Web Scraping 101

Module 1 - No-code Tools

Mar 2025

Our Goals Today

After this workshop, you will be able to

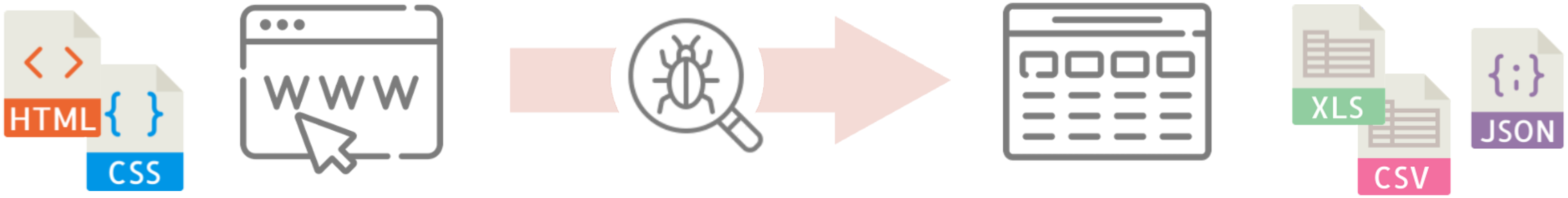
- Understand the basic concepts and principles of web scraping
- Identify available web scraping tools and their appropriate uses
- Use **Power Query** and **Web Scraper** to extract content of interest from the web
- Make wise decisions when collecting data from the web

Web Scraping Basics

What is Web Scraping

> the process of **extracting information** from the web

> **automate**, **rerun** the process



Websites

Identify your target website

Scraping

Extract info from the web

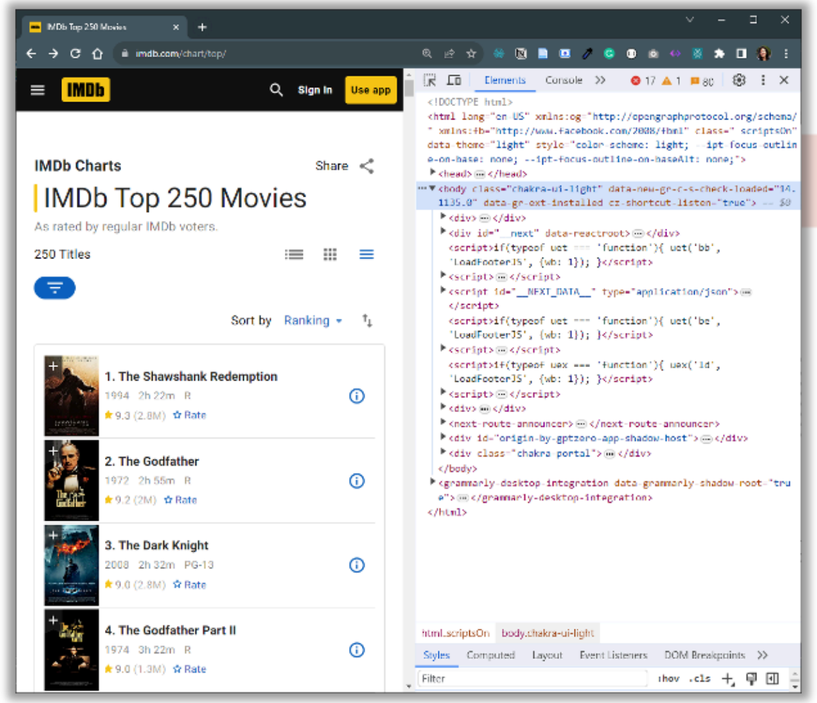
Structured Data

Store it in a structured format / database

What is Web Scraping

> the process of **extracting information** from the web

> **automate**, **rerun** the process



Scraping

Extract info
from the web

	E	F	G	H	I	J	K
1	title	year	imdb-ratin	genre	director	stars	description
2	Dances with Wolves	1990	8.0/10	AdventureDra	Kevin Costner	Kevin Costne	Lieutenant John D
3	Dersu Uzala	1975	8.2/10	AdventureBio	Akira Kurosawa	Maksim Mur	The Russian army
4	Aladdin	1992	8.0/10	AnimationAdv	Ron Clements	Ron Clement	A kindhearted stre
5	The Help	2011	8.1/10	Drama	Tate Taylor	Viola DavisE	An aspiring author
6	The Iron Giant	1999	8.1/10	AnimationActi	Brad Bird	Tim McCanli	A young boy befrie
7	Life of Brian	1979	8.0/10	Comedy	Terry Jones	Graham Cha	Born on the origin
8	Persona	1966	8.1/10	DramaThriller	Ingmar Bergman	Bibi Anderss	A nurse is put in c
9	It Happened One Nigh	1934	8.1/10	ComedyRoma	Frank Capra	Clark GableC	A renegade report
10	The 400 Blows	1959	8.1/10	CrimeDrama	François Truffaut	Jean-Pierre L	A young boy, left v
11	The Sound of Music	1965	8.1/10	BiographyDra	Robert Wise	Georg Hurda	A young novice is
12	The Handmaiden	2016	8.1/10	DramaRoman	Park Chan-wook	Kim Min-hee	A woman is hired
13	Cool Hand Luke	1967	8.1/10	CrimeDrama	Stuart Rosenberg	Paul Newma	A laid back Southe
14	Rebecca	1940	8.1/10	DramaFilm-Nc	Alfred Hitchcock	Daphne Du M	A self-conscious w
15	Amores Perros	2000	8.1/10	DramaThriller	Alejandro G. Iñár	Emilio Echev	A horrific car accic
16	My Father and My Son	2005	8.2/10	DramaFamily	Çagan Irmak	Çetin Tekind	The family of a lef
17	Jai Bhim	2021	8.8/10	CrimeDramaM	T.J. Gnanavel	SuriyaLijo M	When a tribal man
18	The Battle of Algiers	1966	8.1/10	DramaWar	Gillo Pontecorvo	Brahim Hadj	In the 1950s, fear
19	The Grapes of Wrath	1940	8.1/10	Drama	John Ford	Henry Fonda	An Oklahoma fami
20	Hachi: A Dog's Tale	2009	8.1/10	BiographyDra	Lasse Hallström	Richard Gere	A college professo
21	Pather Panchali	1955	8.2/10	Drama	Satyajit Ray	Kanu Banner	Impoverished pries
22	Pirates of the Caribbe	2003	8.1/10	ActionAdventu	Gore Verbinski	Ted ElliottTe	Blacksmith Will Tu
23	To Be or Not to Be	1942	8.2/10	ComedyRoma	Ernst Lubitsch	Carole Lomb	During the Nazi oc

Things to consider before you scrape

1. Can I get the data in an easier way?

Is the data **already available** somewhere?

Adapted from: DeVito, N. J., Richards, G. C., & Inglesby, P. (2020). How we learnt to stop worrying and love web scraping. *Nature*, 585(7826), 621–622. https://doi.org/10.1038/d41586-020-02558-0

Things to consider

before you scrape

Adapted from: DeVito, N. J., Richards, G. C., & Inglesby, P. (2020). How we learnt to stop worrying and love web scraping. *Nature*, 585(7826), 621–622. https://doi.org/10.1038/d41586-020-02558-0

1. Can I get the data in an easier way?

Is the data **already available** somewhere?

Find raw data

Google Dataset Search

開放數據平台 DATA.GOV.HK

Hugging Face (data + code)

kaggle (data + code)

Library databases, e.g.

Passport (consumer data)

CEIC (economy data)

Find research data

zenodo

figshare

DataSpace@HKUST (HKUST data repository)

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES (find data repository by subject)

Things to consider
before you scrape

2. Can this website be scraped, technically?

- Any restrictions, e.g. CAPTCHAS (“I’m not a robot” test)?
- Any rules to follow, e.g. no. of requests per min?

Adapted from: DeVito, N. J., Richards, G. C., & Inglesby, P. (2020). How we learnt to stop worrying and love web scraping. *Nature*, 585(7826), 621–622. https://doi.org/10.1038/d41586-020-02558-0

Things to consider before you scrape

2. Can this website be scraped, technically?

- Any restrictions, e.g. CAPTCHAS (“I’m not a robot” test)?
- Any rules to follow, e.g. no. of requests per min?

3. Can this website be scraped, legally?

- Any restrictions on copyright, or privacy issues in scraping or sharing?

Adapted from: DeVito, N. J., Richards, G. C., & Inglesby, P. (2020). How we learnt to stop worrying and love web scraping. *Nature*, 585(7826), 621–622. https://doi.org/10.1038/d41586-020-02558-0

Things to consider before you scrape

2. Can this website be scraped, technically?

- Any restrictions, e.g. CAPTCHAS (“I’m not a robot” test)?
- Any rules to follow, e.g. no. of requests per min?

3. Can this website be scraped, legally?

- Any restrictions on copyright, or privacy issues in scraping or sharing?

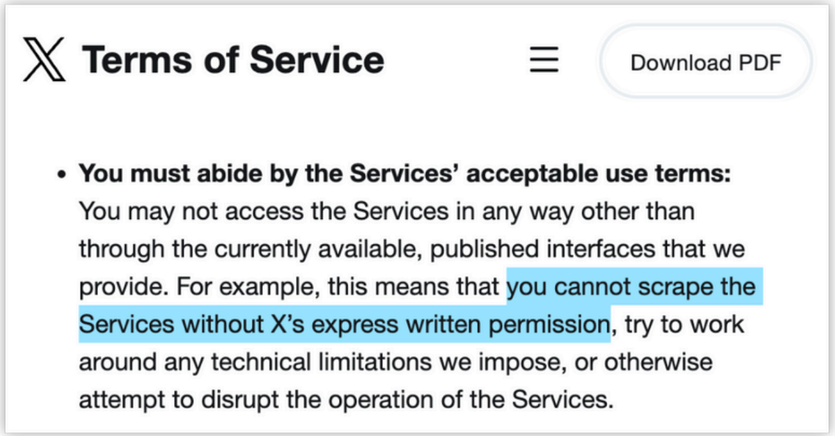
Check “Terms of Use”, “robots.txt”, any official API available, copyright / data sharing policies.

Adapted from: DeVito, N. J., Richards, G. C., & Inglesby, P. (2020). How we learnt to stop worrying and love web scraping. *Nature*, 585(7826), 621–622. https://doi.org/10.1038/d41586-020-02558-0

Things to consider before you scrape

Example

https://x.com/en/tos



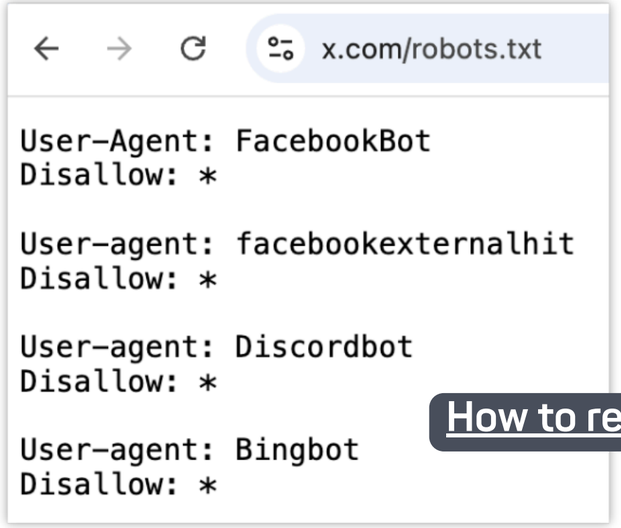
Terms of Service

Download PDF

- **You must abide by the Services' acceptable use terms:** You may not access the Services in any way other than through the currently available, published interfaces that we provide. For example, this means that you cannot scrape the Services without X's express written permission, try to work around any technical limitations we impose, or otherwise attempt to disrupt the operation of the Services.

to give instructions to web crawlers or bots about which parts of a website they are allowed or disallowed to access

https://x.com/**robots.txt**



x.com/robots.txt

```
User-Agent: FacebookBot
Disallow: *

User-agent: facebookexternalhit
Disallow: *

User-agent: Discordbot
Disallow: *

User-agent: Bingbot
Disallow: *
```

How to read this

More examples:

https://arxiv.org/**robots.txt** (allow, with crawl delay)

https://www.amazon.com/**robots.txt** (block GPTbot)

https://stackoverflow.com/**robots.txt** (block all bots)

Adapted from: DeVito, N. J., Richards, G. C., & Inglesby, P. (2020). How we learnt to stop worrying and love web scraping. *Nature*, 585(7826), 621–622. https://doi.org/10.1038/d41586-020-02558-0

Web Scraping Tools

Not all scraping tasks need programming!

Pick the tool that fits your needs.

No-code tools

(Browser plug-in, e.g. WebScraper)

- “Point-and-click”
- Static and simple websites
- Small scale (a few pages)

Programming

(e.g. Python selenium, scrapy)

- Dynamic websites
- Large scale projects
- Integrate with databases, build complex workflows

Not all scraping tasks need programming!

Pick the tool that fits your needs.

No-code tools

Power Query

Web Scraper

Programming

(e.g. Python selenium, scrapy)

- Dynamic websites
- Large scale projects
- Integrate with databases, build complex workflows

Power Query

A built-in Excel tool, for connecting and manipulating data from various sources, including web sources.

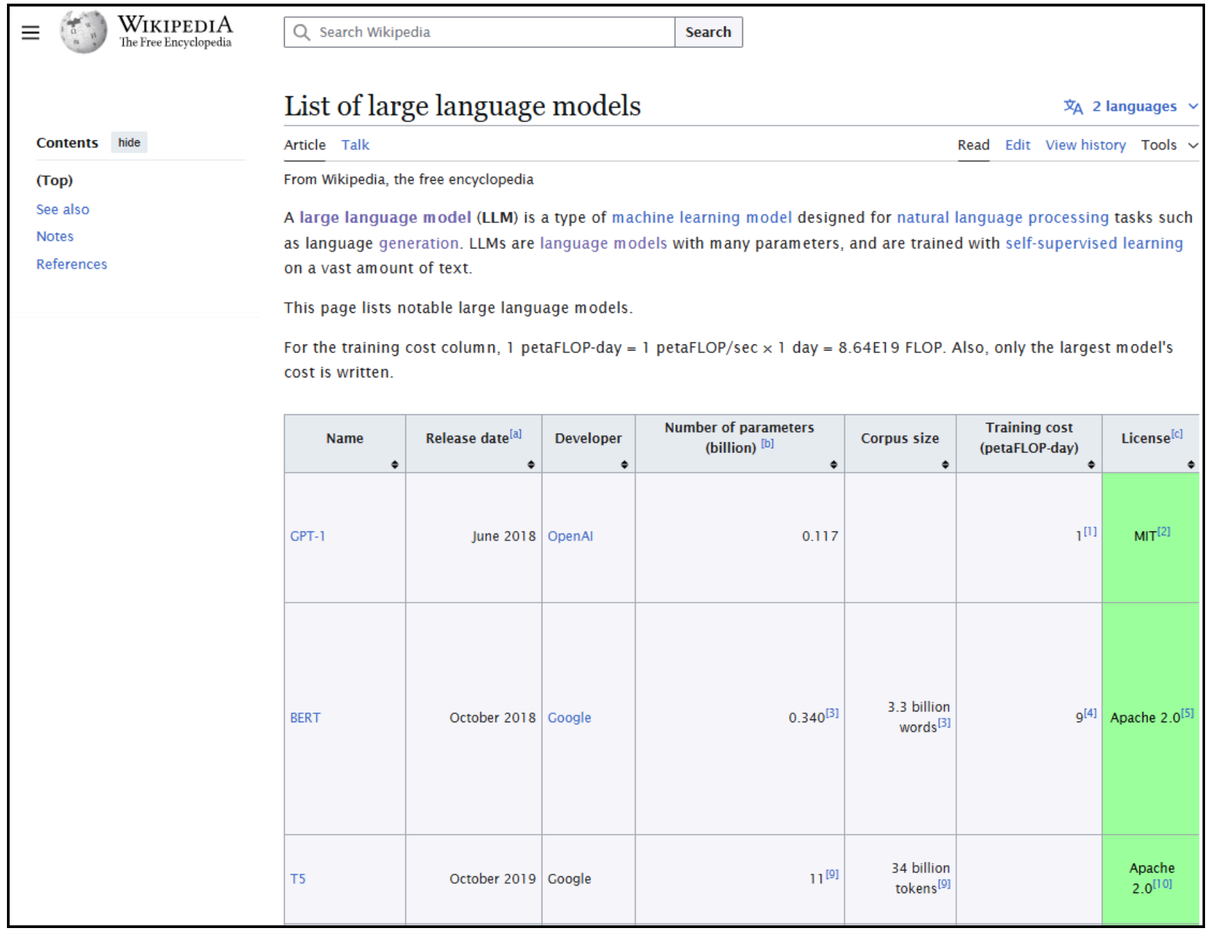
1. Get the page URL
2. Select the info you want
3. Extract the data
4. Save in Excel (for further data manipulation)

Note: Some features (e.g. get data from Web) **are NOT supported in Mac** (learn more). So, use Windows PCs to run Power Query.

Demo

Extract info from a Wikipedia page

https://en.wikipedia.org/wiki/List_of_large_language_models



WIKIPEDIA
The Free Encyclopedia

Search Wikipedia | Search

List of large language models

2 languages

Contents hide

(Top)
See also
Notes
References

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

A **large language model** (**LLM**) is a type of machine learning model designed for natural language processing tasks such as language generation. LLMs are language models with many parameters, and are trained with self-supervised learning on a vast amount of text.

This page lists notable large language models.

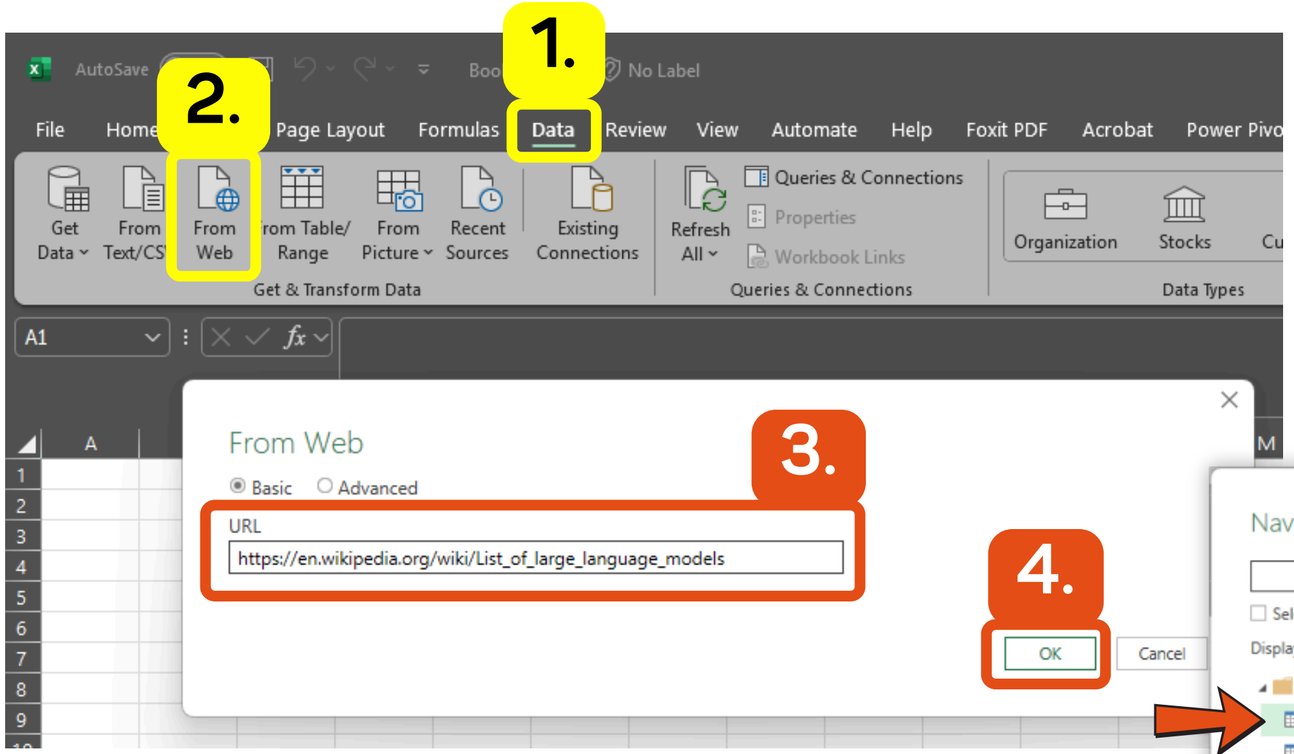
For the training cost column, 1 petaFLOP-day = 1 petaFLOP/sec × 1 day = 8.64E19 FLOP. Also, only the largest model's cost is written.

Name	Release date[a]	Developer	Number of parameters (billion) [b]	Corpus size	Training cost (petaFLOP-day)	License[c]
GPT-1	June 2018	OpenAI	0.117		1[1]	MIT[2]
BERT	October 2018	Google	0.340[3]	3.3 billion words[3]	9[4]	Apache 2.0[5]
T5	October 2019	Google	11[9]	34 billion tokens[9]		Apache 2.0[10]

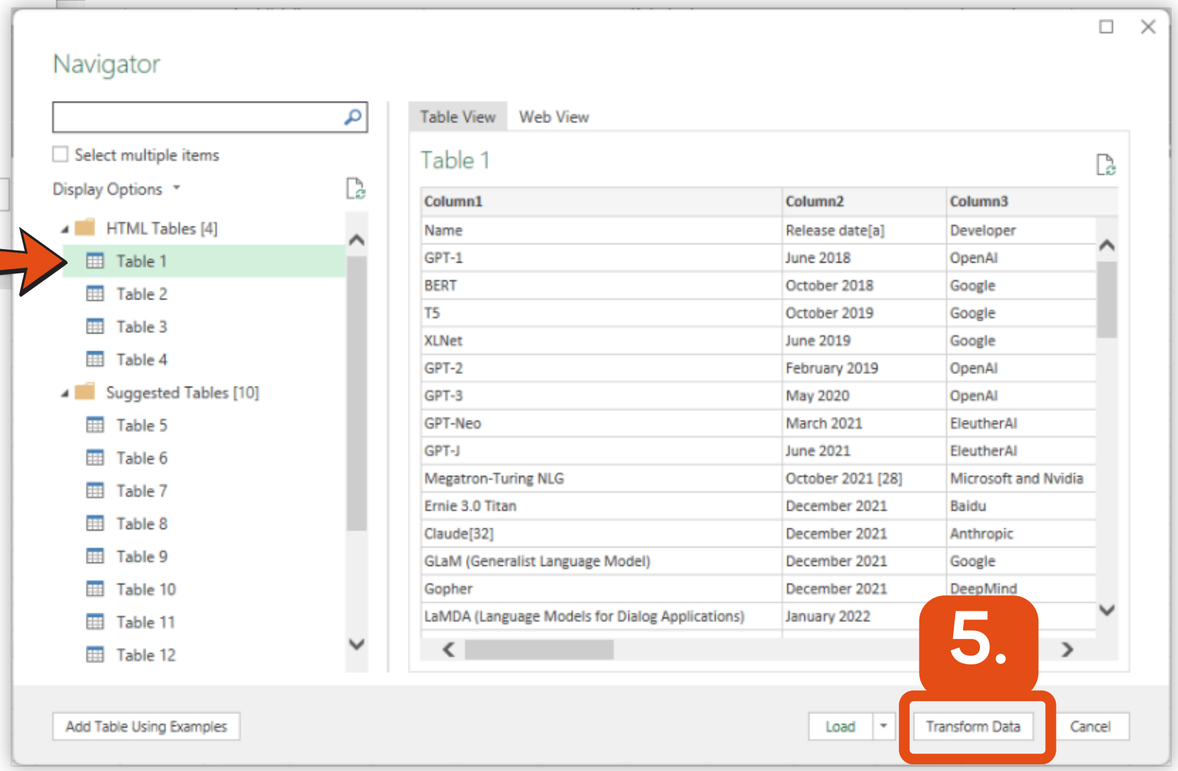
What we want:

- A cleaned list of Large Language Models
- Saved in Excel

Power Query



- **Data > From Web > Type in URL > OK**
- **Select target table > Transform Data**



Target webpage:
https://en.wikipedia.org/wiki/List_of_large_language_models

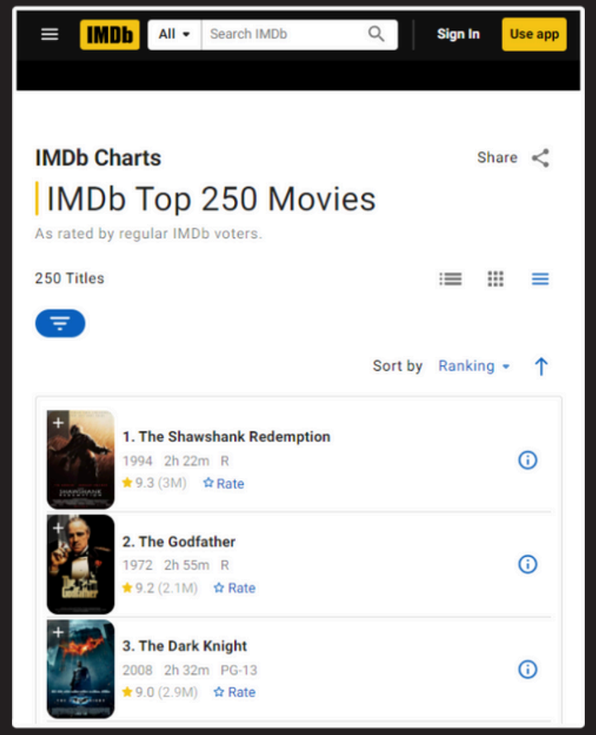
Transform Data

Hands-on (1)

Power Query

Get the list of IMDb Top 250 Movies

https://www.imdb.com/chart/top/?ref_=nv_mv_250



	A	B	C
1	Rank	Title	Rating
2	1	The Shawshank Redemption	9.3
3	2	The Godfather	9.2
4	3	The Dark Knight	9
5	4	The Godfather Part II	9
6	5	12 Angry Men	9
7	6	The Lord of the Rings: The Return of the King	9
8	7	Schindler's List	9
9	8	Pulp Fiction	8.9
10	9	The Lord of the Rings: The Fellowship of the Ring	8.9
11	10	The Good, the Bad and the Ugly	8.8
12	11	Forrest Gump	8.8
13	12	The Lord of the Rings: The Two Towers	8.8
14	13	Fight Club	8.8
15	14	Inception	8.8
16	15	Star Wars: Episode V - The Empire Strikes Back	8.7
17	16	The Matrix	8.7
18	17	Goodfellas	8.7
19	18	One Flew Over the Cuckoo's Nest	8.7
20	19	Interstellar	8.7
21	20	Se7en	8.6
22	21	It's a Wonderful Life	8.6
23	22	Seven Samurai	8.6
24	23	The Silence of the Lambs	8.6
25	24	Saving Private Ryan	8.6

What we want:

List of movies

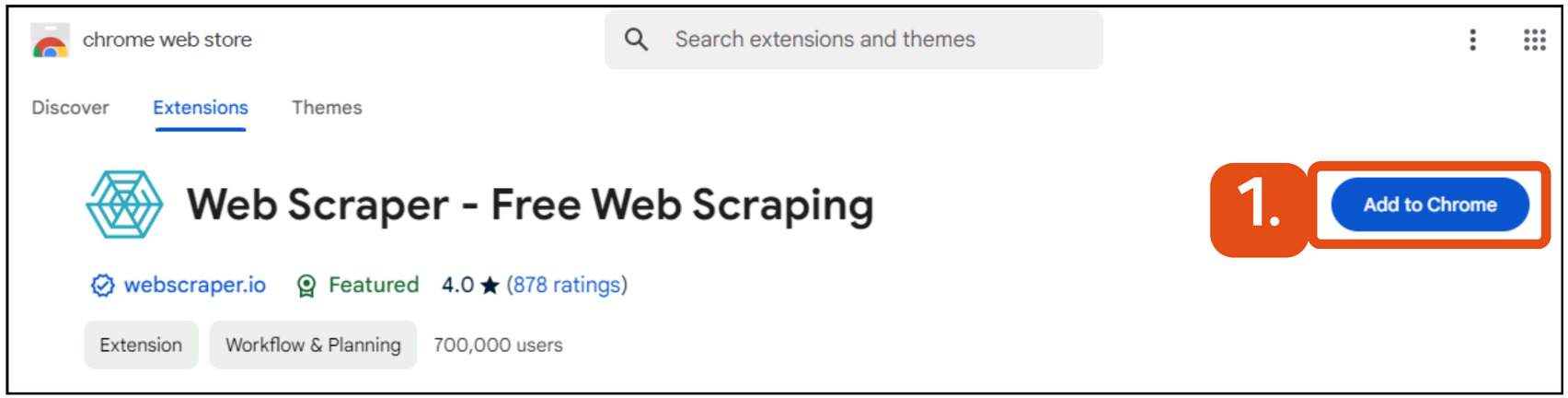
- Rank
- Title
- Rating

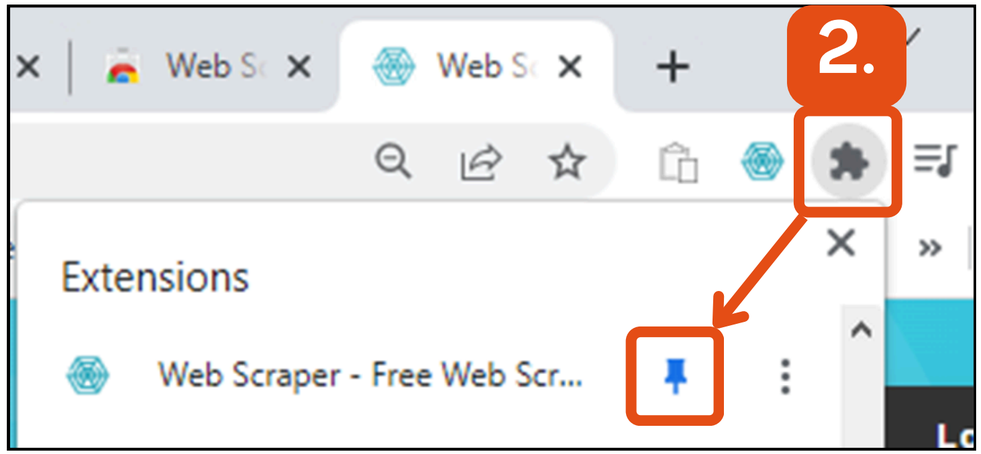
Web Scraper

A browser plug-in for web scraping.

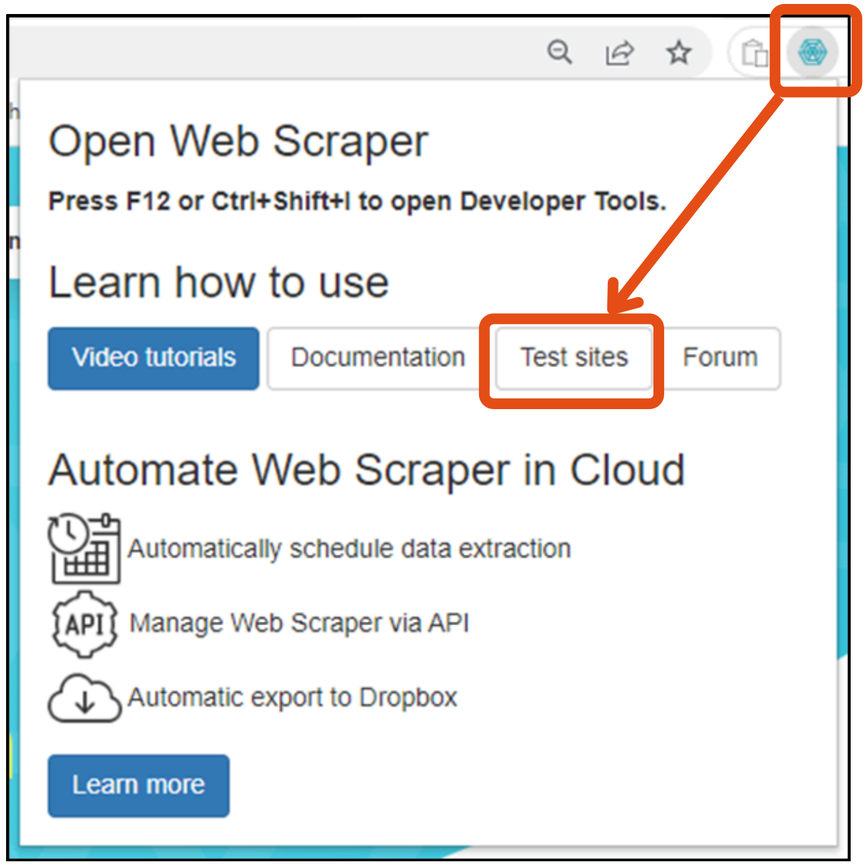
1. Understand basic structure of the page
2. Point and click the “field”
3. Extract the data
4. Download in .csv or .xlsx files

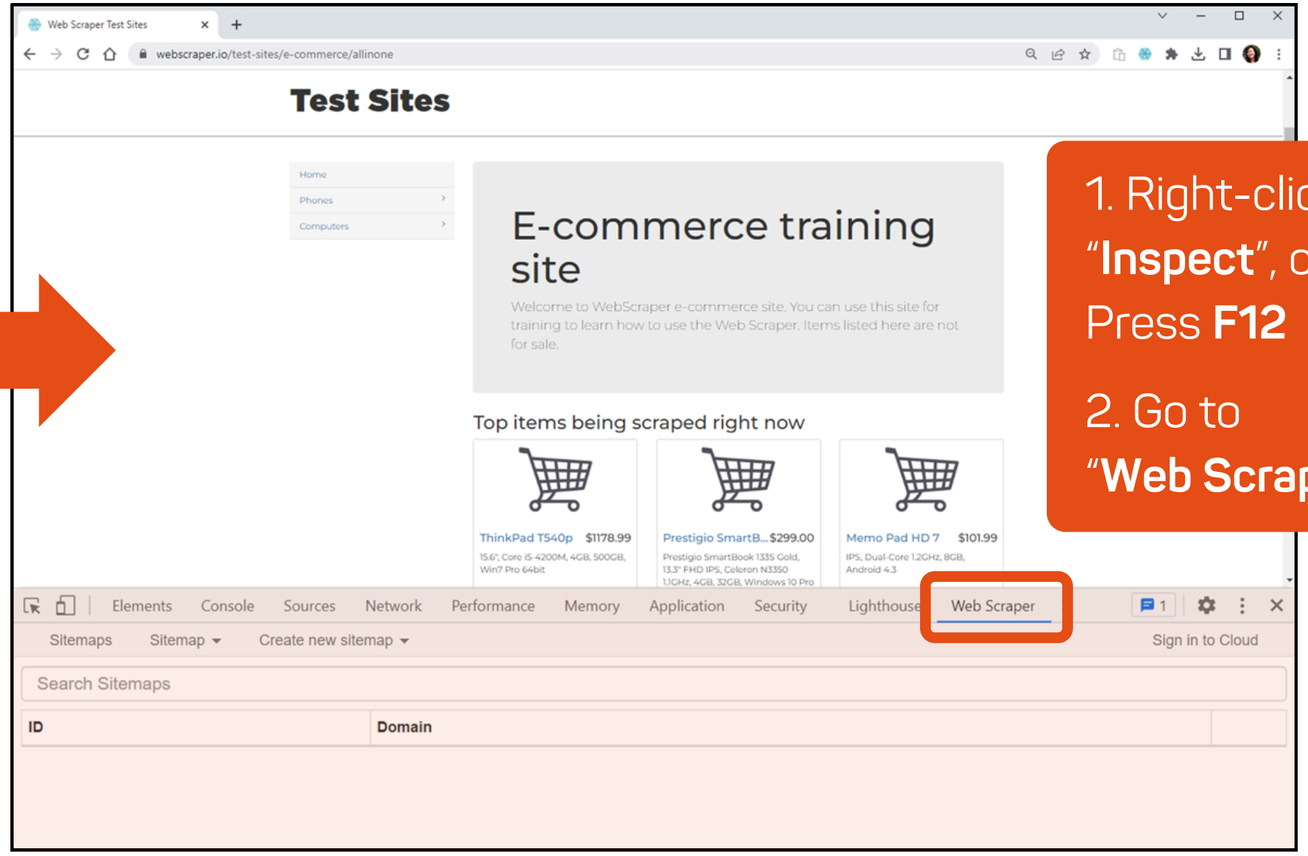
Install Web Scraper plug-in





Launch Web Scraper





1. Right-click and "**Inspect**", or Press **F12**

2. Go to "**Web Scraper**"

Some HTML Basics

HTML - **Structure** of a webpage

```
HTML
1  <html>
2
3  <head>
4    <title>Title of the page</title>
5  </head>
6
7  <body>
8    <div>
9      <h1>Header</h1>
10     <p>This is a paragraph.</p>
11     <p>This is another paragraph.</p>
12   </div>
13
14   <div>
15     <p>Here is a link:</p>
16     <a href="https://library.hkust.edu.hk/">Library homepage</a>
17   </div>
18 </body>
19
20 </html>
CSS
```

Tags

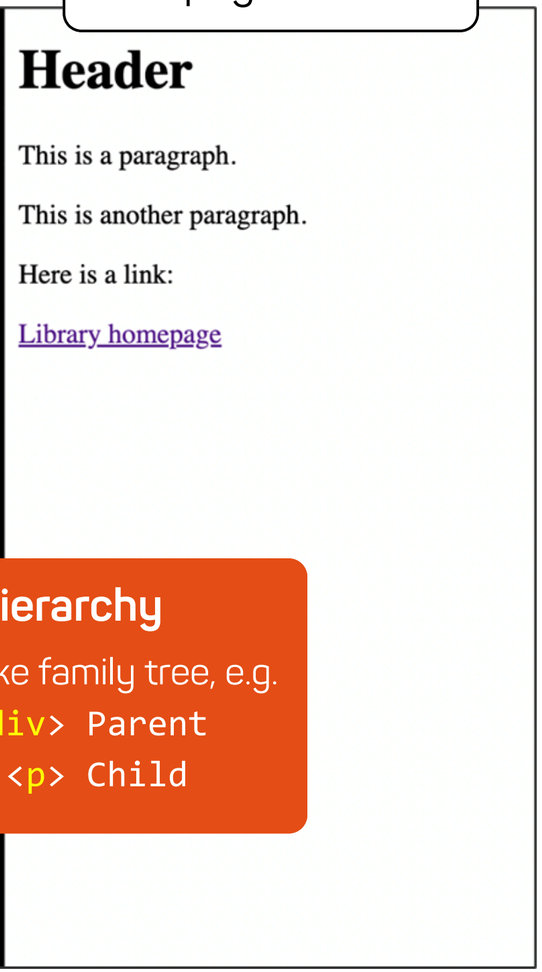
```
<body>…</body>
<div>…</div>
<h1>…</h1>
<p>…</p>
```

Hierarchy

Like family tree, e.g.

```
<div> Parent
  <p> Child
```

Webpage Preview



https://codepen.io/asterzhao/pen/oNaXYKw?editors=1100

Some HTML Basics

HTML - **Structure** of a webpage

```html
<body>
  <div class="some-class">
    <h1>Header</h1>
    <p>This is a paragraph.</p>
    <p id="unique-id">This is ano
  </div>

  <div>
    <p>Here is a link:</p>
    <a href="https://library.hkust.edu.hk/">Library homepage</a>
  </div>
</body>
```

Attributes

<class = “...”>
<id = “...”>
<href = “...”>

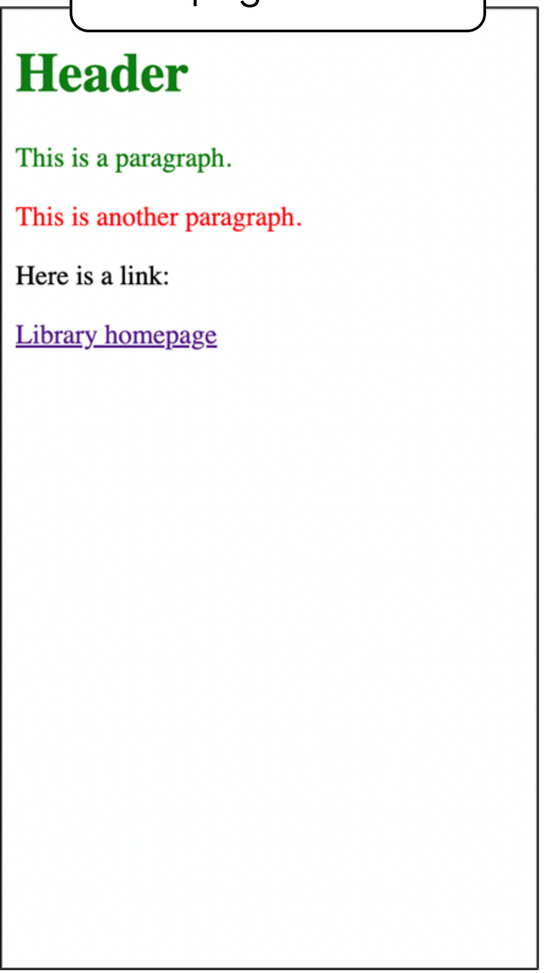
Useful for scrapping!

CSS - **Style** of a webpage

```css
.some-class {
  color: green
}

#unique-id {
  color: red
}
```

Webpage Preview



https://codepen.io/asterzhao/pen/oNaXYKw?editors=1100

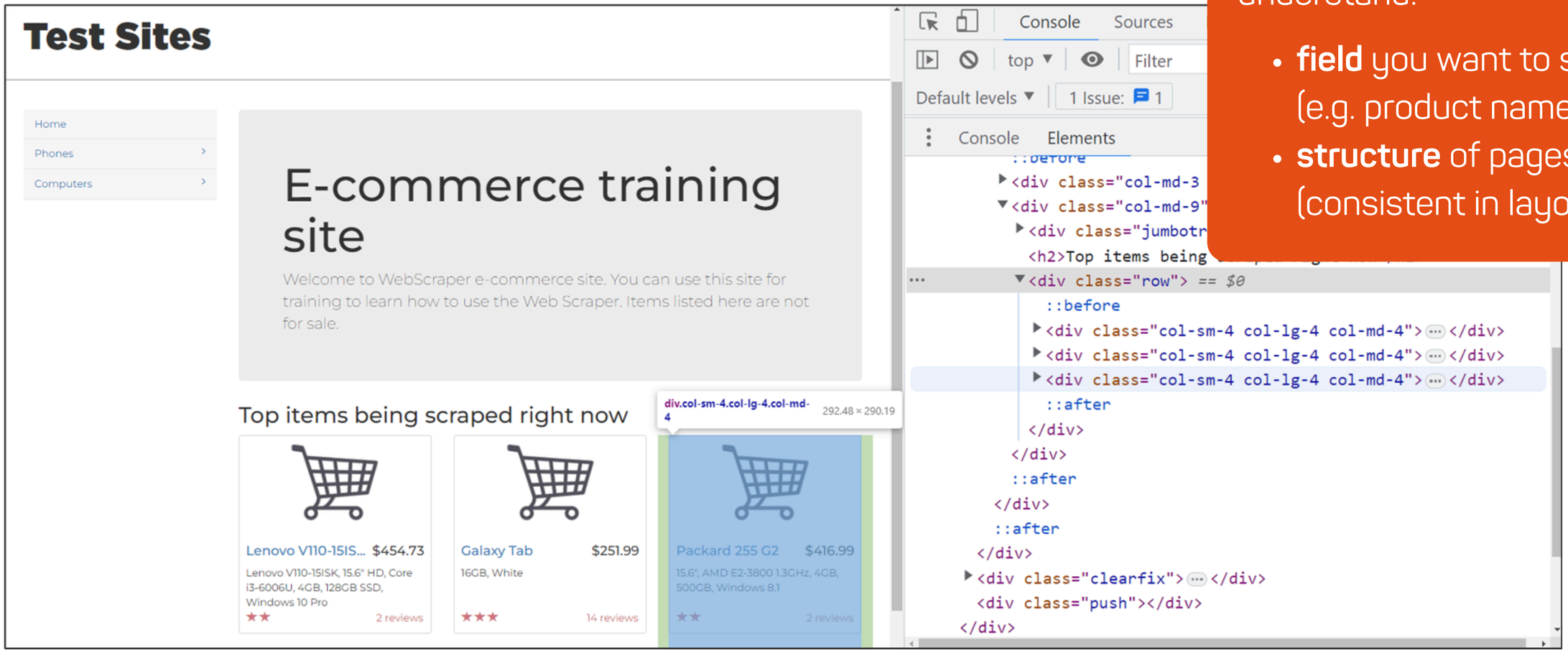
Demo

WebScraper - Test site

https://webscraper.io/test-sites/e-commerce/allinone

"Inspect" the page helps you understand:

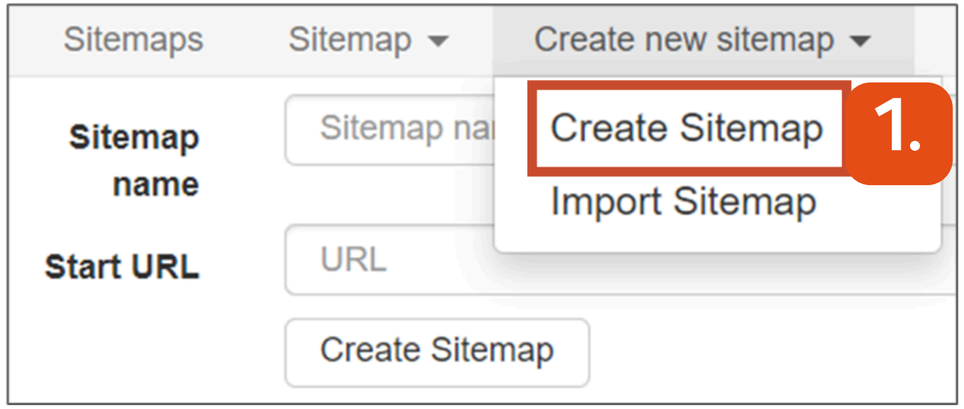
- **field** you want to scrape (e.g. product name, price)
- **structure** of pages (consistent in layout?)



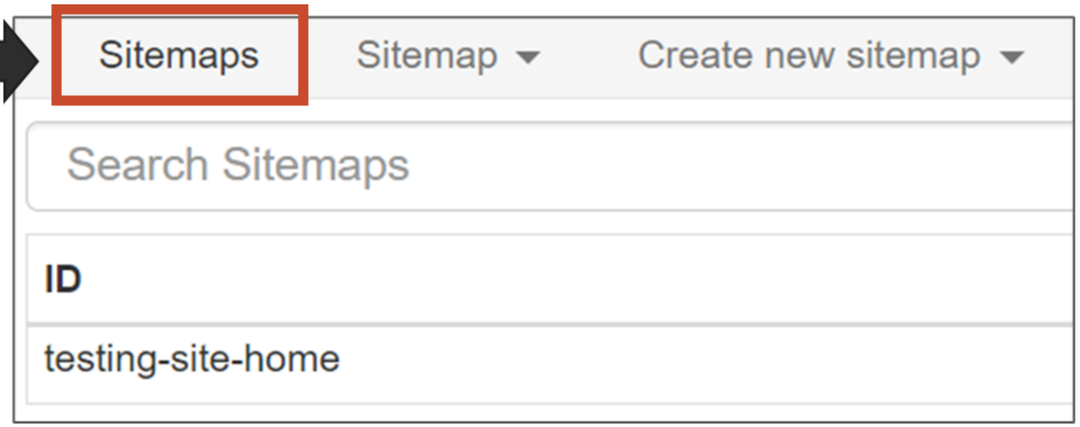
Demo WebScraper - Test site

Steps:

1. Create a sitemap



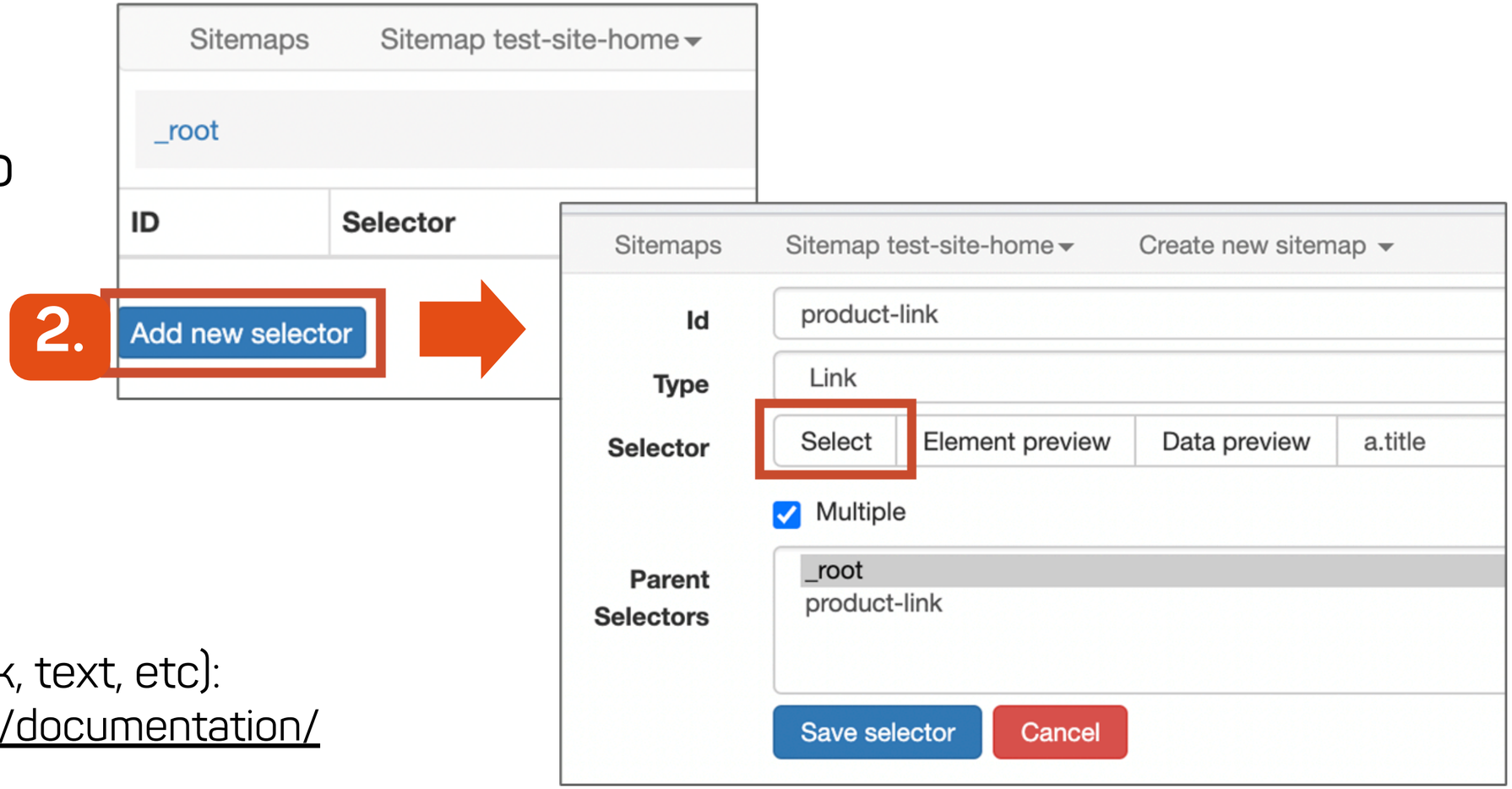
View all created Sitemaps



Demo WebScraper - Test site

Steps:

1. Create a sitemap
2. **Add selectors**



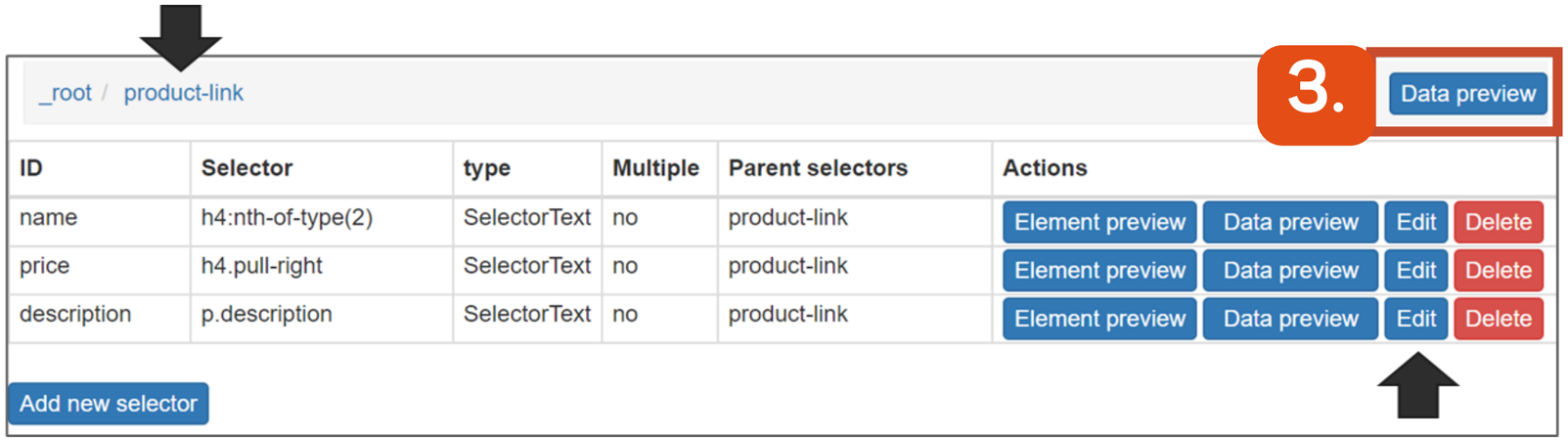
More about Selectors (link, text, etc):
https://www.webscraper.io/documentation/selectors

Demo WebScraper - Test site

Steps:

1. Create a sitemap
2. Add selectors
3. **Preview data**

Navigate to upper level



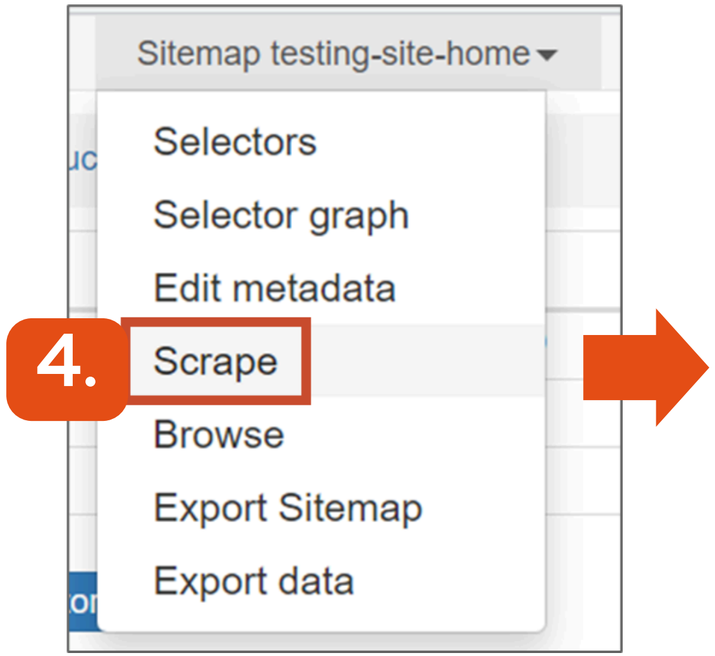
Edit selectors

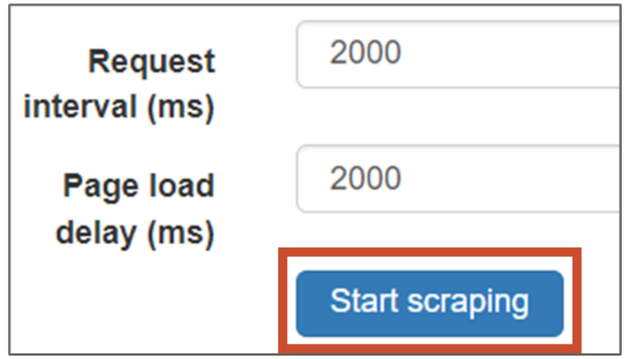
Demo

WebScraper - Test site

Steps:

1. Create a sitemap
2. Add selectors
3. Preview data
4. **Scrape**





Suggest using the default delay

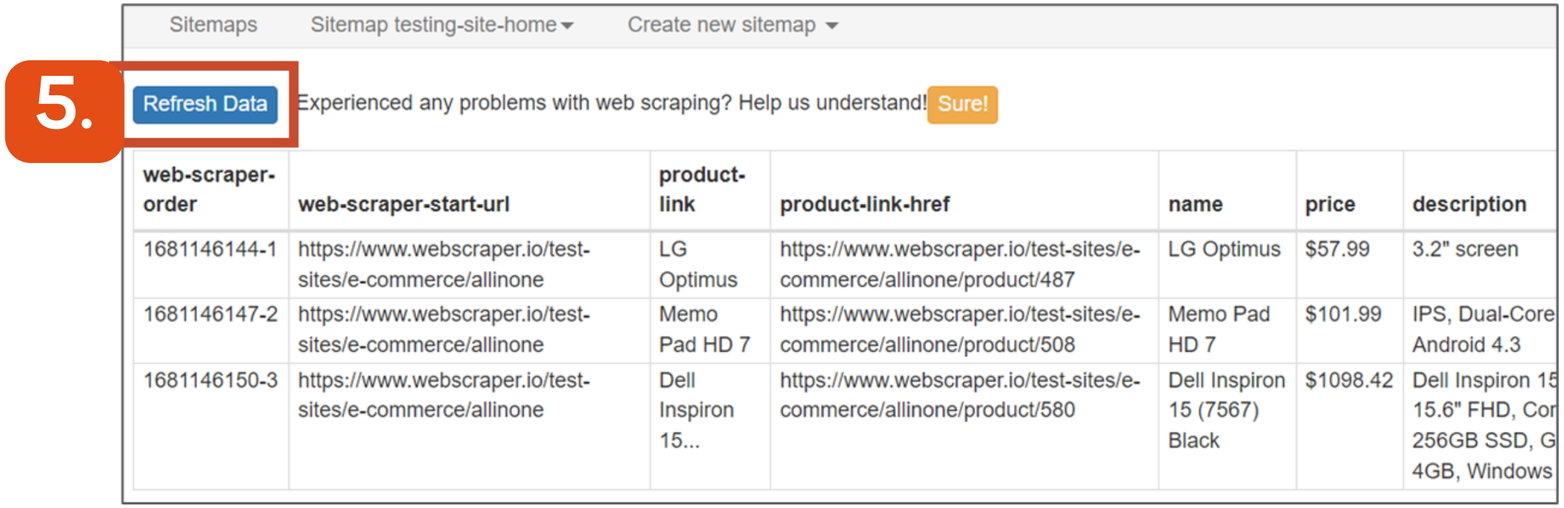
(**2000 ms** - can be longer but NOT shorter)

Demo

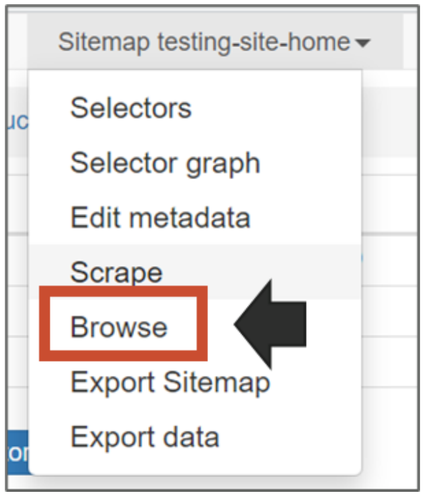
WebScraper - Test site

Steps:

1. Create a sitemap
2. Add selectors
3. Preview data
4. Scrape
5. **Refresh to check results**



web-scraper-order	web-scraper-start-url	product-link	product-link-href	name	price	description
1681146144-1	https://www.webscraper.io/test-sites/e-commerce/allinone	LG Optimus	https://www.webscraper.io/test-sites/e-commerce/allinone/product/487	LG Optimus	$57.99	3.2" screen
1681146147-2	https://www.webscraper.io/test-sites/e-commerce/allinone	Memo Pad HD 7	https://www.webscraper.io/test-sites/e-commerce/allinone/product/508	Memo Pad HD 7	$101.99	IPS, Dual-Core Android 4.3
1681146150-3	https://www.webscraper.io/test-sites/e-commerce/allinone	Dell Inspiron 15...	https://www.webscraper.io/test-sites/e-commerce/allinone/product/580	Dell Inspiron 15 (7567) Black	$1098.42	Dell Inspiron 15 15.6" FHD, Cor 256GB SSD, G 4GB, Windows

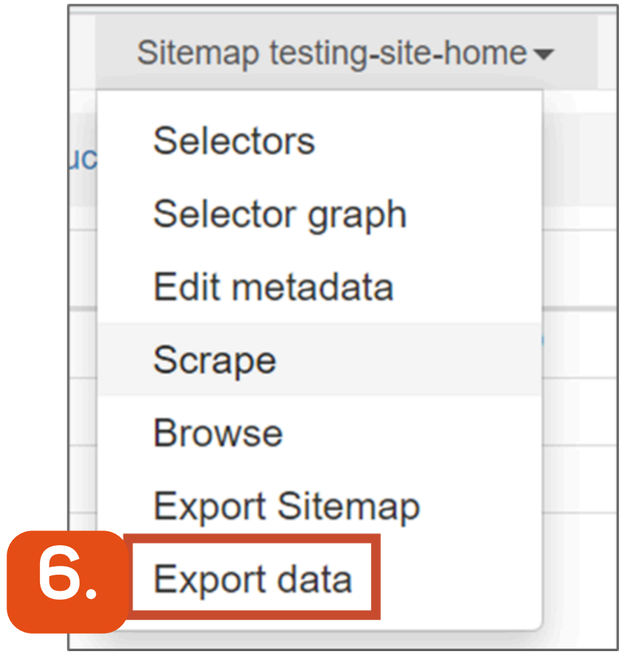


If you left the page and wish to go back to result list, click "**Browse**".

Demo WebScraper - Test site

Steps:

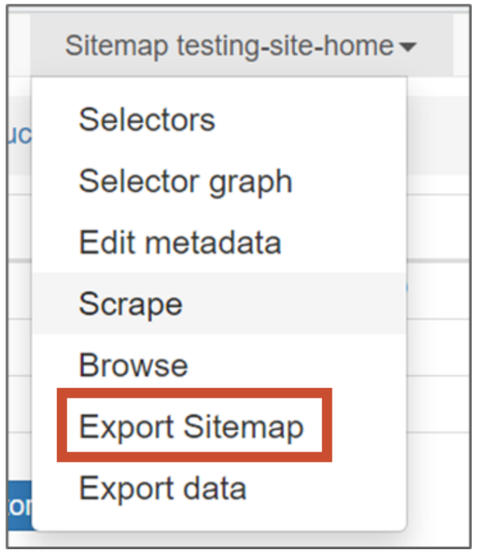
1. Create a sitemap
2. Add selectors
3. Preview data
4. Scrape
5. Refresh to check results
6. **Export data** (.csv or .xlsx)



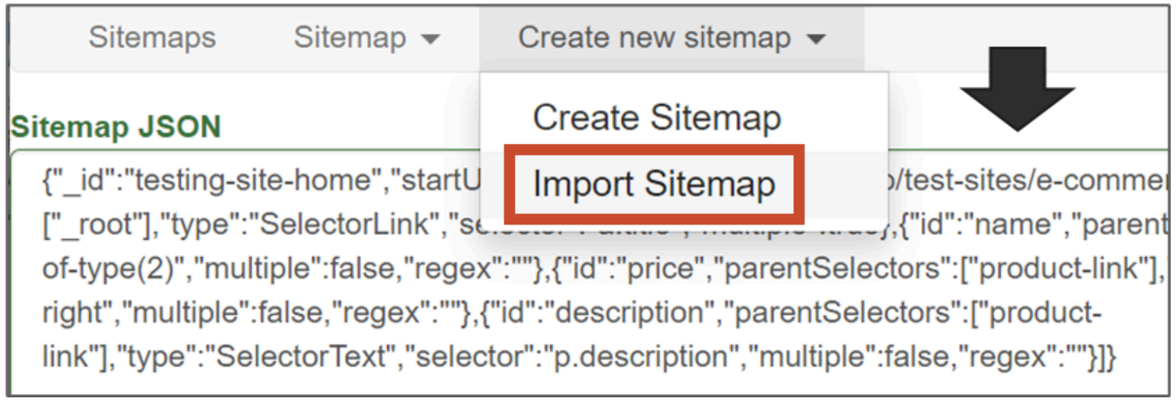
Demo WebScraper - Test site

Optional:

Export Sitemap for backup
(to share or reuse in another device)



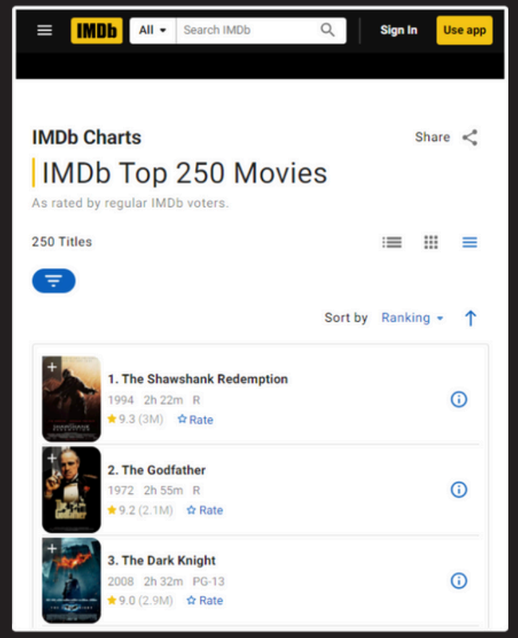
Sitemap is in **JSON** format

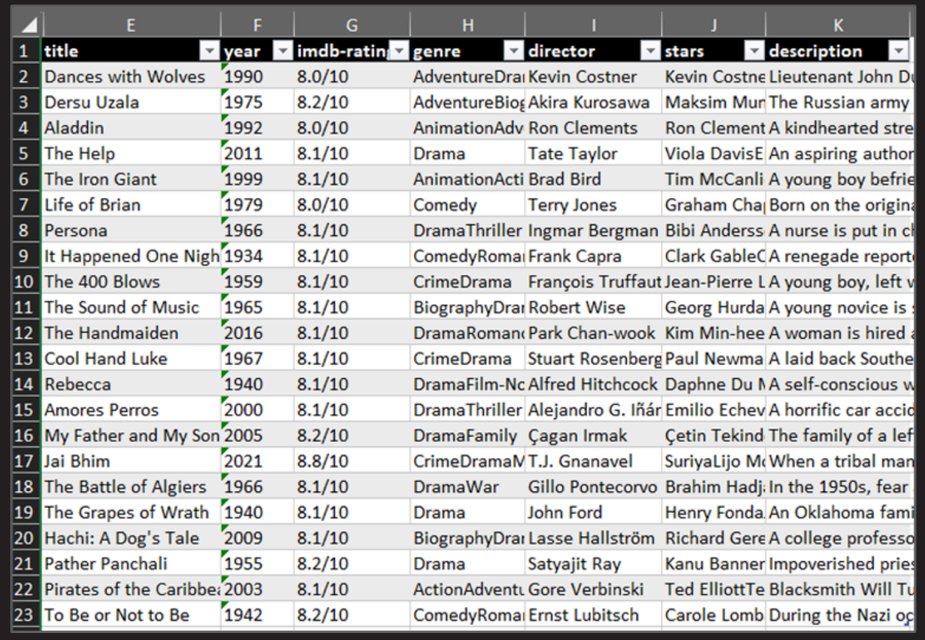


Hands-on (2)

Get the IMDb Top 250 Movies - with more details

https://www.imdb.com/chart/top/?ref_=nv_mv_250





	E	F	G	H	I	J	K
1	title	year	imdb-ratin	genre	director	stars	description
2	Dances with Wolves	1990	8.0/10	AdventureDra	Kevin Costner	Kevin Costne	Lieutenant John D
3	Dersu Uzala	1975	8.2/10	AdventureBio	Akira Kurosawa	Maksim Mur	The Russian army
4	Aladdin	1992	8.0/10	AnimationAdv	Ron Clements	Ron Clement	A kindhearted stre
5	The Help	2011	8.1/10	Drama	Tate Taylor	Viola DavisE	An aspiring author
6	The Iron Giant	1999	8.1/10	AnimationActi	Brad Bird	Tim McCanli	A young boy befrie
7	Life of Brian	1979	8.0/10	Comedy	Terry Jones	Graham Cha	Born on the origina
8	Persona	1966	8.1/10	DramaThriller	Ingmar Bergman	Bibi Anderss	A nurse is put in ch
9	It Happened One Nigh	1934	8.1/10	ComedyRoma	Frank Capra	Clark GableC	A renegade report
10	The 400 Blows	1959	8.1/10	CrimeDrama	François Truffaut	Jean-Pierre L	A young boy, left v
11	The Sound of Music	1965	8.1/10	BiographyDra	Robert Wise	Georg Hurda	A young novice is s
12	The Handmaiden	2016	8.1/10	DramaRoman	Park Chan-wook	Kim Min-hee	A woman is hired a
13	Cool Hand Luke	1967	8.1/10	CrimeDrama	Stuart Rosenberg	Paul Newma	A laid back Southe
14	Rebecca	1940	8.1/10	DramaFilm-No	Alfred Hitchcock	Daphne Du M	A self-conscious w
15	Amores Perros	2000	8.1/10	DramaThriller	Alejandro G. Iñár	Emilio Echev	A horrific car accid
16	My Father and My Son	2005	8.2/10	DramaFamily	Çagan Irmak	Çetin Tekind	The family of a lef
17	Jai Bhim	2021	8.8/10	CrimeDramaM	T.J. Gnanavel	SuriyaLijo M	When a tribal man
18	The Battle of Algiers	1966	8.1/10	DramaWar	Gillo Pontecorvo	Brahim Hadj	In the 1950s, fear
19	The Grapes of Wrath	1940	8.1/10	Drama	John Ford	Henry Fonda	An Oklahoma fami
20	Hachi: A Dog's Tale	2009	8.1/10	BiographyDra	Lasse Hallström	Richard Gere	A college professo
21	Pather Panchali	1955	8.2/10	Drama	Satyajit Ray	Kanu Banner	Impoverished pries
22	Pirates of the Caribbe	2003	8.1/10	ActionAdventu	Gore Verbinski	Ted ElliottTe	Blacksmith Will Tu
23	To Be or Not to Be	1942	8.2/10	ComedyRoma	Ernst Lubitsch	Carole Lomb	During the Nazi oc

What we want:

List of movies

- Rank
- Title
- Rating
- **Genre**
- **Director**
- **Stars**
- **Description**

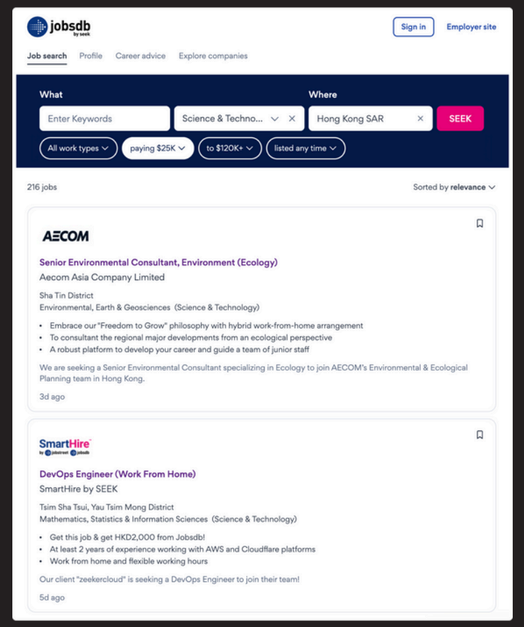
Check before scrape: https://www.imdb.com/robots.txt

Conditions of Use: https://www.imdb.com/conditions?ref_=ft_cou

Hands-on (3)

Get the list of jobs from JobsDB in an interested industry

https://hk.jobsdb.com/jobs/in-Hong-Kong-SAR



	A	B	C	D	E
1	job-link	company	location	subject	desc
2	Field Application Specialist	Medikonia Limited	Mong Kok, Yau Tsim Mong District	Biological & Biomedical Sciences (Science & Technology)	What You'll DoWe are looking for enthusiastic people to join our scientific support team. We are passionate about bringing cutting-edge and affordable solutions to the scientists, helping Hong Kong move forward in the life sciences sectors.Your day-to-day life will be focusing on providing guidance on biological education, applications, troubleshooting and product selection to respond to our clients' scientific and technical needs. As a Field
3	Data Scientist - Recommendation Engine & AI Innovation	TVB New Media Group Limited	Tseung Kwan O Industrial Estate, Sai Kung District	Mathematics, Statistics & Information Sciences (Science & Technology)	We are seeking a motivated and curious entry-level Data Scientist to join our dynamic data team. We prioritize growth mindset over years of experience and offer ample opportunities for learning, hands-on projects, and mentorship in cutting-edge machine learning and AI technologies.ResponsibilitiesCollaborate on designing, implementing, and improving Recommendation Engine systems in productionExplore and integrate Generative AI (GenAI) and Large
4	Chemist / Assistant Chemist		Sheung Wan, Central and Western District	Laboratory & Technical Services (Science & Technology)	The position will be responsible for:Conduct laboratory testing and inspection according to prescribed procedures/standards.Prepare test sample/ standards/ working solution and operate analytical instruments accurately. Operate various instruments and hand-on experience on equipment maintenance & calibrationAssist to update quality documents and technical documents according to HOKLAS/CNAS & ISO/IEC 17025 requirements,
5	Senior Engineer	GRST Holdings Limited	Sha Tin District	Chemistry & Physics (Science & Technology)	Ref. No.: GRST-(R&D Centre)-2025004GRST Holdings Limited, a company within the GRST Group, is an energy technology company with its headquarters located at Hong Kong Science Park. Through our state-of-the-art research facilities, we develop green and sustainable technology for the manufacturing and recycling of batteries and other energy storage systems. Alongside its research and development capabilities, the company also possesses a
6	Data Scientist/AI Engineer	ST Partnership Limited	Central, Central and Western District	Mathematics, Statistics & Information Sciences (Science & Technology)	Join a leading bank's strong data team as an AI Engineer, collaborating with cross-functional teams to build and integrate machine learning models and AI systems with a focus on practical application. Key Responsibilities:Design and implement AI models to address challenges and improve system capabilities.Collaborate with data scientists, stakeholders, and product managers to develop and deploy AI solutions, ensuring clear communication of technical
7	Senior Project Engineer - Testing and Certification	TUV SUD Hong Kong Limited	Yuen Long, Yuen Long District	Laboratory & Technical Services (Science & Technology)	Are you ready to embark on an exciting professional journey where innovation, collaboration, and excellence are at the heart of everything we do? Look no further! We are seeking passionate individuals like you to join our dynamic team. This position is responsible for daily chemical project handling, report review and supporting sales for quoting price and technical inquiry. **What You'll Do** Manage project with reasonable manner and reports to project
8	Senior Engineer - Materials Scientist	SAE Magnetics (Hong Kong) Ltd	Science Park, Tai Po District	Materials Sciences (Science & Technology)	Successful candidate will join the Materials Development Department to develop the latest magnetic recording heads in the R&D and prototype stages within a big cross-functional team. ResponsibilityWork closely with material science lab on material characterization, suggest possible improvement to achieve in-depth knowledge of material properties;Work with the teams of material characterization, reliability, process and wafer design to understand the
9	Biomedical Officer I/II/III (Ref: JDB/20240126/EST_BOIII)	Union Medical Centre Ltd	Sha Tin District	Laboratory & Technical Services (Science & Technology)	Job DutiesTo oversee day to day management of all medical equipment.To perform emergency repair & quality assurance check for equipment To monitor the maintenance scheduleTo implement T&C works on new installation and equipmentTo perform any ad hoc responsibilities as assigned by supervisors. Job RequirementsHigher Diploma or above with major in Electronics/Electrical/Biomedical EngineeringAt least 1 year's relevant working experience and
10	Biomedical Engineer (Ref: MRC2025-007)	Multi-Scale Medical Robotics Center Limited	Sha Tin District	Biological & Biomedical Sciences (Science & Technology)	Multi-Scale Medical Robotics CenterMulti-Scale Medical Robotics Center Limited (MRC) was established by the Faculty of Medicine and the Faculty of Engineering of the Chinese University of Hong Kong in collaboration with ETH Zürich, Imperial College London, Johns Hopkins University and the University of Hong Kong. MRC focusing on technological innovation with a strong emphasis on clinical translation and direct patient benefits, serves as a

What we want:

List of jobs

- Position
- Company
- Location
- Industry
- Job description

Check before scrape: https://hk.jobsdb.com/robots.txt

Conditions of Use: https://hk.jobsdb.com/terms

Learn more

- **Common Scenarios:**
  - Scrape from **individual subpages** (covered in class)
  - Scrape **multiple items** from **one single page** (how-to video)
  - **Pages with paginations**
    - “Next” button (how-to video)
    - Page number buttons (how-to video)
- **Documentation:**
  - https://webscraper.io/documentation/scraping-a-site

Take-away message

- **“When to use what”**
  - Simple static webpage -> **Power Query**
  - Static & a few webpages (w/ pagination) -> **Web Scraper plug-in**
  - Dynamic webpage, large scale projects -> **Programming, API**
- **Check before you scrape:**
  - Is the data available somewhere?
  - Technical limitations
  - Legal and ethical restrictions