Using Selenium to do $fancy$ scrapes

Bots and Scraping

Bots and Scraping

Most modern websites are designed to prevent automated tools from extracting their information.

Additionally, some modern pages employ tools such as infinite scrolling (like social media platforms, etc.)

Selenium

Selenium is a toolkit that can be used both for testing website behavior and for interacting with websites. We can use it to:

- Load a page in browser
- Access information on that page
- Scroll
- Click
- Do other user-like actions

Getting set up

You must have a **local** install (on your computer) of Python to follow this tutorial. If you need to download Python, I recommend using Anaconda Python

In the terminal:

```
pip install selenium
```

On your computer:

- Have Chrome/Chromium installed
- Download the driver from this page

Chrome driver

Stable

Version: `118.0.5993.70` (`r1192594`)

Binary	Platform	URL	HTTP status
chrome	**linux64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/linux64/chrome-linux64.zip	200
chrome	**mac-arm64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/mac-arm64/chrome-mac-arm64.zip	200
chrome	**mac-x64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/mac-x64/chrome-mac-x64.zip	200
chrome	**win32**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/win32/chrome-win32.zip	200
chrome	**win64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/win64/chrome-win64.zip	200
chromedriver	**linux64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/linux64/chromedriver-linux64.zip	200
chromedriver	**mac-arm64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/mac-arm64/chromedriver-mac-arm64.zip	200
chromedriver	**mac-x64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/mac-x64/chromedriver-mac-x64.zip	200
chromedriver	**win32**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/win32/chromedriver-win32.zip	200
chromedriver	**win64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/win64/chromedriver-win64.zip	200
chrome-headless-	**linux64**	https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/118.0.5993.70/linux64/chrome-	200

Chrome driver placement

Put the downloaded file into your script directory (the folder from which your code will be run)

For Windows:

- The file is named `chromedriver.exe`

For Mac/Linux

- The file is named `chromedriver`

Beginning our scrape

We will try scraping some news articles from resources hosted by the UNO library.

YOU MUST USE YOUR UNO ACCOUNT TO LOG IN

Launching Chrome

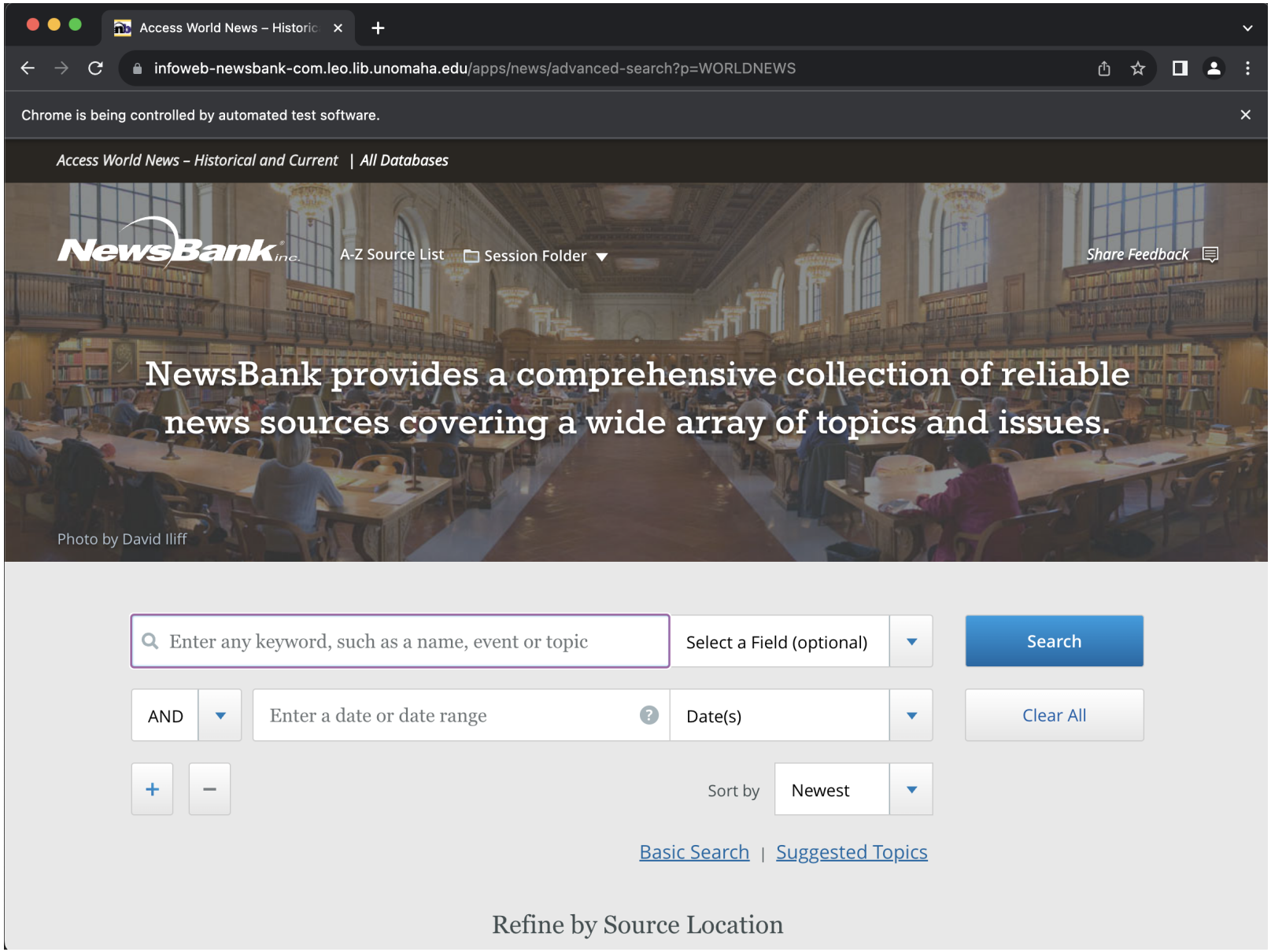
```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By

# Mac/Linux
driver = webdriver.Chrome(executable_path="chromedriver")

# Windows
# driver = webdriver.Chrome(executable_path="chromedriver.exe")
```

When this code has completed you should have an empty web browser open on your computer. It should have a notice that states "Chrome is being controlled by automated test software."

Open a website

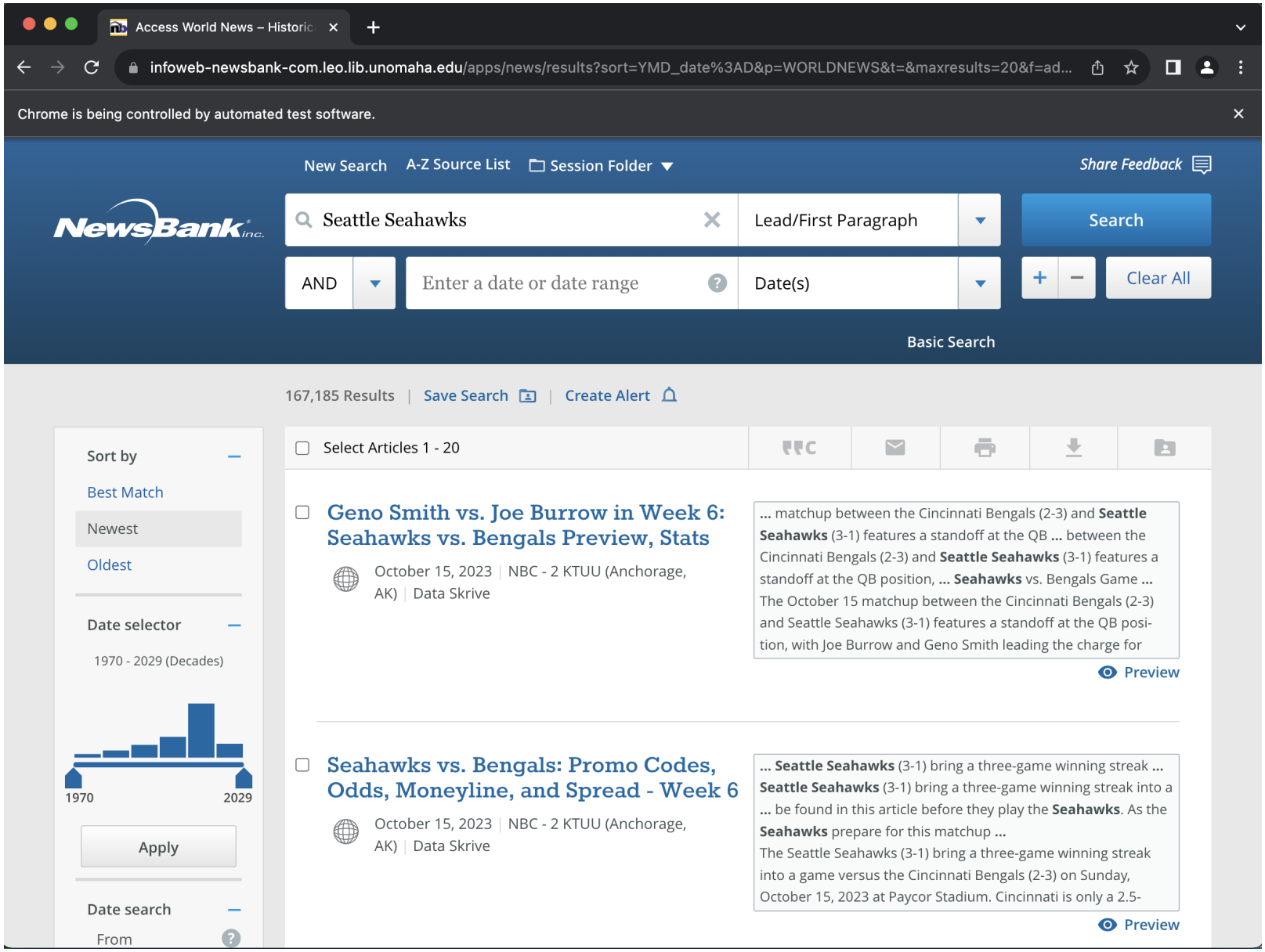
We want to open a news aggregator that we can search through and scrape from. The code below will open the link to the NewsBank database. You will then need to log in with your UNO credentials and two-factor authentication.

```
driver.get("https://infoweb-newsbank-com.leo.lib.unomaha.edu/apps/news/advanced-search?p=WORLDNEWS")
```


Access World News – Historic
infoweb-newsbank-com.leo.lib.unomaha.edu/apps/news/advanced-search?p=WORLDNEWS
Chrome is being controlled by automated test software.
Access World News – Historical and Current | All Databases
NewsBank inc.
A-Z Source List
Session Folder
Share Feedback
NewsBank provides a comprehensive collection of reliable news sources covering a wide array of topics and issues.
Photo by David Iliff
Enter any keyword, such as a name, event or topic
Select a Field (optional)
Search
AND
Enter a date or date range
Date(s)
Clear All
Sort by
Newest
Basic Search | Suggested Topics
Refine by Source Location

Manual search

For our purposes, we can simply initiate a manual search based on our desired criteria. I'll try looking for news articles where the "Seattle Seahawks" are mentioned in the lead.


Access World News – Historic
infoweb-newsbank-com.leo.lib.unomaha.edu/apps/news/results?sort=YMD_date%3AD&p=WORLDNEWS&t=&maxresults=20&f=ad...
Chrome is being controlled by automated test software.
New Search
A-Z Source List
Session Folder
Share Feedback
NewsBank inc.
Seattle Seahawks
Lead/First Paragraph
Search
AND
Enter a date or date range
Date(s)
Clear All
Basic Search
167,185 Results | Save Search | Create Alert
Sort by
Best Match
Newest
Oldest
Date selector
1970 - 2029 (Decades)
1970
2029
Apply
Date search
From
Select Articles 1 - 20
Geno Smith vs. Joe Burrow in Week 6: Seahawks vs. Bengals Preview, Stats
October 15, 2023 | NBC - 2 KTUU (Anchorage, AK) | Data Skrive
... matchup between the Cincinnati Bengals (2-3) and Seattle Seahawks (3-1) features a standoff at the QB ... between the Cincinnati Bengals (2-3) and Seattle Seahawks (3-1) features a standoff at the QB position, ... Seahawks vs. Bengals Game ... The October 15 matchup between the Cincinnati Bengals (2-3) and Seattle Seahawks (3-1) features a standoff at the QB position, with Joe Burrow and Geno Smith leading the charge for
Preview
Seahawks vs. Bengals: Promo Codes, Odds, Moneyline, and Spread - Week 6
October 15, 2023 | NBC - 2 KTUU (Anchorage, AK) | Data Skrive
... Seattle Seahawks (3-1) bring a three-game winning streak ... Seattle Seahawks (3-1) bring a three-game winning streak into a ... be found in this article before they play the Seahawks. As the Seahawks prepare for this matchup ...
The Seattle Seahawks (3-1) bring a three-game winning streak into a game versus the Cincinnati Bengals (2-3) on Sunday, October 15, 2023 at Paycor Stadium. Cincinnati is only a 2.5-
Preview

Finding the article links

Next, I want to find all of the articles on the page:

```
# Identify the <body> of the html
body = driver.find_element(By.TAG_NAME, "body")

# Identify <article> tags
art_tags = body.find_elements(By.TAG_NAME,'article')

# Search for the <h3> tags containing the links
#   and then extract the href attributes
links = [i.find_element(
    By.TAG_NAME, "h3").find_element(
        By.TAG_NAME, 'a').get_attribute("href") for i in art_tags]

# Find the link to the next page to keep searching
next_page = driver.find_element(
    By.CLASS_NAME, "pager__item--next").find_element(
        By.TAG_NAME, 'a').get_attribute("href")
```

Pull article text

Next, I want to follow each link on that page to extract the article text:

```
# Container for the data
articles = []

# Loop over links
for i in links:
    # Fetch the page
    driver.get(i)

    # Make a single string containing all paragraphs of text
    single = " ".join(
        [i.text for i in driver.find_element(
            By.CLASS_NAME, "document-view__body").find_elements(
                By.TAG_NAME, 'p')])

    # Stick it into the container list
    articles.append(single)
```

To Be Continued?

If I wanted to keep going, I could write code very similar to our Lego scraping code to advance to the next page of results, and extract those articles as well!

In the end, we have our own custom text data set that we can use in our research!

When finished

We can close down our browser with a single line of code:

```
driver.close()
```

How do we feel?