped.

Google uses search engine bots to index websites and comparison websites use bots to check prices across multiple websites. These are both automated ways of accessing those websites. So in effect they are web scraping.

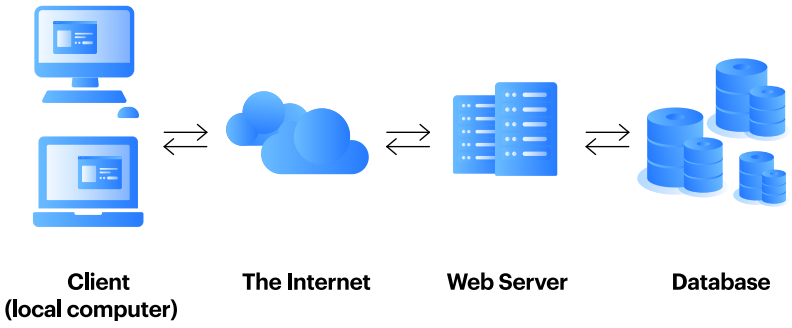
How does the web work?

Before you start getting into the world of web scraping, it might help to understand more about how the Internet and the web work.

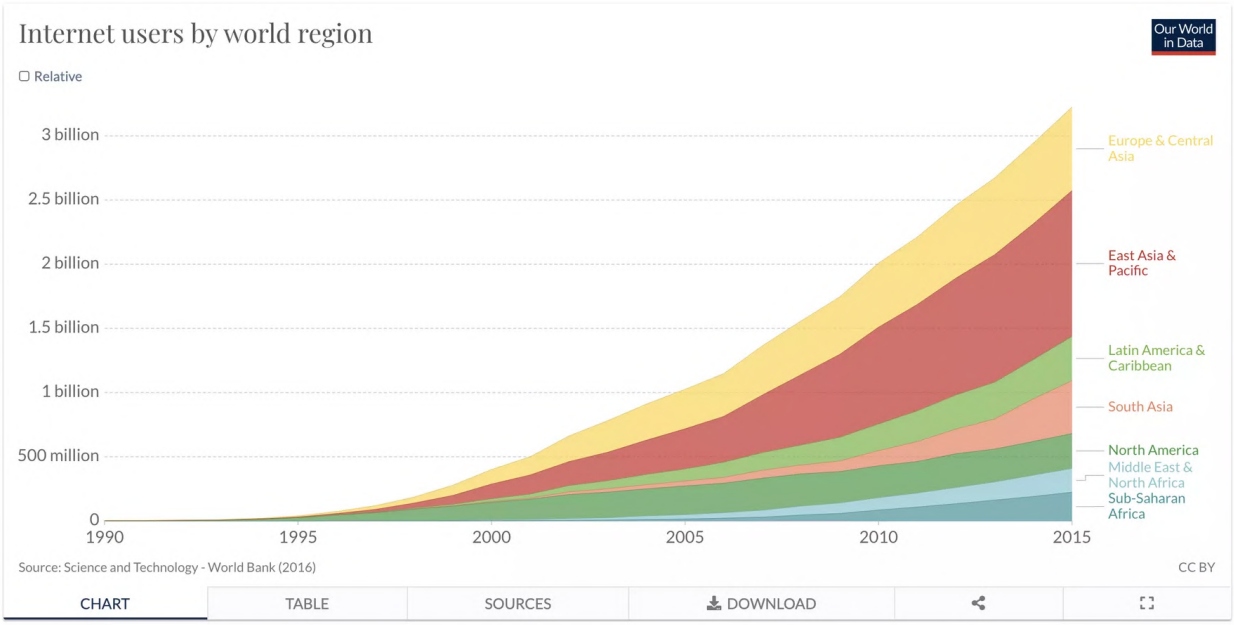
The Internet was born during the Cold War in the 1960s, but the web came into being many years later when Sir Tim Berners-Lee proposed a networked hypertext system to his boss at CERN.

That idea eventually led Berners-Lee to create three important technologies:

- Hypertext Transfer Protocol (HTTP). This enables computers to retrieve linked resources across the web.
- Hypertext Markup Language (HTML). The markup language of the web. Allows text to be formatted so that it can be displayed correctly.
- Uniform Resource Locator (URL). Otherwise known as a "web address". Used to identify all the resources on the web.



Put those together and you have the vital building blocks of what eventually became known as the World Wide Web.



Decentralization was fundamental to the early web as envisaged by Berners-Lee, as was universal compatibility and making it simple to share information. Over time, standards were established through a transparent and participatory process by the World Wide Web Consortium (W3C). These open standards are one of the cornerstones that have made it possible for the web to grow.

Berners-Lee still firmly believes that it is vital to "defend and advance the open web as a public good and a basic right" and created the World Wide Web Foundation just over ten years ago to ensure digital equality and transparency for everyone.

That vision of an open web is just as important now as it was then. And making data accessible to everyone is part of keeping the web open.

What is a web browser?

You're using a web browser to view this web page. A web browser is just software, or a computer program, that enables you to access, view and interact with web pages.

Did you know?

Think the Internet and World Wide Web mean the same thing? Nope, the Internet is a network of computers, while the World Wide Web is a bridge for accessing and sharing information across it.

How do web browsers work?

Your browser retrieves information from the web and displays it on your computer or mobile device.

It uses the Hypertext Transfer Protocol (HTTP) to retrieve the content of websites and Hypertext Markup Language (HTML) to determine how to render the content.

The final result is that you see a web page on your device, and you can interact with that web page. Underlying the web page can be a multitude of other technologies, such as HTML, CSS, JavaScript, etc.

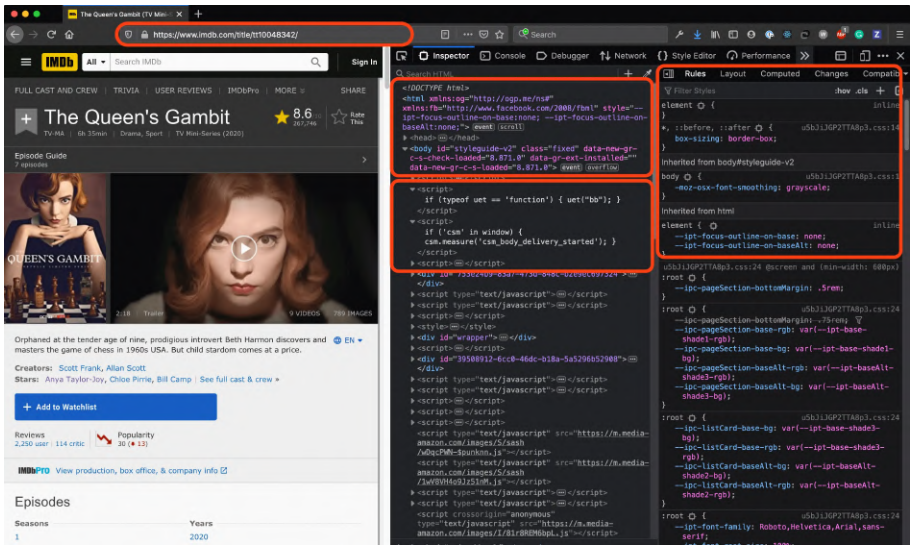
Try it yourself

You can easily see the source code of a website:

1. Open any page in a browser on a Mac or PC. For example, you could open the IMDb page for The Queen's Gambit.
2. Then right-click and select **Inspect** at the bottom of the menu.
3. The code that created the page will be displayed.

In the image below:

- the website is shown in the left-hand panel,
- in the middle are the source code (HTML and JavaScript),
- the right-hand panel shows the code used to style the page (Cascading Style Sheets, or CSS).



How can I start web scraping?

We find that web scraping works best if you pause and ask yourself these three questions before you start coding or ordering a solution:

1

What information are you looking for?

What data do you want to get?

2

Where can you find the data?

What's the website and what's the URL?

3

What will you do with the data?

What format do you need it in and how should you extract it?

Once you've answered these questions, you can start thinking about how you will scrape the data you want.

Basic scraping terminology

Web scraping

The process of automatically extracting data from websites. Also known as screen scraping, web data extraction, web harvesting.

Web scrapping

This is just a really common and easy-to-make typo!

Web crawling

Web crawlers are spiders or spider bots that systematically browse the web and index it. Search engines use these bots to make it easier for us to search the web.

Structured data

Information that is organized and formatted in such a way that it is easy for computers to read and store in databases. A spreadsheet is a good example of how data can be organized in a structured way.

Hypertext Transfer Protocol (HTTP)

Enables computers to retrieve linked resources across the web.

Hypertext Markup Language (HTML)

The markup language of the web. Allows text to be formatted so that it can be displayed correctly.

Uniform Resource Locator (URL)

A “web address”. Used to identify all the resources on the web.

Cascading Style Sheets (CSS)

The design language of the web. It enables web page authors to style content and control presentation across an entire website.

JavaScript

A programming language used all over the Internet to control the behavior of websites and enable complicated interaction between user and web page.

IP address

An Internet Protocol address is a number assigned to every device on the Internet. These numbers allow devices to communicate with each other.

Proxy

A proxy server is a device that acts as an intermediary between other devices on the Internet. Proxies are commonly used to hide the geographical location of a particular device, often for privacy reasons.

Application Programming Interface (API)

A computing interface that makes it possible for multiple different applications to communicate with each other. An API operates as a set of rules to tell the software what requests or instructions can be exchanged and how data are to be transmitted. Apify got its name from API ☺

Software Development Kit (SDK)

A package that enables developers to create applications on a particular platform. An SDK can include programming libraries, APIs, debugging tools and utilities designed to make it easy for a developer to use the platform. Apify has its own SDK.

Spot quiz

What’s the difference between web scraping and web crawling?

Web scraping companies and tools

So you want to start web scraping, you know what you want to scrape, and you've decided to explore the ways you can start.

There are lots of methods and companies out there involved in web scraping. To help you choose, let's split the web scraping world into four different categories.

Enterprise consulting companies

These provide high-end turnkey "data-as-a-service" solutions to large companies. They will carry out scraping at any scale, but at a price.

Examples: Import.io, Mozenda, Apify.

Point-and-click tools

Allow you to go to a website and just click on the elements you want to scrape. These are good enough for simple use cases, but not so good for more complicated projects.

Examples: Dexi.

Programming platforms

A platform is designed for developers and offers a lot of flexibility. Instead of building the infrastructure for scraping, you use an existing system that was specifically designed for the task.

Examples: Zyte, Apify.

AI knowledge extractors

These companies take an AI approach and attempt to extract data from websites automatically. It works for standardized pages, but is not flexible enough to cover a variety of use cases.

Examples: DiffBot.

> ☆ **You have plenty of options, but we believe that you should use Apify for your web scraping needs 😁**
>
> We've built a versatile and fast web scraping and automation platform that works for beginners, developers, and enterprise customers. Our goal from the outset was to create an organic ecosystem of scrapers and automation tools that would develop and grow with the needs of its users.
>
> *Read on to see why Apify has the best web scraping tools in the business.*

Web scraping with Apify

Apify offers several different ways to scrape. You can start from scratch with your own solution, build upon existing tools, use ready-made tools, or get a solution created for you.

Enterprise solution

Enterprise customers can order a more specialized web scraping or automation solution at any scale from a dedicated Apify data expert. We will work with you all the way to project completion and can continue to provide maintenance once it is up and running.

Tell us more about your project

You can use this form or click on the chat bubble in the bottom-right of the screen to chat with an Apify expert!

Order a custom solution

Apify Marketplace enables you to submit your project specifications to a pool of Apify-approved developers. These developers then send you their proposals, and you can select the best offer. This is the fast track to getting the data you need and means you don't need to learn any coding yourself.

It's easy to request a custom solution on Apify Marketplace.

Just fill in the form.

Use a ready-made tool

Apify Store has existing solutions for popular sites. This is the quickest way to get your data as the tools are already optimized for particular use cases. Our tools are designed to be easy for even those with no previous coding experience and our support team is always ready to help.

Try it yourself

When it comes to Apify's ready-made tools, a lot of the web scraping code you need has already been written by a developer. So you just have to decide what information you want to extract. Okay, it's time for a real-world example, so let's get some data from IMDb about the recent Netflix hit series, The Queen's Gambit.

1. Go to Apify's IMDb Scraper and click **Try me**.
2. Fill in the URL for The Queen's Gambit in the input field.
3. Click on **Save and Run**.

The **output data** will contain the following information about each movie or series that you have listed in the input schema of the IMDb scraper:

JSON Copy


```
[
  {
    title: "The Queen's Gambit",
    original title: "",
    runtime: 395,
    certificate: "TV-MA",
    year: "",
    rating: "8.6",
    ratingcount: "250392",
    description: "Orphaned at the tender age of nine, prodigious
    introvert Beth Harmon discovers and masters the game of
    chess in 1960s USA. But child stardom comes at a price.",
    stars: "Anya Taylor-Joy, Chloe Pirrie, Bill Camp",
    director: "",
    genre: "Drama, Sport",
    country: "USA",
    url: "https: //www.imdb.com/title/tt10048342"
  }
]
```


Code it yourself

You can use our generic scrapers and customize them with just a bit of JavaScript. Or you can use Apify SDK to create your own scraping solution.

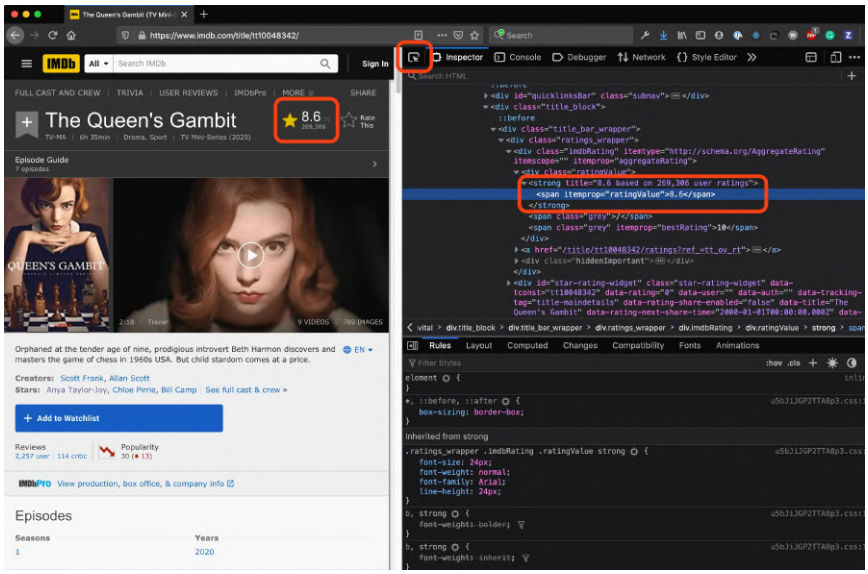
Try it yourself

Let's try a more complicated version of our example from above, where we used Apify's IMDb Scraper to get information about The Queen's Gambit. This time, we'll go with a universal web scraping tool, Apify's Swiss Army Knife of web scraping, our Web Scraper.

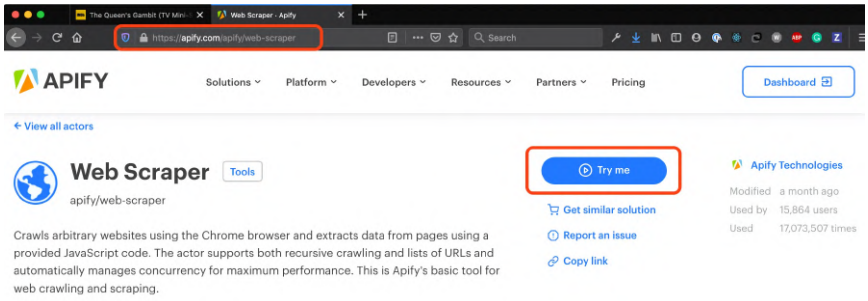
Just follow the steps and scrape the rating of The Queen's Gambit from IMDb.com with your own JavaScript-powered scraper.

1. Inspect the source of your data, in other words this link (remember that you just have to right-click on the page and select "Inspect" at the bottom of the menu), and find and select the information you want to scrape. For our example, the code will look like this:

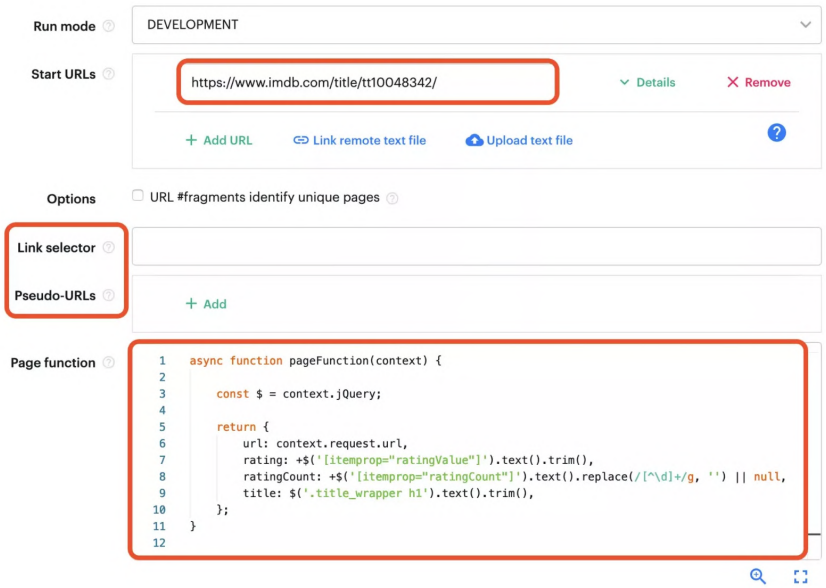
```
<span itemprop="ratingValue">8.6</span>
```



2. Create a task for Web Scraper on the Apify platform by clicking on **Try me**.



3. Paste the URL to the Queen's Gambit IMDb page into the Start URLs field and replace the code in the **Page function** field with the code below. Remove the **Link selector** and **Pseudo-URLs** fields.



```
async function pageFunction(context) {
  const $ = context.jQuery;
  return {
    url: context.request.url,
    rating: +$('[itemprop="ratingValue"]').text().trim(),
    ratingCount: +$('[itemprop="ratingCount"]').text().replace(/[^\d]+/g, '') || null,
    title: $('.title_wrapper h1').text().trim(),
  };
}
```

4. Click **Save and run** and then check the dataset with the final result.

JSON Copy

```
{
  url: "https: //www.imdb.com/title/tt10048342"
  rating: "8.6",
  ratingcount: "250392",
  title: "The Queen's Gambit",
}
```

Tip: for a more detailed explanation, check out our extensive tutorial for this scraper.

If you still can't decide which option is right for you, read more on choosing the right solution or just email us at hello@apify.com for free expert advice on your use case.

Learn web scraping

Now that you know the basics of web scraping, you might want to explore the topic further. To save you time, we've collected a few courses and tutorials suitable for all levels. We recommend these as a great way to quickly get up to speed on web scraping.

Courses for beginners

Udemy has a course for beginners to introduce you to web scraping in 60 minutes.

Pluralsight has a course on web scraping with Python for more experienced beginners.

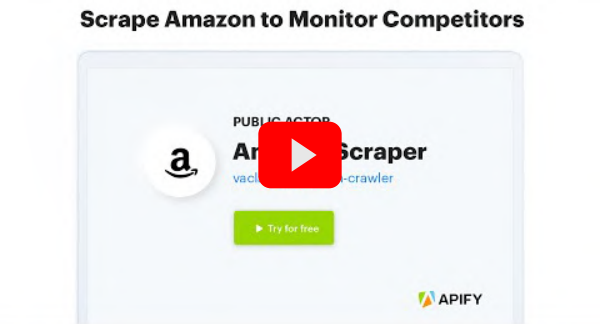
Coursera has a guided project on scraping with Python and Beautiful Soup, for much more advanced users.

Guides for beginners

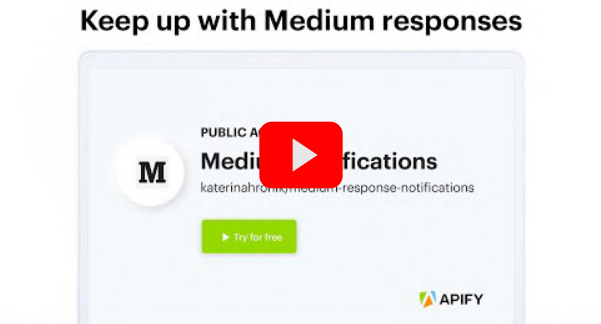
Our own Apify blog has general articles to inspire you and also several step-by-step guides to scraping popular websites.

- What's the difference between web scraping and crawling?
- How to scrape any website for absolute beginners.
- How to scrape Facebook pages.
- Scaping Google Maps locations.

How to scrape Amazon to monitor your competitors (web scraping).

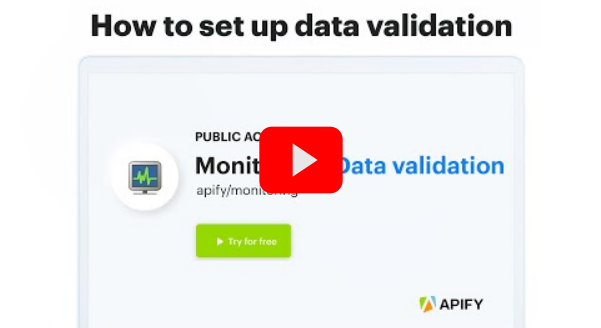


Scrape Medium publication notifications: keep up with all responses (process automation).



How to set up monitoring for your Apify projects (web scraping automation).

Monitoring: How to set up data validation.



Top web scraping tips from Apify devs

Vaclav
Apify developer

"Don't always try to make your scraper as fast as possible - you might break the website! Always check how the website behaves under heavy load before running your scraper at scale."

Interesting technical reading on our blog

These are the most popular technical posts on the Apify blog.

[Bypassing web scraping protection: get the most out of your proxies with shared IP address emulation]

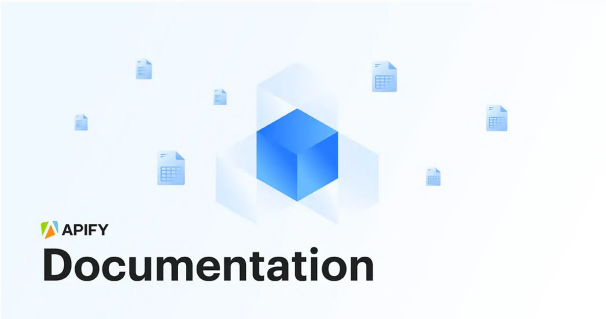
Learn about modern web scraping protection techniques from Petr and how to bypass them. Scrape up to three times more pages by combining IP address rotation with shared IP address emulation.

APIFY

[Using a man-in-the-middle proxy to scrape data from a mobile app API]

Petr will show you how to set up a man-in-the-middle proxy and install a self-signed certificate on your mobile phone in order to intercept HTTPS communication between any mobile app and its backend API.

Want to make your own web scrapers?



Learn more about Apify and what we do by reading the extensive Apify documentation. Get familiar with the platform and get all the technical advice you need from our top developers.

Explore Apify SDK, the scalable web crawling and scraping library for JavaScript/Node.js. Enables development of data extraction and web automation jobs with headless Chrome, Puppeteer, and Playwright.

Share on Twitter