

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

Scraping!

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

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

The image shows the JupyterLab interface. On the left, there's a sidebar with 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). In the center, there's a code editor window titled "Untitled2.ipynb" containing Python code for web scraping using Playwright. A large pink banner at the bottom reads "Download, install and open JupyterLab".

```
[ ]: 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()
```

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 **scraping**

<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 **scraping**

```
<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

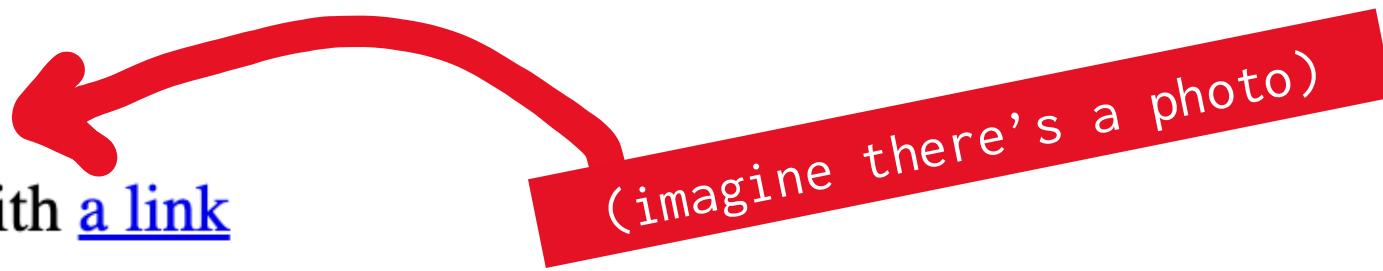
And an even smaller headline

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 **scraping**

<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 **scraping**

<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>

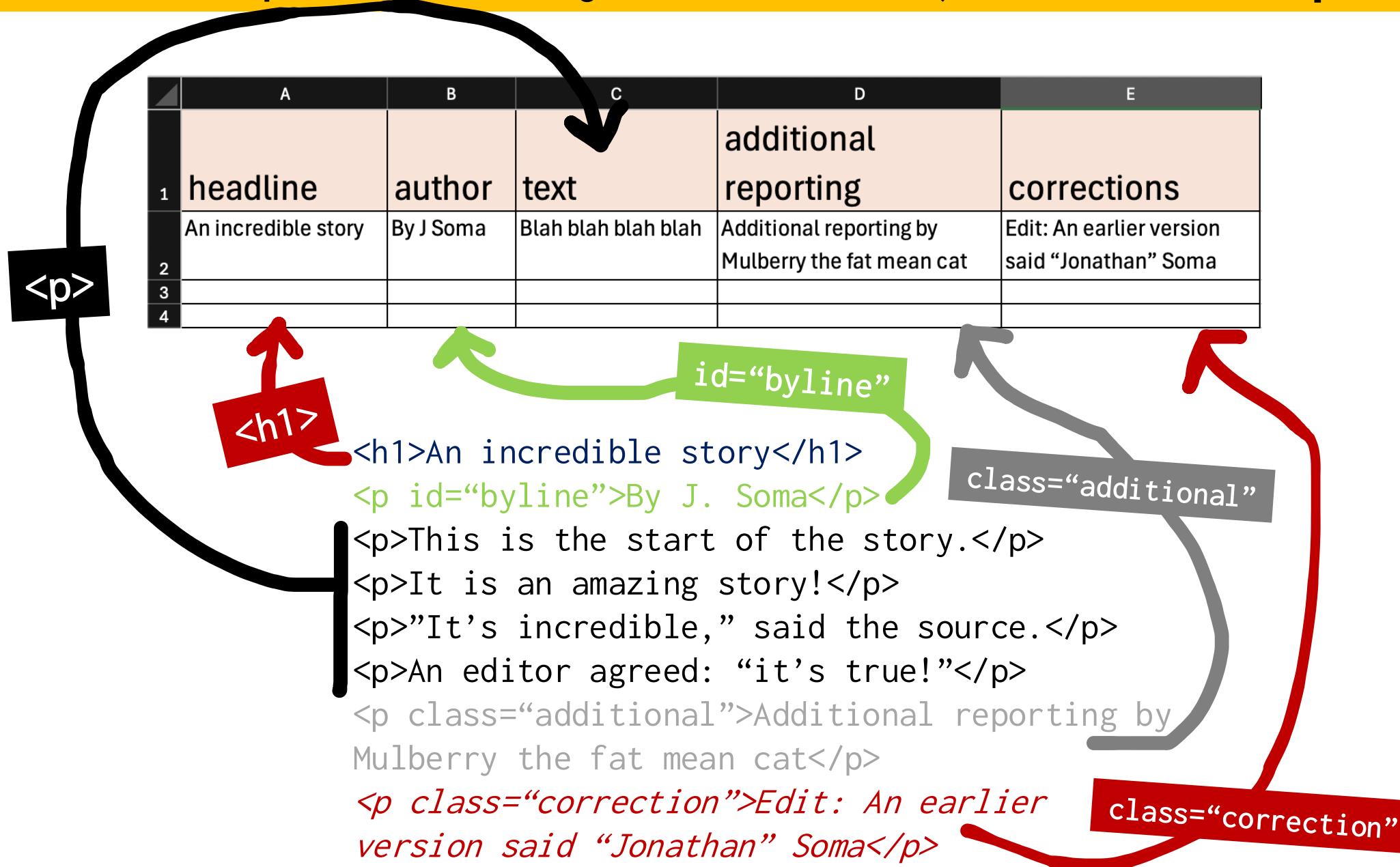
Visit <https://bit.ly/birn-data>, click **scraping**

```
<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.com

BBC Home - Breaking News

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

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



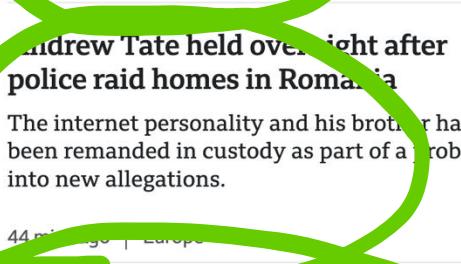
Divers find five bodies in wreck of sail yacht



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.

Data sets - Open data

data.gov.rs/sr/datasets/

Data sets

2435 sets in total

Download data→

What will be a row in our spreadsheet?

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...

 **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...

 **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-zemljshta-u-urbanim-sredinama/> training providers

resources 1
use 0

Data sets - Open data

data.gov.rs/sr/datasets/

Data sets

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

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

resources 1 use 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...



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

Right click (or command+click) and Inspect

Data.gov.rs

Претрага података...

Распоред према... Организације Ознаке Формати ...

ресурса 1
употреба 0
подржавања 0

ресурса 22
употреба 0
подржавања 0

ресурса 6
употреба 0
подржавања 0

ресурса 1
употреба 0
подржавања 0

Look Up "КОНТАМИНИРАНИХ"
Copy
Copy Link to Highlight
Search Google for "КОНТАМИНИРАНИХ"
Print...
Open in Reading Mode
Translate Selection to English
Block element...
Inspect
Speech
Services

Катастар контаминираних ло
земљиштима и он је саставни
животне средине. Државни о
локалне самоуправе...

Стање земљишта у урбани
Подаци показују степен угрож
2005-2023. године. Испитива
саобраћајница, индустријски
станова, рекреац...

Редовни програм Реализација
Скуп података о извршеним услугама Програма делатности зоохигијене.

Евиденција акредитованих спроводилаца обука НАЈУ
Евиденција акредитованих спроводилаца обука НАЈУ За потребе акредитације спроводилаца програма
обуке, Национална академија, по правилу, два пута годишње објављује јавни позив за поједине
области стручног усавршавања. Јавни позив објављује се на интернет страници Националне академије,
министарства н



Find the HTML for our “row”

[Elements](#) [Console](#) [Sources](#) [Network](#) [Performance](#)

preview_url : /tabular/preview/?url=https://opendata.stat.gov.rs%2fodata%2fwcfjs
onRestService.Service1.svc%2Fdataset%2F2201IND01%2F1%2Fcsv", "published": "2024-03-20T1
5:56:43.643000", "schema": {}, "title": "\u0414\u043e\u043b\u0430\u0441\u0446\u0438 \u0
442\u0443\u0440\u0438\u0441\u0442\u0430 - \u0433\u043e\u0434\u0438\u0448\u045a\u0438 \u
043f\u043e\u0434\u0430\u0446\u0438", "type": "api", "url": "https://opendata.stat.gov.r
s/WcfJsonRestService.Service1.svc/dataset/2201IND01/1/csv"}], "slug": "dolasti-tu
-godishnji-podatsi", "spatial": null, "tags": ["statistika", "ugostiteljstvo-i-tur
42\u0443\u0440\u0438\u0441\u0442\u0430 - \u0433\u043e\u0434\u0438\u0448\u045a\u0438
3f\u043e\u0434\u0430\u0446\u0438", "uri": "https://data.gov.rs/api/1/datasets/dolas
tsi-turista-godishnji-podatsi/"}}] data-v-2e2e7c>
▼<ul data-v-2e2e7c class=>
▼<li 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-kontamin
iranikh-lokatsija/" slug="katastar-kontaminiranih-lokatsija" tags uri="https://
/data.gov.rs/api/1/datasets/katastar-kontaminiranih-lokatsija/" data-v-c9e660
-2e2e7c>
class="card-logo search-logo-card" data-v-c9e660>...</div> flex
class="card-data search-logo-card" data-v-c9e660>
class="card-title" data-v-c9e660>КАТАСТАР КОНТАМИНИРАНИХ ЛОКАЦИЈА</h4>
div class="card-description mt-xs" data-v-c9e660> == \$0
"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...

КАТАСТАР КОНТАМИНИРАНИХ ЛОКАЦИЈА

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

Стање земљишта у урбаним срединама

Подаци показују степен угрожености земљишта од хемијског загађења у урбаним срединама у периоду 2005-2023. године. Испитивање земљишта вршено је на локацијама у непосредној близини прометних саобраћајница, индустријских близини депонија, изворишта водоснабдевања, педагошких установа, рекреац

Click this

Search + click here

METHOD TWO

:hov .cls +,

amazon.com
search results

Sales at Lidl

List of amphibians
from Germany's Red
List Center

Headlines from
Nikkei Asia

Now try it.
Anywhere!!!!

School Board Minutes
from Grand Island Public
Schools in Nebraska

datasets on
data.gov.ro

Wikipedia's
Fictional Big
Cats

Scraping is just
connecting to some software
to do that for us!

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



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



Some sites need interaction

Chochkees Asso

Ducks Toy Duckies for Kids
and Toddlers, Bath Birthday
Baby Showers Classroom,...
 1,155

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

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

Company



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 Moon 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.

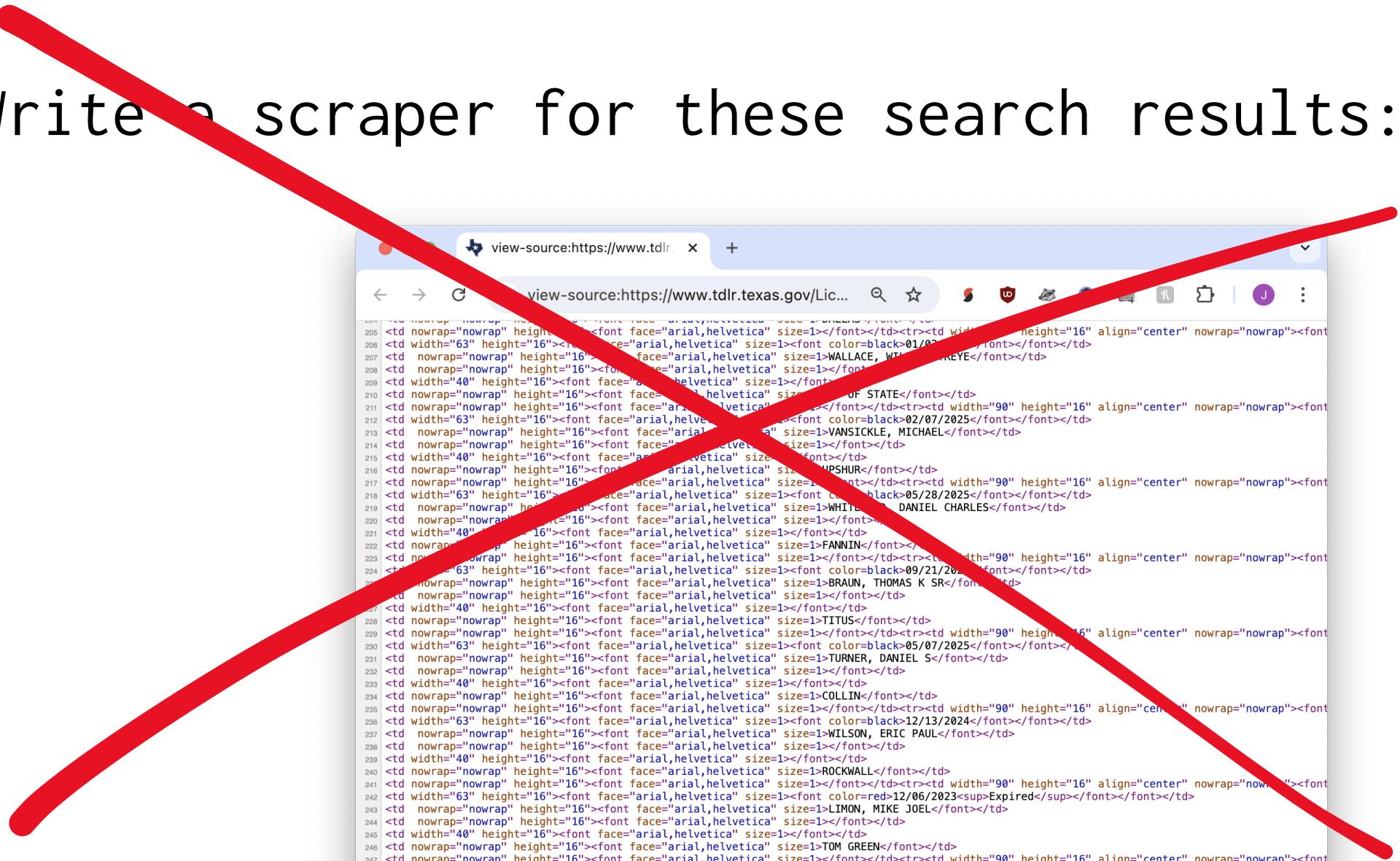
Playwright + ChatGPT +
pasting samples of HTML

=

infinite scrapers



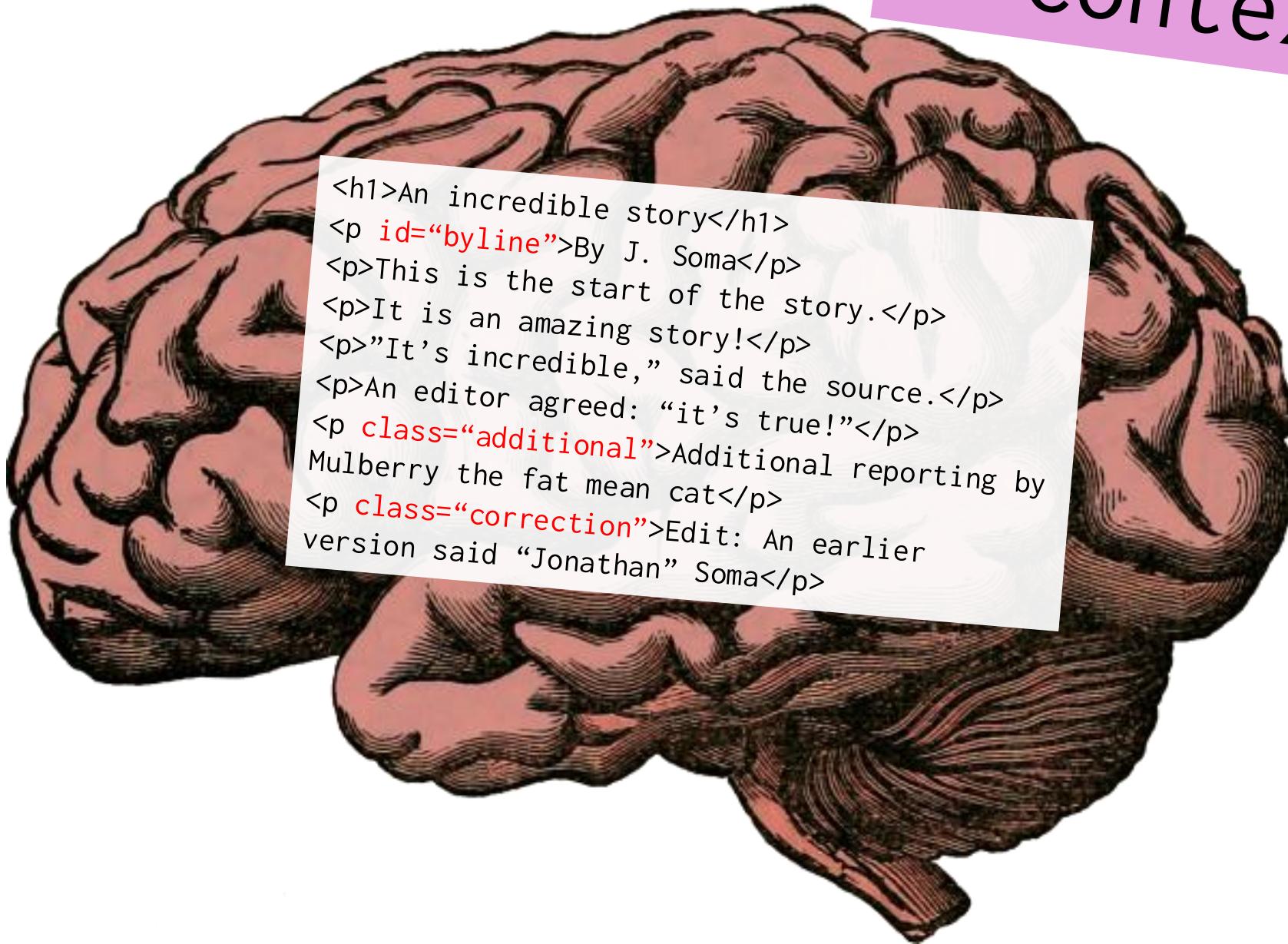
“Write a scraper for these search results:”

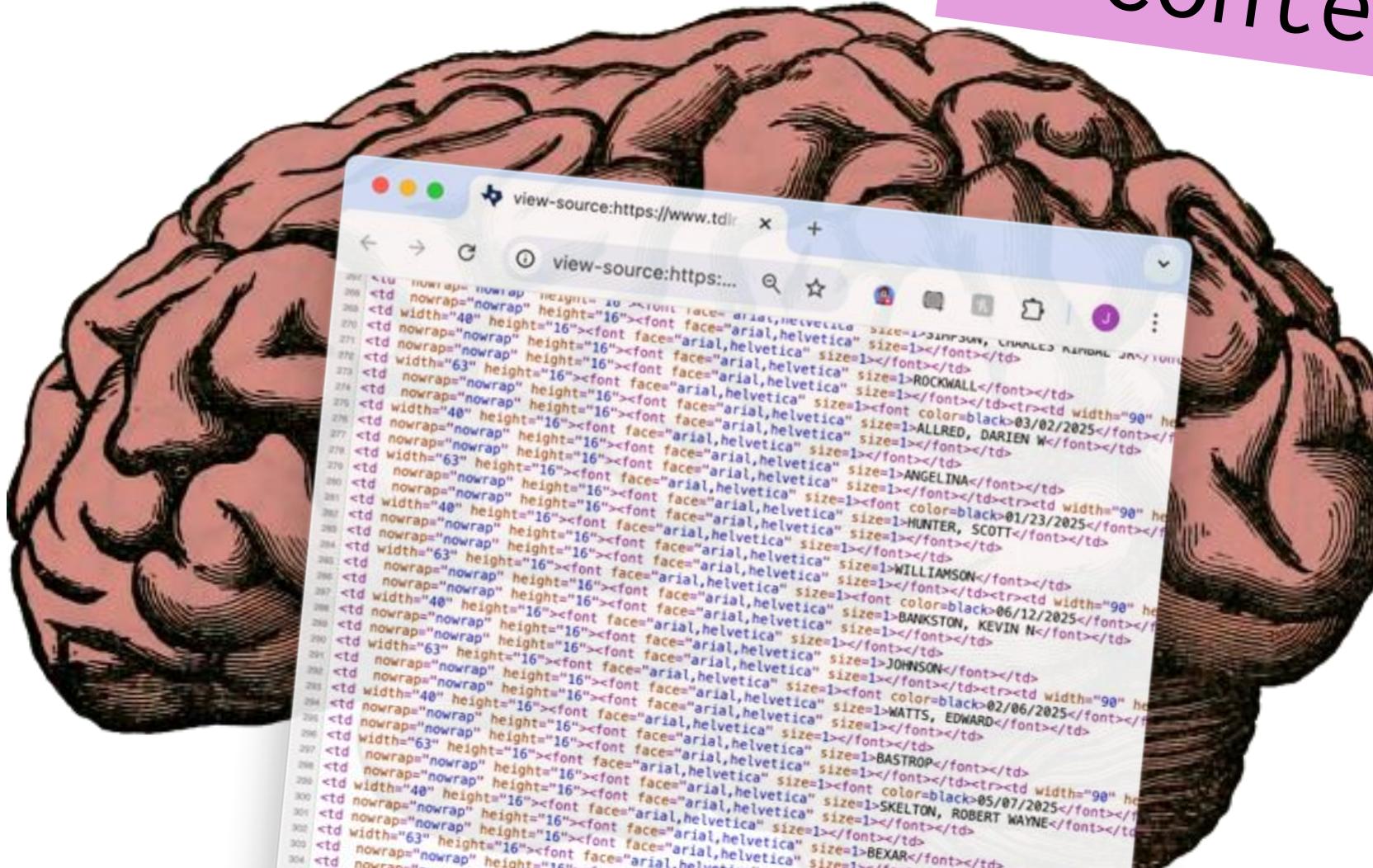


A screenshot of a web browser window titled "view-source:https://www.tdlr.texas.gov/Lic...". The page content is the raw HTML source code of a search results page. The code consists of numerous `<td>` elements, each containing text such as names and dates. A large red 'X' is drawn across the entire screenshot.

```
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>WALLACE, WILSON, RAYE</font></td>
<td width="40" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>OF STATE</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>02/07/2025</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>VANSICKLE, MICHAEL</font></td>
<td width="40" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>WIESE, RICHARD L</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>05/28/2025</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>WHITE, DANIEL CHARLES</font></td>
<td width="40" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>FANNIN</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>09/21/2025</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>BRAUN, THOMAS K SR</font></td>
<td width="40" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>TITUS</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>05/07/2025</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>TURNER, DANIEL S</font></td>
<td width="40" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>COLLIN</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>12/13/2024</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>WILSON, ERIC PAUL</font></td>
<td width="40" height="16" align="center" nowrap="nowrap"><font face="arial,helvetica" size=1>ROCKWALL</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1>12/06/2023</font><sup>Expired</sup></font></td>
<td width="63" height="16" align="center" nowrap="nowrap">><font color="red" face="arial,helvetica" size=1>12/06/2023</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1>LIMON, MIKE JOEL</font></td>
<td width="40" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1>TOM GREEN</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1>07/13/2025</font></td>
<td width="63" height="16" align="center" nowrap="nowrap">><font color="black" face="arial,helvetica" size=1>MOEHL, JAMES C</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1>HARRIS</font></td>
<td width="40" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1>05/07/2025</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1>WIESE, RICHARD L</font></td>
<td width="40" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1>DENTON</font></td>
<td nowrap="nowrap" height="16" align="center" nowrap="nowrap">><font face="arial,helvetica" size=1></font></td>
```

“Context window”





“Context window”

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

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!

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](#)

Viewing 1-25 of 148

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

[Next Page ▶](#)

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/birn-data>, click **scraping**

Scraping!

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