BSDG

BSDG Discord!

Join our community to stay engaged between events, and help drive the content of future meetings

Web Scraping with Playwright & HAR Files

By **Tom Day**

InstaGram Giveaway!

BEARDEDBOARBBQ

Playwright

“an open-source framework for *reliable* end-to-end testing for modern web apps” - Microsoft

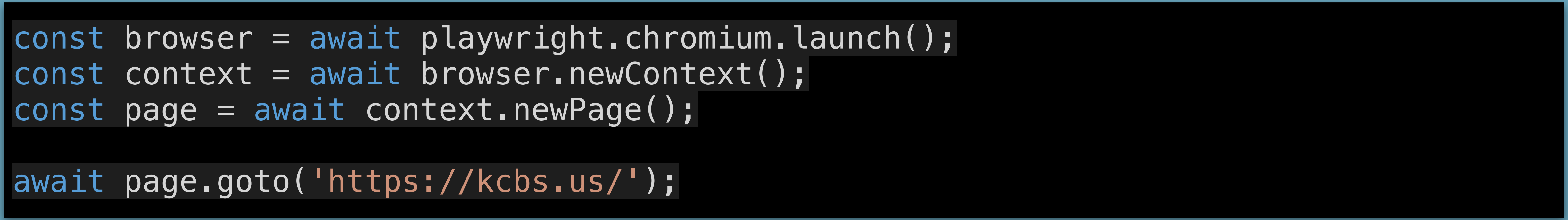
https://playwright.dev

https://2023.stateofjs.com/en-US/libraries/testing/#testing_ratios

https://learn.microsoft.com/en-us/training/modules/build-with-playwright/

Playwright Navigation

https://playwright.dev/docs/navigations



```
const browser = await playwright.chromium.launch();
const context = await browser.newContext();
const page = await context.newPage();

await page.goto('https://kcbs.us/');
```

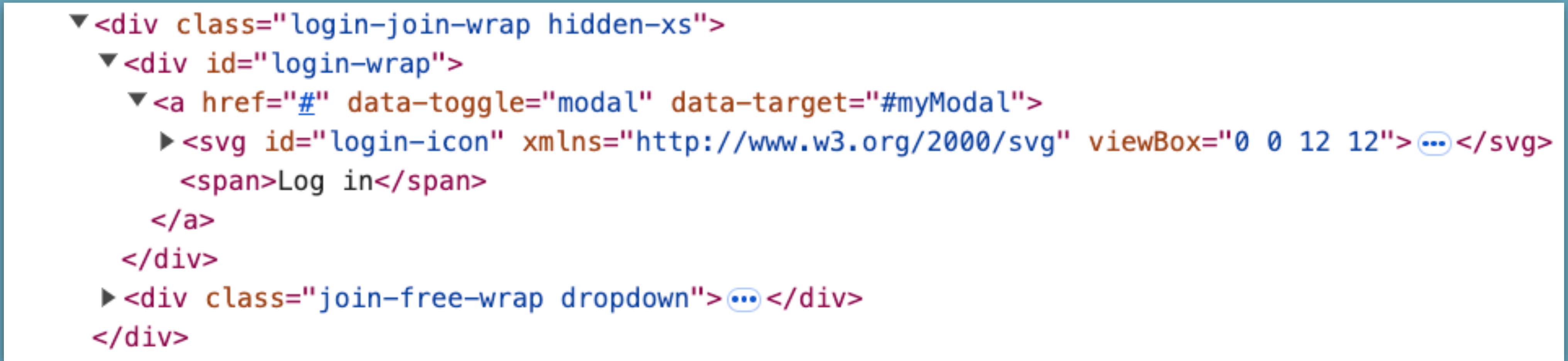
The code above loads the page and waits for the web page to fire the load event. The load event is fired when the whole page has loaded, including all dependent resources such as stylesheets, scripts, iframes, and images.

Playwright Locators

https://playwright.dev/docs/locators

```
await page.getByLabel('User Name').fill('John');
await page.getByLabel(‘Password').fill('secret-password');
await page.getByRole('button', { name: 'Sign in' }).click();
await expect(page.getByText('Welcome, John!')).toBeVisible();
```

Playwright Locators



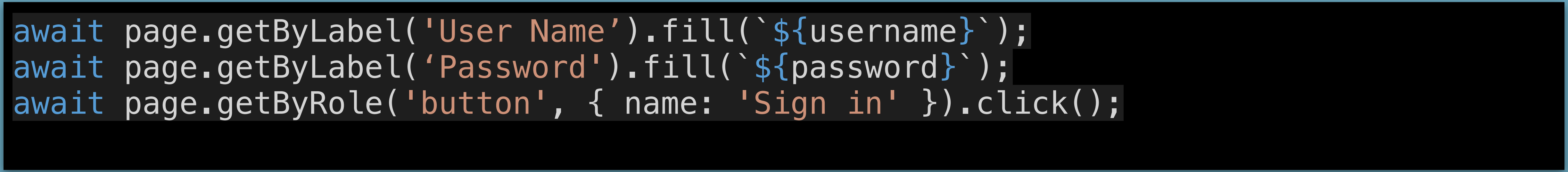
```
▼<div class="login-join-wrap hidden-xs">
  ▼<div id="login-wrap">
    ▼<a href="#" data-toggle="modal" data-target="#myModal">
      ▶<svg id="login-icon" xmlns="http://www.w3.org/2000/svg" viewBox="0 0 12 12">…</svg>
       <span>Log in</span>
     </a>
   </div>
  ▶<div class="join-free-wrap dropdown">…</div>
 </div>
```

```
await page.locator(‘#login-wrap > a‘).click();
```

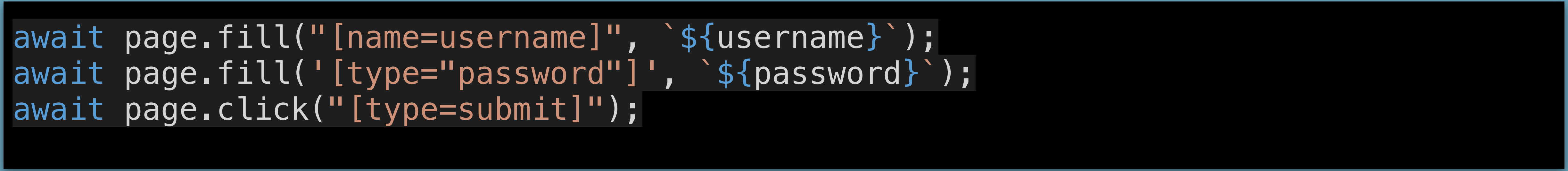
Playwright Locators Example

scrape-kcbs-toy.ts Walkthrough

Playwright Forms



```
await page.getByLabel('User Name’).fill(`${username}`);
await page.getByLabel(‘Password').fill(`${password}`);
await page.getByRole('button', { name: 'Sign in' }).click();
```



```
await page.fill("[name=username]", `${username}`);
await page.fill('[type="password"]', `${password}`);
await page.click("[type=submit]");
```

Playwright Forms Example

scrape-kcbs-events.ts Walkthrough

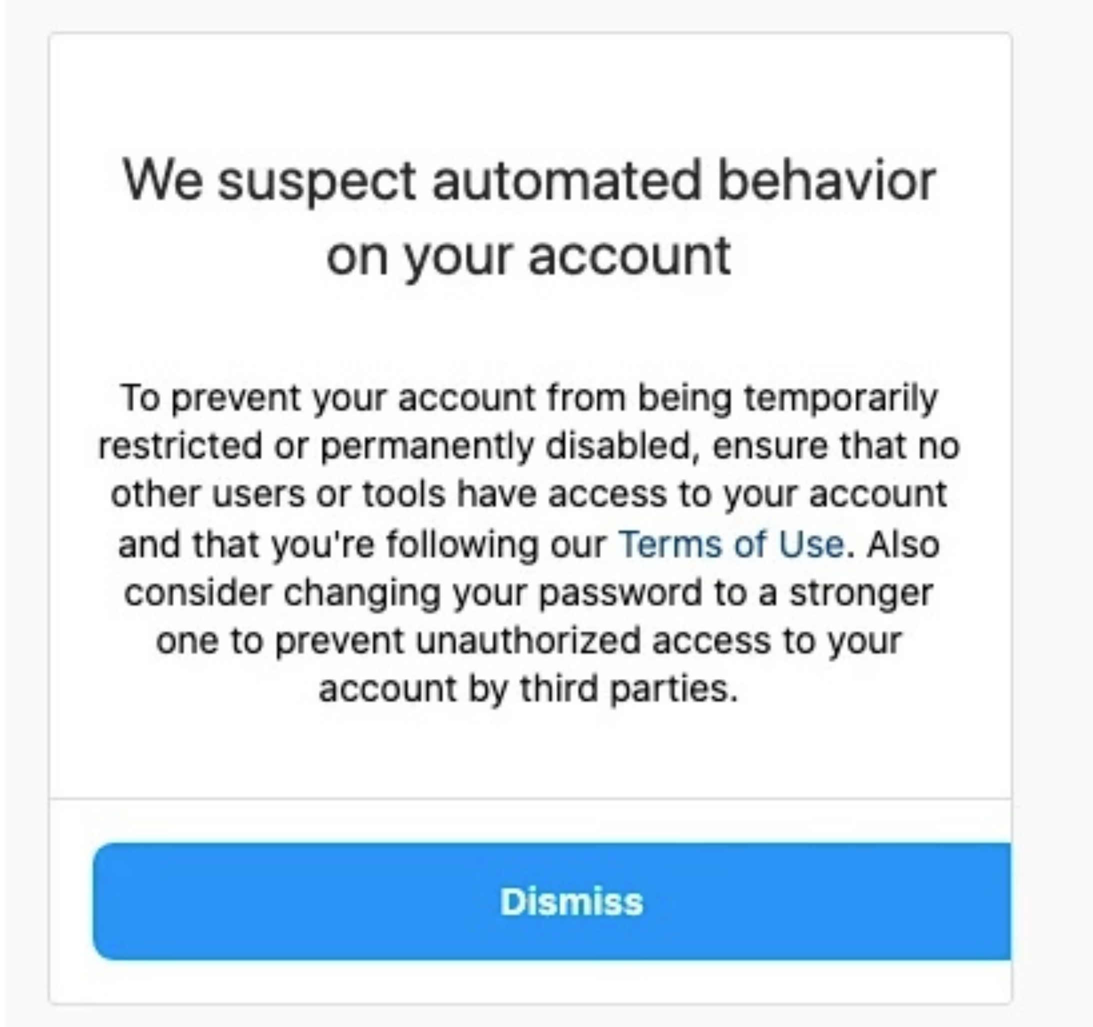
Playwright Scroll Example

```
// Locate with RegEx
const followersLink = page.getByText(/[0-9,.KM]+ followers/i);

await followersLink.click();

// Locate the second dialog
const followersDialog = page.getByRole(‘dialog’).nth(1);

// Locate within another locator
await followersDialog
      .locator(‘div:visible’)
      .last()
      .hover();

// Scroll down (x, y)
await page.mouse.wheel(0, 100);
```

Playwright Interaction Example

scrape-ig-followers.ts Walkthrough

It Really Works!



We suspect automated behavior on your account

To prevent your account from being temporarily restricted or permanently disabled, ensure that no other users or tools have access to your account and that you're following our Terms of Use. Also consider changing your password to a stronger one to prevent unauthorized access to your account by third parties.

Playwright Connect Through CDP

https://playwright.dev/docs/api/class-browsertype#browser-type-connect-over-cdp

- Attaches Playwright to an existing browser instance using the Chrome DevTools Protocol.
- Must start browser with debugger enabled.

Playwright CDP Example

scrape-ig-followers-cdp.ts Walkthrough

HAR Files

“The **HTTP Archive** format, or **HAR**, is a JSON-formatted archive file format for logging of a web browser's interaction with a site. The common extension for these files is **.har**.” - Wikipedia

https://en.wikipedia.org/wiki/HAR_(file_format)

HAR File Example

Capture IG followers interaction.

Questions

https://github.com/MeshSoftware/playwright-web-scraping