Visit https://bit.ly/lede-ai-playwright for material

Magical Scraping!

stealing (??) data from the internet with the magic of Playwright + AI

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>
<img src=“photo.jpg”>
<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

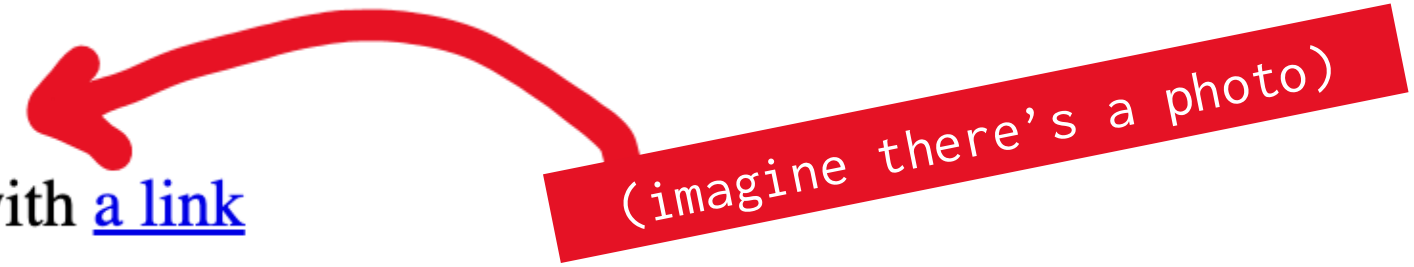
This is a smaller headline

And an even smaller headline

This is a paragraph

This is a paragraph with a link

This is ANYTHING
Anything! Anything in the world



```
<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!”

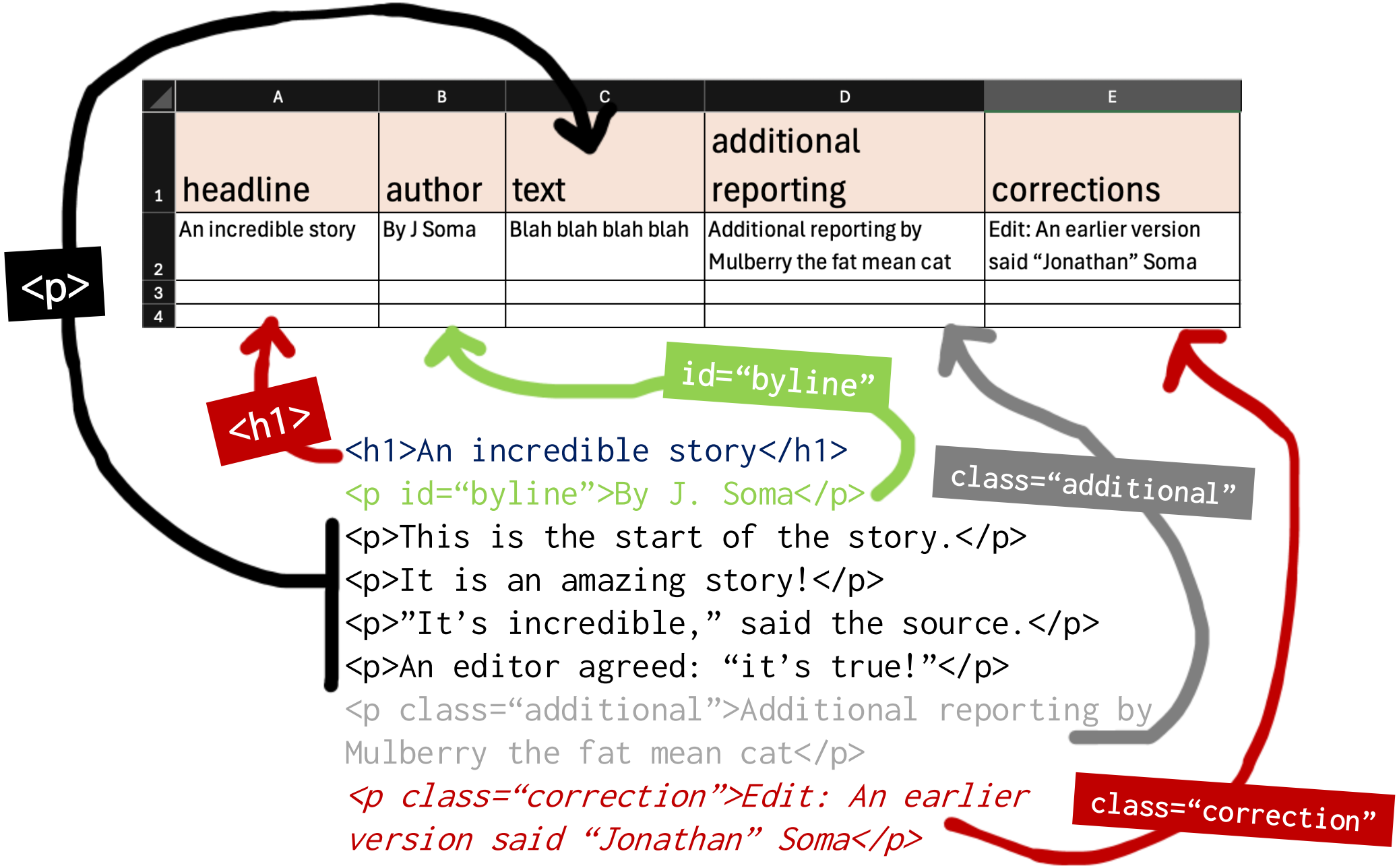
Additional reporting by Mulberry the fat mean cat

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>
```

```
<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>
```


A
B
C
D
E
1
headline
author
text
additional reporting
corrections
2
An incredible story
By J Soma
Blah blah blah blah
Additional reporting by Mulberry the fat mean cat
Edit: An earlier version said “Jonathan” Soma
3
4

<h1>
id=“byline”
class=“additional”
class=“correction”
<h1>An incredible story</h1>
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

BBC

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.

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

...ife and six children were ...

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

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

BBC

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 alongsid[e] [P]residential nominee Kamala Harris for N[o]vember's US election.

Three things the Democrats have avoided s[o] far at the DNC

What they have tried to avoid says as much about their weaknesses as what they c[h]ose 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

[D]ivers find five bodies in wreck [o]f [S]icily yacht

Gaz[a] [n]urse says whole fam[i]ly, i[n]cluding quadruplets, kille[d] in air [s]trike

Ashraf El Attar's wife and six childr[e]n were killed when an air strike destroyed [t]heir home in Deir al-Balah.

[5] hrs ago | Middle East

[A]ndrew Tate held ove[r]night after police raid homes in Roma[n]ia

The internet personality and his broth[e]r have been remanded in custody as part of a [p]robe into new allegations.

44 mi[n]s 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.

メルカリ - 日本最大のフリマサー

jp.mercari.com/search?category_id=79

mercari

楽器・機材

ログイン 会員登録 出品

What will be a row in our spreadsheet?

絞り込み クリア

除外キーワー

カテゴリー

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

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

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 みぎて

¥1,500 バイオリン用　顎当て　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…

¥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

mercari

楽器・機材

ログイン　会員登録　出品

絞り込み　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

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

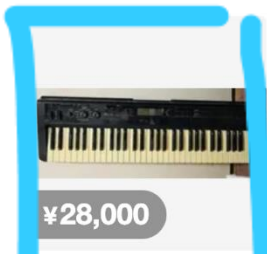
商品の状態

楽器・機材 の検索結果

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

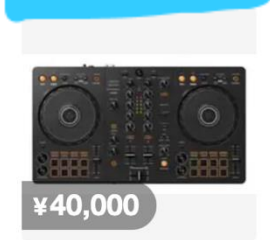

¥28,000
korg kross61 電池駆動軽量化シンセ

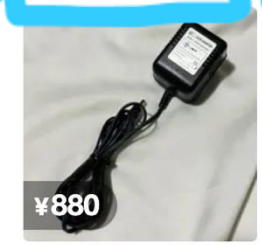

¥3,800
iQ7 ステレオマイク Lightning接続 (修理品)

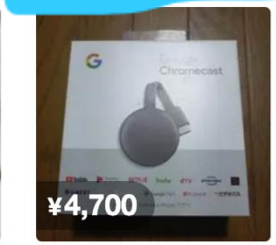

¥990
ギターストラップ　未使用


¥700
音楽之友社 うたとピアノの絵本 1 みぎて

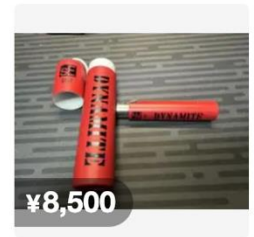

¥1,500
バイオリン用　顎当て　黒檀　訳あり

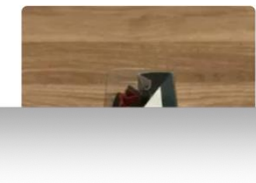

¥40,000
YUCKY様専用 Pioneer DDJ-FLX4

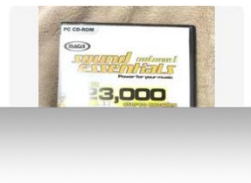

¥880
ACアダプター JH35P1200150D


¥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...


¥8,500
SE ELECTRONICS DM1 DYNAMITE



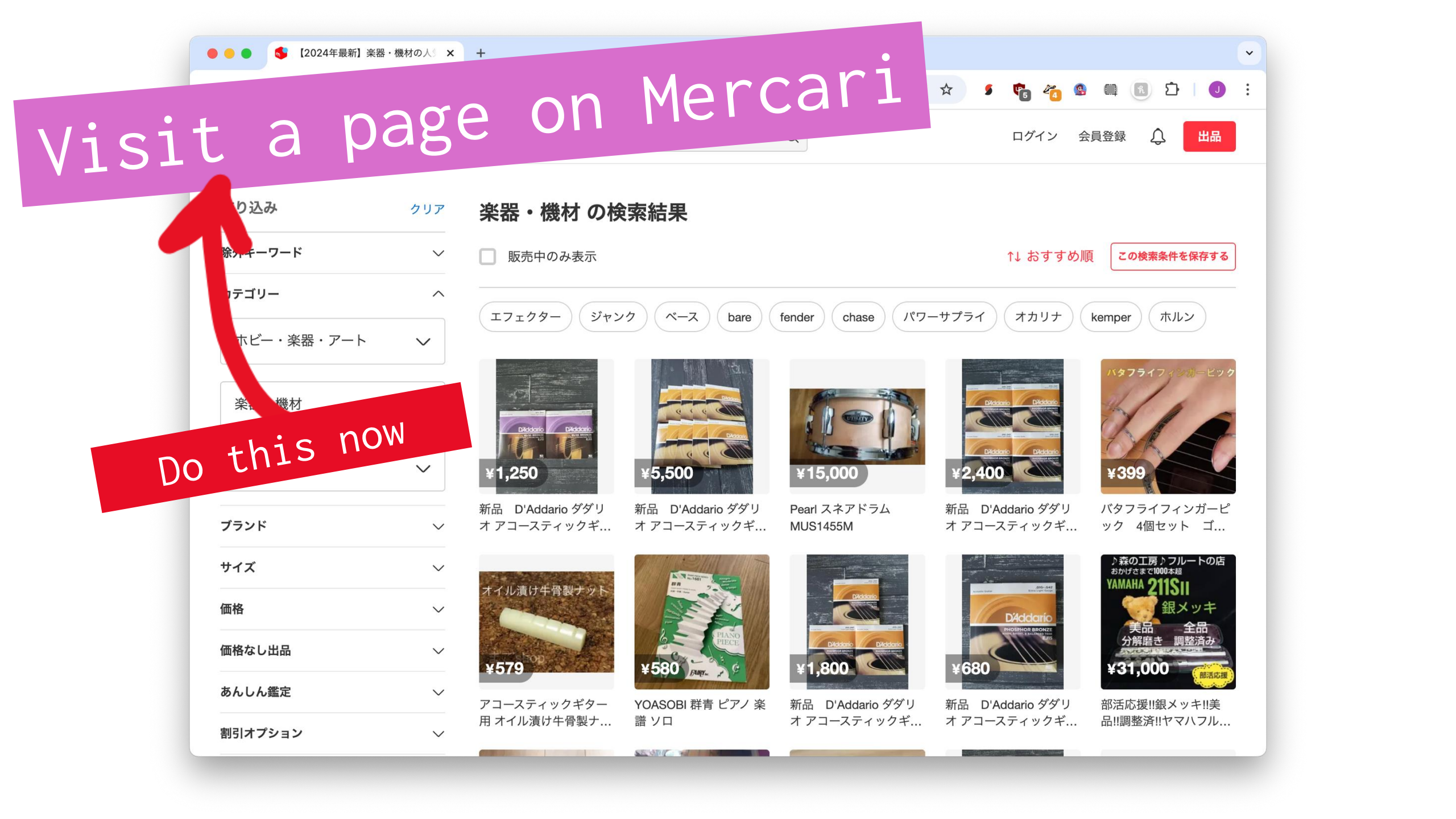


https://jp.mercari.com/item/m97630098490

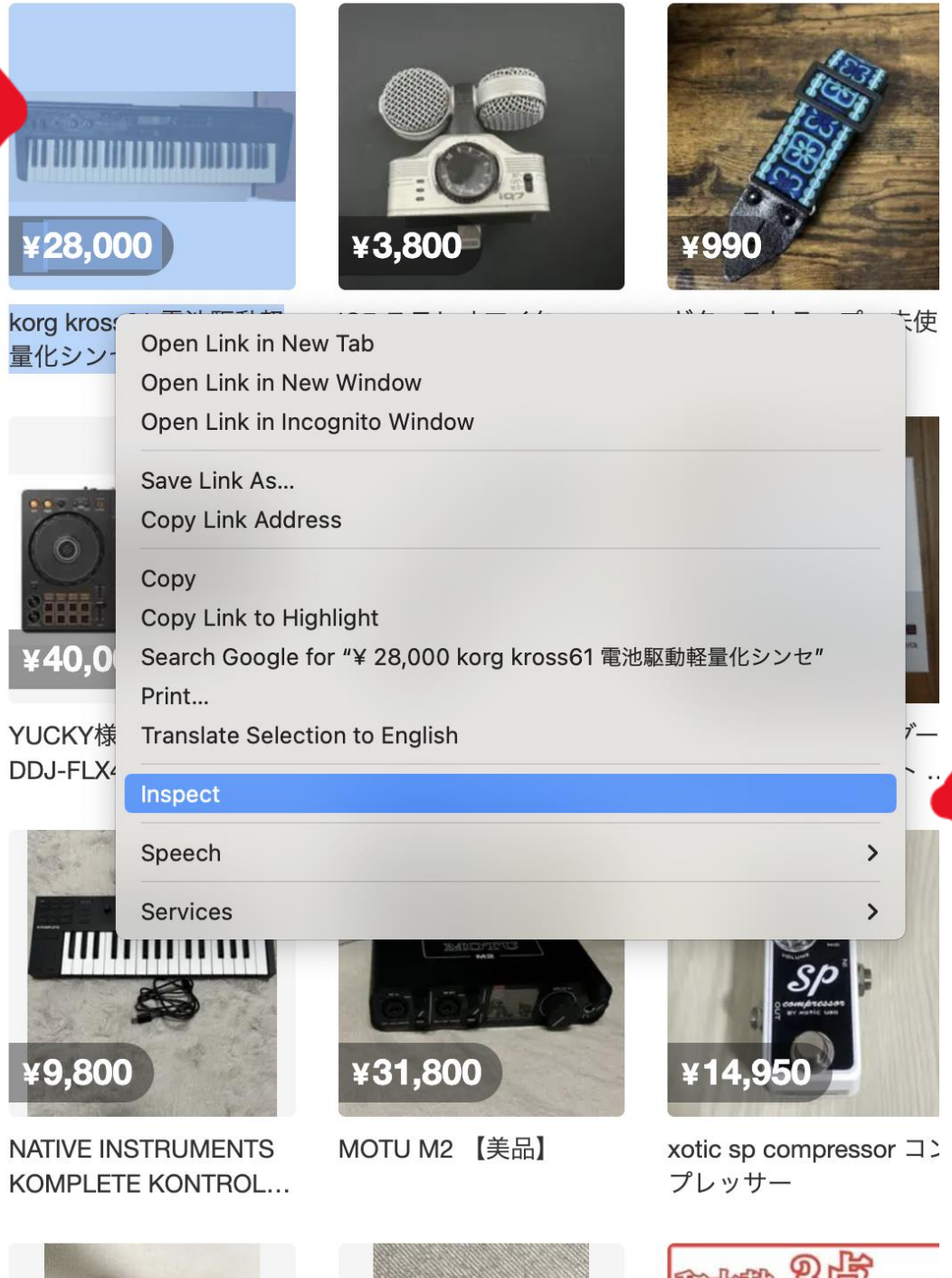
What will be a column of data?

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

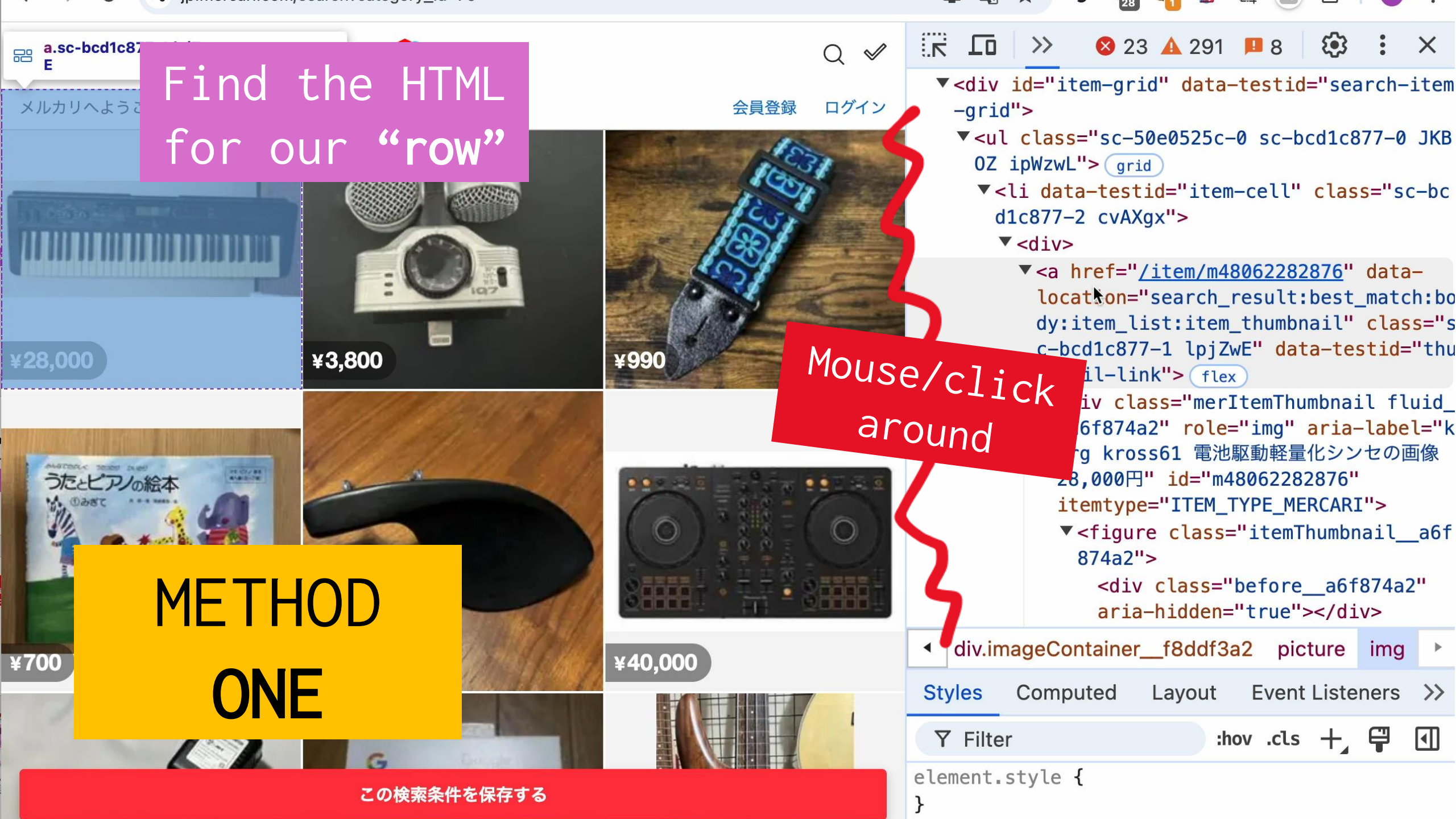
HTML code


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
出品
り込み
クリア
除外キーワード
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
楽器・機材 の検索結果
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
¥1,250
新品　D'Addario ダダリオ アコースティックギ...
¥5,500
新品　D'Addario ダダリオ アコースティックギ...
¥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...
バタフライフィンガーピック
¥399
バタフライフィンガーピック　4個セット　ゴ...
オイル漬け牛骨製ナット
¥579
アコースティックギター用 オイル漬け牛骨製ナ...
¥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...
¥680
新品　D'Addario ダダリオ アコースティックギ...
♪森の工房♪フルートの店
おかげさまで1000本超
YAMAHA 211SII
銀メッキ
美品　全品
分解磨き　調整済み
¥31,000
部活応援
部活応援!!銀メッキ!!美品!!調整済!!ヤマハフル...

Right click (or command+click) and Inspect



検証

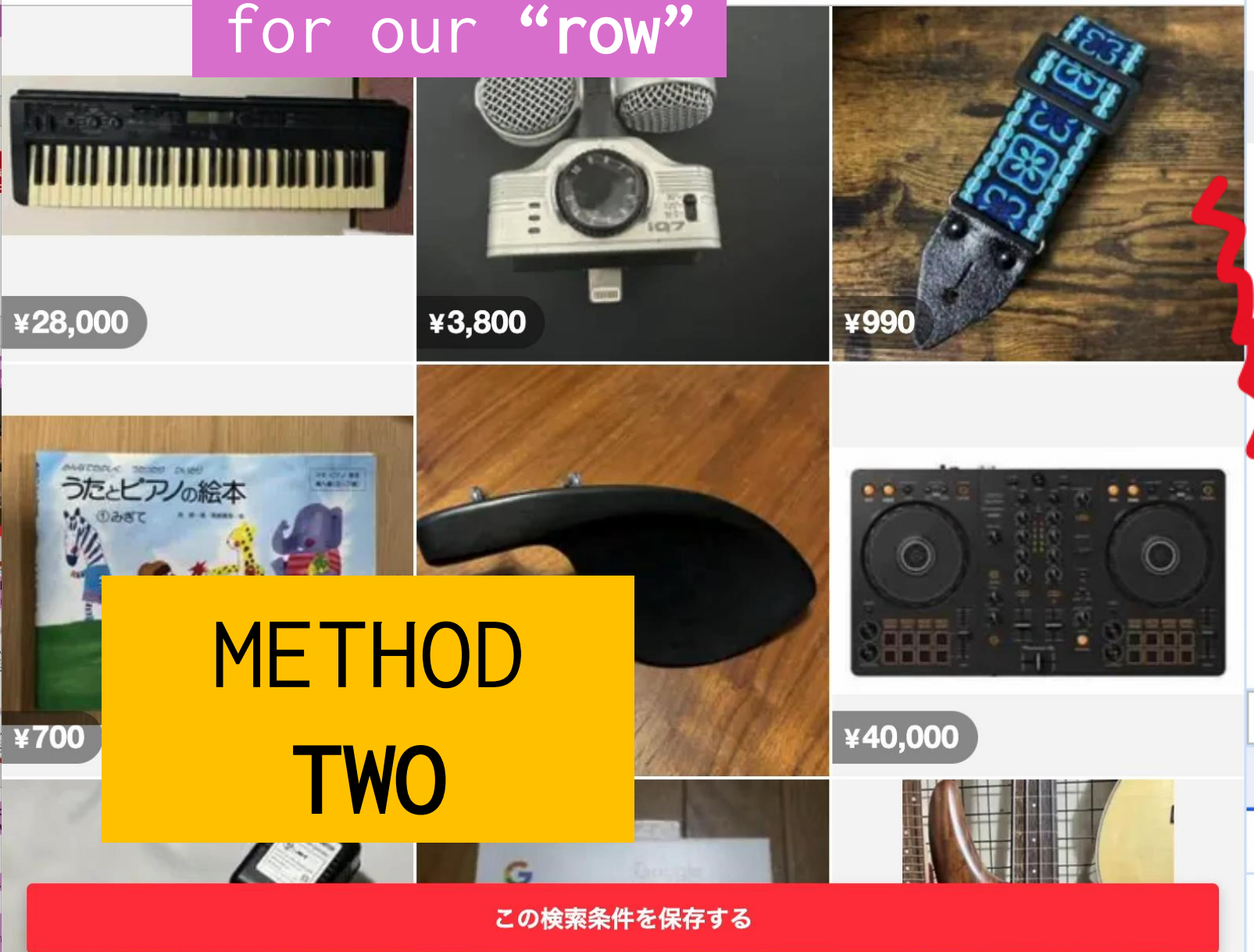

Find the HTML for our “row”
Mouse/click around
METHOD ONE
会員登録
ログイン
¥28,000
¥3,800
¥990
¥700
¥40,000
この検索条件を保存する

<ul class="sc-50e0525c-0 sc-bcd1c877-0 JKBOZ ipWzwL">
grid
<li data-testid="item-cell" class="sc-bcd1c877-2 cvAXgx">

<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_ 6f874a2" role="img" aria-label="k rg kross61 電池駆動軽量化シンセの画像 28,000円" id="m48062282876" itemtype="ITEM_TYPE_MERCARI">
<figure class="itemThumbnail__a6f874a2">

div.imageContainer__f8ddf3a2
picture
img
Styles
Computed
Layout
Event Listeners
Filter
:hov
.cls
element.style {
}

Find the HTML for our “row”



METHOD TWO

Click this

Search + click here

amazon.com search results

Headlines from Nikkei Asia

items on Mercari

Now try it.

Sales at Lidl

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 connecting to columns in your spreadsheet

But we need some software to do that for us!

What software?

All Images Videos News Books Web Finance Tools

Scraping libraries

From sources across the web

- Selenium
- Beautiful Soup
- Scrapy
- Lxml
- Requests
- MechanicalSoup
- Playwright
- Puppeteer
- Urllib3
- ZenRows S.L.
- Cheerio

[–] **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

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

Look into learning and understanding HTML, then BeautifulSoup and Selenium.

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

From my own research, run away from Selenium. The right direction is CDP (Chrome Developers

```
all_links = soup.find_all(
```

Do python on them

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

for small projects use built in module

for bigger projects go with scrapy, its

permalink source embed save save-RES

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

Google "Selenium find_element(By.XPATH, '/XPATH/')"

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

P.S.

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

Selenium is **garbage!**

…but BeautifulSoup

can’t scrape all sites.

Amazon.com : rubber ducks

amazon.com/s?k=rubber+ducks&crid=126R7EQI3ZJ1A&sprefix=rubber+ducks%2Caps%2C268&ref=nb_sb_noss_1

Top discounts for you

Page 1 of 3

Continental Grand Prix 5000 Rennrad Faltreifen // 28-622 (700x28C)
★★★★½ 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…
★★★★½ 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"…
★★★★½ 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…
★★★★½ 569
4K+ viewed in past month
$21.89
prime **Overnight by 8:00 AM**

ArtCreativity Assorted Rubber Ducks Jeep Ducking (100Pack) - Rubber Ducki…
★★★★½ 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…
★★★★½ 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**

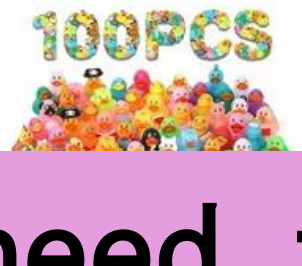
Bulk savings to consider

Page 1 of 7

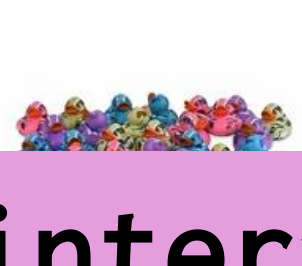
Chochkees Asso… Ducks Toy Duckies for Kids and Toddlers, Bath Birthday Baby Showers Classroom,…
★★★★½ 1,155

Rubber Ducks Jeep Ducking (50 Pack) - Rubber Duckie…
★★★★½ 982
100+ viewed in past month

Set, Mini Colorful Rubber Duckies Bath Toy for Child,Float & Squeak Tiny…
★★★★½ 763



Bulk, Jeep Ducks for Ducking, Assorted Rubber Ducks Jeep Ducking, Bab…
★★★★½ 569

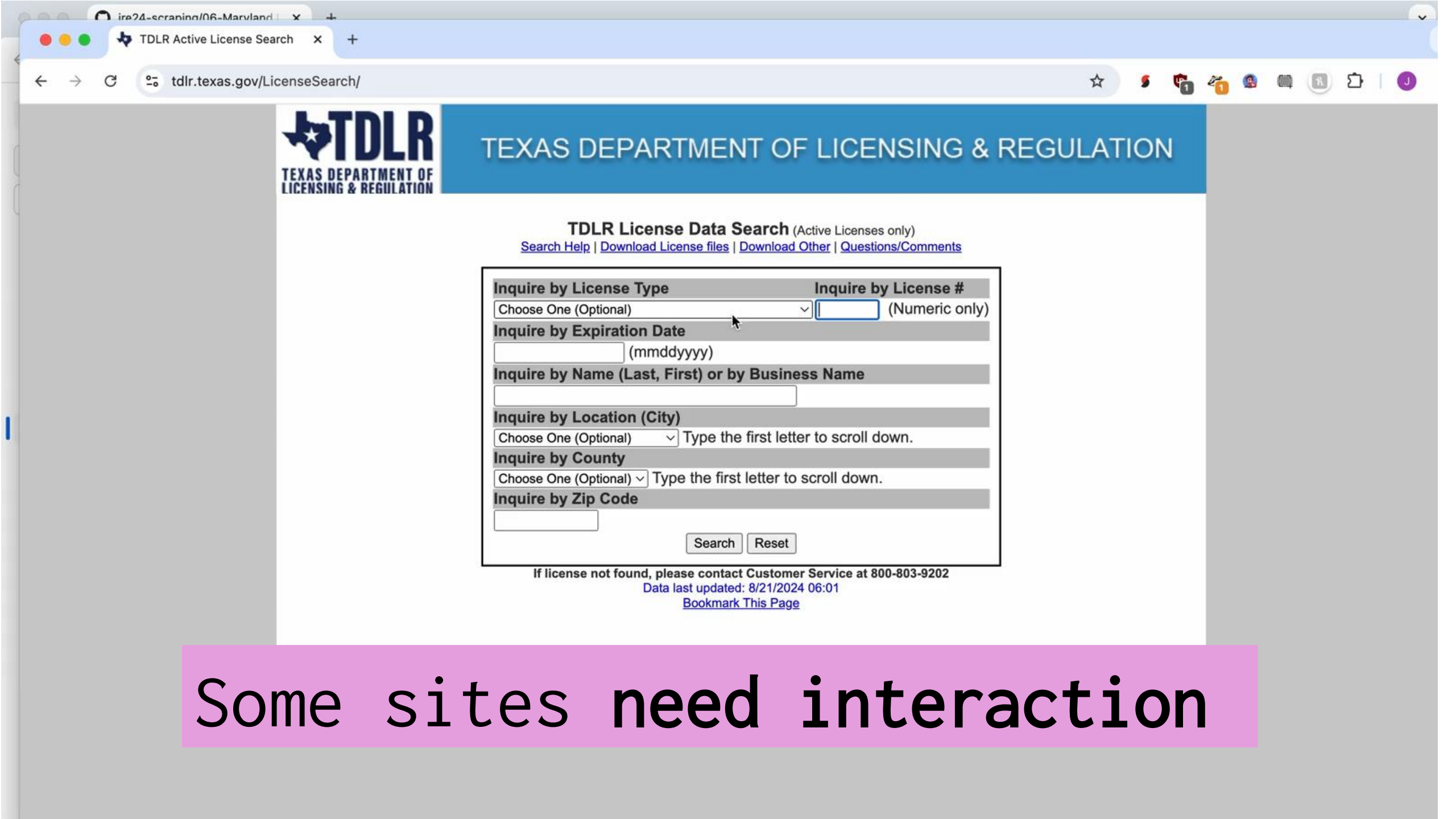


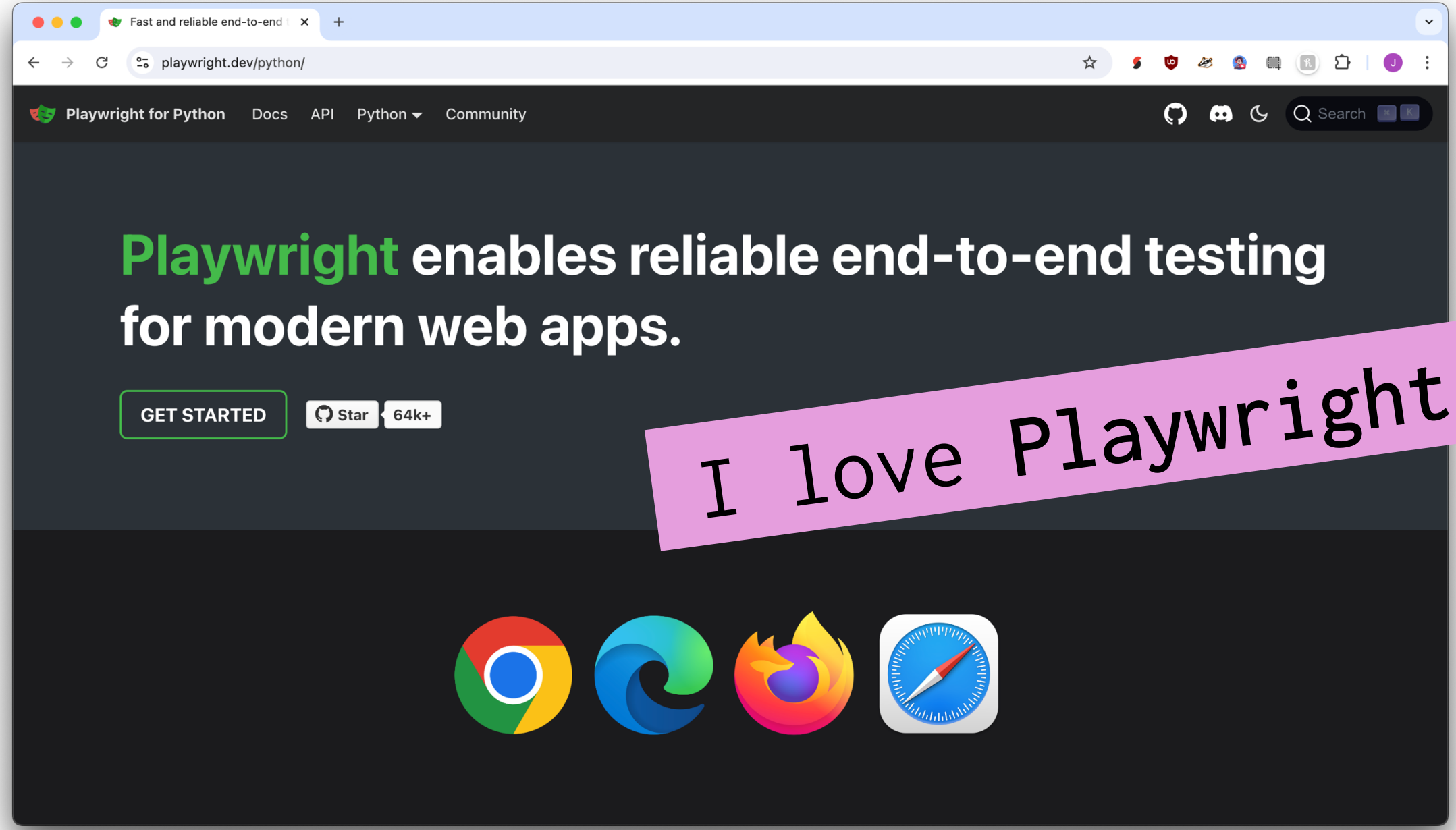
Ducks: Fun Unique Military-Inspired Bath Toys for Jeep Ducking or Play - 2 inches
800+ viewed in past month

Bulk, Jeep Ducks for Ducking, Assorted Rubber Ducks for Jeeps, Bath Toy…
★★★★½ 569

Company Glitter Rubber Duck Toy Assortment Duckies for Kids, Bath Birthday Gifts…
★★★★½ 1,094

Some sites need interaction


TDLR Active License Search
tdlr.texas.gov/LicenseSearch/
TDLR
TEXAS DEPARTMENT OF LICENSING & REGULATION
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)
(Numeric only)
Inquire by Expiration Date
(mmddyyyy)
Inquire by Name (Last, First) or by Business Name
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
Search
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
Search
Playwright enables reliable end-to-end testing for modern web apps.
GET STARTED
Star
64k+
I love Playwright!

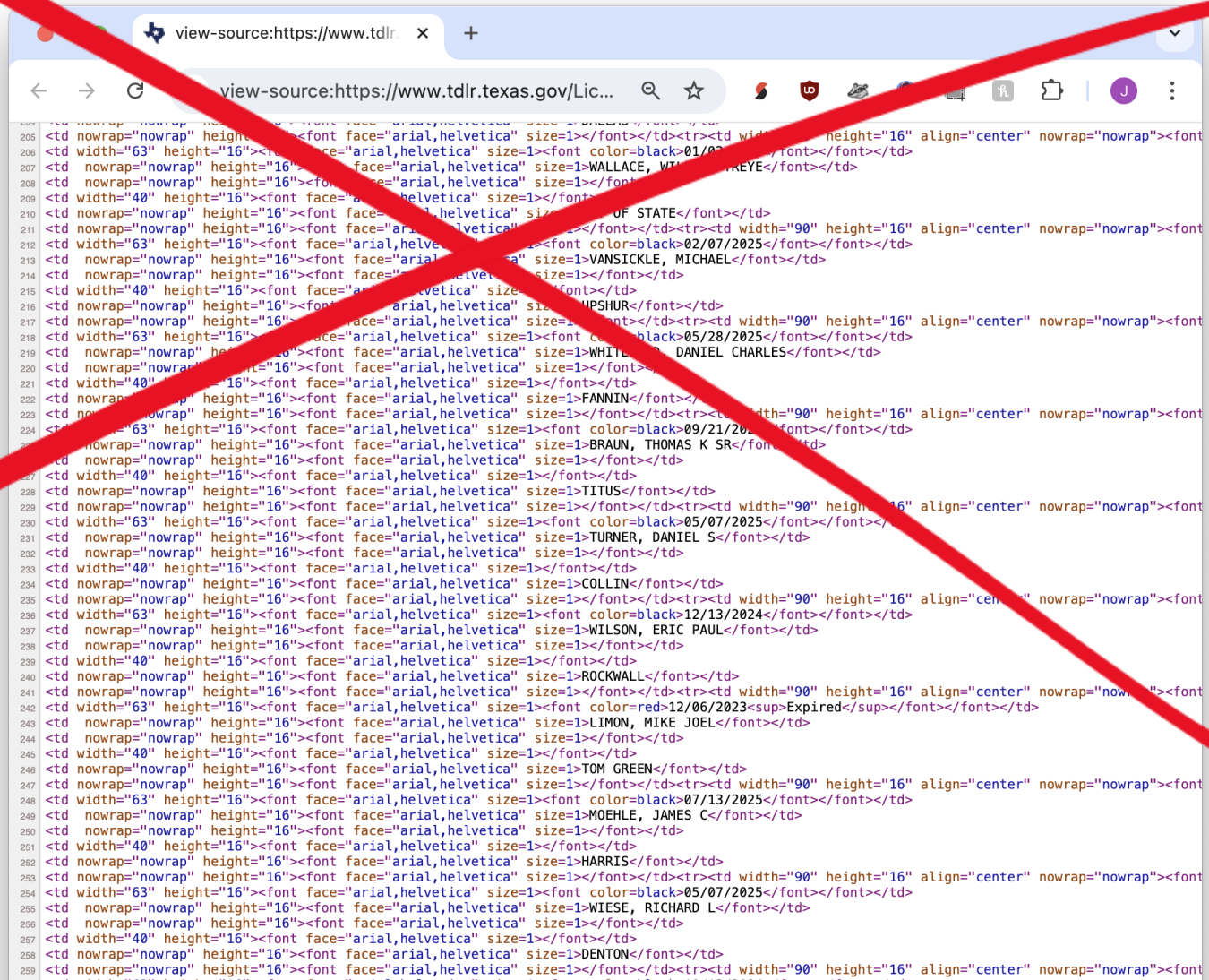
Playwright is **perfect**!
But! It's **new**, so
ChatGPT isn't very good
at it. *But!* We try.

Playwright + ChatGPT + pasting samples of HTML

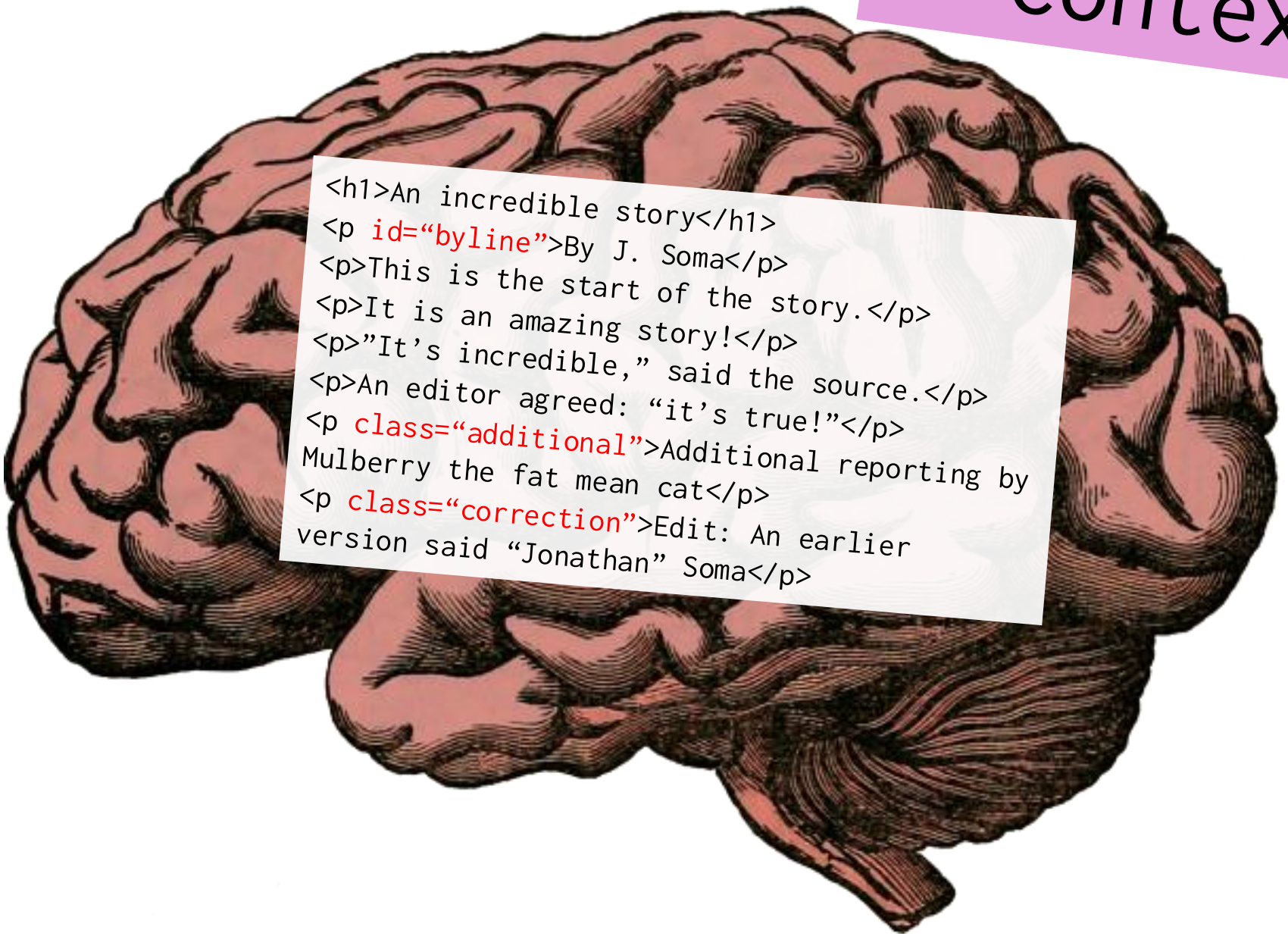
=

infinite scrapers 

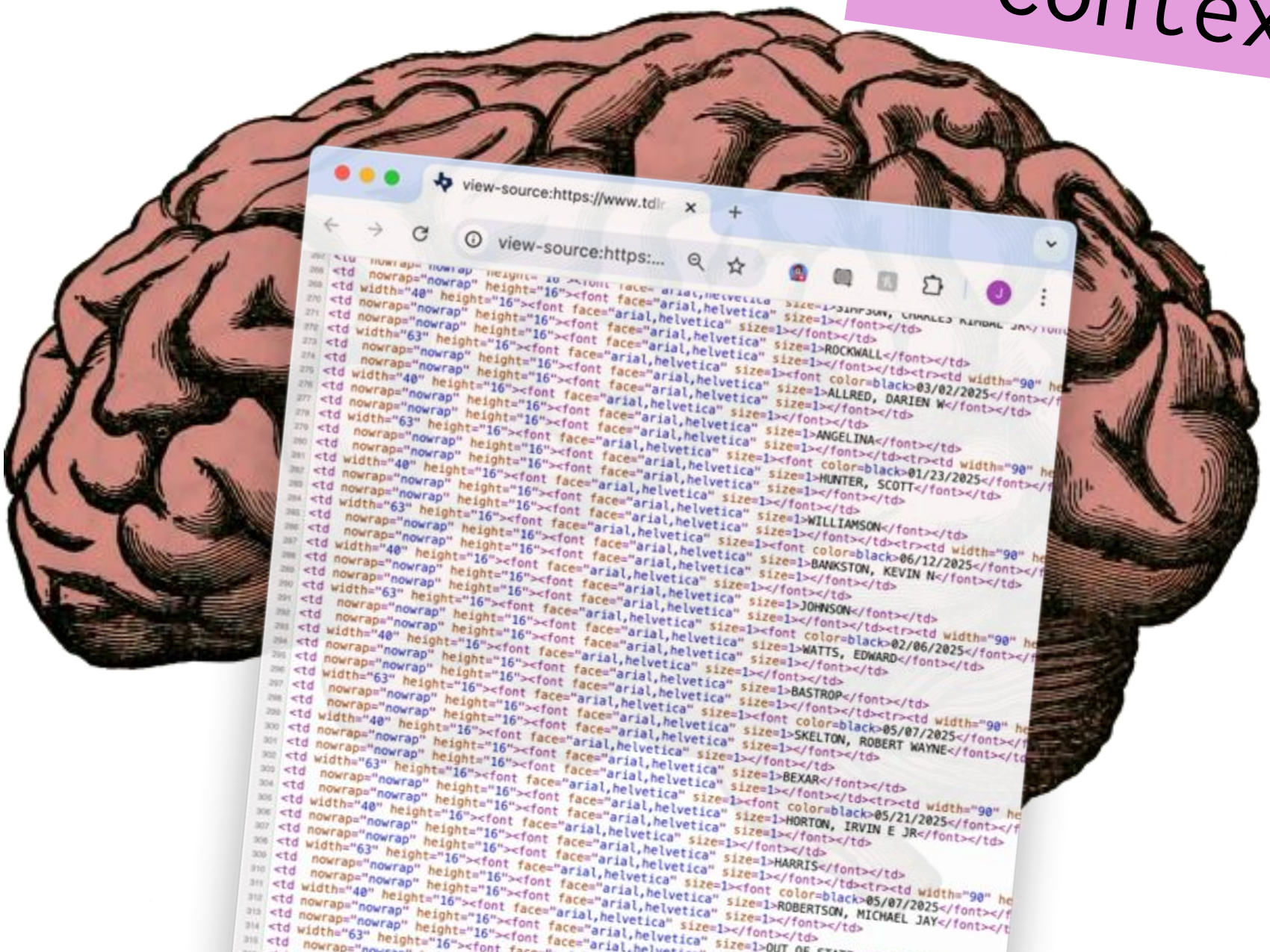
“Write a scraper for these search results:”



“Context window”



```
<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”

What we might need:

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

Visit https://bit.ly/lede-ai-playwright for material

Magical Scraping!

stealing (??) data from the internet with the magic of Playwright + AI