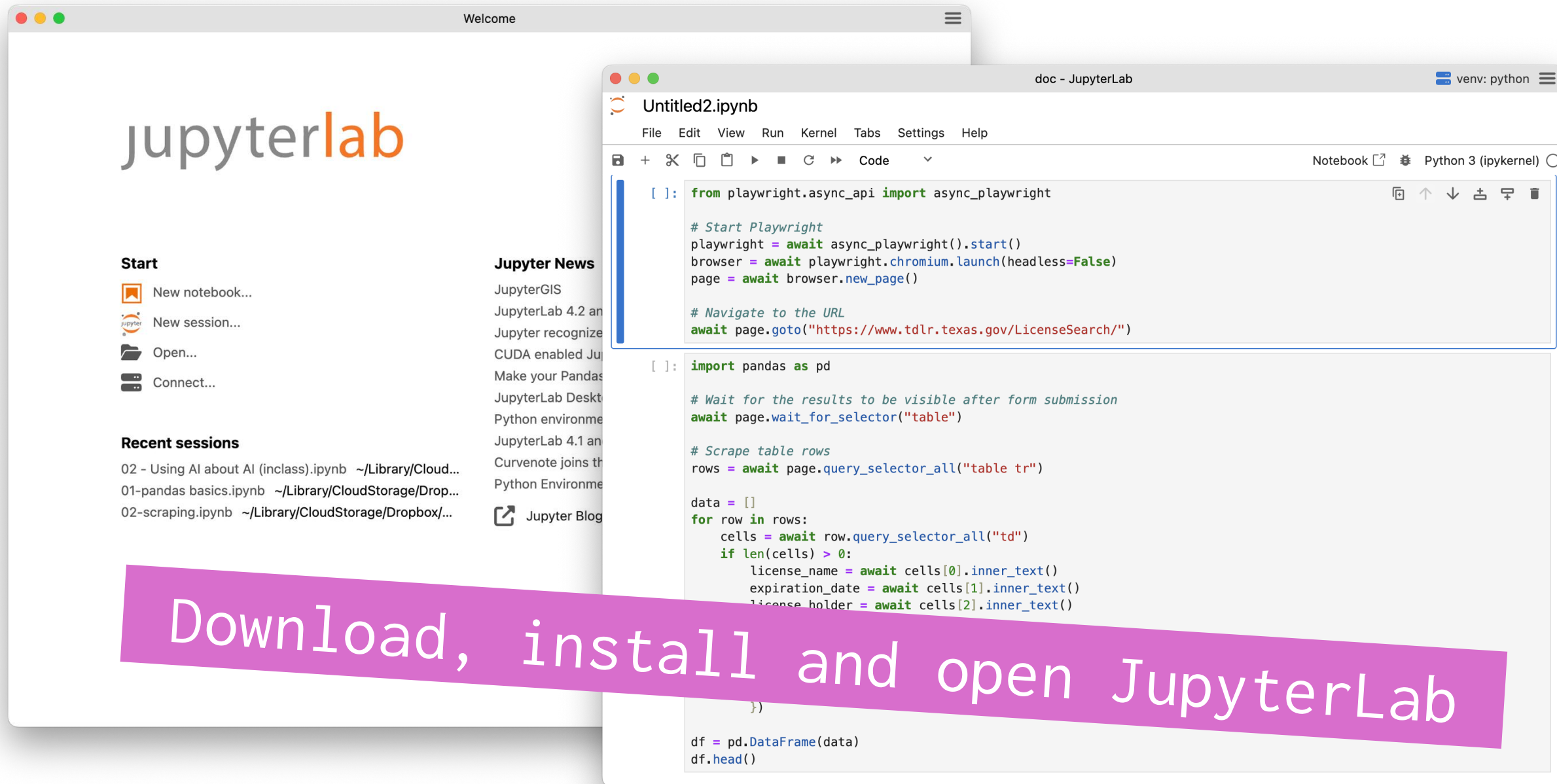


Visit <https://bit.ly/birn-data>, click **scrapping**

# Scrapping!

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

Visit <https://bit.ly/birn-data>, click scraping



The image displays the JupyterLab web interface. On the left, the 'Welcome' page shows the 'jupyterlab' logo, a 'Start' section with options like 'New notebook...', 'New session...', 'Open...', and 'Connect...', and a 'Recent sessions' list. The main area shows a notebook titled 'Untitled2.ipynb' with the following code:

```
[ ]: 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()

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

A large purple banner at the bottom of the image contains the text: Download, install and open JupyterLab

Visit <https://bit.ly/birn-data>, click scraping

# HTML

the language that all web  
sites are built with

Visit <https://bit.ly/birn-data>, click **scrapping**

<h1>This is a headline</h1>

<h2>This is a smaller headline</h2>

<h3>And an even smaller headline</h3>

Visit <https://bit.ly/birn-data>, click scraping

**This is a headline**

**This is a smaller headline**

**And an even smaller headline**

Visit <https://bit.ly/birn-data>, click **scrapping**

```
<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>
```

Visit <https://bit.ly/birn-data>, click scraping

# **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)

Visit <https://bit.ly/birn-data>, click **scrapping**

<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>



Visit <https://bit.ly/birn-data>, click scraping

# 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

Visit <https://bit.ly/birn-data>, click **scrapping**

<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>

Visit <https://bit.ly/birn-data>, click scraping

<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>

Visit <https://bit.ly/birn-data>, click scraping

<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>

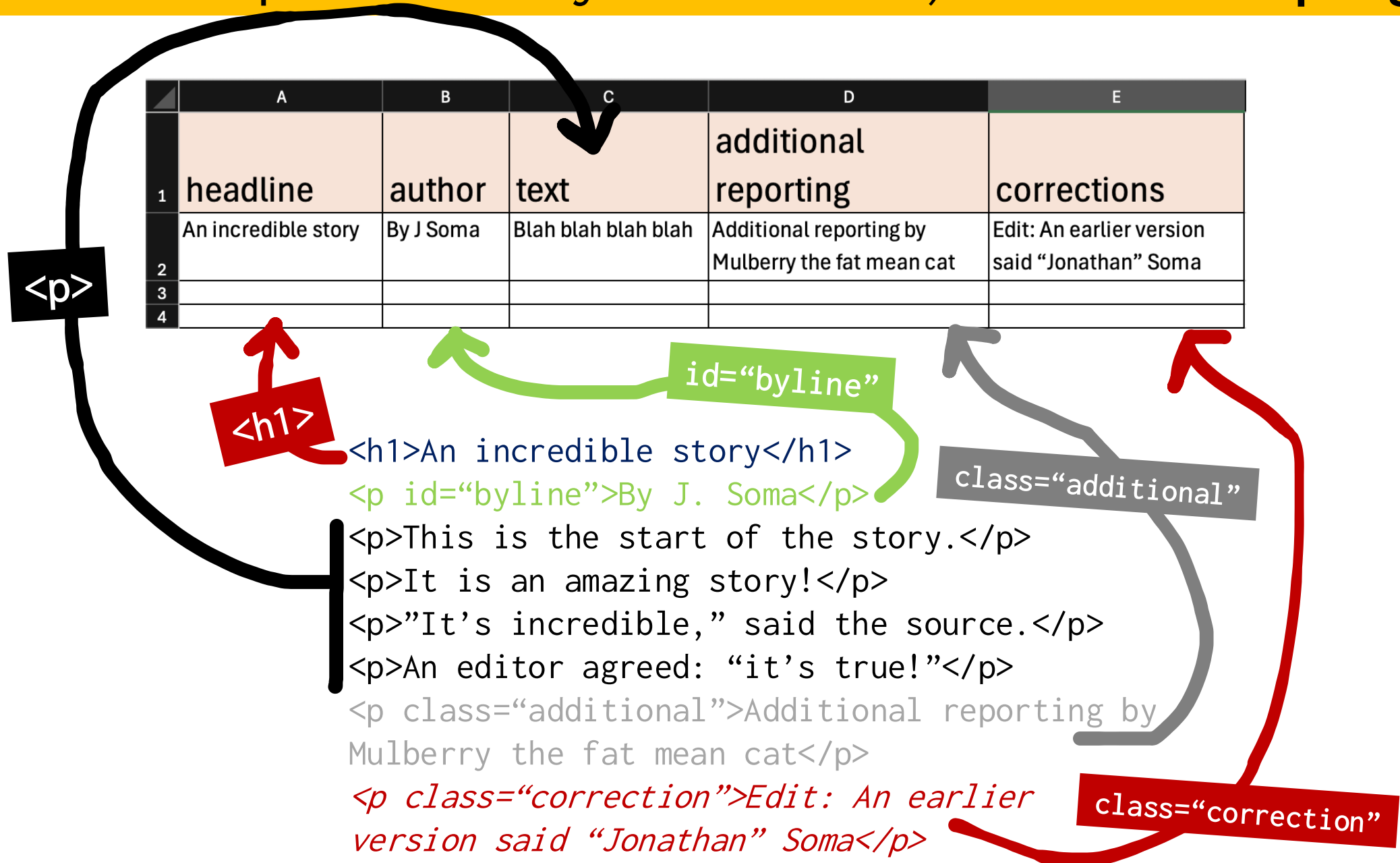
<p>"It's incredible," said the source.</p>

<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>

Visit <https://bit.ly/birn-data>, click scraping



Visit <https://bit.ly/birn-data>, click scraping

Let's see it on  
the internet!


BBC Home - Breaking News, x +

bbc.com

Register Sign In


Home News Sport Business Innovation Culture Travel Earth Video Live

**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.



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


5 hrs ago



**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




**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**



**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, x +

bbc.com


Register Sign In

Home News Sport Business Innovation Culture Travel Earth Video Live



**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



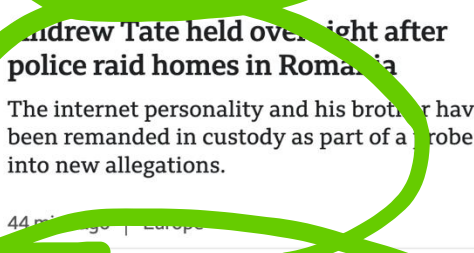
**Divers find five bodies in wreck of family yacht**



**Gaz 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



📺 **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.

2435 sets in total

Ex

Download data→



## Data search.

Schedule according to...

## Organizations



## CATALOG OF CONTAMINATED LOCATIONS

The cadastre of contaminated sites represents a set of data on endangered, polluted and degraded lands and it is an integral part of the Land Information System managed by the Environmental Protection Agency. State bodies, i.e. organizations, bodies of autonomous provinces, local self-government units...

use 0

supporting 0



## Soil condition in urban areas

The data show the degree of soil endangerment from chemical pollution in urban areas in the period 2005-2023. year. Soil testing was carried out at locations in the immediate vicinity of traffic roads, industrial zones, near landfills, water supply sources, educational institutions, recreation...

|            |    |
|------------|----|
| resources  | 22 |
| use        | 0  |
| supporting | 0  |



## Realization regular program

A set of data on the performed services of the Animal Hygiene Program.

resources 6  
use 0  
supporting 0

<https://data.gov.rs/sr/datasets/stanje-zemljišta-u-urbanim-sredinama/> **training providers**

```
resources 1
          use 0
```

2435 sets in total

### Examples of use→

[Download data](#)→

🔍 Data search...

Schedule according to...

## Organizations

## Labels

## Formats



## CATALOG OF CONTAMINATED LOCATIONS

The cadastre of contaminated sites represents a set of data on endangered, polluted and degraded lands and it is an integral part of the Land Information System managed by the Environmental Protection Agency. State bodies, i.e. organizations, bodies of autonomous provinces, local self-government units...

|            |   |
|------------|---|
| use        | 0 |
| supporting | 0 |



## Soil condition in urban areas

The data show the degree of soil endangerment from chemical pollution in urban areas in the period 2005-2023. year. Soil testing was carried out at locations in the immediate vicinity of traffic roads, industrial zones, near landfills, water supply sources, educational institutions, recreation...

|            |    |
|------------|----|
| resources  | 22 |
| use        | 0  |
| supporting | 0  |



## Realization regular program



A set of data on the performed services of the Animal Hygiene Program.

|            |   |
|------------|---|
| resources  | 6 |
| use        | 0 |
| supporting | 0 |



## CATALOG OF CONTAMINATED LOCATIONS

The cadastre of contaminated sites represents a set of data on endangered, polluted and degraded lands and it is an integral part of the Land Information System managed by the Environmental Protection Agency. State bodies, i.e. organizations, bodies of autonomous provinces, local self-government units...

resources 1   
use 0   
supporting 0 

What will be a column of data?

Visit <https://bit.ly/birn-data>, click scraping

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

HTML code







Катастар контаминираних локација представља скуп података о угроженим, загађеним и деградираним земљиштима и он је саставни део Информационог система земљишта који води Агенција за заштиту животне средине. Државни органи, односно организације, органи аутономних покрајина, јединице локалне самоуправе...

|            |   |
|------------|---|
| ресурса    | 1 |
| употреба   | 0 |
| подржавања | 0 |



Подаци показују степен угрожености земљишта од хемијског загађења у ур срединама у периоду 2005–2023. године. Испитивање земљишта вршено је локацијама у непосредној близини прометних саобраћајница, индустријски близини депонија, изворишта водоснабдевања, педагошких установа, рекре

## METHOD ONE

rhythm.less:66

Mouse/click  
around

```
Elements Console Network Performance >> [X] [5] [Settings] [Close]
preview_url : /tabular/preview/?url=https%3A%2F%2Fopendata.stat.gov.rs%2Fdata%2FWCJJSonRestService.Service1.svc%2Fdataset%2F2201IND01%2F1%2Fcsv", "published": "2024-03-20T15:56:43.643000", "schema": {}, "title": "\u0414\u043e\u043b\u0430\u0441\u0442\u0438-\u0442\u0443\u0440\u0438\u0441\u0442\u0430-\u0433\u043e\u0434\u0438\u0448\u043d\u0438\u043c\u0430-\u043f\u043e\u0434\u0430\u0442\u0438 - \u0434\u043e\u043b\u0430\u0441\u0442\u0438-\u0442\u0443\u0440\u0438\u0441\u0442\u0430-\u0433\u043e\u0434\u0438\u0448\u043d\u0438\u043c\u0430-\u043f\u043e\u0434\u0430\u0442\u0438", "type": "api", "url": "https://opendata.stat.gov.rs/data/WcfJsonRestService.Service1.svc/dataset/2201IND01/1/csv"}], "slug": "dolastsi-turista-godishnji-podatsi", "spatial": null, "tags": ["statistika", "ugostiteljstvo-i-turizam"], "temporal_coverage": null, "title": "\u0414\u043e\u043b\u0430\u0441\u0442\u0438-\u0442\u0443\u0440\u0438\u0441\u0442\u0430-\u0433\u043e\u0434\u0438\u0448\u043d\u0438\u043c\u0430-\u043f\u043e\u0434\u0430\u0442\u0438 - \u0434\u043e\u043b\u0430\u0441\u0442\u0438-\u0442\u0443\u0440\u0438\u0441\u0442\u0430-\u0433\u043e\u0434\u0438\u0448\u043d\u0438\u043c\u0430-\u043f\u043e\u0434\u0430\u0442\u0438", "uri": "https://data.gov.rs/api/1/datasets/dolastsi-turista-godishnji-podatsi/"}]" data-v-2e2e7c>
<ul data-v-2e2e7c class=...>
  <li data-v-2e2e7c>
    <a class="unstyled w-10 block" data-v-2e2e7c>
      <article class="dataset-card height-auto dataset-search-result py-xs" badges created_at="2024-05-29T13:28:17.964000" extras="[object Object]" frequency_date="2024-05-29T00:00:00" id="66571151a561adbb2191d89e" last_modified="2024-06-06T08:22:33.848000" last_update="2024-05-29T00:00:00" license="public_domain" page="https://data.gov.rs/sr/datasets/katastar-kontaminiranikh-lokatsija/" slug="katastar-kontaminiranikh-lokatsija" tags uri="https://data.gov.rs/api/1/datasets/katastar-kontaminiranikh-lokatsija/" data-v-c9e660 data-v-2e2e7c>
        <div class="card-logo search-logo-card" data-v-c9e660> ... </div> flex
        <div class="card-data search-logo-card" data-v-c9e660>
          <h4 class="card-title" data-v-c9e660>КАТАСТАР КОНТАМИНИРАНИХ ЛОКАЦИЈА</h4>
          <div class="card-description mt-xs" data-v-c9e660> ... </div> == $0
        </div>
        <ul class="card-hover-data" data-v-c9e660> ... </ul> flex
      </article>
```



Find the HTML  
for our "row"

Click  
this

Search +  
click here

METHOD TWO

amazon.com  
search results

Headlines from  
Nikkei Asia

datasets on  
data.gov.ro

Sales at Lidl

Now try it.

Anywhere!!!!

Wikipedia's  
Fictional Big  
Cats

List of amphibians  
from Germany's Red  
List Center

School Board Minutes  
from Grand Island Public  
Schools in Nebraska

Scraping is just

conn


But we need some software  
to do that for us!

to columns  
in your spreadsheet

What software?

# 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  
we  
Do

[-] **NerdvanaNC** 1 point 2 years ago

Look into learning and understanding HTML, then BeautifulSoup and Selenium. sites  
so Google for that. Oh and bonus tip, there's (official?)  
which made me very happy (I like Docker!)

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

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

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

Do python on them until

Bea

Req

P.S.

[-] **riisen** 3 points 2 years ago

for small projects use built in module req  
for bigger projects go with scrapy, its am

permalink source embed save save-RES repo

[-] **ned334** 5 points 2 years ago

Google "Selenium find\_element(By.XPATH, '/XPath/')" this

All elements have an XPath that you can copy from chrome by  
Inspect -> right click on code block -> copy full Xpath.

Scraping solved

permalink source embed save save-RES report reply

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  
✓prime 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)  
✓prime 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  
✓prime 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  
✓prime 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  
✓prime 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  
✓prime 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  
✓prime Today by 10:00 PM

Bulk savings to consider

Page 1 of 7

Chochkees Assorted Rubber Ducks Toy Duckies for Kids and Toddlers, Bath Birthday Baby Showers Classroom, ...  
★★★★☆ 1,155

Rubber Ducks Jeep Ducking (50 Pack) - Rubber Duckie...  
★★★★☆ 982  
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

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

Some sites need interaction





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

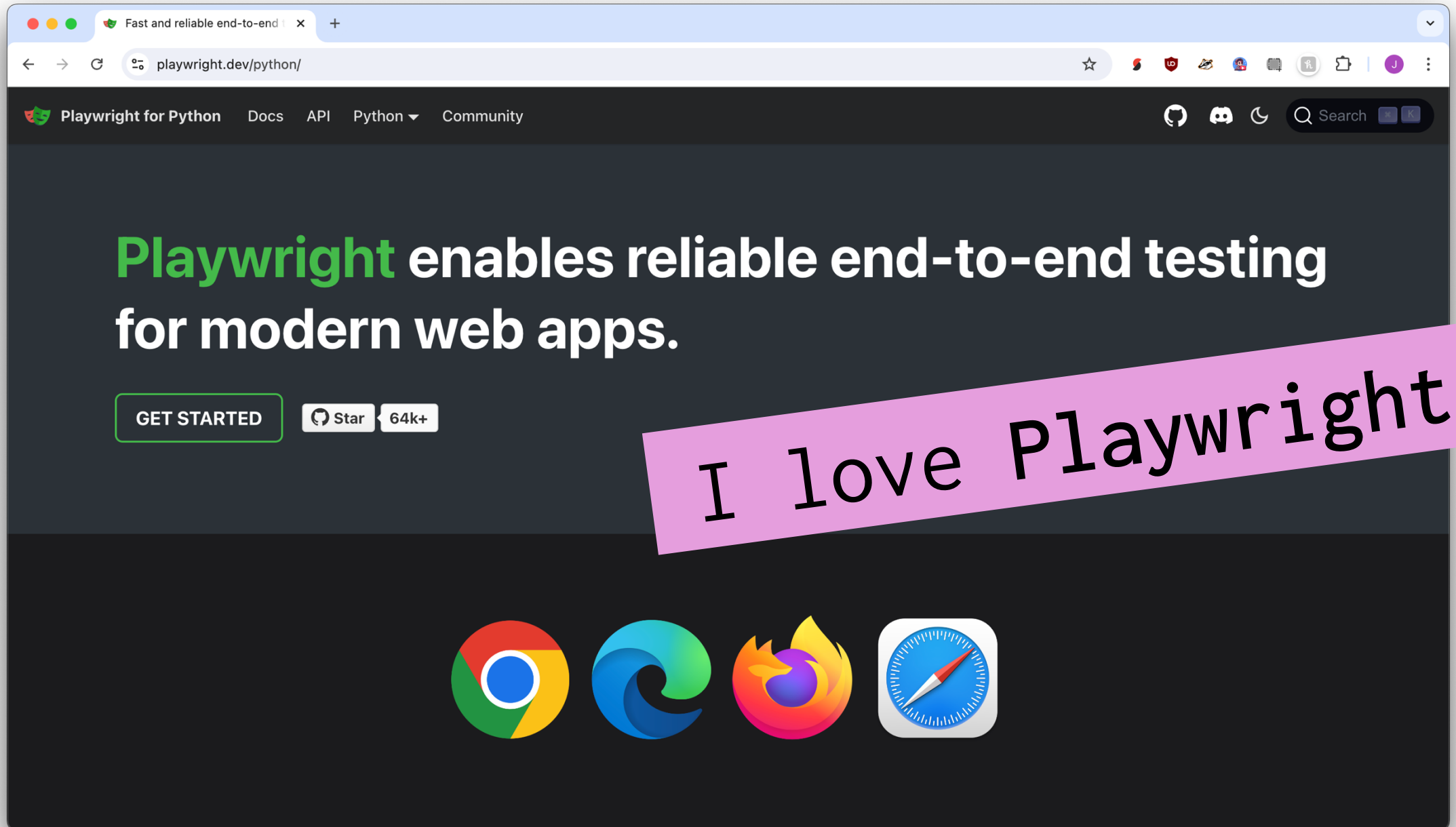
If license not found, please contact Customer Service at 800-803-9202

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

[Bookmark This Page](#)

Some sites need interaction





Playwright is perfect!

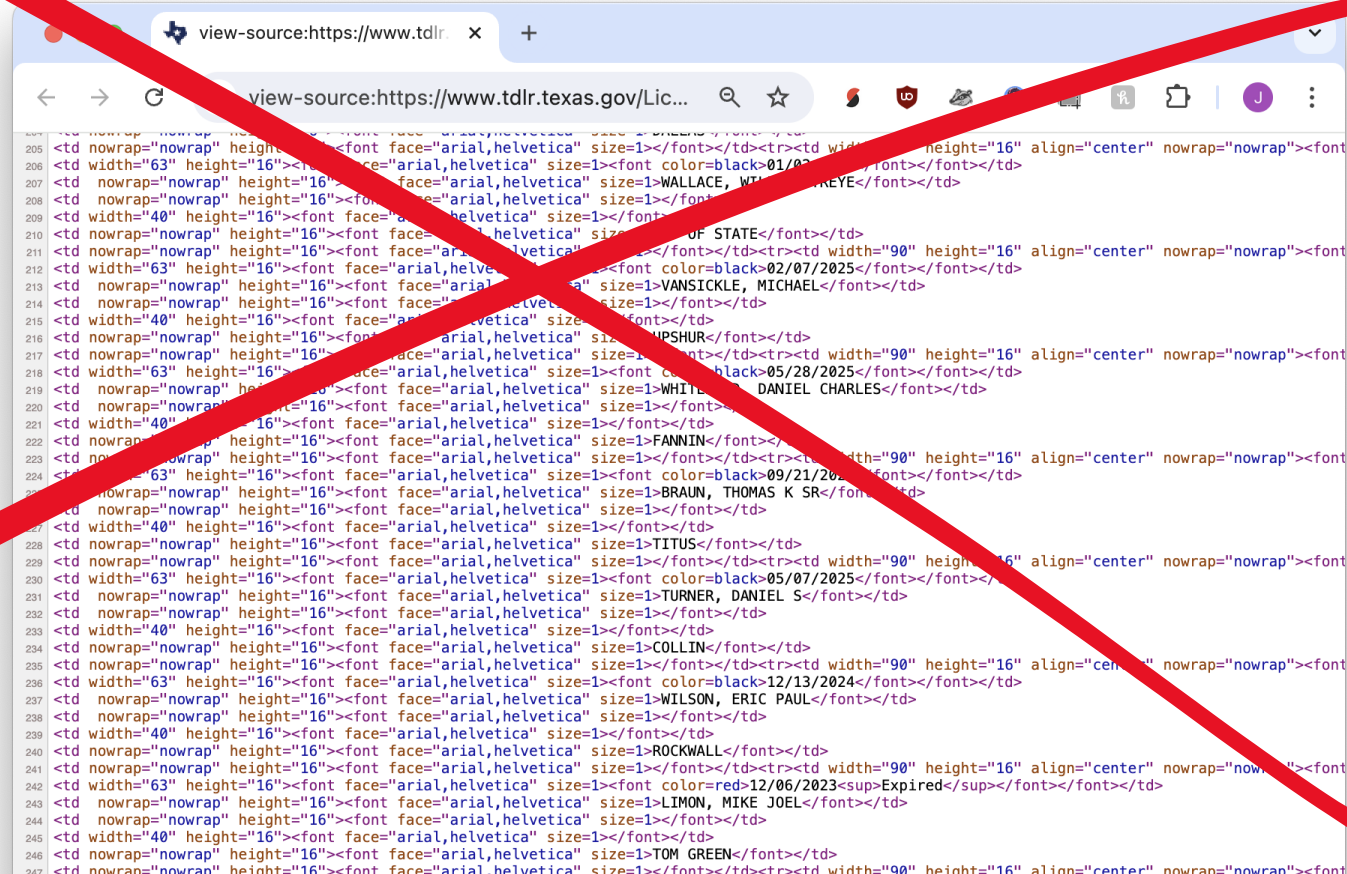
*But!* It's new, so

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:”

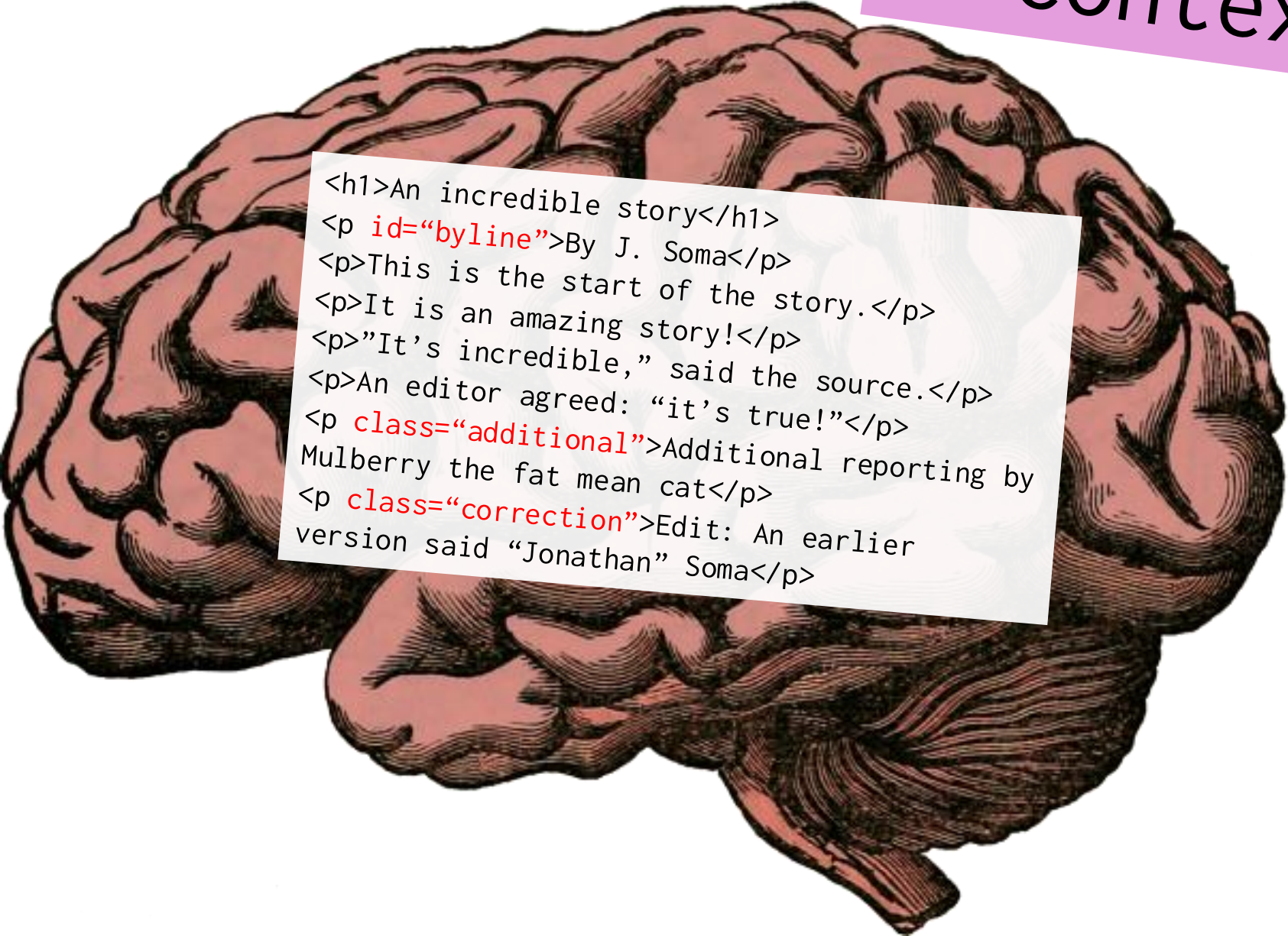


view-source:https://www.tdlr. x +

view-source:https://www.tdlr.texas.gov/Lic...

```
205 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
206 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=black>01/02/2025</font></font></td>
207 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>WALLACE, WYATT RAYMOND</font></td>
208 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
209 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
210 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>OF STATE</font></td>
211 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
212 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=black>02/07/2025</font></font></td>
213 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>VANSICKLE, MICHAEL</font></td>
214 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
215 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
216 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>HUPSHUR</font></td>
217 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
218 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=black>05/28/2025</font></font></td>
219 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>WHITFIELD, DANIEL CHARLES</font></td>
220 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
221 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
222 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>FANNIN</font></td>
223 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
224 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=black>09/21/2024</font></font></td>
225 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>BRAUN, THOMAS K SR</font></td>
226 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
227 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
228 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>TITUS</font></td>
229 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
230 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=black>05/07/2025</font></font></td>
231 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>TURNER, DANIEL S</font></td>
232 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
233 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
234 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>COLLIN</font></td>
235 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
236 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=black>12/13/2024</font></font></td>
237 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>WILSON, ERIC PAUL</font></td>
238 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
239 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
240 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>ROCKWALL</font></td>
241 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="now. "><font
242 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=red>12/06/2023</font></font></td><sup>Expired</sup></font></td>
243 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>LIMON, MIKE JOEL</font></td>
244 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
245 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
246 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>TOM GREEN</font></td>
247 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
248 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=black>07/13/2025</font></font></td>
249 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>MOEHLE, JAMES C</font></td>
250 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
251 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
252 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>HARRIS</font></td>
253 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
254 <td width="63" height="16"><font face="arial,helvetica" size=1><font color=black>05/07/2025</font></font></td>
255 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>WIESE, RICHARD L</font></td>
256 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font>
257 <td width="40" height="16"><font face="arial,helvetica" size=1><font></font>
258 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font>DENTON</font></td>
259 <td nowrap="nowrap" height="16"><font face="arial,helvetica" size=1><font></font></td><tr><td width="90" height="16" align="center" nowrap="nowrap"><font
```

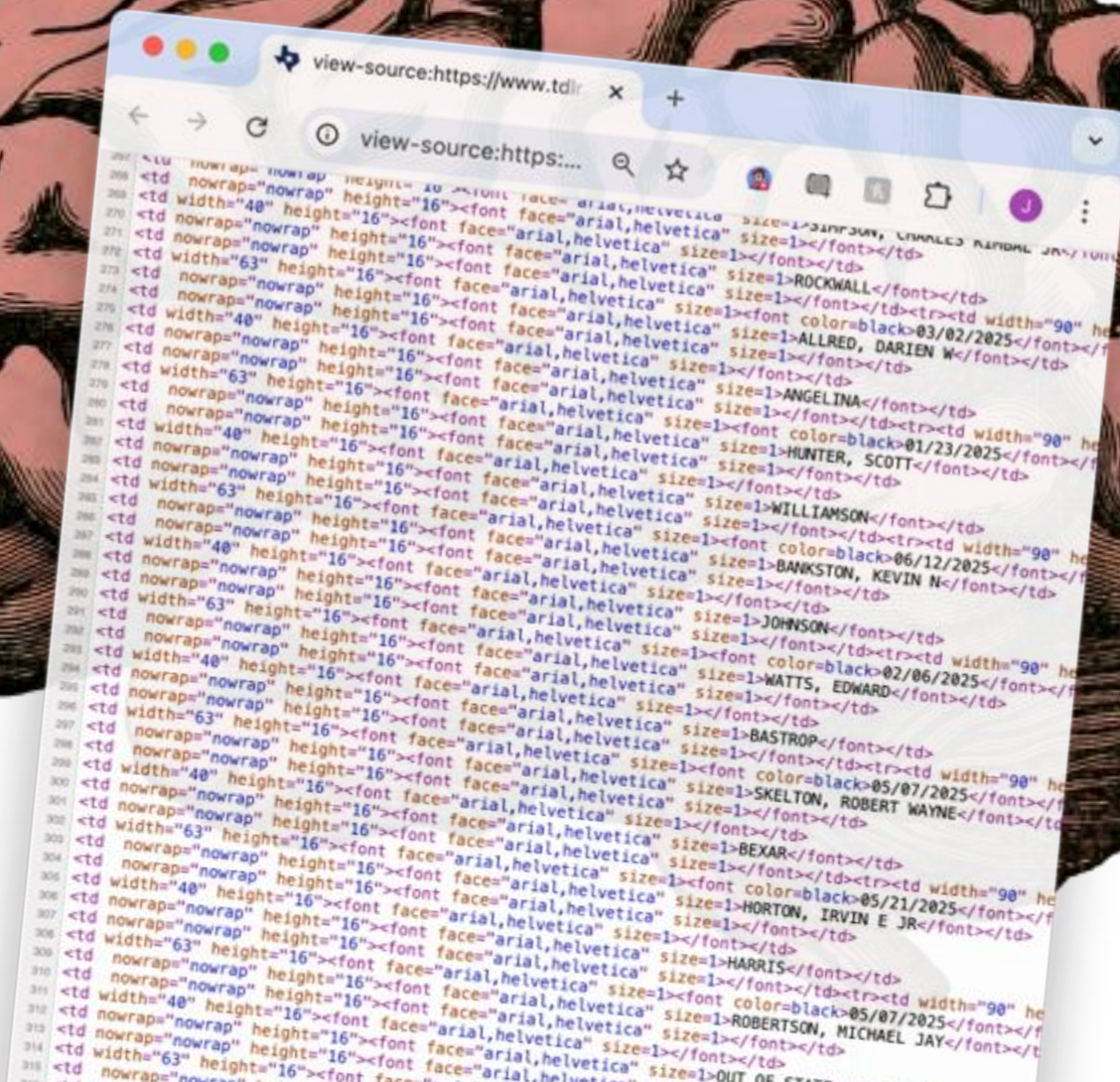
# “Context window”



```
<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>
```



# “Context window”



Visit <https://bit.ly/birn-data>, click **scrapping**

## 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/birn-data>

Open a new Jupyterlab  
notebook and let's  
start scraping!



# 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"

Viewing 1-25 of 148  
[Download Results](#) | [Email Results](#)

[Next Page](#) »

[Previous Page](#) «

| Number | Licensee                             | License Type                              | Licensing Board              | License Status | ContactLastNameGroup |
|--------|--------------------------------------|---|------------------------------|----------------|----------------------|
| 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/birn-data>, click **scrapping**

# Scrapping!

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