

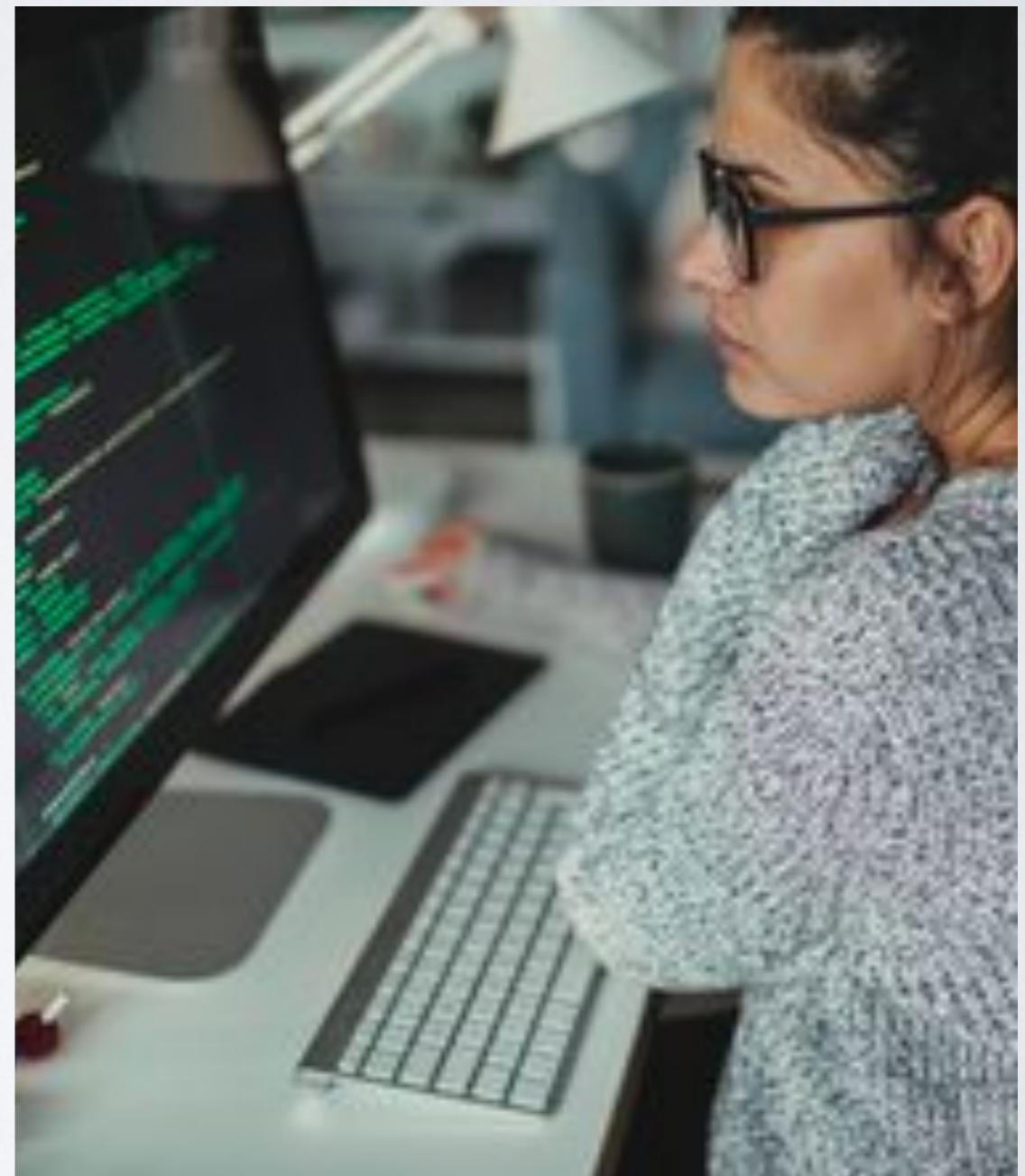
INTRODUCTION TO **SECURE WEB SCRAPING** WITH PYTHON



Lecturer: Jason Townes French

COURSE OVERVIEW

- A hands-on overview to writing scripts to extract data from websites
- An exploration of challenges and concerns regarding web scraping



WHAT YOU'LL NEED

- A basic understanding of Python
- A computer running Python 2.7
- A basic understanding of Git and GitHub.com
- Chrome
- MySQL Workbench

WHO AM I?

- Jason Townes French
- Brown University, Computer Science
- ex-Apple Software Engineer
- Entrepreneur from Silicon Valley
- Game developer



Raytheon
Space and Airborne Systems



XPERT

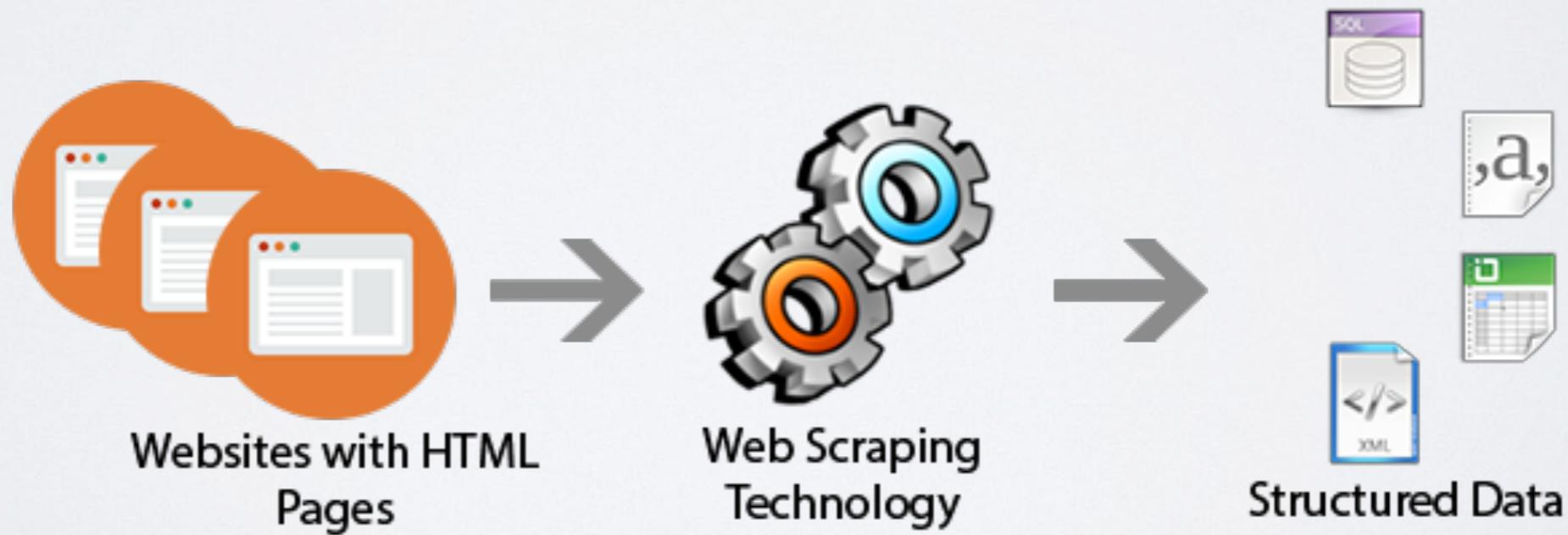


WHAT IS “SCRAPING”?



WHAT IS WEB SCRAPING?

- Extracting data from a website
- Storing extracted data in structured format for further processing

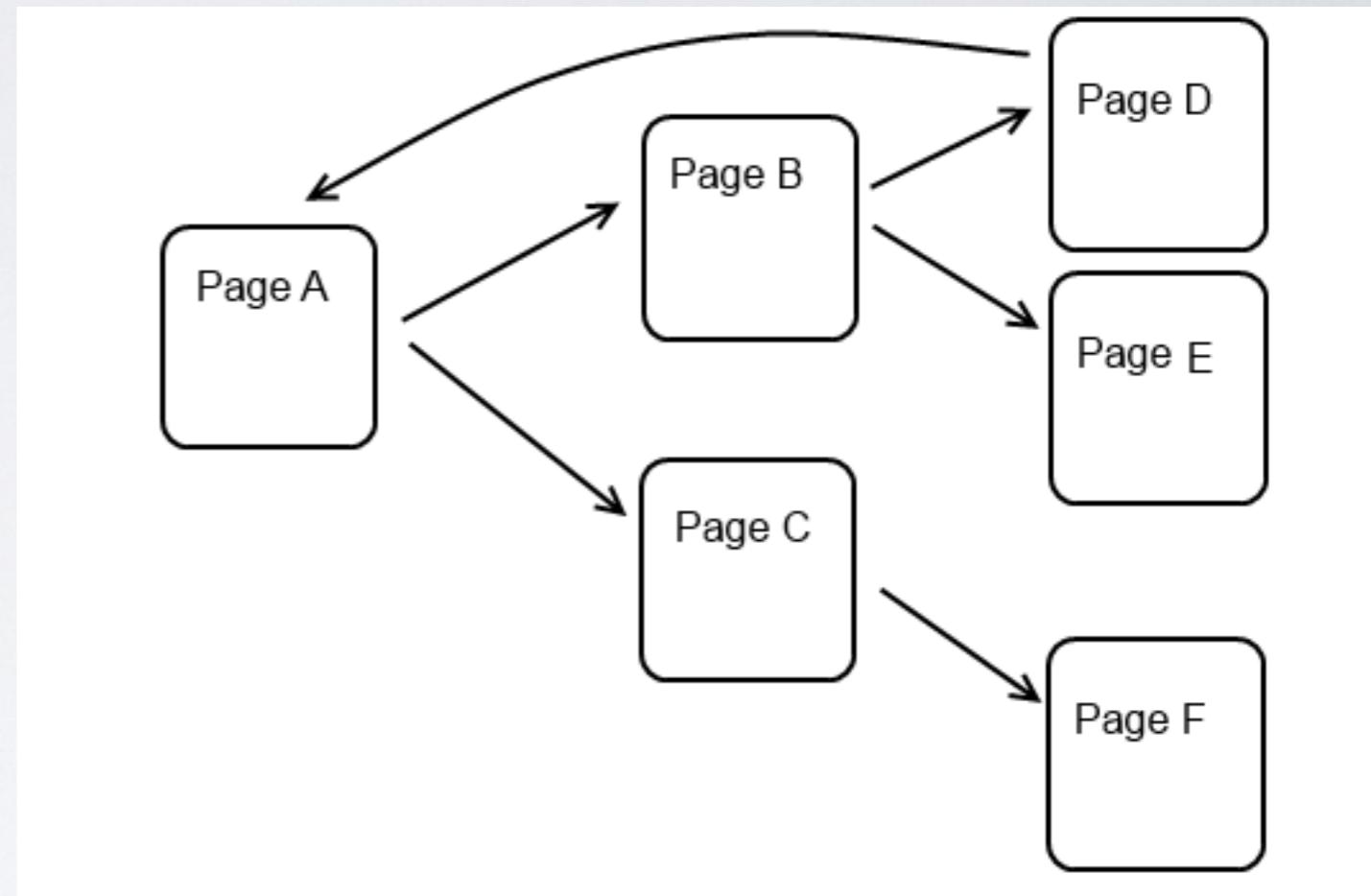


SCRAPING VS. CRAWLING



SCRAPING VS. CRAWLING

- Crawlers start at a page, visiting & indexing each link on the page, repeating until there are no more links to traverse
- Scrapers visit a specific page, extracting specific data for further processing



SCRAPING VS. CRAWLING

Google = CRAWLER

COMMON SCRAPING PROBLEMS

- Terms of Service
- robots.txt
- Interaction Throttling
- Tracking

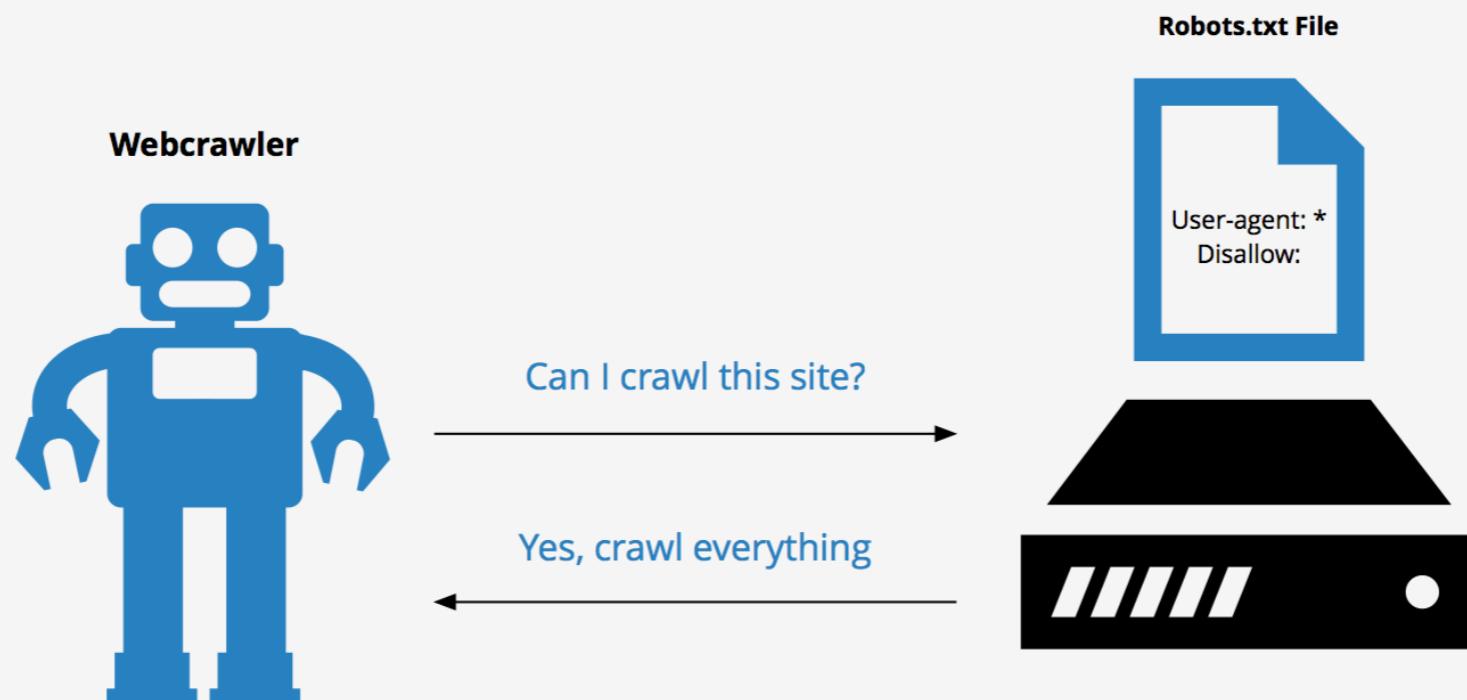
COMMON SCRAPING PROBLEMS

1. **Description of Service:** The Service provides consumer content regarding insurance products and services, for informational purposes only. The Service does not sell any type of insurance and is not an insurer or insurance broker, nor does it recommend, support, or endorse any particular insurance plan. The Service enables you to request to receive insurance or discount program quotes from a network of insurance companies, agents, brokers, discount program representatives and other providers (the "Insurance Representatives"). Through the Service, you choose to provide information about yourself and your insurance preferences ("User Information") which is in turn used to attempt to match you with Insurance Representatives who may be able to follow-up on your request. If you use the Service, we cannot guarantee that any of the Insurance Representatives to whom we forward your information will contact you or agree to offer you coverage. We also cannot guarantee the carrier affiliation of any Insurance Representatives who may contact you. We have no responsibility whatsoever for the conduct of any of the Insurance Representatives with whom you may communicate.
2. **User Obligations:** You shall not (a) use the Service in a manner that, while using the Service, violates the laws of any person or entity; (b) submit false or misleading information; (c) exploit the Service for any commercial purpose; (d) engage in spamming or other similar activity; (e) restrict or inhibit others from using the Service; (f) use an application that would interfere with our systems or negatively impact them; (g) load a virus or other harmful code onto our systems or networks; (h) use the Service to send unsolicited emails through the Service; (i) take any other actions that would damage, overburden, interfere with, or otherwise harm the Service or the use of the Service by any other party. You acknowledge and agree that you are solely responsible for the accuracy of the User Information you submit. We may delete or destroy User Information at any time and we reserve the right to refuse to post or to remove any User Information, in whole or in part, that we, in our sole discretion, believe to be inappropriate.
3. **Third Party Information and Trademarks:** All content provided on this site about third parties (including companies and brokers) is provided for informational purposes only. Such content is not an endorsement of or a recommendation for any third party. It does not imply, directly or indirectly, any sponsorship or affiliation with such third parties, and no guarantees regarding the same are made herein. All third party trademarks are the property of their respective

Terms of Services

Did you ever read them?

COMMON SCRAPING PROBLEMS



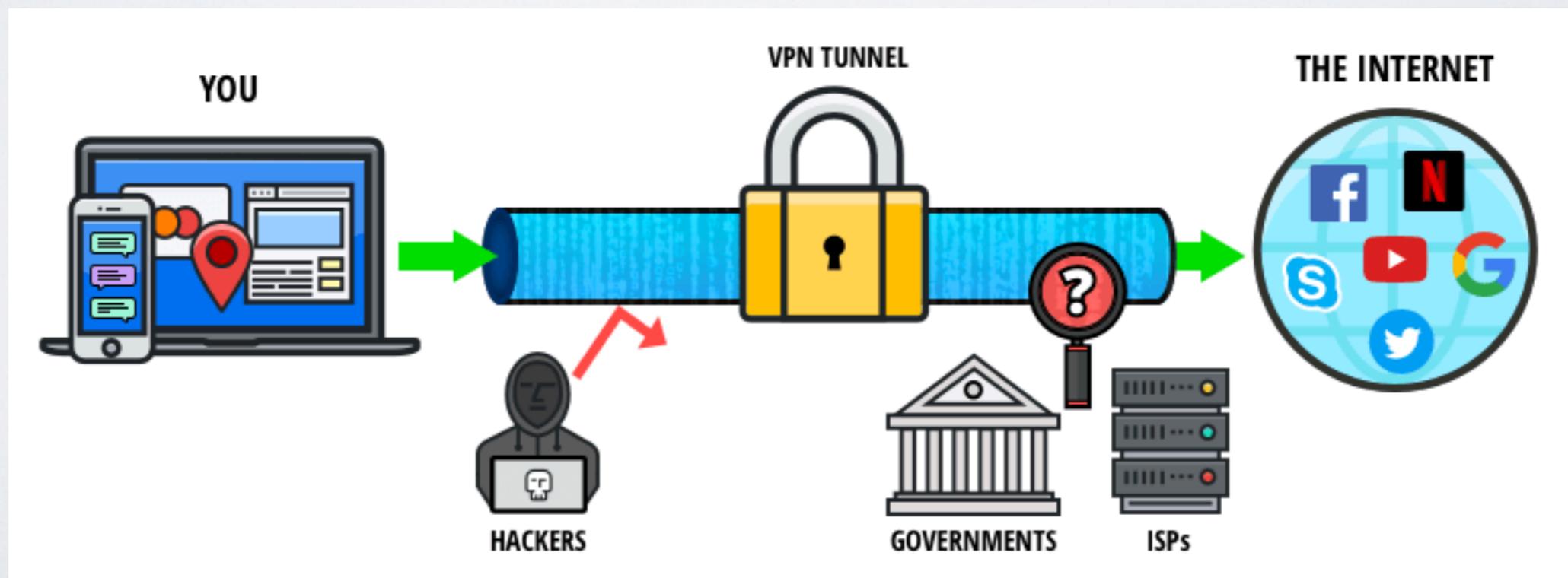
What Is a Robots.txt File

COMMON SCRAPING PROBLEMS



COMMON SCRAPING PROBLEMS

||| DONT TRACK US |||



SECTION OVERVIEW: WHAT IS SCRAPING?

- Extracting structured data from websites
- Different from crawling
- Watch out for Terms of Service, robots.txt, and speed throttling
- Best to use a VPN

INTRODUCTION TO SELENIUM



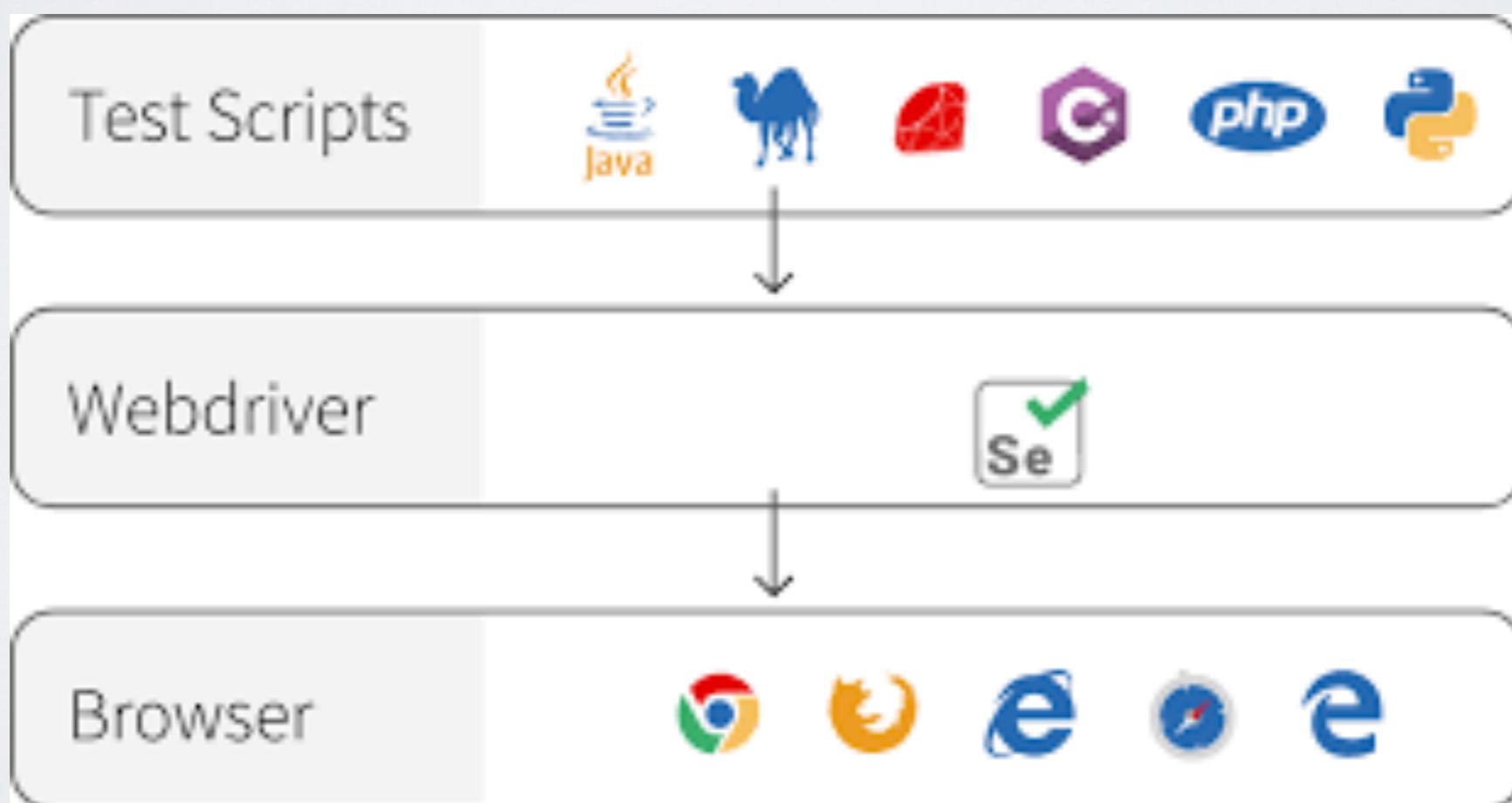
WHAT IS SELENIUM?

- A tool for automating interaction with a web browser
- Useful for many types of testing but also scraping



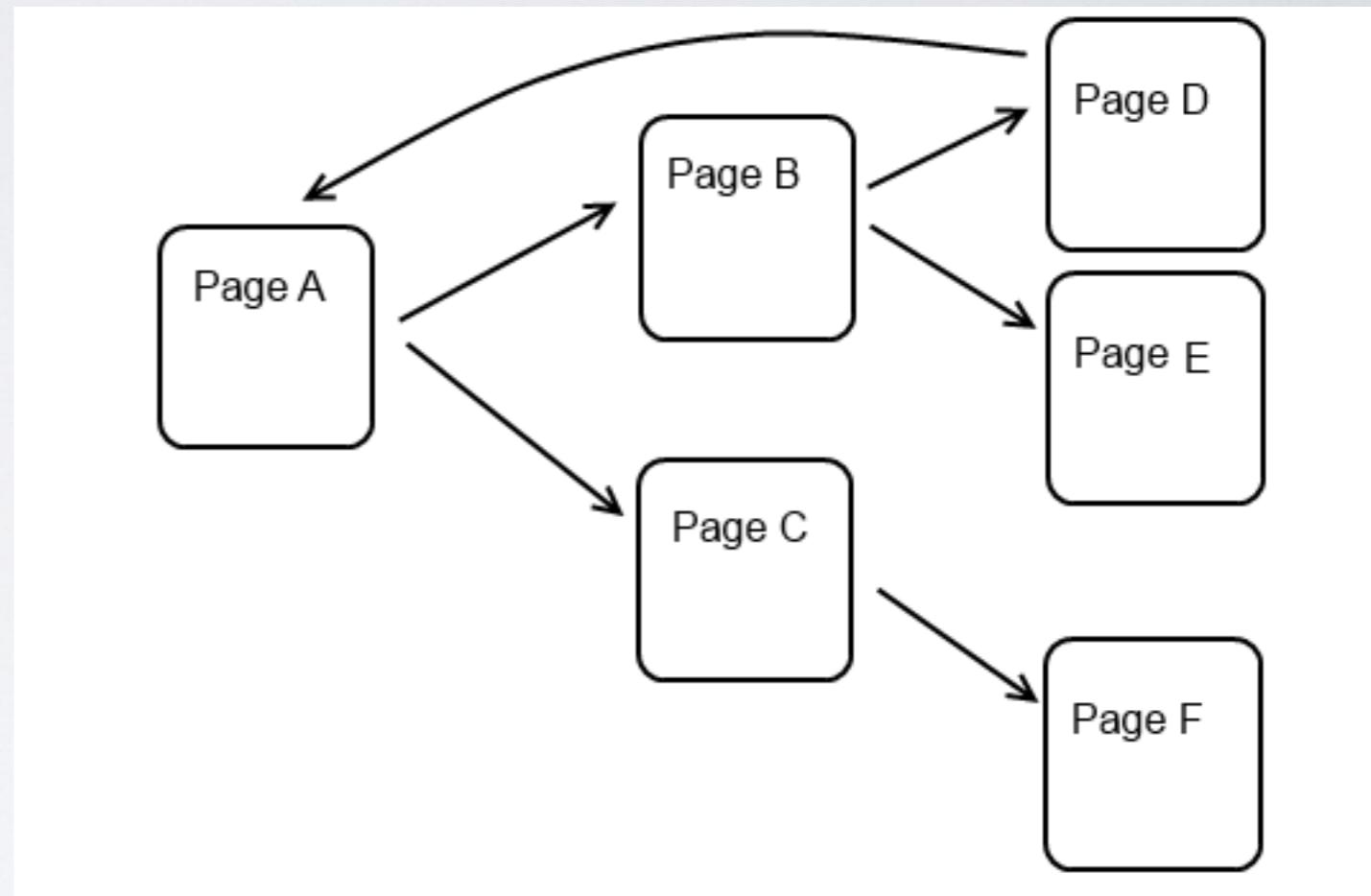
HOW IT WORKS

- You write scripts which talk to the Web Driver
- The Web Driver talks to the Browser
- The Browser does what your script said



CORE FUNCTIONALITY

- Crawlers start at a page, visiting & indexing each link on the page, repeating until there are no more links to traverse
- Scrapers visit a specific page, extracting specific data for further processing



EXERCISE I:

INSTALL ENVIRONMENT

- Checkout git repository:

```
git clone https://github.com/DigitalSolomon/ftw-secure-web-scraping.git
```

- Install Python Virtual Environment:

```
sudo pip install virtualenv
```

- Create virtual environment:

```
virtualenv env
```

- Activate virtual environment

```
source env/bin/activate
```

CYGWIN

← → C https://www.cygwin.com ☆ b 🔍 1 🔥 🎯

Apps EDUKASYON Jira Staging | Internshi... | Ot

Cygwin

- Install Cygwin
- Update Cygwin
- Search Packages
- Licensing Terms

Cygwin/X

Community

- Reporting Problems
- Mailing Lists
- Newsgroups
- IRC channels
- Gold Stars
- Mirror Sites
- Donations

Documentation

- FAQ
- User's Guide
- API Reference
- Acronyms

Contributing

- Snapshots
- Source in Git
- Cygwin Packages

Related Sites

Cygwin

Get that [Linux](#) feeling - on Windows

This is the home of the Cygwin project

What...

...is it?

Cygwin is:

- a large collection of GNU and Open Source tools which provide functionality similar to a [Linux distribution](#) on Windows.
- a DLL (cygwin1.dll) which provides substantial POSIX API functionality.

...isn't it?

Cygwin is not:

- a way to run native Linux apps on Windows. You must rebuild your application *from source* if you want it to run on Windows.
- a way to magically make native Windows apps aware of UNIX® functionality like signals, ptys, etc. Again, you need to build your apps *from source* if you want to take advantage of Cygwin functionality.

The Cygwin DLL currently works with all recent, commercially released x86 32 bit and 64 bit versions of Windows, starting with Windows Vista. For more information see the [FAQ](#).

Cygwin version

The most recent version of the Cygwin DLL is [3.0.7](#).

Installing Cygwin

Install Cygwin by running [setup-x86_64.exe](#) (64-bit installation) or [setup-x86.exe](#) (32-bit installation)

Use the setup program to perform a [fresh install](#) or to [update](#) an existing installation.

Keep in mind that individual packages in the distribution are updated separately from the DLL so the Cygwin DLL version is not useful as a general Cygwin distribution release number.

EXERCISE 2:

INSTALL SELENIUM

- Install Selenium

```
pip install selenium
```

- Download Web Driver

```
https://sites.google.com/a/chromium.org/chromedriver/downloads
```

- Write first script
 - Fill out exercise2.py

EXERCISE2.PY

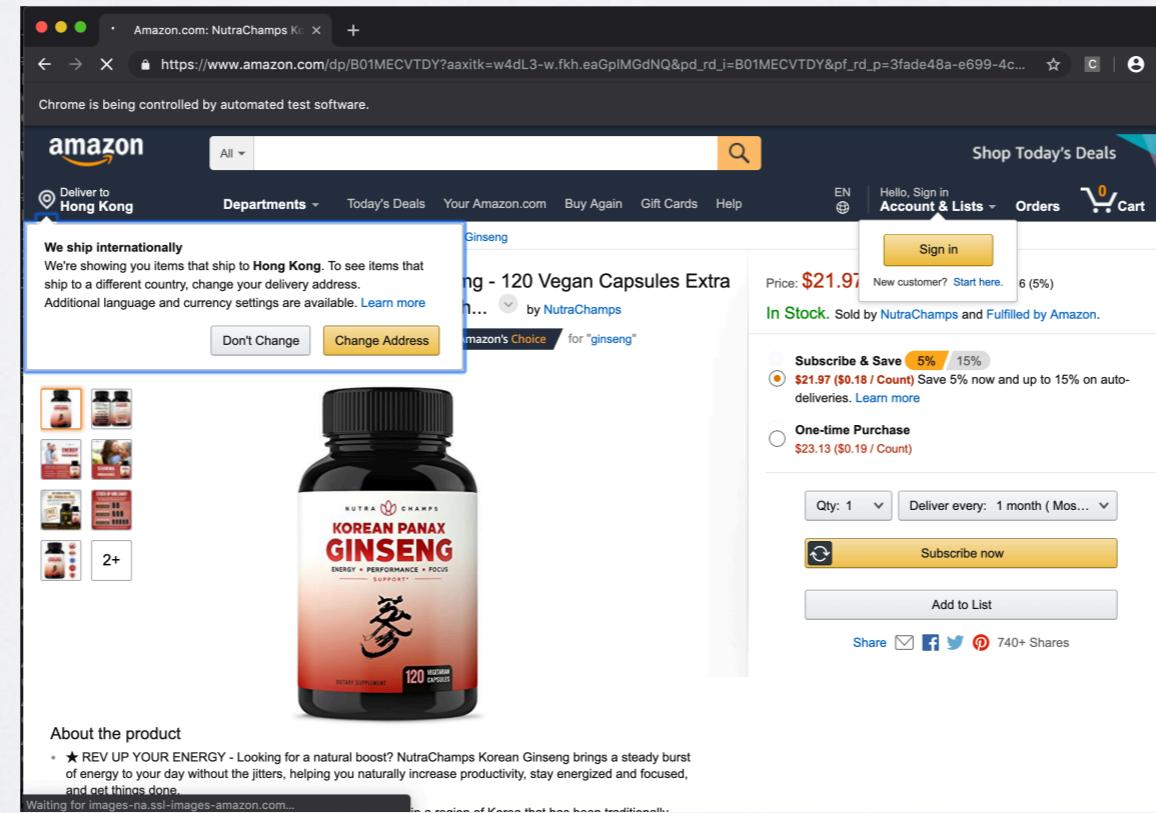
```
1 from selenium import webdriver  
2 import os  
3  
4  
5 dirpath = os.getcwd()  
6 filepath = dirpath + '/chromedriver'  
7 print('Path to Driver: ' + filepath)  
8 browser = webdriver.Chrome(executable_path = filepath)  
9 # browser.get('...')  
10
```

EXERCISE 3: CLICK REVIEWS

- exercise3.py
- Access Amazon Product Listing

```
browser.get('AmazonListingURLHere')
```

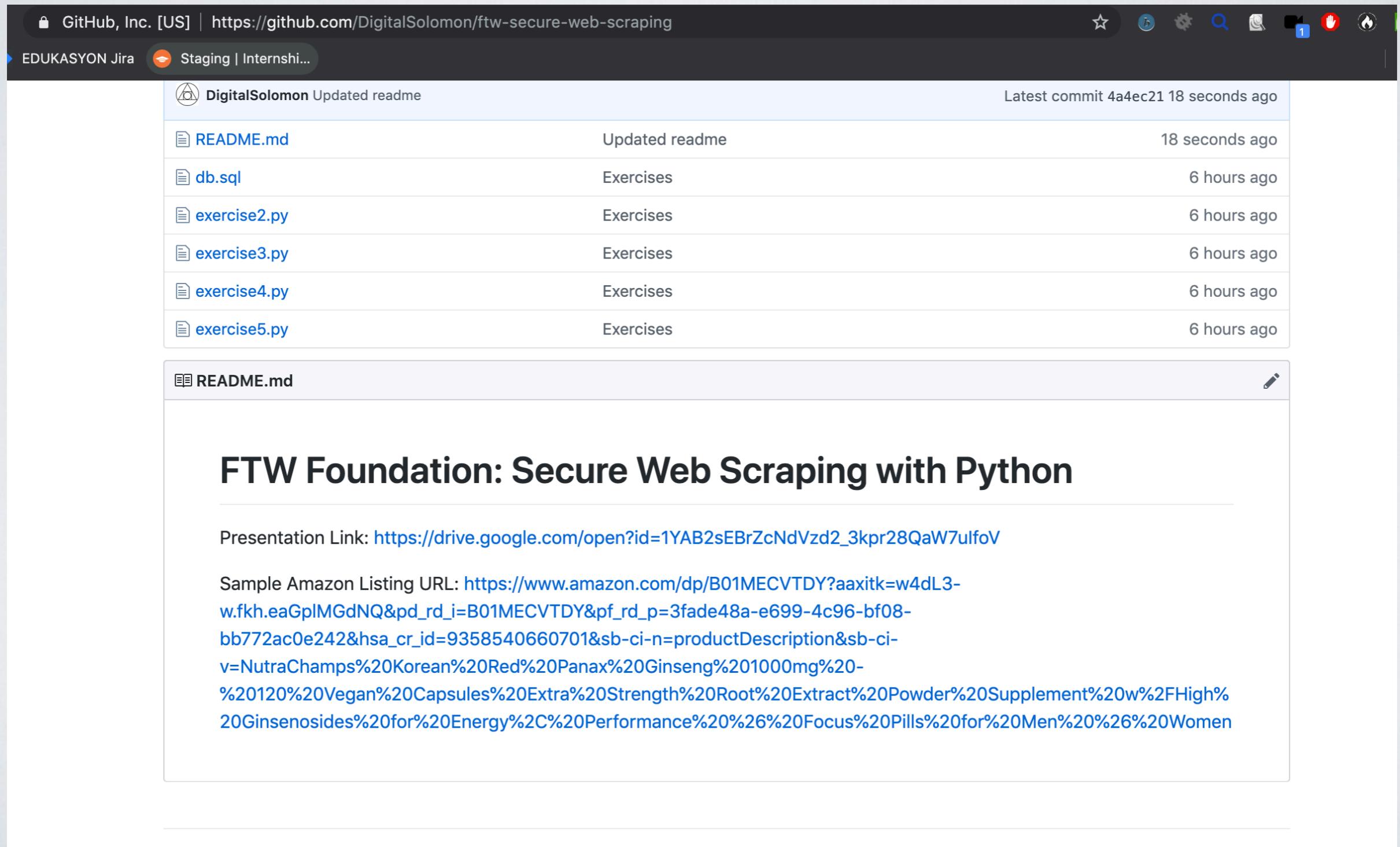
- Close Popup & Click on Reviews



EXERCISE3.PY

```
1  from selenium import webdriver
2  from selenium.webdriver.support.ui import WebDriverWait
3  from selenium.webdriver.support import expected_conditions as EC
4  from selenium.webdriver.common.by import By
5  from selenium.common.exceptions import TimeoutException
6  import os
7
8  dirpath = os.getcwd()
9  filepath = dirpath + '/chromedriver'
10 print('Path to Driver: ' + filepath)
11 browser = webdriver.Chrome(executable_path = filepath)
12 browser.get('AmazonListingURLHere')
13
14 try:
15     # Dismiss initial popup
16     element = WebDriverWait(browser,5).until(
17         EC.presence_of_element_located((By.XPATH, 'FillThisIn')))
18
19     element.click()
20
21     # Click on Reviews
22     element = WebDriverWait(browser,5).until(
23         EC.presence_of_element_located((By.XPATH, 'FillThisIn')))
24
25     element.click()
26
27
28 except TimeoutException:
29     print("Failed to load element")
30 finally:
31     browser.quit()
```

SAMPLE AMAZON URL



The screenshot shows a GitHub repository page for `DigitalSolomon/ftw-secure-web-scraping`. The repository contains several files:

File	Description	Last Commit
<code>README.md</code>	Updated readme	18 seconds ago
<code>db.sql</code>	Exercises	6 hours ago
<code>exercise2.py</code>	Exercises	6 hours ago
<code>exercise3.py</code>	Exercises	6 hours ago
<code>exercise4.py</code>	Exercises	6 hours ago
<code>exercise5.py</code>	Exercises	6 hours ago

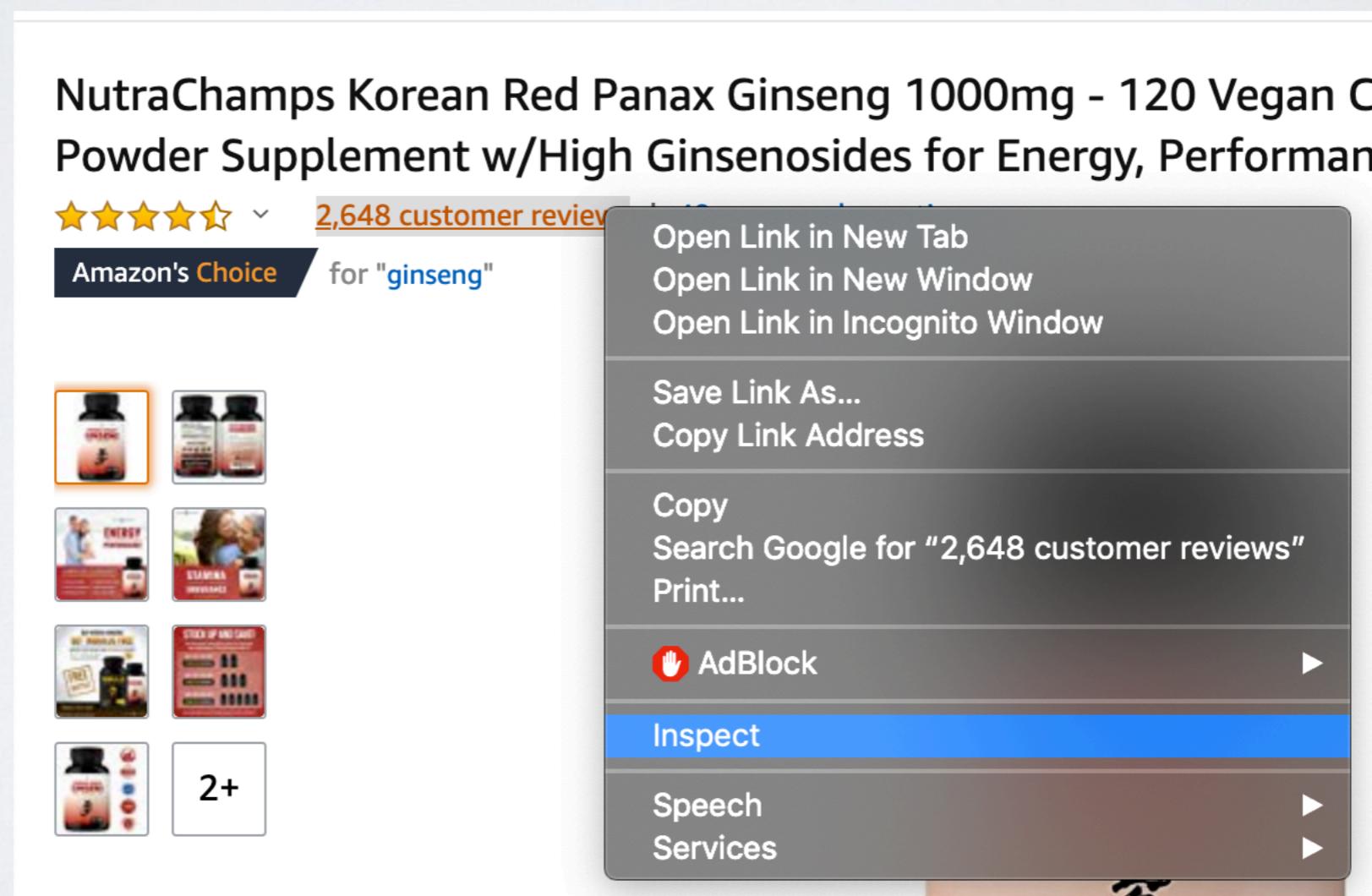
A modal window titled "FTW Foundation: Secure Web Scraping with Python" is open, containing the following information:

Presentation Link: https://drive.google.com/open?id=1YAB2sEBrZcNdVzd2_3kpr28QaW7ulfoV

Sample Amazon Listing URL: https://www.amazon.com/dp/B01MECVTDY?aaxitk=w4dL3-w.fkh.eaGpIMGdNQ&pd_rd_i=B01MECVTDY&pf_rd_p=3fade48a-e699-4c96-bf08-bb772ac0e242&hsa_cr_id=9358540660701&sb-ci-n=productDescription&sb-ci-v=NutraChamps%20Korean%20Red%20Panax%20Ginseng%201000mg%20-%20120%20Vegan%20Capsules%20Extra%20Strength%20Root%20Extract%20Powder%20Supplement%20w%2FHigh%20Ginsenosides%20for%20Energy%2C%20Performance%20%26%20Focus%20Pills%20for%20Men%20%26%20Women

EXERCISE 3: CLICK REVIEWS

- Use “Inspect Element” to copy the XPath



EXERCISE 3: CLICK REVIEWS

- Use “Inspect Element” to copy the XPath



```
> <script type="text/javascript">...</script>
> <style type="text/css">...</style>
> <div id="ppd-top" class="a-section a-spacing-none burj">...</div>
> <div id="actionPanelContainer" class="a-section burj">...</div>
> <div id="leftCol" class="a-section burj">
>   <div id="companyCompliancePolicies_feature_div" class="feature" data-feature-name="companyCompliancePolicies_feature_div">...</div>
>   <div id="instantOrderUpdate_feature_div" class="feature" data-feature-name="instantOrderUpdate_feature_div">...</div>
>   <div id="title_feature_div" class="feature" data-feature-name="title" data-cel-widget="title">...</div>
>   <div id="averageCustomerReviews_feature_div" class="feature" data-feature-name="averageCustomerReviews_feature_div">
>     <style type="text/css">...</style>
>     <div id="averageCustomerReviews" class="a-spacing-none" data-asin="B01MECVTDY" data-ref="...>
>       <span class="a-declarative" data-action="acrStarsLink-click-metrics" data-acrstarslink="...>
>         <span class="a-letter-space"></span>
>       <span class="a-declarative" data-action="acrLink-click-metrics" data-acrlink-click-metrics="...>
>         <a id="acrCustomerReviewLink" href="#" data-customerReviews="#customerReviews" data-customerReviewsCount="2648" data-customerReviewsLabel="2,648 customer reviews">...</a>
>       <span id="acrCustomerReviewLink" data-customerReviews="2648" data-customerReviewsLabel="2,648 customer reviews">...</span>
>     </div>
>     <script type="text/javascript">...</script>
>     <script type="text/javascript">...</script>
>   </div>
> </div>
> <div id="ask_feature_div" class="a-section a-spacing-none burj">...</div>
> <div id="primeExclusiveBadge_feature_div" class="a-section a-spacing-none burj">...</div>
> <div id="acBadge_feature_div" class="a-section a-spacing-none burj">...</div>
> <div id="zeitgeistBadge_feature_div" class="a-section a-spacing-none burj">...</div>
> <div id="socialFabric_feature_div" class="a-section a-spacing-none burj">...</div>
> <div id="scenes_stage" class="a-section a-spacing-none burj">
>   <div id="hover-zoom-end" style="display: none;">...</div>
> </div>
> <div id="newerVersion_feature_div" class="a-section a-spacing-none burj">...</div>
> <div id="featurebullets_feature_div" class="a-section a-spacing-none burj">...</div>
```

Copy

Add attribute
Edit attribute
Edit as HTML
Delete element

Cut element
Copy element
Paste element

Copy outerHTML
Copy selector
Copy JS path
Copy XPath

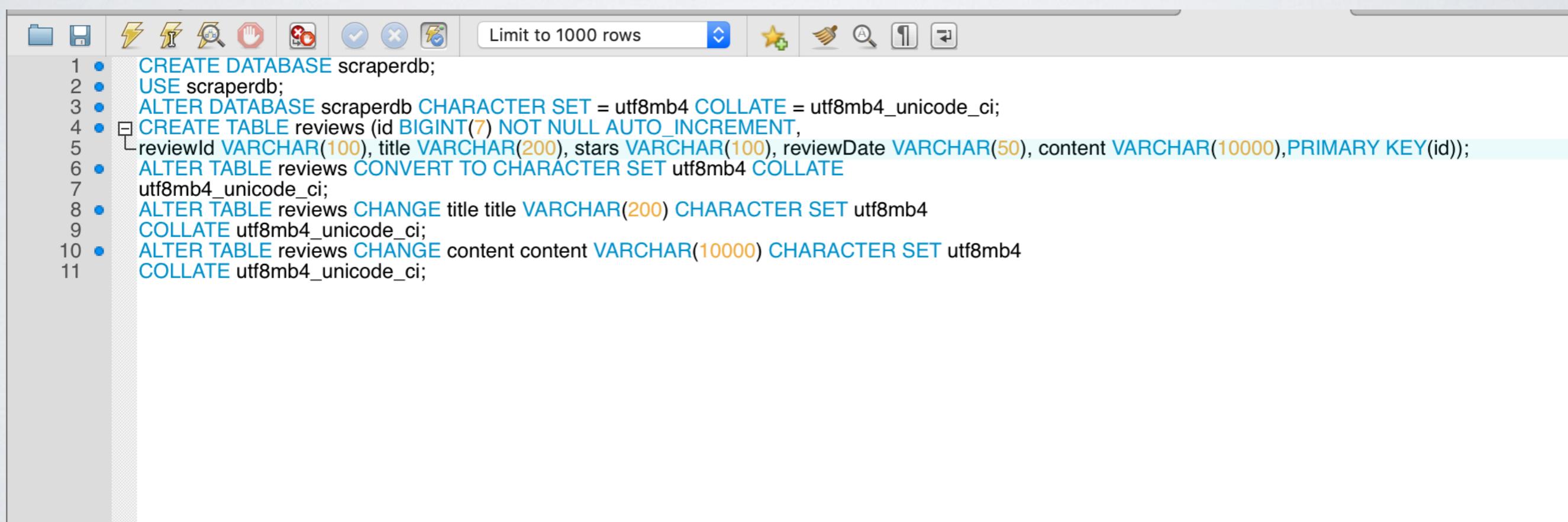
EXERCISE 4: PROCESS REVIEWS

- Open exercise4.py
- Find the element which contains all of the reviews
- Find the class name of each individual review
- Hint: Use the Inspector!

EXERCISE 5:

STORE IN DATABASE

- Copy/paste db.sql into MySQL Workbench and execute it to create your database and tables



The screenshot shows the MySQL Workbench interface with the SQL editor open. The code in the editor is as follows:

```
1 • CREATE DATABASE scraperdb;
2 • USE scraperdb;
3 • ALTER DATABASE scraperdb CHARACTER SET = utf8mb4 COLLATE = utf8mb4_unicode_ci;
4 • CREATE TABLE reviews (id BIGINT(7) NOT NULL AUTO_INCREMENT,
5     reviewId VARCHAR(100), title VARCHAR(200), stars VARCHAR(100), reviewDate VARCHAR(50), content VARCHAR(10000), PRIMARY KEY(id));
6 • ALTER TABLE reviews CONVERT TO CHARACTER SET utf8mb4 COLLATE
7     utf8mb4_unicode_ci;
8 • ALTER TABLE reviews CHANGE title title VARCHAR(200) CHARACTER SET utf8mb4
9     COLLATE utf8mb4_unicode_ci;
10 • ALTER TABLE reviews CHANGE content content VARCHAR(10000) CHARACTER SET utf8mb4
    COLLATE utf8mb4_unicode_ci;
```

EXERCISE 5:

STORE IN DATABASE

- Install PyMySQL

```
pip install PyMySQL
```

- Import PyMySQL in Python script

```
import pymysql
```

- Fill out Database Connection Details

```
pymysql.connect(host='127.0.0.1', user='root', passwd = 'YourPasswordHere', db = 'mysql',  
charset = 'utf8')
```

- Complete Database Insert Function

```
cur.execute('INSERT INTO reviews (reviewId, title, stars, reviewDate, content)  
VALUES ' + ("'%s','%s','%s','%s','%s') ', (id, title, stars, date, content))
```

- Call your Database Insert Function

```
store(reviewId, date, stars, title, reviewContent)
```