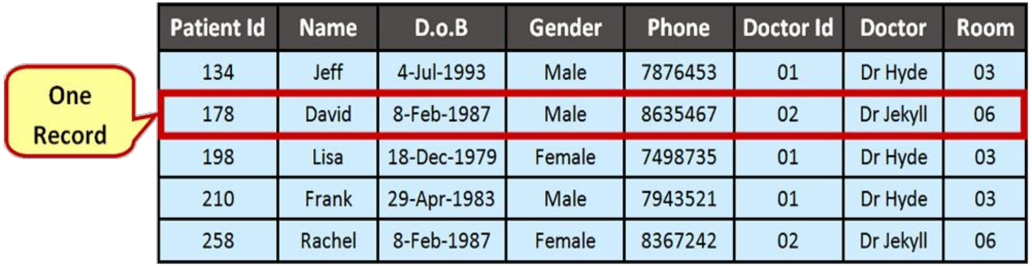
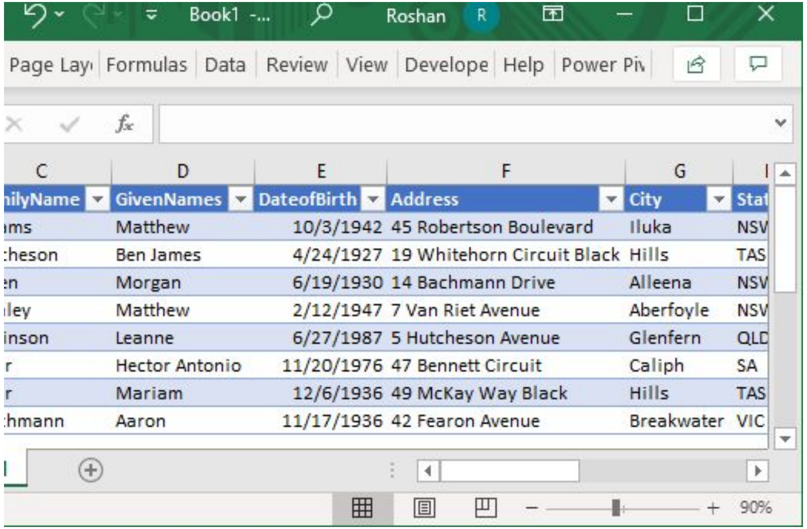
Web Scraping

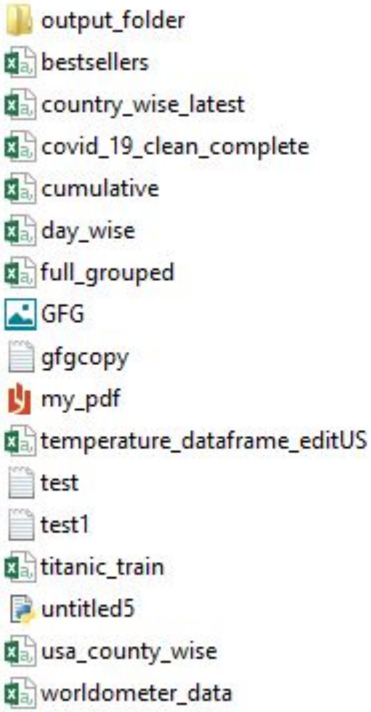


One Record

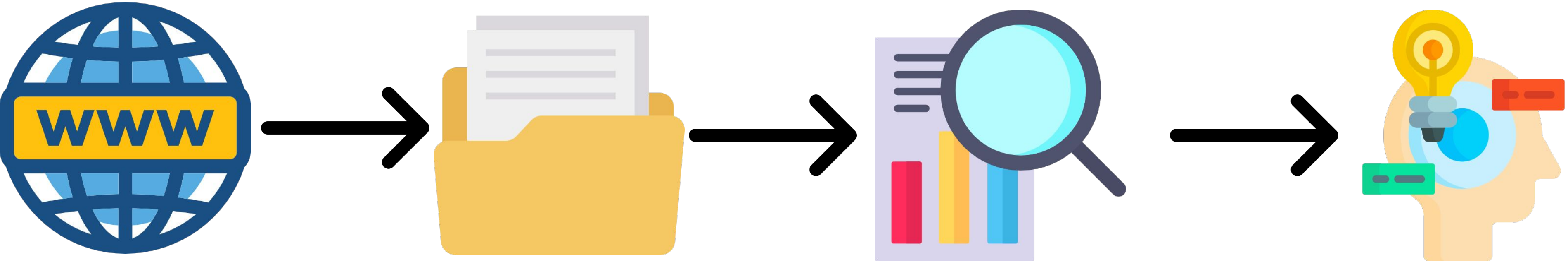
Patient Id	Name	D.o.B	Gender	Phone	Doctor Id	Doctor	Room
134	Jeff	4-Jul-1993	Male	7876453	01	Dr Hyde	03
178	David	8-Feb-1987	Male	8635467	02	Dr Jekyll	06
198	Lisa	18-Dec-1979	Female	7498735	01	Dr Hyde	03
210	Frank	29-Apr-1983	Male	7943521	01	Dr Hyde	03
258	Rachel	8-Feb-1987	Female	8367242	02	Dr Jekyll	06



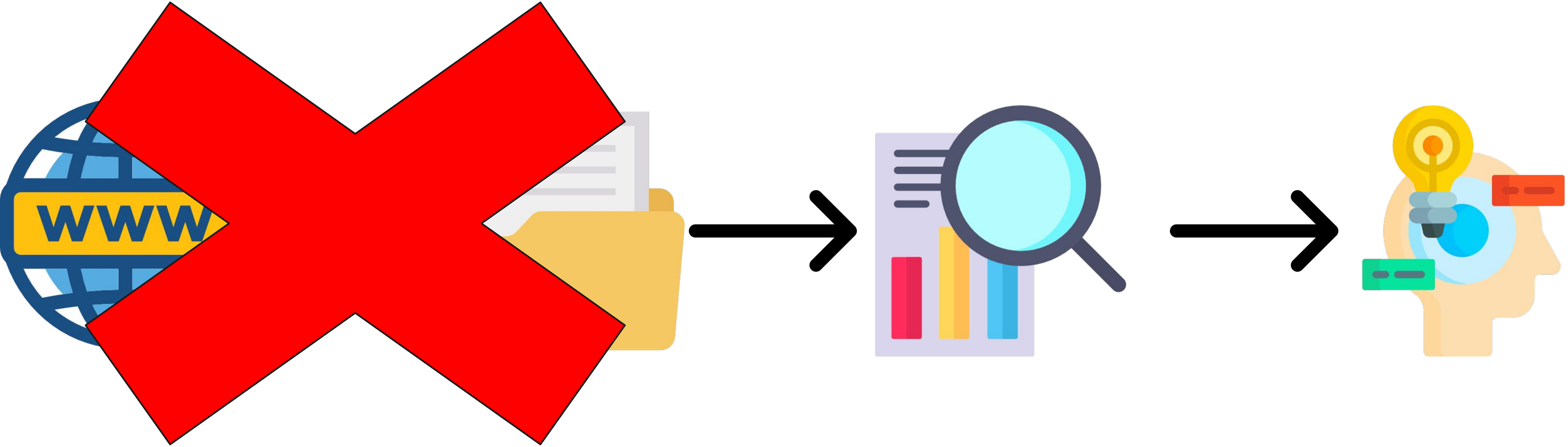
C	D	E	F	G	I
ilyName	GivenNames	DateofBirth	Address	City	Stat
ıms	Matthew	10/3/1942	45 Robertson Boulevard	Iluka	NSV
:heson	Ben James	4/24/1927	19 Whitehorn Circuit Black	Hills	TAS
ın	Morgan	6/19/1930	14 Bachmann Drive	Alleena	NSV
ıley	Matthew	2/12/1947	7 Van Riet Avenue	Aberfoyle	NSV
inson	Leanne	6/27/1987	5 Hutcheson Avenue	Glenfern	QLD
r	Hector Antonio	11/20/1976	47 Bennett Circuit	Caliph	SA
r	Mariam	12/6/1936	49 McKay Way Black	Hills	TAS
:hmann	Aaron	11/17/1936	42 Fearon Avenue	Breakwater	VIC



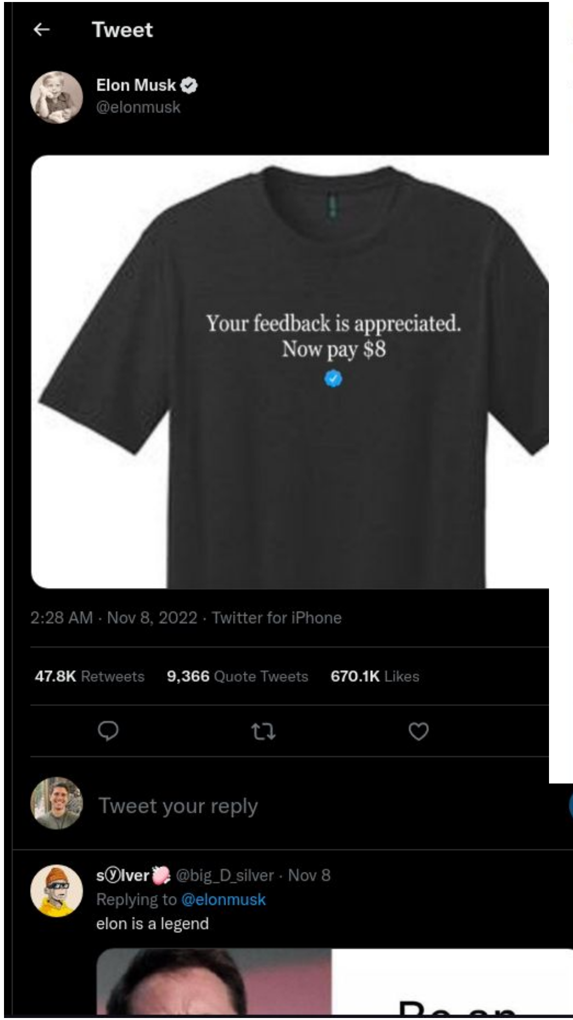
- output_folder
- bestsellers
- country_wise_latest
- covid_19_clean_complete
- cumulative
- day_wise
- full_grouped
- GFG
- gfgcopy
- my_pdf
- temperature_dataframe_editUS
- test
- test1
- titanic_train
- untitled5
- usa_county_wise
- worldometer_data



WWW



WWW

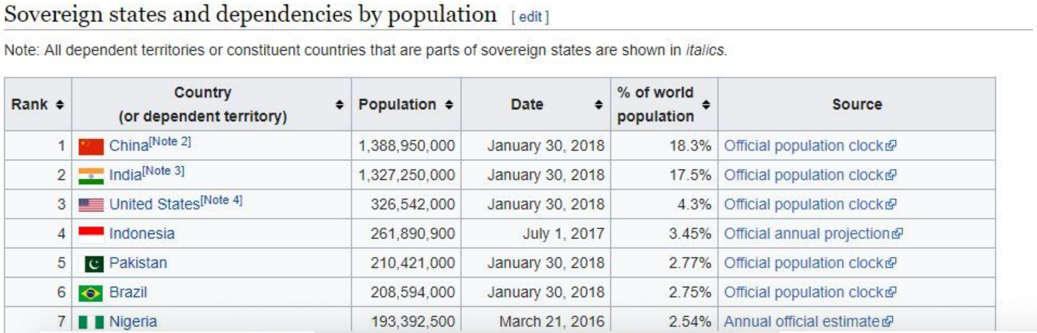


Elon Musk @elonmusk

2:28 AM · Nov 8, 2022 · Twitter for iPhone

47.8K Retweets 9,366 Quote Tweets 670.1K Likes

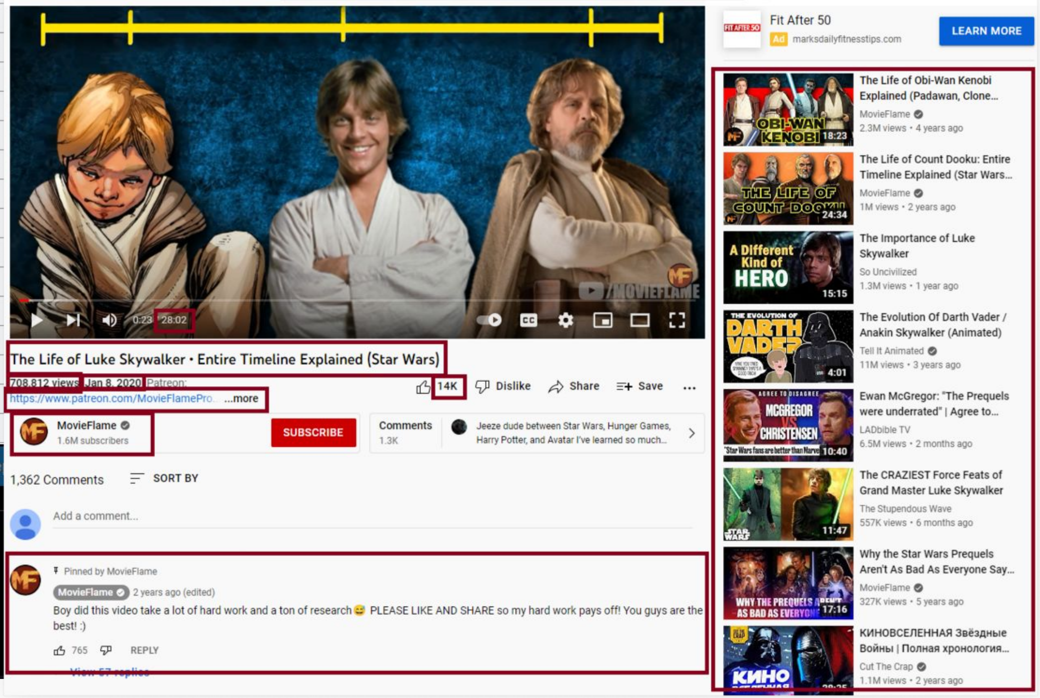
s🅈lver @big_D_silver · Nov 8
Replying to @elonmusk
elon is a legend



Sovereign states and dependencies by population [edit]

Note: All dependent territories or constituent countries that are parts of sovereign states are shown in *italics*.

Rank	Country (or dependent territory)	Population	Date	% of world population	Source
1	China[Note 2]	1,388,950,000	January 30, 2018	18.3%	Official population clock
2	India[Note 3]	1,327,250,000	January 30, 2018	17.5%	Official population clock
3	United States[Note 4]	326,542,000	January 30, 2018	4.3%	Official population clock
4	Indonesia	261,890,900	July 1, 2017	3.45%	Official annual projection
5	Pakistan	210,421,000	January 30, 2018	2.77%	Official population clock
6	Brazil	208,594,000	January 30, 2018	2.75%	Official population clock
7	Nigeria	193,392,500	March 21, 2016	2.54%	Annual official estimate


MOVIEFLAME
0:23 28:02
Dislike
Share
Save
SUBSCRIBE
SORT BY
REPLY
LEARN MORE
OBI-WAN KENOBI 18:23
THE LIFE OF COUNT DOOKU 24:34
A Different Kind of HERO 15:15
THE EVOLUTION OF DARTH VADER 4:01
AGREE TO DISAGREE MCGREGOR VS CHRISTENSEN "Star Wars fans are better than Marvel 10:40
STAR WARS 11:47
WHY THE PREQUELS AREN'T AS BAD AS EVERYONE 17:16
КИНО

The Life of Luke Skywalker • Entire Timeline Explained (Star Wars)

708,812 views Jan 8, 2020 Patreon:
https://www.patreon.com/MovieFlamePro... ...more

14K

MovieFlame
1.6M subscribers

Comments 1.3K
Jeeze dude between Star Wars, Hunger Games, Harry Potter, and Avatar I've learned so much...

1,362 Comments

Add a comment...

Pinned by MovieFlame
MovieFlame 2 years ago (edited)
Boy did this video take a lot of hard work and a ton of research 😅 PLEASE LIKE AND SHARE so my hard work pays off! You guys are the best! :)
765

View 57 replies

The Life of Obi-Wan Kenobi Explained (Padawan, Clone...
MovieFlame
2.3M views • 4 years ago

The Life of Count Dooku: Entire Timeline Explained (Star Wars...
MovieFlame
1M views • 2 years ago

The Importance of Luke Skywalker
So Uncivilized
1.3M views • 1 year ago

The Evolution Of Darth Vader / Anakin Skywalker (Animated)
Tell It Animated
11M views • 3 years ago

Ewan McGregor: "The Prequels were underrated" | Agree to...
LADbible TV
6.5M views • 2 months ago

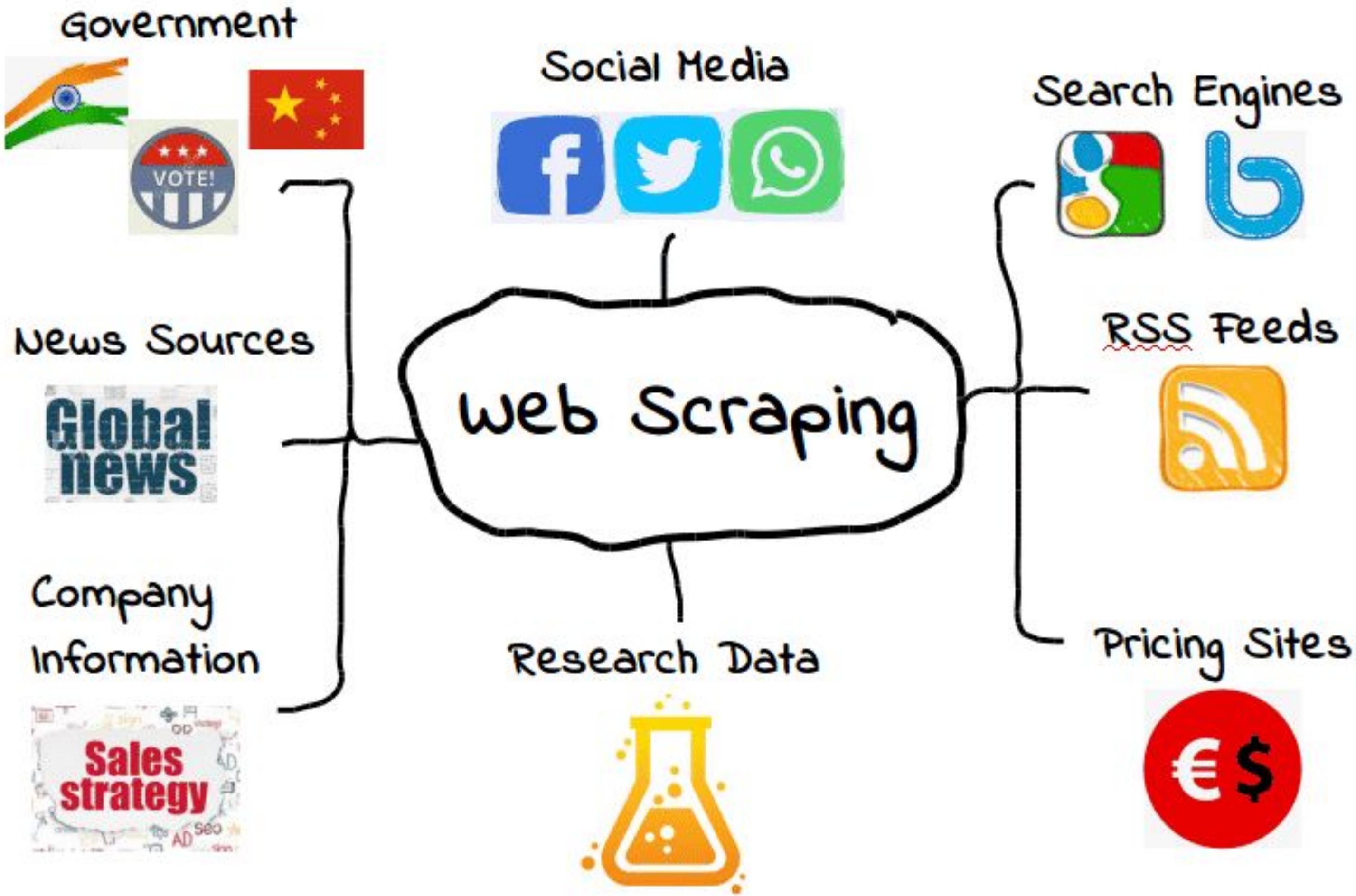
The CRAZIEST Force Feats of Grand Master Luke Skywalker
The Stupendous Wave
557K views • 6 months ago

Why the Star Wars Prequels Aren't As Bad As Everyone Say...
MovieFlame
327K views • 5 years ago

КИНОВСЕЛЕННАЯ Звёздные Войны | Полная хронология...
Cut The Crap
1.1M views • 2 years ago

Web Scraping

The activity of gathering data from the web both manually or automatically.



Web Scraping

Government

Social Media

Search Engines

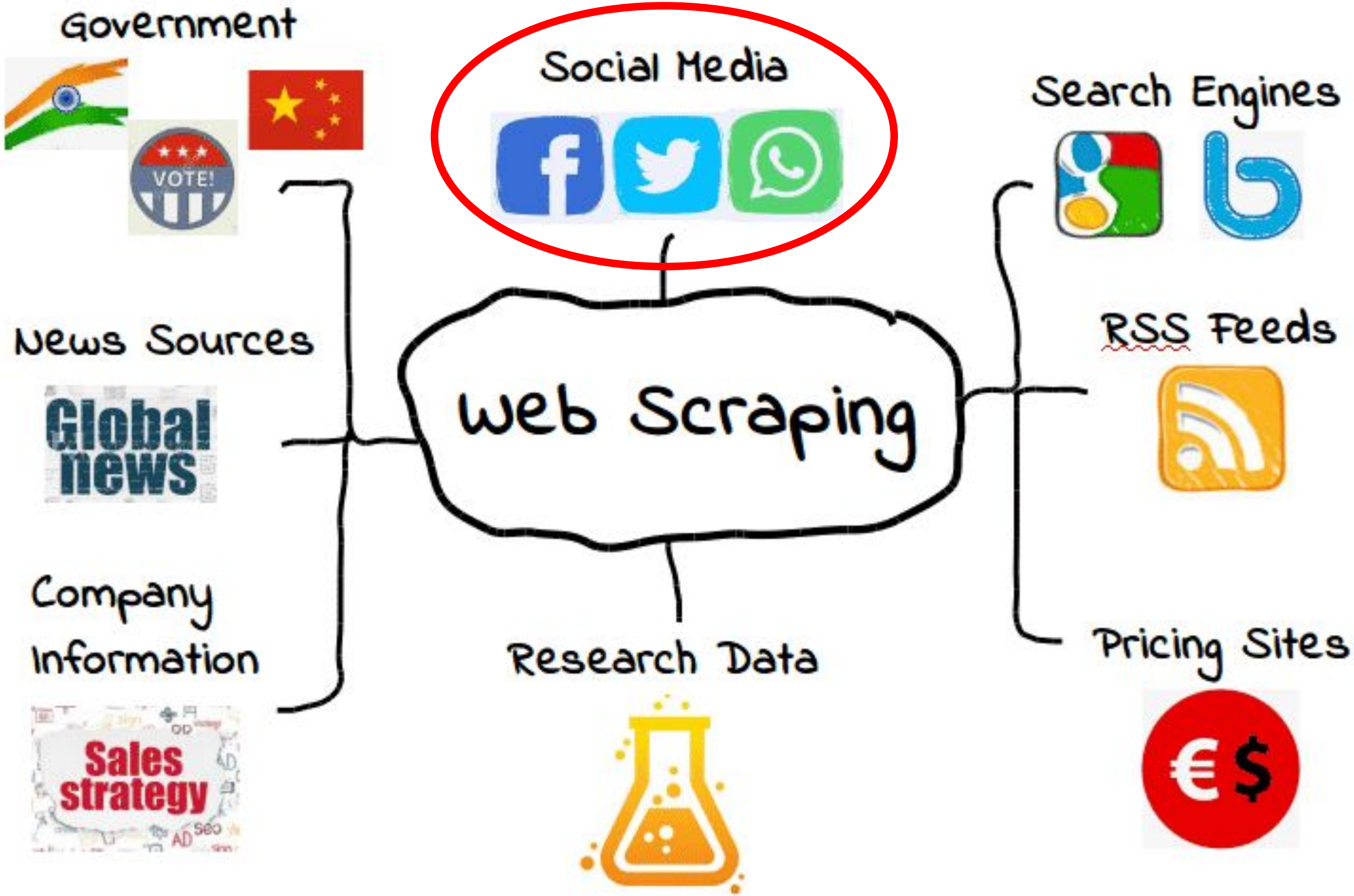
News Sources

RSS Feeds

Company Information

Research Data

Pricing Sites



Web Scraping

Social Media

Search Engines

RSS Feeds

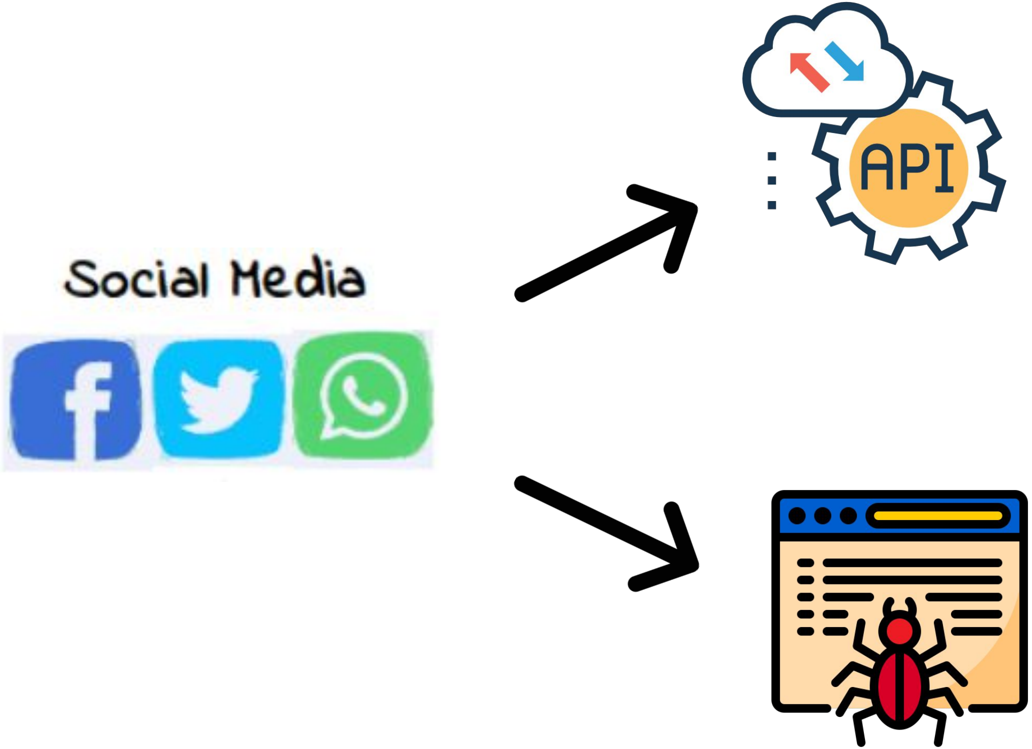
Pricing Sites

Research Data

Company Information

News Sources

Government



Social Media

Crawling through APIs

Obtaining data through Web API

An **Application Programming Interface (API)** is a framework to communicate with an external entity (human or robot) with the goal of performing operations of different kind.

Obtaining data through Web API

An **Application Programming Interface (API)** is a framework to communicate with an external entity (human or robot) with the goal of performing operations of different kind.

Advantages:

- Introduces an abstraction to the callers, hiding them technical details
- Defines a response format that is common among the different calls

Obtaining data through Web API

Nel nostro caso ci concentreremo sulle **Web API**, riconoscibili tramite URL che possono essere contattate attraverso appositi parametri.
We are going to focus on the **Web APIs,** recognizable by an URL that can be called using proper parameters

A typical API structure is the following:

http://example.com/the-api-route/parameters

Examples:

- *https://twitter.com/elonmusk/status/1590755506112823296*
- *https://mastodon.uno/users/billgates/followers*
- *https://ipstack.com/ipstack_api.php?ip=151.100.179.62*

Demo: obtaining information regarding our IP address through Web API

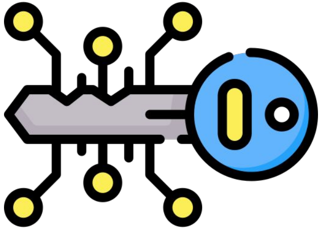
Web API and social media

Exposing APIs that can be queried from the outside is a state of the art from many online services.

Social media offers these kind of features. Specifically:

- **Facebook/Instagram:** offers a service named Crowdtangle which allows the download of posts from both platforms
- **Twitter:** runs a Developer Program that offers access to their API to collect several kind of information
- **YouTube:** via YouTube Data API we can obtain metadata about video and users
- **Reddit:** offers a developer API framework to collect information about comments and posts

Web API and social media



These kind of API requires a **authentication** and a **configuration** phase before performing the requests

- Calling the APIs by only using the URL will result in a 403 error
- These kind of APIs are usually subjected to limitations that affects the number of possible request per minute

Twitter Developer Platform

Twitter is a social media which is known for its data transparency and its willingness to collaborate with developers and researchers

- Communicating with Twitter requires to sign up to Twitter Developer Platform, which offers documentation and a test environment about their API
- Usually, these APIs are called by using a programming language that creates requests from scratch or that employs an external library

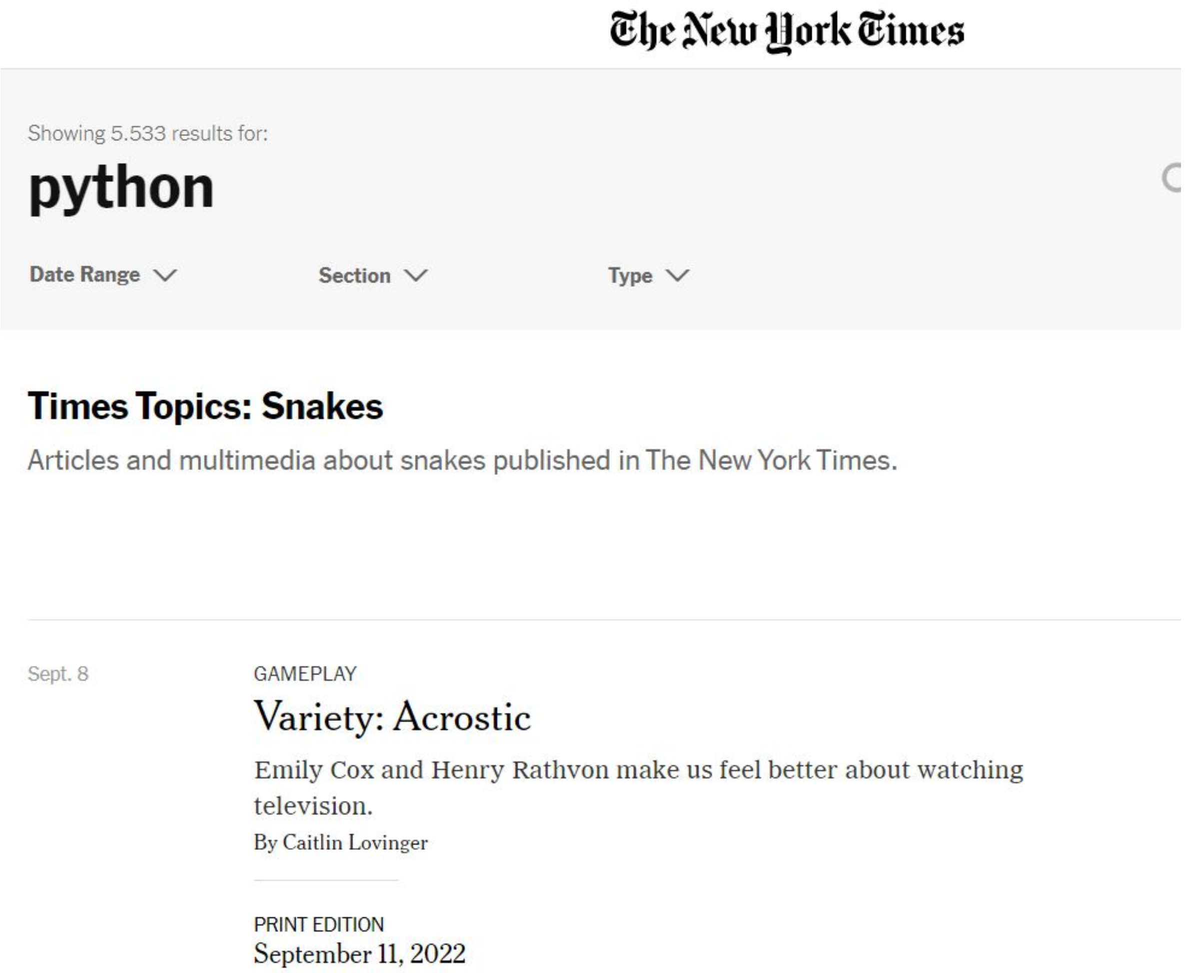
In our examples we are going to use ***tweepy*** package for Python

Demo: accessing Twitter APIs through tweepy library

Perform requests to undocumented APIs

Sometimes, the API we would like to interact with does not provide a proper documentation.

For example:
https://www.nytimes.com/search?dropmab=false&query=python&sort=best provides all the articles that include the term *python*, but there is no documentation that explains its usage.



The New York Times

Showing 5.533 results for:

python

Times Topics: Snakes

Articles and multimedia about snakes published in The New York Times.

Sept. 8

GAMEPLAY

Variety: Acrostic

Emily Cox and Henry Rathvon make us feel better about watching television.

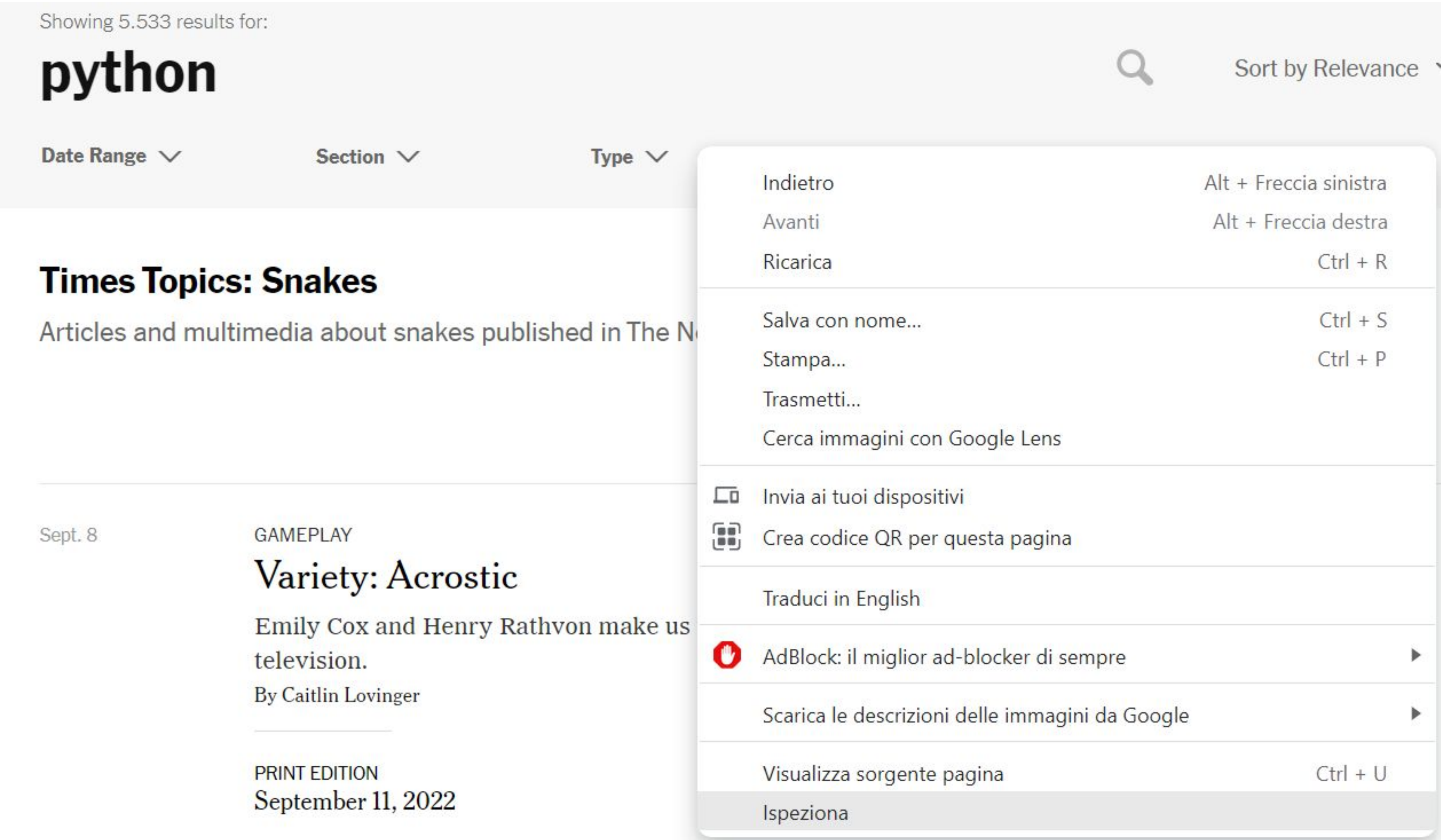
By Caitlin Lovinger

PRINT EDITION
September 11, 2022

Perform requests to undocumented APIs

To discover the API requested under the hood, we usually refer to the **inspector tool** provided by Web Browsers.

On Google Chrome, we can access them using the *Inspect* function or by pressing F12.



Showing 5.533 results for:

python

Times Topics: Snakes

Articles and multimedia about snakes published in The N

Sept. 8

GAMEPLAY

Variety: Acrostic

Emily Cox and Henry Rathvon make us

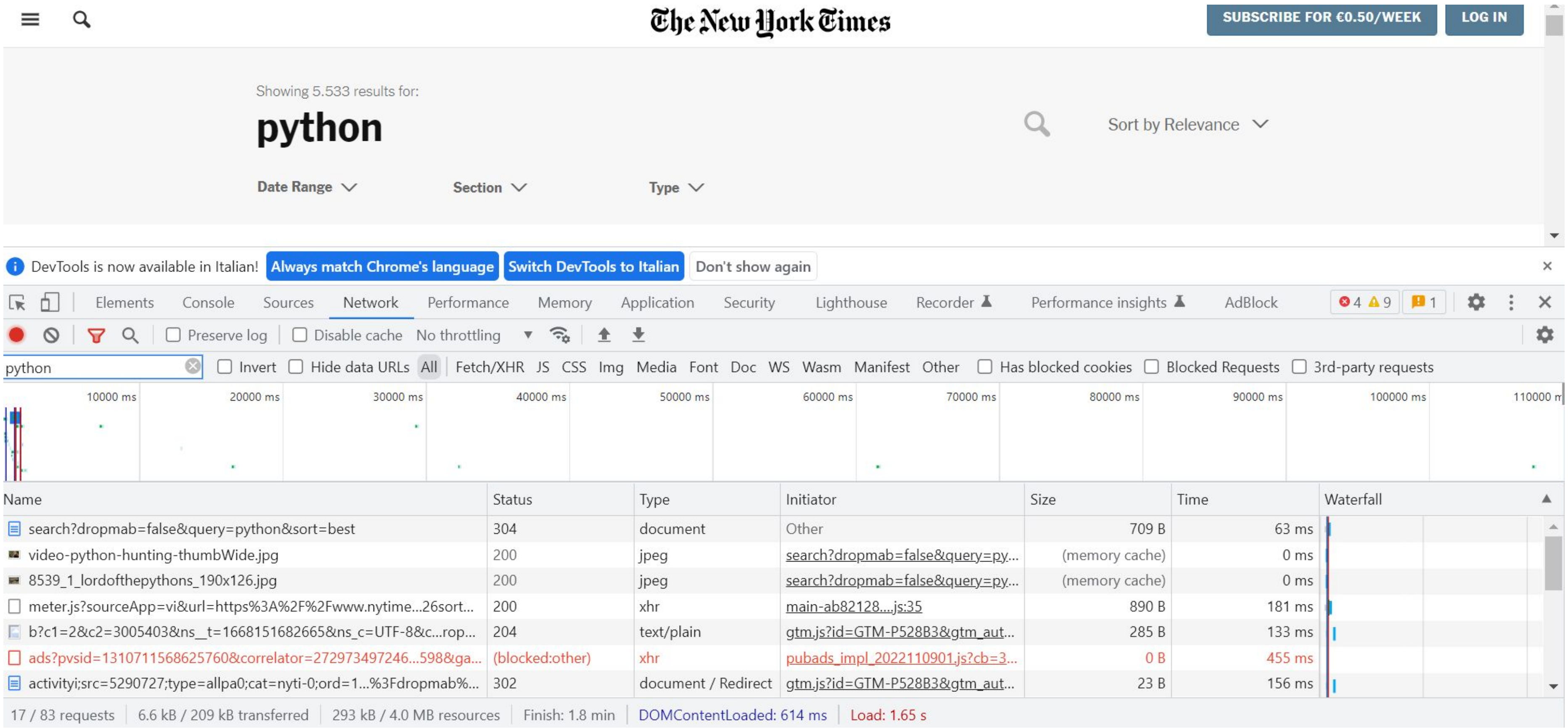
television.

By Caitlin Lovinger

PRINT EDITION

September 11, 2022

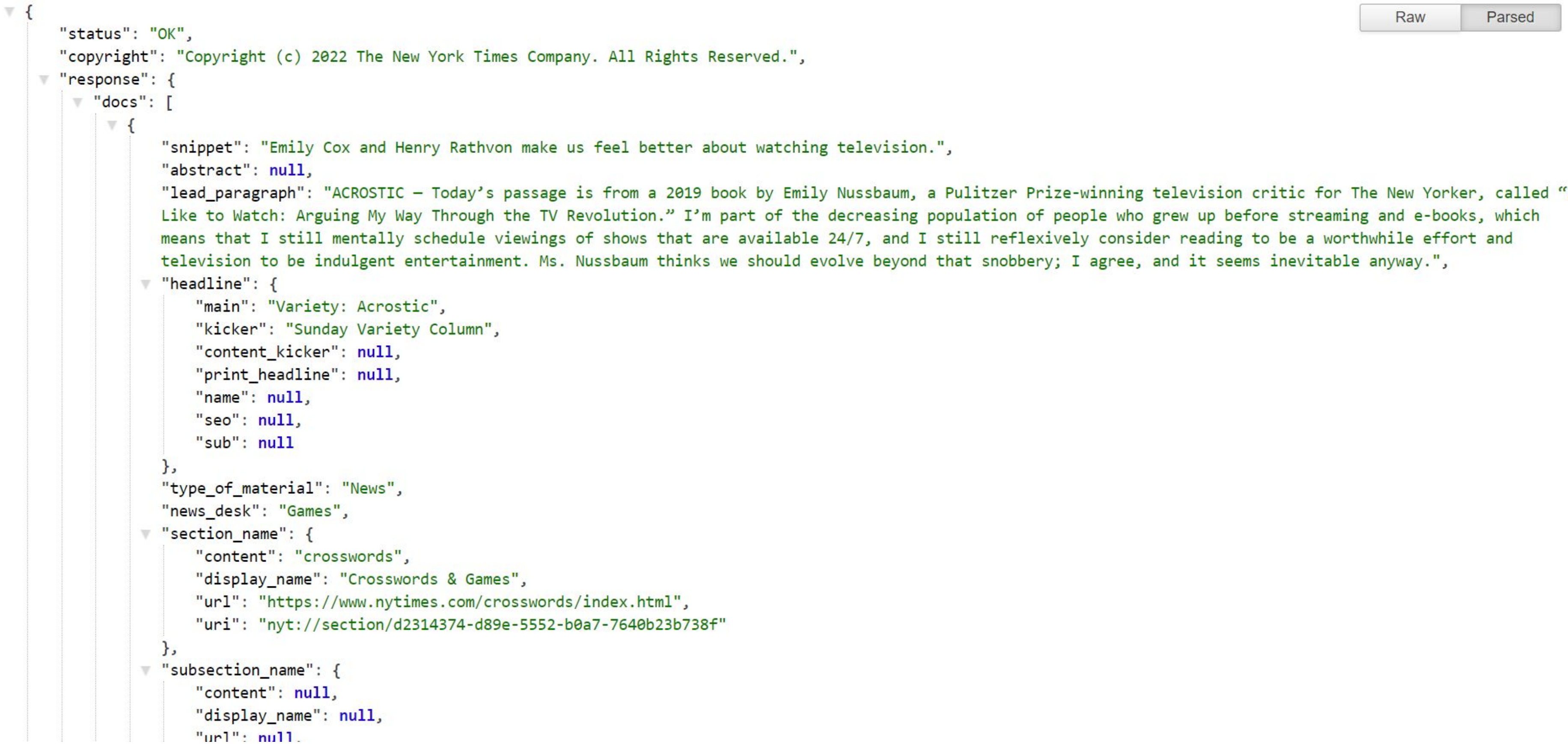
Sort by Relevance
Date Range
Section
Type
Indietro Alt + Freccia sinistra
Avanti Alt + Freccia destra
Ricarica Ctrl + R
Salva con nome... Ctrl + S
Stampa... Ctrl + P
Trasmetti...
Cerca immagini con Google Lens
Invia ai tuoi dispositivi
Crea codice QR per questa pagina
Traduci in English
AdBlock: il miglior ad-blocker di sempre
Scarica le descrizioni delle immagini da Google
Visualizza sorgente pagina Ctrl + U
Ispeziona



Name	Status	Type	Initiator	Size	Time	Waterfall
search?dropmab=false&query=python&sort=best	304	document	Other	709 B	63 ms	
video-python-hunting-thumbWide.jpg	200	jpeg	search?dropmab=false&query=py...	(memory cache)	0 ms	
8539_1_lordofthepythons_190x126.jpg	200	jpeg	search?dropmab=false&query=py...	(memory cache)	0 ms	
meter.js?sourceApp=vi&url=https%3A%2F%2Fwww.nytime...26sort...	200	xhr	main-ab82128....js:35	890 B	181 ms	
b?c1=2&c2=3005403&ns_t=1668151682665&ns_c=UTF-8&c...rop...	204	text/plain	gtm.js?id=GTM-P528B3&gtm_aut...	285 B	133 ms	
ads?pvsid=1310711568625760&correlator=272973497246...598&ga...	(blocked:other)	xhr	pubads_impl_2022110901.js?cb=3...	0 B	455 ms	
activityi;src=5290727;type=allpa0;cat=nyti-0;ord=1...%3Fdropmab%...	302	document / Redirect	gtm.js?id=GTM-P528B3&gtm_aut...	23 B	156 ms	

17 / 83 requests | 6.6 kB / 209 kB transferred | 293 kB / 4.0 MB resources | Finish: 1.8 min | DOMContentLoaded: 614 ms | Load: 1.65 s

https://www.nytimes.com/svc/add/v1/sitesearch.json?q=python&spotlight=true&facet=true



```
{
  "status": "OK",
  "copyright": "Copyright (c) 2022 The New York Times Company. All Rights Reserved.",
  "response": {
    "docs": [
      {
        "snippet": "Emily Cox and Henry Rathvon make us feel better about watching television.",
        "abstract": null,
        "lead_paragraph": "ACROSTIC — Today’s passage is from a 2019 book by Emily Nussbaum, a Pulitzer Prize-winning television critic for The New Yorker, called “I Like to Watch: Arguing My Way Through the TV Revolution.” I’m part of the decreasing population of people who grew up before streaming and e-books, which means that I still mentally schedule viewings of shows that are available 24/7, and I still reflexively consider reading to be a worthwhile effort and television to be indulgent entertainment. Ms. Nussbaum thinks we should evolve beyond that snobbery; I agree, and it seems inevitable anyway.",
        "headline": {
          "main": "Variety: Acrostic",
          "kicker": "Sunday Variety Column",
          "content_kicker": null,
          "print_headline": null,
          "name": null,
          "seo": null,
          "sub": null
        },
        "type_of_material": "News",
        "news_desk": "Games",
        "section_name": {
          "content": "crosswords",
          "display_name": "Crosswords & Games",
          "url": "https://www.nytimes.com/crosswords/index.html",
          "uri": "nyt://section/d2314374-d89e-5552-b0a7-7640b23b738f"
        },
        "subsection_name": {
          "content": null,
          "display_name": null,
          "url": null,
```

Gab

Gab is a social media whose community mainly belongs to the far-right electors from America. It is known for promoting free speech and providing no censorship, but in the end it enforces the circulation of harmful and conspiracy narratives.

Gab

Gab is a social media whose community mainly belongs to the far-right electors from America. It is known for promoting free speech and providing no censorship, but in the end it enforces the circulation of harmful and conspiracy narratives.

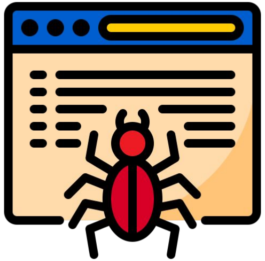
Gab is an example of a social media whose API are not documented, but they are still accessible once authenticated to the website

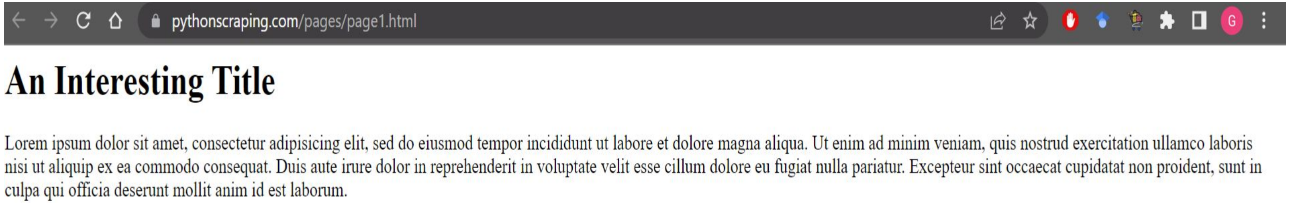
Demo: communicate with Gab APIs

Writing Web Crawlers

Obtaining web data through crawlers

Let’s suppose we are interested in obtaining the content of the following page:



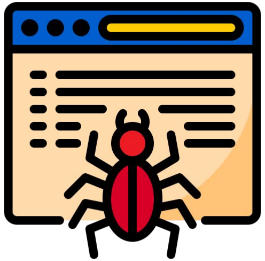


pythonscraping.com/pages/page1.html

An Interesting Title

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Obtaining web data through crawlers



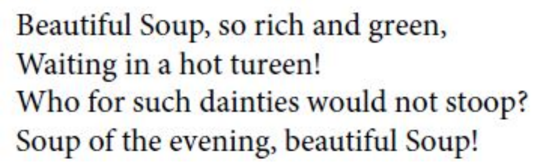
To our current knowledge, we would look for the API that provides the content on the website.

However, the page content is static and it purely relies on the HTML.

How can we overcome this problem? By **crawling and analyzing the HTML**, of course.

Demo: crawling HTML content from a webpage

BeautifulSoup



Beautiful Soup, so rich and green,
Waiting in a hot tureen!
Who for such dainties would not stoop?
Soup of the evening, beautiful Soup!

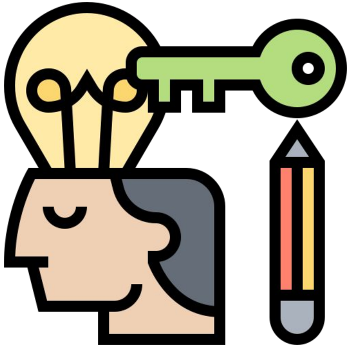
BeautifulSoup is a Python library that helps with the crawling and management of HTML elements, providing a structured way of extracting information from a desired web page.

We will be using the BeautifulSoup4 library, which can be installed with the following command:

pip install beautifulsoup4

Trademarks, Copyrights and Patents: Intellectual Property 101

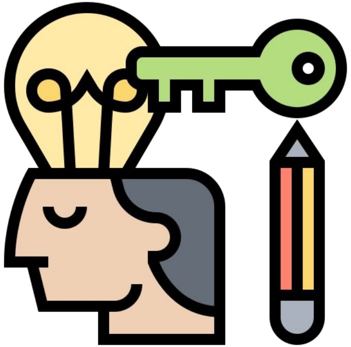
Trademarks, Copyrights, Patents, Oh My!



There are three basic types of Intellectual Property:

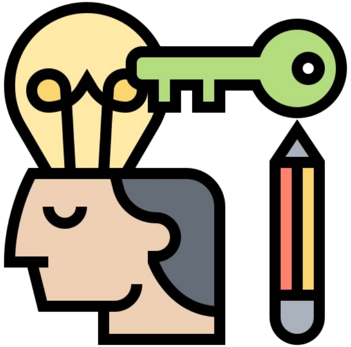
1. **Trademarks** ™: word, phrase, symbol, and/or design that identifies and distinguishes the source of the goods of one party from those of others. Ex: the Coca-Cola logo
2. **Copyrights** ©: every piece of material you create is automatically subject to copyright law as soon as you bring it into existence.
3. **Patents**: used to declare ownership and inventions only (no foto, text or any information itself)

Trademarks, Copyrights, Patents, Oh My!



In the context of web scraping, we need to take into account the Trespass to chattels tort, whereby the infringing party has intentionally interfered with another person's lawful possession of a chattel (movable personal property).

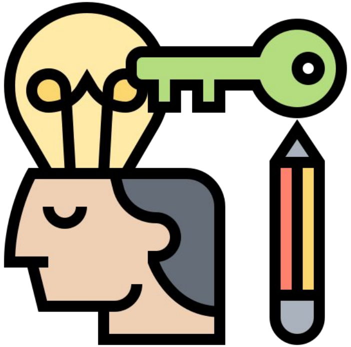
Trademarks, Copyrights, Patents, Oh My!



Three criteria need to be met for a web scraper to violate trespass to chattels:

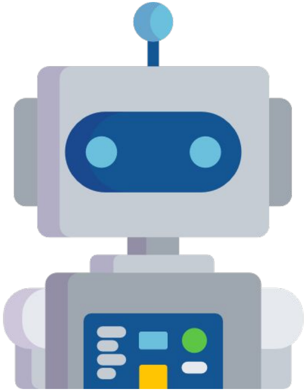
1. **Lack of consent:** because web servers are open to everyone, they are generally “giving consent” to web scrapers as well. However, many websites’ Terms of Service agreements specifically prohibit the use of scrapers.
2. **Actual harm:** if our scrapers take a website down, or limit its ability to serve other users, this can add to the “harm” we cause
3. **Intentionally:** we know what our scraper does, so whenever we obtain data, we do it intentionally

Trademarks, Copyrights, Patents, Oh My!

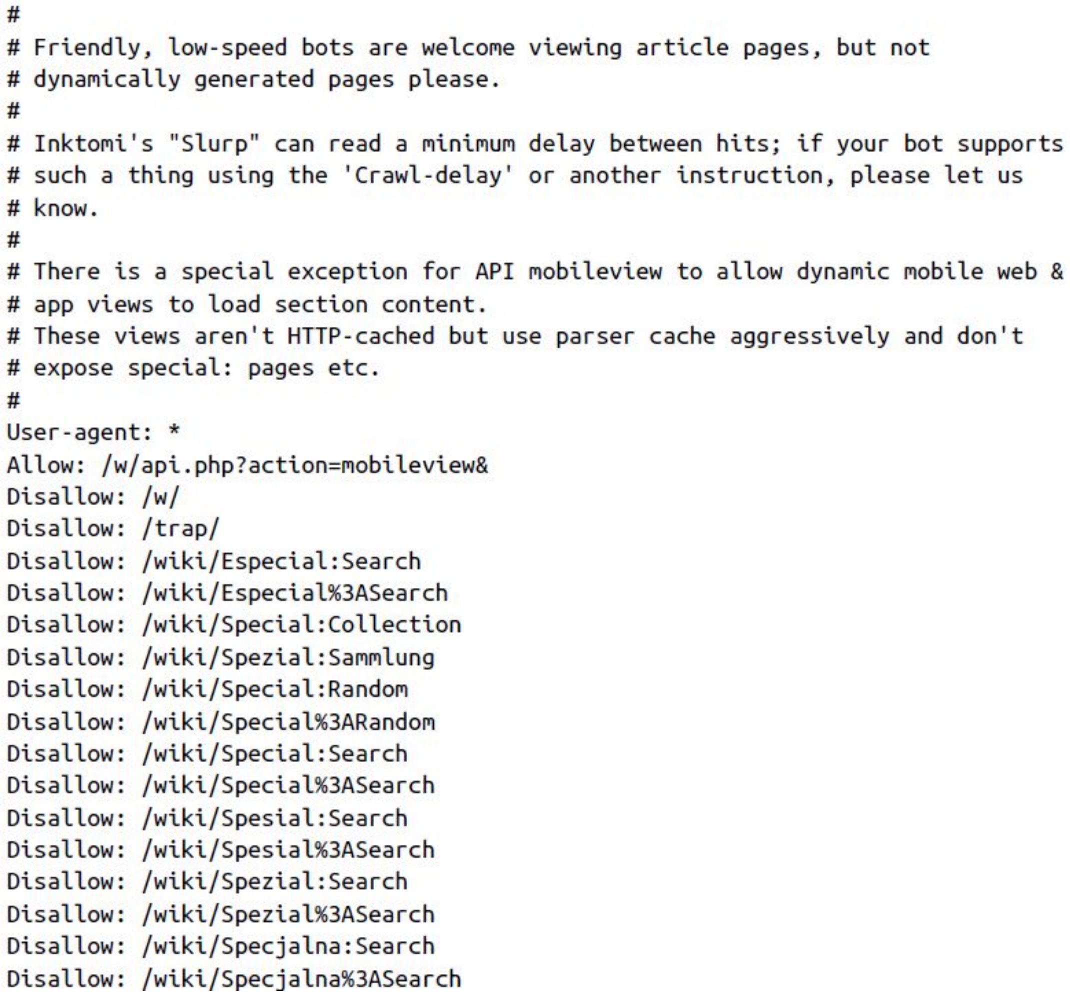


We must meet all three of these criteria for trespass to chattels to apply. However, if we are violating a Terms of Service agreement, but not causing actual harm, we might be violating the copyright law, the DMCA, the Computer Fraud and Abuse Act

robots.txt and Terms of Service

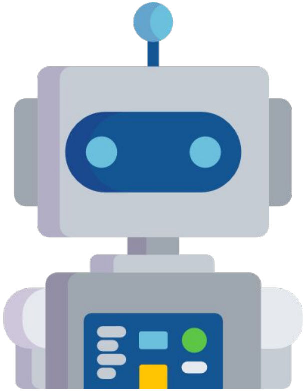


A **robots.txt** file tells search engine crawlers which URLs the crawler can access on your site. This is used mainly to avoid overloading your site with requests; **it is not a mechanism for keeping a web page out of Google**. To keep a web page out of Google, block indexing with noindex or password-protect the page



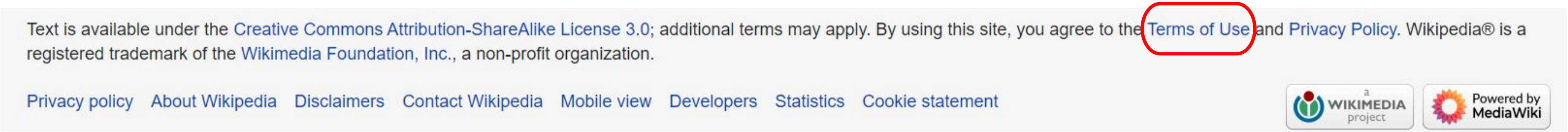
```
#
# Friendly, low-speed bots are welcome viewing article pages, but not
# dynamically generated pages please.
#
# Inktomi's "Slurp" can read a minimum delay between hits; if your bot supports
# such a thing using the 'Crawl-delay' or another instruction, please let us
# know.
#
# There is a special exception for API mobileview to allow dynamic mobile web &
# app views to load section content.
# These views aren't HTTP-cached but use parser cache aggressively and don't
# expose special: pages etc.
#
User-agent: *
Allow: /w/api.php?action=mobileview&
Disallow: /w/
Disallow: /trap/
Disallow: /wiki/Especial:Search
Disallow: /wiki/Especial%3ASearch
Disallow: /wiki/Special:Collection
Disallow: /wiki/Spezial:Sammlung
Disallow: /wiki/Special:Random
Disallow: /wiki/Special%3ARandom
Disallow: /wiki/Special:Search
Disallow: /wiki/Special%3ASearch
Disallow: /wiki/Spesial:Search
Disallow: /wiki/Spesial%3ASearch
Disallow: /wiki/Spezial:Search
Disallow: /wiki/Spezial%3ASearch
Disallow: /wiki/Specjalna:Search
Disallow: /wiki/Specjalna%3ASearch
```

robots.txt and Terms of Service

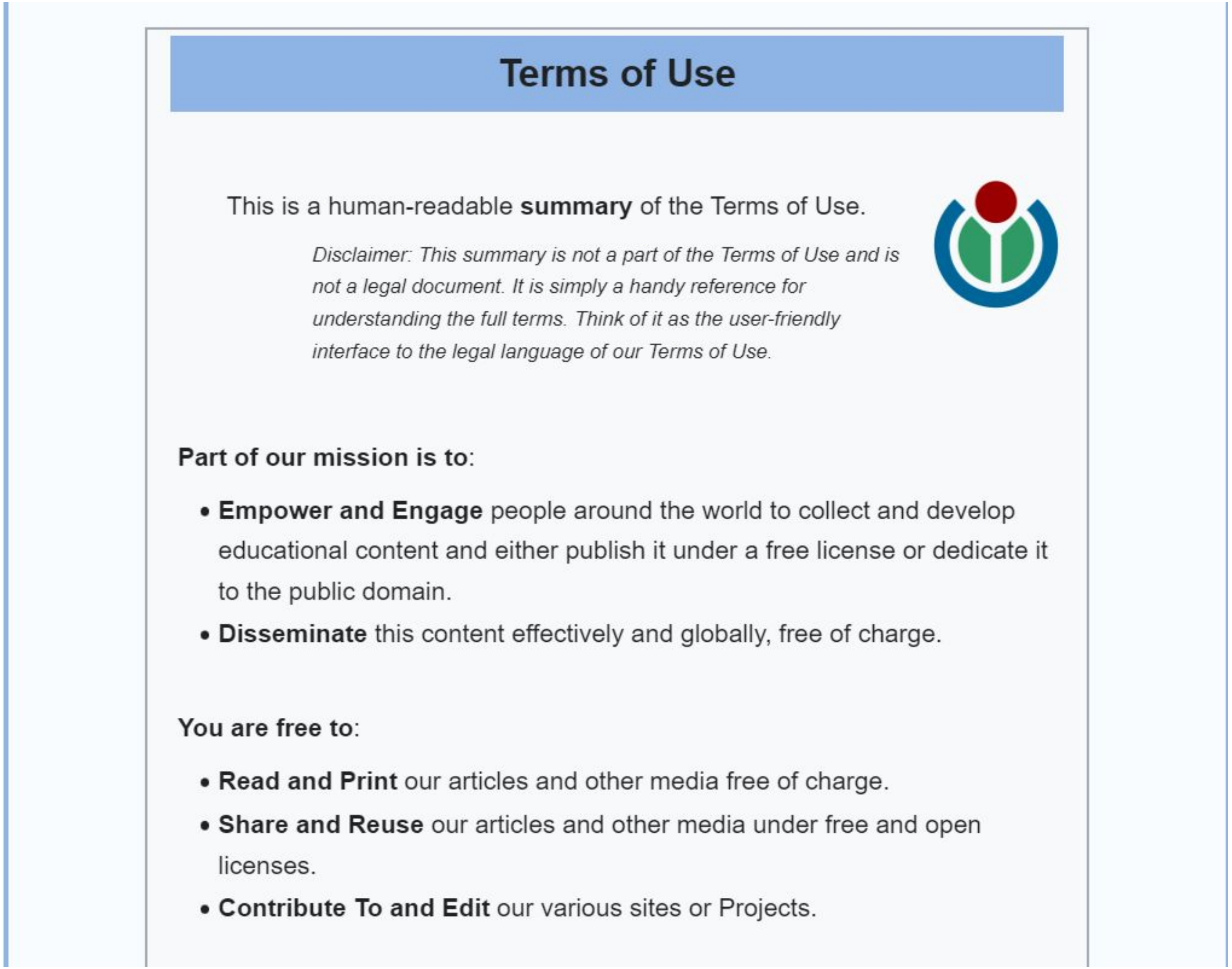


Most sites have a link to their **Terms of Service (TOS)** in the footer on every page.

The TOS contains more than just the rules for web crawlers and automated access; it often has information about what kind of information the website collects, what it does with it,
and usually a legal disclaimer that the services provided by the website come without any express or implied warranty.



Text is available under the Creative Commons Attribution-ShareAlike License 3.0; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.



Terms of Use

This is a human-readable **summary** of the Terms of Use.

Disclaimer: This summary is not a part of the Terms of Use and is not a legal document. It is simply a handy reference for understanding the full terms. Think of it as the user-friendly interface to the legal language of our Terms of Use.

Part of our mission is to:

- **Empower and Engage** people around the world to collect and develop educational content and either publish it under a free license or dedicate it to the public domain.
- **Disseminate** this content effectively and globally, free of charge.

You are free to:

- **Read and Print** our articles and other media free of charge.
- **Share and Reuse** our articles and other media under free and open licenses.
- **Contribute To and Edit** our various sites or Projects.