

Visit <https://bit.ly/ds-dojo-2024> for material

Unstructured data

Scraping and data cleaning

A small change!

Today we will do **more pandas** in the afternoon, so you can finish homework from yesterday. Also a little AI.

We will do more AI tomorrow!

homework

is for the rest of your life!

You do not need to finish it
today, or tomorrow, or this
weekend!

Structured

```
Pretty-print ✓
    "url": "https://pokeapi.co/api/v2/version/30/"
    }
    }
    ]
},
"id": 143,
"is_default": true,
"location_area_encounters": "https://pokeapi.co/api/v2/pokemon/143/encounters",
"moves": [
{
"move": {
"name": "mega-punch",
"url": "https://pokeapi.co/api/v2/move/5/"
},
"version_group_details": [
{
"level_learned_at": 0,
"move_learn_method": {
"name": "machine",
"url": "https://pokeapi.co/api/v2/move-learn-method/4/"
},
"version_group": {
"name": "red-blue",
"url": "https://pokeapi.co/api/v2/version-group/1/"
}
},
{
"level_learned_at": 0,
"move_learn_method": {
"name": "egg",
"url": "https://pokeapi.co/api/v2/move-learn-method/13/"
}
}
]
```

Unstructured

Website

APIs

How to find data

- There is no secret answer!
- Search and search and search
- Talk to other people
- Visit websites

Visit <https://bit.ly/ds-dojo-2024> for material

Scraping!

stealing (??) data from
the internet

HTML

the language that all web
sites are built with

```
<h1>This is a headline</h1>
<h2>This is a smaller headline</h2>
<h3>And an even smaller headline</h3>
```

This is a headline

This is a smaller headline

And an even smaller headline

```
<h1>This is a headline</h1>
<h2>This is a smaller headline</h2>
<h3>And an even smaller headline</h3>
<p>This is a paragraph</p>

<p>This is a paragraph with
<a href="google.com">a link</a></p>
<div>This is ANYTHING</div>
<div>Anything! Anything in the world</div>
```

This is a headline

This is a smaller headline

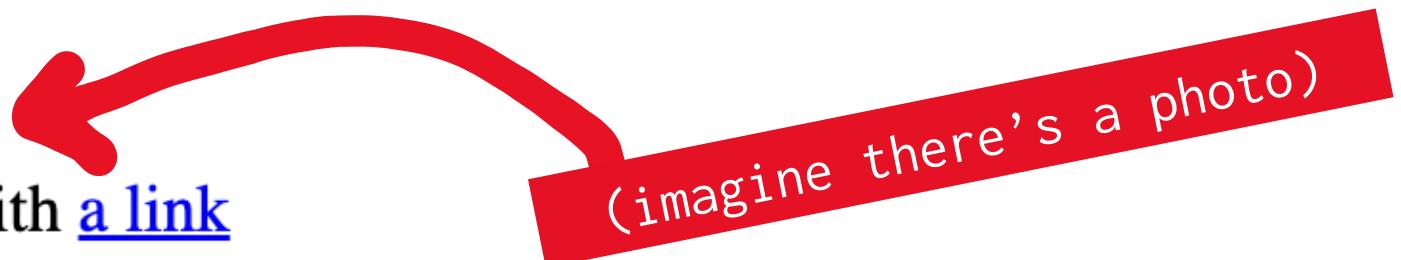
And an even smaller headline

This is a paragraph

This is a paragraph with [a link](#)

This is ANYTHING

Anything! Anything in the world



(imagine there's a photo)

<h1>An incredible story</h1>

<p>By J. Soma</p>

<p>This is the start of the story.</p>

<p>It is an amazing story!</p>

<p>”It’s incredible,” said the source.</p>

<p>An editor agreed: “it’s true!”</p>

<p>Additional reporting by Mulberry the fat
mean cat</p>

<p>Edit: An earlier version said “Jonathan”
Soma</p>

An incredible story

By J. Soma

This is the start of the story.

It is an amazing story!

"It's incredible," said the source.

An editor agreed: "it's true!"

Additional reporting by Mulberry the fat mean cat

Edit: An earlier version said "Jonathan" Soma

<h1>An incredible story</h1>

<p>By J. Soma</p>

<p>This is the start of the story.</p>

<p>It is an amazing story!</p>

<p>”It’s incredible,” said the source.</p>

<p>An editor agreed: “it’s true!”</p>

<p>Additional reporting by Mulberry the fat
mean cat</p>

<p>Edit: An earlier version said “Jonathan”
Soma</p>

```
<h1>An incredible story</h1>
<p id="byline">By J. Soma</p>
<p>This is the start of the story.</p>
<p>It is an amazing story!</p>
<p>"It's incredible," said the source.</p>
<p>An editor agreed: "it's true!"</p>
<p class="additional">Additional reporting by
Mulberry the fat mean cat</p>
<p class="correction">Edit: An earlier version
said "Jonathan" Soma</p>
```

An incredible story

By J. Soma

This is the start of the story.

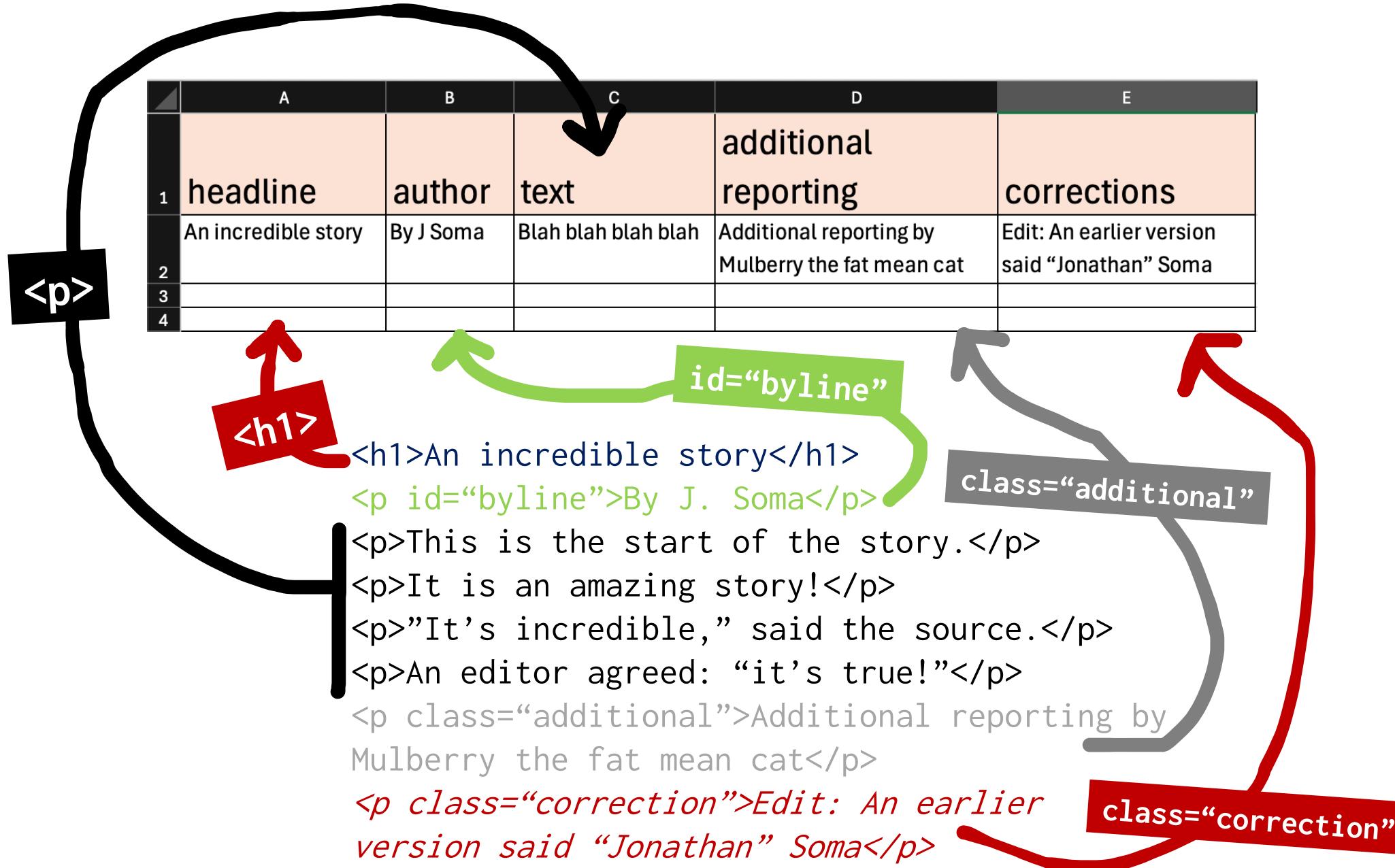
It is an amazing story!

”It’s incredible,” said the source.

An editor agreed: “it’s true!”

Additional reporting by
Mulberry the fat mean cat

*Edit: An earlier version said “Jonathan”
Soma*



Let's see it on
the internet!

BBC Home - Breaking News, +

bbc.com

Home News Sport Business Innovation Culture Travel Earth Video Live

Register Sign In

What will be a row in our spreadsheet?

 **LIVE VP pick Walz to speak as Democrats deploy star guests**

The Minnesota governor will run alongside presidential nominee Kamala Harris for November's US election.

 **Three things the Democrats have avoided so far at the DNC**

What they have tried to avoid says as much about their weaknesses as what they choose to highlight, writes Anthony Zurcher.

2 hrs ago | US & Canada

- Day three of Democratic Convention features party's rising stars
- Obamas mock Trump over 'black jobs' and crowd sizes

 **Divers find five bodies in wreck of Sicily yacht**

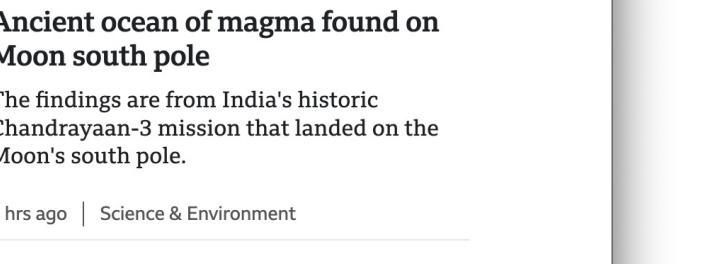
 **Gaza nurse says whole family, including quadruplets, killed in air strike**

5 hrs ago | Middle East & Africa

 **Andrew Tate held overnight after police raid homes in Romania**

The internet personality and his brother have been remanded in custody as part of a probe into new allegations.

44 mins ago | Europe

 **Ancient ocean of magma found on Moon south pole**

The findings are from India's historic Chandrayaan-3 mission that landed on the Moon's south pole.

6 hrs ago | Science & Environment

► **German Navy blasts out Darth Vader theme on Thames**

BBC Home - Breaking News, +

bbc.com

Home News Sport Business Innovation Culture Travel Earth Video Live

Register Sign In



● LIVE VP pick Walz to speak as Democrats deploy star guests

The Minnesota governor will run alongside presidential nominee Kamala Harris for November's US election.



Three things the Democrats have avoided so far at the DNC

What they have tried to avoid says as much about their weaknesses as what they choose to highlight, writes Anthony Zurcher.

2 hrs ago | US & Canada

- Day three of Democratic Convention features party's rising stars
- Obamas mock Trump over 'black jobs' and crowd sizes



Divers find five bodies in wreck of Saily yacht



Gaza nurse says whole family, including quadruplets, killed in air strike

Ashraf El Attar's wife and six children were killed when an air strike destroyed their home in Deir al-Balah.

5 hrs ago | Middle East



Andrew Tate held overnight after police raid homes in Romania

The internet personality and his brother have been remanded in custody as part of a probe into new allegations.

44 mins ago | Europe



Ancient ocean of magma found on Moon south pole

The findings are from India's historic Chandrayaan-3 mission that landed on the Moon's south pole.

6 hrs ago | Science & Environment

► German Navy blasts out Darth Vader theme on Thames

What will be a column of data?

Andrew Tate held overnight after police raid homes in Romania

The internet personality and his brother have been remanded in custody as part of a probe into new allegations.

44 mins ago

| Europe



[L] **LIVE** VP pick Walz to speak as Democrats deploy star guests

[R] The Minnesota governor will run alongside presidential nominee Kamala Harris for November's US election.

メルカリ - 日本最大のフリマサー × +

jp.mercari.com/search?category_id=79

楽器・機材 ×

ログイン 会員登録 ベル 出品

絞り込み クリア 楽器・機材 の検索結果 ↑↓ おすすめ順 この検索条件を保存する

除外キーワード カテゴリー オカリナ kemper ホルン

ホビー・楽器・アート

楽器・機材

すべて

ブランド

サイズ

価格

価格なし出品

あんしん鑑定

割引オプション

商品の状態

What will be a row in our spreadsheet?

¥28,000 korg kross61 電池駆動軽量化シンセ

¥3,800 iQ7 ステレオマイク Lightning接続 (修理品)

¥990 ギターストラップ 未使用

¥700 音楽之友社 うたとピアノの絵本 1みぎで

¥1,500 バイオリン用 頸当て 黒檀 訳あり

¥40,000 YUCKY様専用 Pioneer DDJ-FLX4

¥880 ACアダプター JH35P1200150D

¥4,700 GoogleChromecast グーグルクロームキャスト ...

¥55,000 日本(フジゲン)製 ibanez SR1000 フレット ...

¥8,500 SE ELECTRONICS DM1 DYNAMITE

<https://jp.mercari.com/item/m97630098490>

メルカリ - 日本最大のフリマサー × +

jp.mercari.com/search?category_id=79

楽器・機材 ×

ログイン 会員登録 ベル 出品

絞り込み クリア

除外キーワード

カテゴリ

ホビー・楽器・アート

楽器・機材

すべて

ブランド

サイズ

価格

価格なし出品

あんしん鑑定

割引オプション

商品の状態

検索結果

販売中のみ表示

↑↓ おすすめ順 この検索条件を保存する

エフェクター ジャンク ベース bare fender chase パワーサプライ オカリナ kemper ホルン

¥28,000 Korg kross61 電池駆動軽量化シンセ

¥3,800 iQ7 ステレオマイク Lightning接続 (修理品)

¥990 ギターストラップ 未使用

¥700 音楽之友社 うたとピアノの絵本 1みぎで

¥1,500 バイオリン用 頸当て 檜 訳あり

¥40,000 YUCKY様専用 Pioneer DDJ-FLX4

¥880 ACアダプター JH35P1200150D

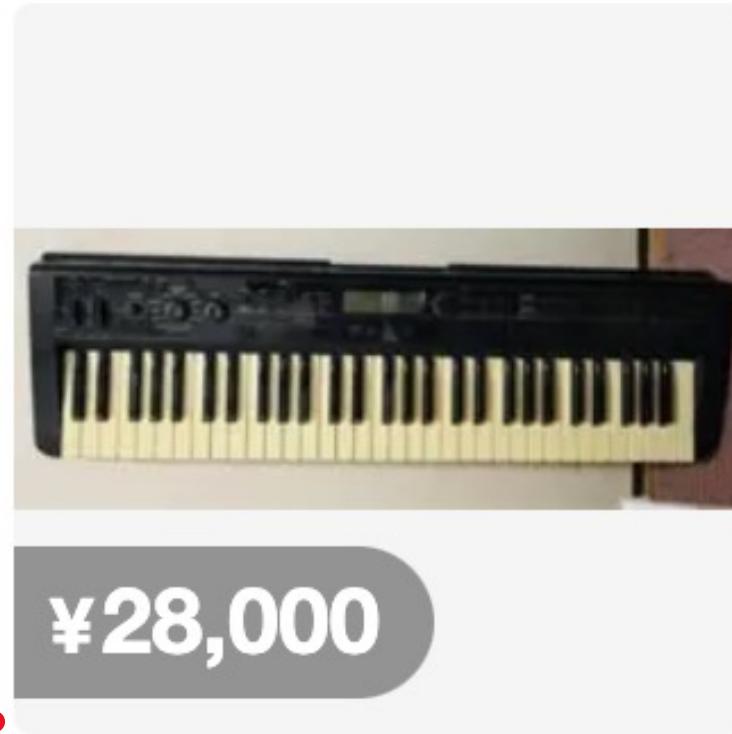
¥4,700 GoogleChromecast グーグルクロームキャスト ...

¥55,000 日本(フジゲン)製 ibanez SR1000 フレット ...

¥8,500 SE ELECTRONICS DM1 DYNAMITE

<https://jp.mercari.com/item/m97630098490>

What will be a column of data?



price

name

korg kross61 電池駆動軽
量化シンセ

image

Computers don't see like us
(usually), we need to see

HTML code

Visit a page on Mercari

Do this now

【2024年最新】楽器・機材の人

ログイン 会員登録 ベル 出品

楽器・機材 の検索結果

クリア

除外キーワード

カテゴリー

ホビー・楽器・アート

楽器・機材

ブランド

サイズ

価格

価格なし出品

あんしん鑑定

割引オプション

¥1,250 新品 D'Addario ダダリオ アコースティックギター弦

¥5,500 新品 D'Addario ダダリオ アコースティックギター弦

¥15,000 Pearl スネアドラム MUS1455M

¥2,400 新品 D'Addario ダダリオ アコースティックギター弦

¥399 バタフライフィンガーピック 4個セット ゴ...
バタフライフィンガーピック 4個セット ゴ...

¥579 オイル漬け牛骨製ナット

¥580 YOASOBI 群青 ピアノ楽譜 ソロ

¥1,800 新品 D'Addario ダダリオ アコースティックギター弦

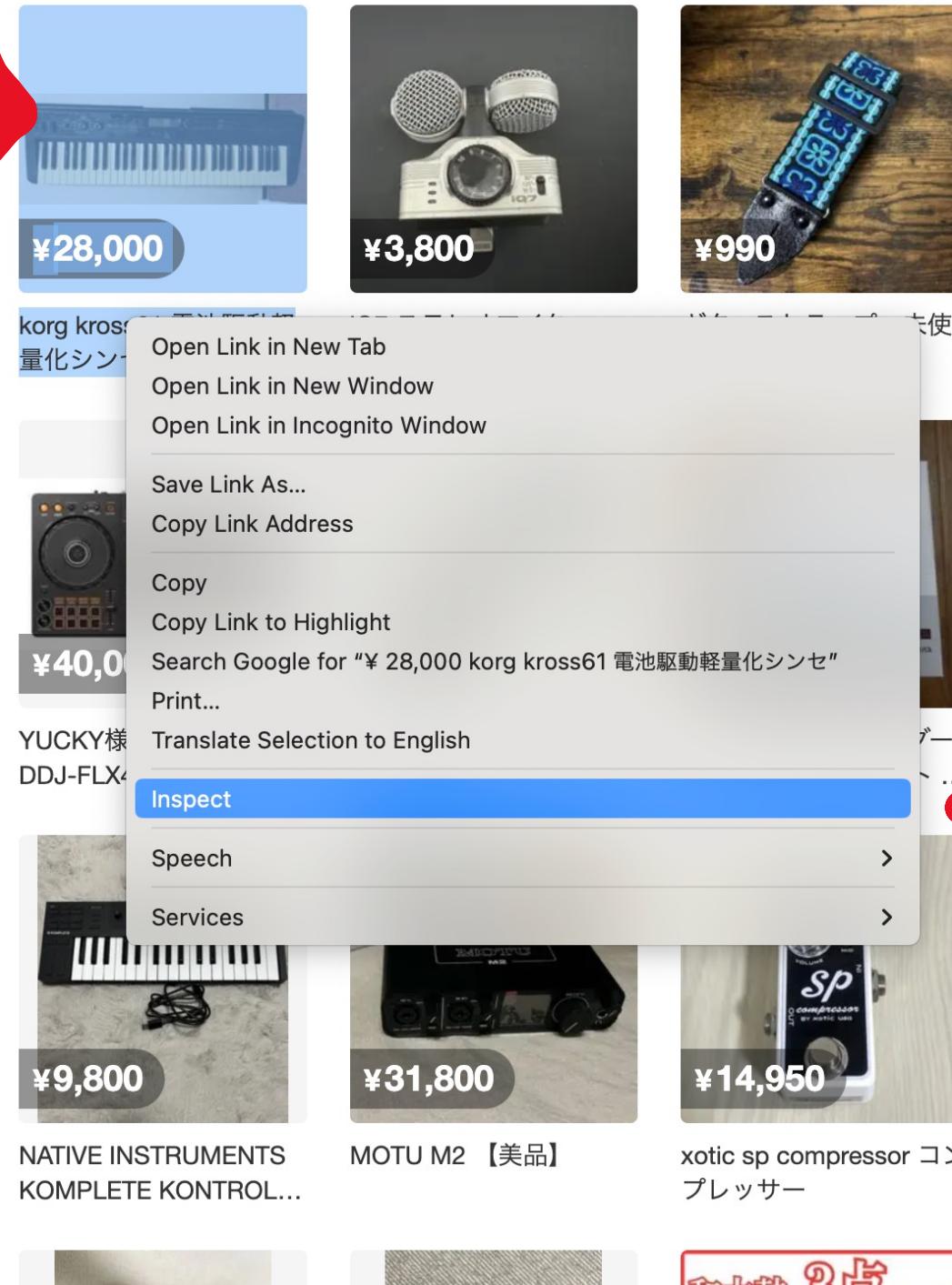
¥680 新品 D'Addario ダダリオ アコースティックギター弦

¥31,000 ♪森の工房♪フルートの店 おかげさまで100本超
YAMAHA 211SII 銀メッキ
美品 分解磨き 調整済み
全品 部活応援!!銀メッキ!!美品!!調整済!!ヤマハフル...

↑↓ おすすめ順 この検索条件を保存する

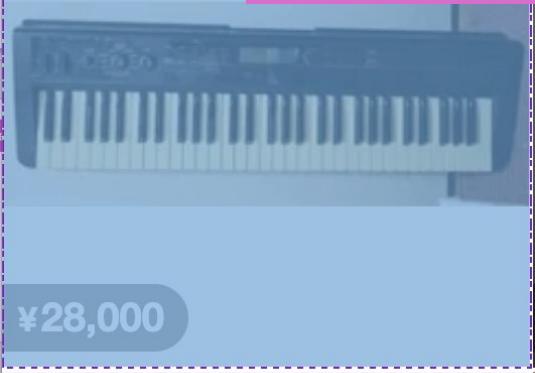
エフェクター ジャンク ベース bare fender chase パワーサプライ オカリナ kemper ホルン

Right click (or
command+click)
and Inspect



検証

Find the HTML
for our “row”

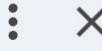


1

METHOD
ONE

この検索条件を保存する

会員登録 ログイン



Mouse/click
around

```
<div id="item-grid" data-testid="search-item-grid">
  <ul class="sc-50e0525c-0 sc-bcd1c877-0 JKB0Z ipWzwL">
    <li data-testid="item-cell" class="sc-bcd1c877-2 cvAXgx">
      <div>
        <a href="/item/m48062282876" data-location="search_result:best_match:body:item_list:item_thumbnail" class="sc-bcd1c877-1 lpjZwE" data-testid="thumbnail-link"> flex
          <div class="merItemThumbnail fluid_a6f874a2" role="img" aria-label="kg kross61 電池駆動軽量化シンセの画像" data-bbox="210 479 415 666" data-testid="itemThumbnail">
            <div class="before_a6f874a2" aria-hidden="true"></div>
          <div class="imageContainer_f8ddf3a2 picture" data-bbox="416 161 621 479" data-testid="imageContainer">
            <img alt="A blue patterned leather belt." data-bbox="416 161 621 479" data-testid="itemThumbnailImage"/>
          </div>
        </div>
      </div>
    </li>
  </ul>
</div>
```

Find the HTML
for our “row”



¥28,000



¥3,800



¥990



¥700



METHOD
TWO

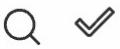


¥40,000



この検索条件を保存する

会員登録 ログイン



Click
this

Search + click
here

```
<div id="item-grid" data-testid="search-item-grid">
  <ul class="sc-50e0525c-0 sc-bcd1c877-0 JKBOZ ipWzwL"> grid
    <li data-testid="item-cell" class="sc-bc...
      <div data-testid="item-list-item_thumbnail" class="sc-bcd1c877-1 lpjZwE" data-testid="item_thumbnail-link"> flex
        <div class="merItemThumbnail fluid_a6f874a2" role="img" aria-label="korg kross61 電池駆動式シンセの画像
          28,000円"
        </div>
      </div>
    </li>
  </ul>
</div>
```

◀ 7-0.JKBOZ.ipWzwL li.sc-bcd1c877-2.cvAXgx ▶

Styles Computed Layout Event Listeners >

Filter

:hov .cls +, □

```
element.style {  
}
```

amazon.com
search results

Sales at Lidl

List of amphibians
from Germany's Red
List Center

Headlines from
Nikkei Asia

items on
Mercari

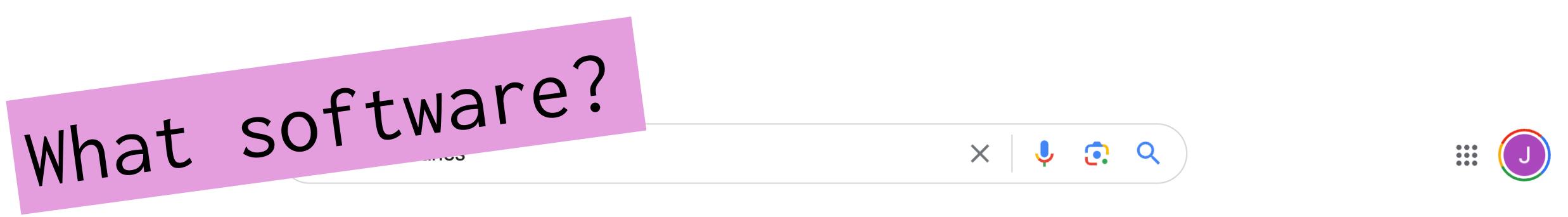
Now try it.

Anywhere!!!!

Wikipedia's
Fictional Big
Cats

School Board Minutes
from Grand Island Public
Schools in Nebraska

Scraping is just connecting to websites
and extracting data.
But we need some software
to do that for us!
So you can add columns in
your spreadsheet



All Images Videos News Books Web Finance

Tools



Scraping libraries

From sources across the web



Selenium



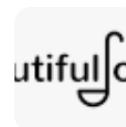
Lxml



Playwright



ZenRows S.L.



Beautiful Soup



Requests



Puppeteer



Cheerio



Scrapy



MechanicalSoup



Urllib3



[\[–\]](#) **Swingbiter**  **70 points** 2 years ago

Learn the basic html elements that build up a website.

[–] **coventous**  **22 points** 2 years ago

I recommend checking

W 1 point 2 years ago

 Nerdvana

Look into learning and understanding Docker, so Google for that. On a bonus tip, there's (official?) Chrome/Chrome/etc) which made me very happy (I like Docker!)

For more information about the study, please contact Dr. John Smith at (555) 123-4567 or via email at john.smith@researchinstitute.org.

item - data - response

[-] luizv4z  **12 points** 2 years ago

From my own research

My own research, run away from Selenium. The right direction is CDP (Chrome Developers Protocol), I could scrap Facebook without getting banned.

```
2 all_links = soup.findAll(name=
```

~~Do python on them until~~

Bea [-] riisen 3 points

▶ for small projects

[more](#) [embed](#) [save](#)

P.S. [permalink](#)

permalink **source** **embed** **save** **save-RES** **report** **reply** **hide child comments**

Selenium is garbage!

...but BeautifulSoup
can't scrape all sites.

Top discounts for you

Page 1 of 3



Continental Grand Prix
5000 Rennrad Faltreifen //
28-622 (700x28C)

85

400+ viewed in past month

\$66.00

FREE Delivery Friday,
Aug 23



Wink Super X-Large Ultra
Thin Lubricated Latex
Condoms, Premium Latex
for Smooth and Natural...

76

100+ viewed in past month

\$14.99 (\$0.30/Count)

FREE Delivery Friday,
Aug 23



Assortment Rubber Ducks
in Bulk, 50-Pack Assorted
Mini Duckies Toy for
Ducking Cruise Ships, 2"

328

5K+ viewed in past month

\$24.99

Overnight by 8:00
AM



50 Pack Rubber Ducks in
Bulk, Jeep Ducks for
Ducking, Assorted Rubber
Ducks Jeep Ducking, Bab...

569

4K+ viewed in past month

\$21.89

Overnight by 8:00
AM



ArtCreativity Assorted
Rubber Ducks Jeep Ducking
(100Pack) - Rubber Ducki...

982

700+ viewed in past month

20% off **Limited time deal**

\$39.98

List: \$49.99

Today by 10:00 PM



Kicko Assorted Rubber
Ducks with Mesh Bag - 50
Ducklings, 2 Inch – Jeep
Ducks for Kids, Baby Bath...

2,543

2K+ viewed in past month

\$28.99

FREE Delivery
Saturday, Aug 24



Glitter Rubber Ducks in
Bulk - (Pack of 50) Assorted
2-inch Duck Toys for Baby
Shower Rubber Duckies,...

308

1K+ viewed in past month

\$26.79

Today by 10:00 PM



Bulk savings to consider

Page 1 of 7



Some sites need interaction

Chockeas Asso

Ducks Toy Duckies for Kids
and Toddlers, Bath Birthday
Baby Showers Classroom,...

1,155

Rubber Ducks Jeep Ducking
(50 Pack) - Rubber Duckie...
100+ viewed in past month

Set, Mini Colorful Rubber
Duckies Bath Toy for
Child,Float & Squeak Tiny...

763

Bulk, Jeep Ducks for
Ducking, Assorted Rubber
Ducks Jeep Ducking, Bab...

569

Ducks: Fun Unique Military-
Inspired Bath Toys for Jeep
Ducking or Play - 2 inches
800+ viewed in past month

569

Bulk, Jeep Ducks for
Ducking, Assorted Rubber
Ducks for Jeeps, Bath Toy...
569

Glitter Rubber Duck Toy
Assortment Duckies for
Kids, Bath Birthday Gifts...
 1,094





TEXAS DEPARTMENT OF LICENSING & REGULATION

TDLR License Data Search (Active Licenses only)

[Search Help](#) | [Download License files](#) | [Download Other](#) | [Questions/Comments](#)

Inquire by License Type	Inquire by License #
Choose One (Optional)	<input type="text"/> (Numeric only)
Inquire by Expiration Date	
<input type="text"/> (mmddyyyy)	
Inquire by Name (Last, First) or by Business Name	
<input type="text"/>	
Inquire by Location (City)	
Choose One (Optional)	Type the first letter to scroll down.
Inquire by County	
Choose One (Optional)	Type the first letter to scroll down.
Inquire by Zip Code	
<input type="text"/>	
<input type="button" value="Search"/> <input type="button" value="Reset"/>	

If license not found, please contact Customer Service at 800-803-9202

Data last updated: 8/21/2024 06:01

[Bookmark This Page](#)

Some sites need interaction

Fast and reliable end-to-end 🎉

playwright.dev/python/

Playwright for Python Docs API Python ▾ Community

GitHub Discord Slack Search

Playwright enables reliable end-to-end testing for modern web apps.

GET STARTED ⚡ Star 64k+

I love Playwright!



Playwright is perfect!

But! It's new, so

ChatGPT isn't very good
at it. *But!* We try.

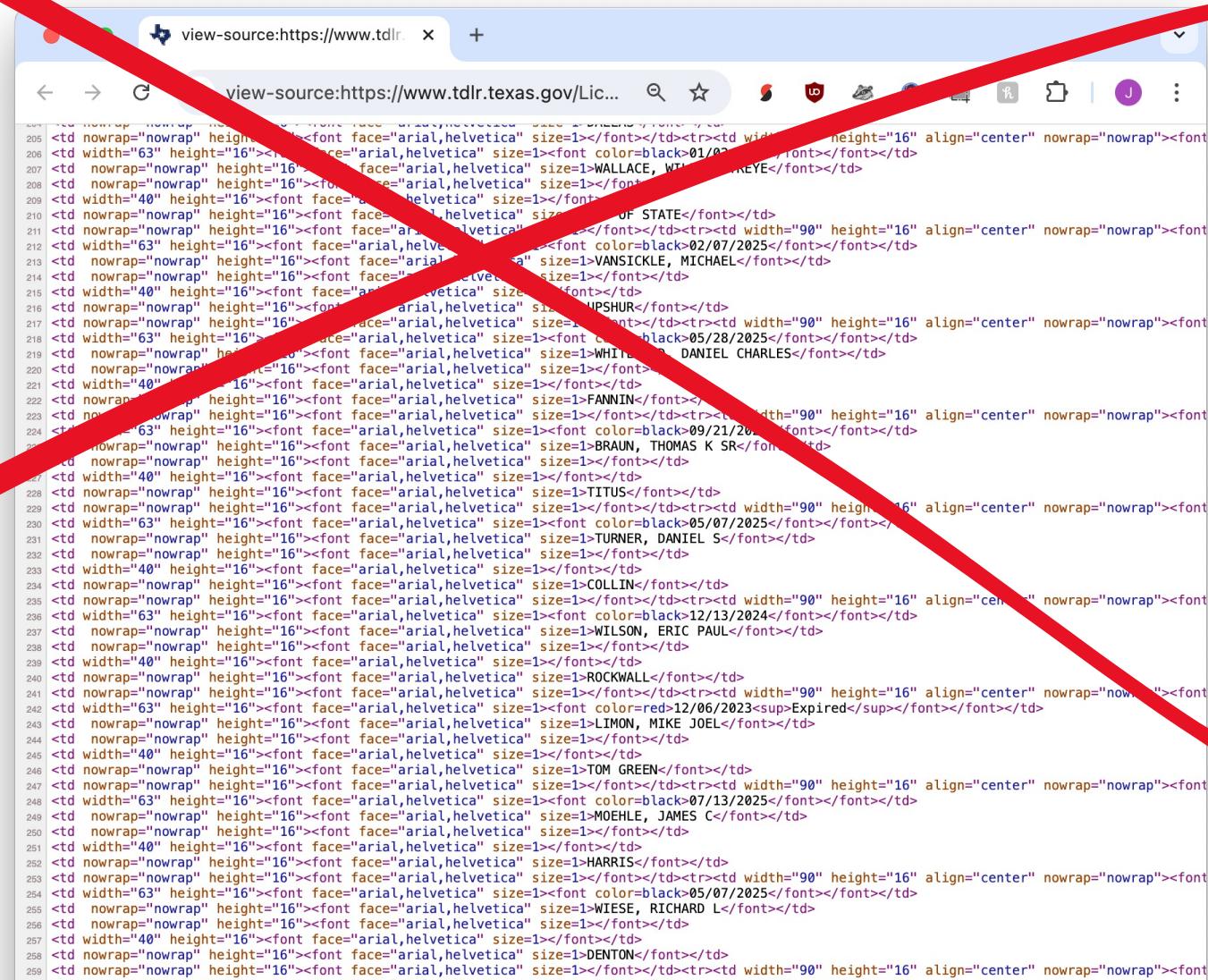
Playwright + ChatGPT +
pasting samples of HTML

=

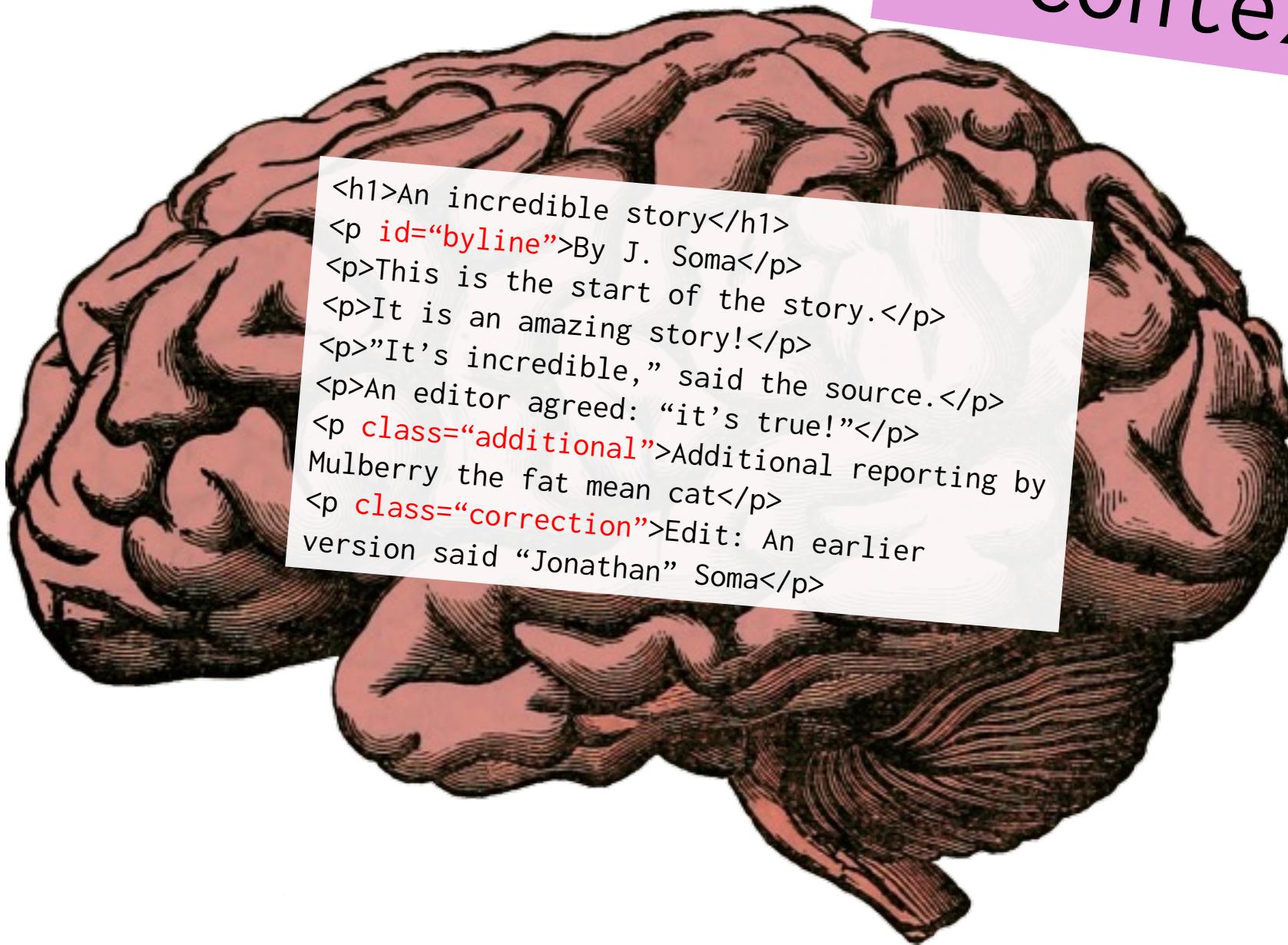
infinite scrapers

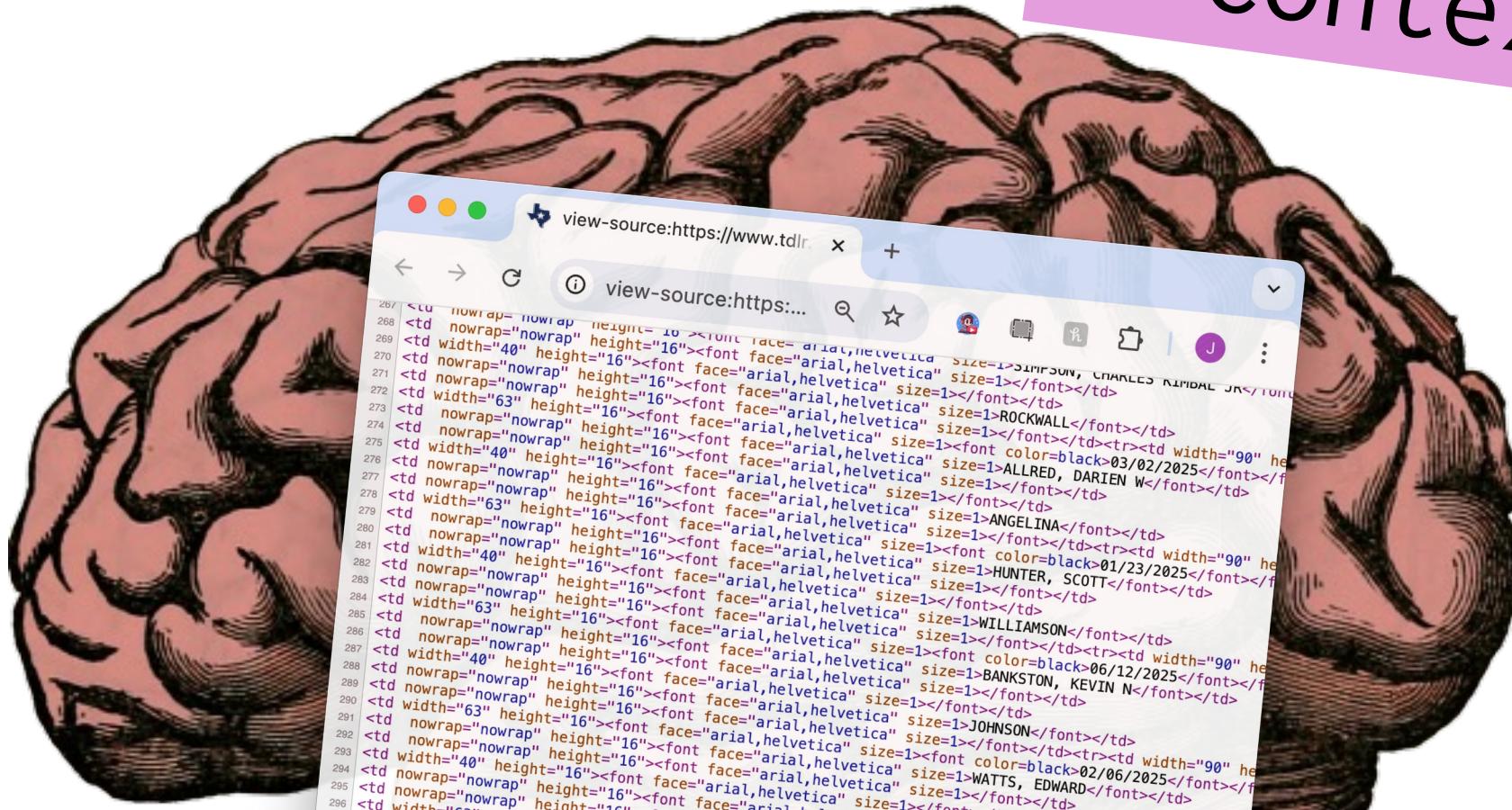


~~“Write a scraper for these search results:”~~



“Context window”





“Context window”

NAME	ADDRESS	CITY	STATE	ZIP CODE
CHARLES KIMBELL JR	1234 JEFFERSON	JACKSON	MS	39201
ROCKWALL	1234 ROCKWALL	ROCKWALL	TX	75087
DARIEN W.	1234 DARIEN	ROCKWALL	TX	75087
ANGELINA	1234 ANGELINA	ANGELINA	TX	75631
WILLIAMSON	1234 WILLIAMSON	WILLIAMSON	TX	78754
JOHNSON	1234 JOHNSON	JOHNSON	TX	78069
BANKSTON, KEVIN N.	1234 BANKSTON	BANKSTON	TX	76421
WATTS, EDWARD	1234 WATTS	WATTS	TX	76477
BASTROP	1234 BASTROP	BASTROP	TX	78601
SKELTON, ROBERT WAYNE	1234 SKELTON	SKELTON	TX	76466
BEXAR	1234 BEXAR	BEXAR	TX	78201
HORTON, IRVIN E JR	1234 HORTON	HORTON	TX	78201
HARRIS	1234 HARRIS	HARRIS	TX	77001
ROBERTSON, MICHAEL JAY	1234 ROBERTSON	ROBERTSON	TX	77001
OUT OF STATE	1234 OUT OF STATE	OUT OF STATE	TX	77001

What we might need:

- Start URL
- Forms to fill out (*optional*)
- “Rows” of our spreadsheet
- Pagination/“next” pages (*optional*)
- A **magical prompt**

<https://bit.ly/ds-dojo-2024>

First, sit and relax
and read the notes
while watching me

IOWA PROFESSIONAL LICENSING

Search License

[License Search / Mailing List](#)

[Course Search](#)

[Submit a Complaint](#)

Note: The license search engine tool operates better if less information is input, i.e. "Last Name" and "Licensing Board"
[Previous Page](#) [Next Page ▶](#)

Viewing 1-25 of 148

[Download Results](#) | [Email Results](#)

Number	Licensee	License Type	Licensing Board	License Status	Contact LastName Group
00045	1st National Appraisal Source, Inc.	Appraisal Management Company Registration	Appraisal Management Company	Lapsed	L-Z
00028	360 Appraisal Group	Appraisal Management Company Registration	Appraisal Management Company	Expired	A-K
00101	AAA APPRAISAL MANAGEMENT COMPANY LLC	Appraisal Management Company Registration	Appraisal Management Company	Active	A-K
00132	Accelerated	Appraisal	Appraisal	Active	L-Z

Visit <https://bit.ly/ds-dojo-2024> for material

The image shows the JupyterLab interface. On the left, a sidebar titled "Welcome" contains sections for "Start" (New notebook..., New session..., Open..., Connect...) and "Recent sessions" (02 - Using AI about AI (inclass).ipynb, 01-pandas basics.ipynb, 02-scraping.ipynb). Below the sidebar is a large pink banner with the text "Open JupyterLab". On the right, a main window titled "Untitled2.ipynb" displays Python code for web scraping using Playwright. The code imports `asyncio` and `async_playwright`, starts a browser, navigates to a URL, waits for a table to appear, scrapes table rows, and creates a DataFrame. The interface includes a toolbar at the top with file operations like File, Edit, View, Run, Kernel, Tabs, Settings, Help, and a bottom toolbar with Notebook, Python 3 (ipykernel), and other icons.

```
[ ]: from playwright.async_api import async_playwright

# Start Playwright
playwright = await async_playwright().start()
browser = await playwright.chromium.launch(headless=False)
page = await browser.new_page()

# Navigate to the URL
await page.goto("https://www.tdlr.texas.gov/LicenseSearch/")

[ ]: import pandas as pd

# Wait for the results to be visible after form submission
await page.wait_for_selector("table")

# Scrape table rows
rows = await page.query_selector_all("table tr")

data = []
for row in rows:
    cells = await row.query_selector_all("td")
    if len(cells) > 0:
        license_name = await cells[0].inner_text()
        expiration_date = await cells[1].inner_text()
        license_holder = await cells[2].inner_text()

    data.append({
        "license_name": license_name,
        "expiration_date": expiration_date,
        "license_holder": license_holder
    })

df = pd.DataFrame(data)
df.head()
```

<https://bit.ly/ds-dojo-2024>

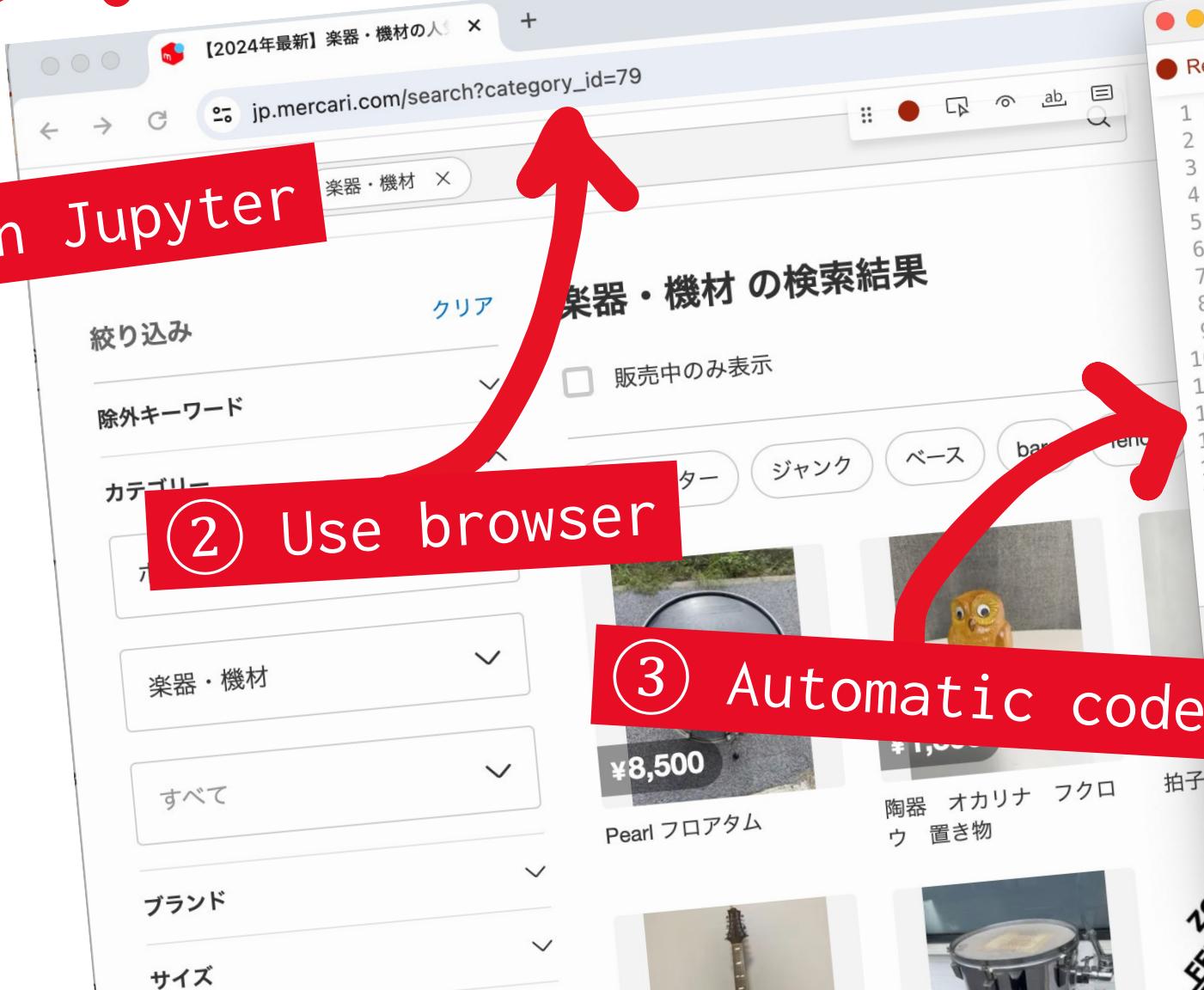
Now open a new notebook
and you try!

In [*]: !playwright codegen

① Run in Jupyter

② Use browser

③ Automatic code



Playwright Inspector

```
1 import re
2 from playwright.sync_api import Playwright, sync_playwright
3
4 def run(playwright: Playwright) -> None:
5     browser = playwright.chromium.launch(headless=False)
6     context = browser.new_context()
7     page = context.new_page()
8     page.goto("https://jp.mercari.com/search?category_id=79")
9
10    # -----
11    # context.close()
12    # browser.close()
13
14
15    with sync_playwright() as playwright:
16        run(playwright)
```

You are screen sharing

This screenshot shows the Playwright Inspector interface. It displays the generated Python code for interacting with the Mercari search results page. A red arrow points from the Mercari search results page above down to this code editor.

In [*]: !playwright codegen

① Run in Jupyter

The screenshot illustrates the workflow for generating a Playwright test script. A red arrow points from the Jupyter cell output at the top to the browser window below. Another red arrow points from the browser window to the generated code in the Jupyter cell.

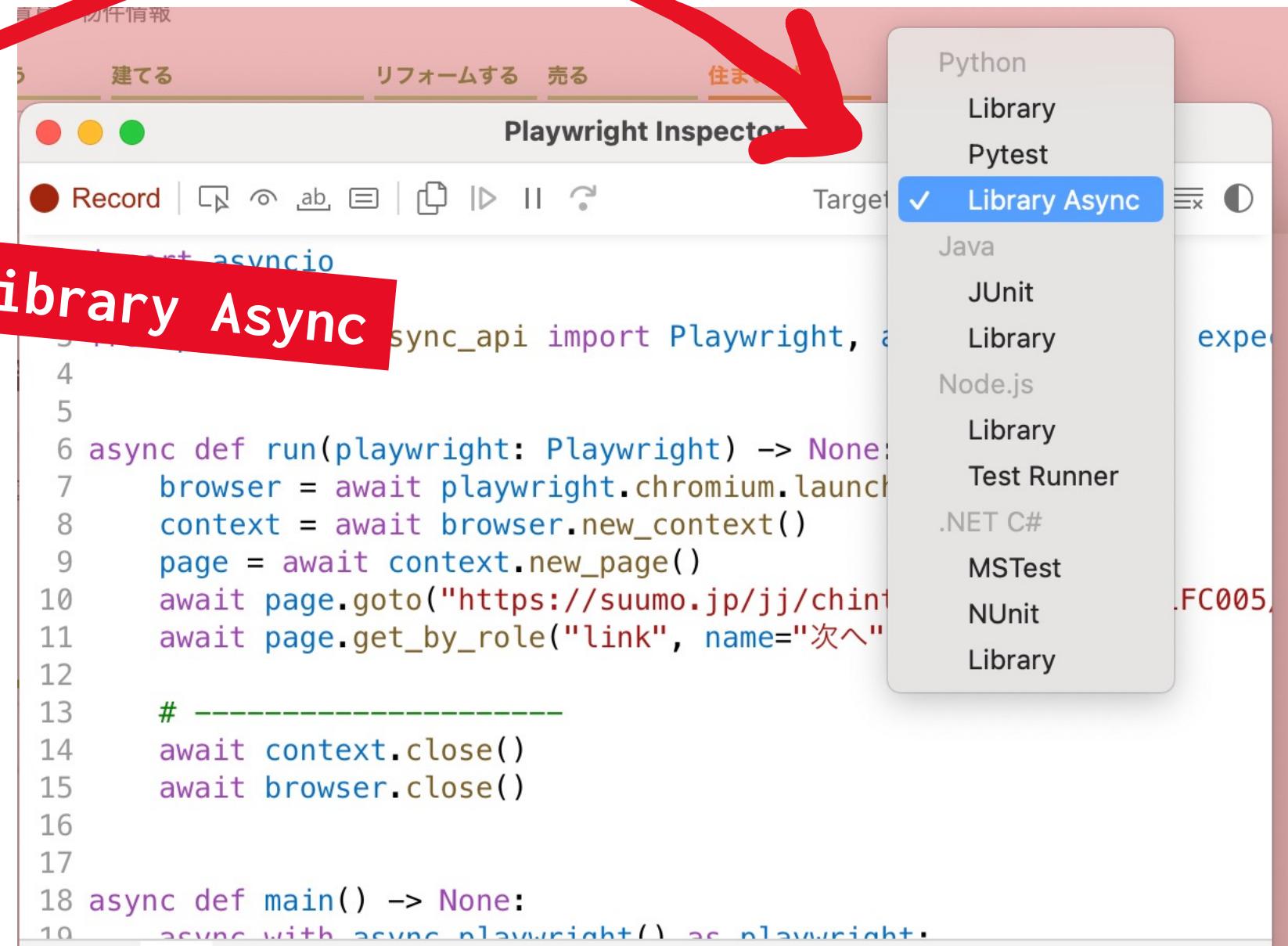
The browser window displays a search results page for rental apartments in Chiyoda-ku, Tokyo, on the SUUMO website. The generated code in the Jupyter cell is a Python script that uses the Playwright library to interact with this page, specifically clicking on the 'Next' button to navigate to the second page of results.

```
1 import asyncio
2 import re
3 from playwright.async_api import Playwright, async_playwright, expect
4
5
6 async def run(playwright: Playwright) -> None:
7     browser = await playwright.chromium.launch(headless=False)
8     context = await browser.new_context()
9     page = await context.new_page()
10    await page.goto("https://suumo.jp/jj/chintai/ichiran/FR301FC005")
11    await page.get_by_role("link", name="次へ").click()
12
13    # -----
14    await context.close()
15    await browser.close()
16
17
18 async def main() -> None:
19     async with async_playwright() as playwright:
20         playwright.chromium.set_user_agent("Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4453.89 Safari/537.36")
21         playwright.chromium.set_default_timeout(10000)
22         await run(playwright)
```

② Use browser

③ Automatic code

Change to Python > Library Async



The screenshot shows a browser window with the title "Playwright Inspector". The window contains a toolbar with various icons and a main area displaying Python code. The code is an example of using the Playwright library to interact with a web page. A red arrow points from the top-left towards the Python library dropdown menu. Another red arrow points from the Python library dropdown menu to the "Library Async" option, which is highlighted with a blue background.

```
async def run(playwright: Playwright) -> None:
    browser = await playwright.chromium.launch()
    context = await browser.new_context()
    page = await context.new_page()
    await page.goto("https://suumo.jp/jj/chintai")
    await page.get_by_role("link", name="次へ")

    # -----
    await context.close()
    await browser.close()

async def main() -> None:
    async with async_playwright() as playwright:
```

Visit <https://bit.ly/ds-dojo-2024> for material



Practice time!

Mercari
SUUMO
you choose a site

Scraping!

stealing data from the
internet for fun
(and investigations!)