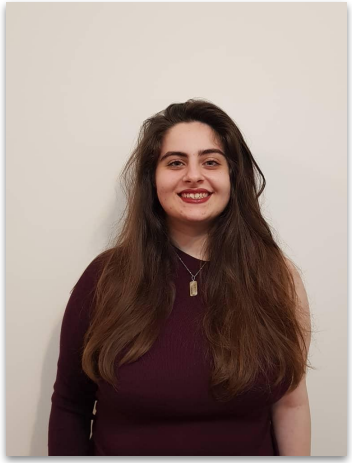
CODING SQUAD

An Introduction to Web Scraping(½)

Seminar

CODING SQUAD

Meet your host!



Negar

Google Developers Student Clubs
(GDSC) Mentorship **Management Team**

Machine Learning Engineer & Python Developer

CODING SQUAD

Agenda

- What is Web Scraping?
- What is XML?
- What are XPATHs?
- Selenium
- Exploring Selenium(in Jupyter)

CODING SQUAD

An Introduction to Web Scraping

Web scraping is a method of extracting data from websites.
It involves using a program or algorithm to download and parse data from the web, typically in HTML or XML format, so that it can be accessed and analyzed.
Web scraping is commonly used to extract data from online sources, such as product information, prices, and reviews, to build databases or to perform data analysis and research.

CODING SQUAD

Tools for Web Scraping

- Beautiful Soup: a Python library that is used for pulling data out of HTML and XML files.
- Scrapy: a powerful Python framework for building web scrapers.
- Selenium: a tool that is used for automating web browsers, which can be used for web scraping.
- ParseHub: a desktop tool that is used for extracting data from websites.
- Import.io: a cloud-based tool that is used for turning web pages into structured data.
- Webhose.io: a platform that provides access to a large amount of web data that can be used for web scraping.

CODING SQUAD

What is XML?

XML (Extensible Markup Language) is a markup language that is used to store and transport data.

XML is based on the idea of tags, which are used to mark the beginning and end of an element, and to identify the type of information that is contained within that element.

XML is often used in conjunction with other technologies, such as XSLT (Extensible Stylesheet Language Transformations) and XPath, to manipulate and transform data.

CODING SQUAD

What is XML?

```xml
<?xml version="1.0" encoding="UTF-8"?>
<library>
  <book genre="fantasy">
    <title>The Lord of the Rings</title>
    <author>J. R. R. Tolkien</author>
    <year>1954</year>
  </book>
  <book genre="sci-fi">
    <title>Dune</title>
    <author>Frank Herbert</author>
    <year>1965</year>
  </book>
</library>
```

CODING SQUAD

What is XPATH?

XPath is a language that is used for navigating through elements and attributes in an XML document.

It is used to find specific elements or attributes in an XML document based on their element name, attribute name, or attribute value.

Some common uses of XPath include searching for specific data in an XML document, extracting data from an XML document, and modifying or deleting data in an XML document.

CODING SQUAD

What is XPATH?

Here is an example of an XPath expression:

```
//book[@genre="fantasy"]/title
```

Selenium

Selenium is a tool that is used for automating web browsers. **It allows users to write scripts and programs that can interact with web pages in the same way that a user would, including clicking buttons, entering text, and navigating to different pages.** Selenium is often used for web scraping, testing web applications, and automating repetitive tasks on the web.

Selenium is a powerful tool that can be used with many different programming languages, including **Python, Java, and C#**. It provides a rich and flexible API that allows users to control web browsers in a variety of ways, including simulating user interactions, extracting data from web pages, and **performing actions on multiple web pages at once**.

CODING SQUAD

Selenium

```
pip install selenium
```

Selenium: WebDrivers

A web driver is a piece of software that provides a bridge between a web browser and an application or tool that is used to automate the web browser. Web drivers are typically used in conjunction with tools such as Selenium, which provide a programmatic interface for controlling the web browser and interacting with web pages.

CODING SQUAD

Selenium: WebDrivers

Headless browsers are web browsers that do not have a graphical user interface (GUI). They are typically used for automated testing or web scraping, where the focus is on the **underlying functionality of the web browser, rather than on the user experience.**

Headless browsers are typically run from the command line, and they allow users to perform actions on web pages, such as **navigating to a URL, clicking buttons, and filling out forms, without the need for a graphical user interface.** This makes them well-suited for automated testing or web scraping tasks, where the goal is to simulate user interactions with a web page in a consistent and repeatable way.

CODING SQUAD

Example of a Headless WebDriver

Copy code

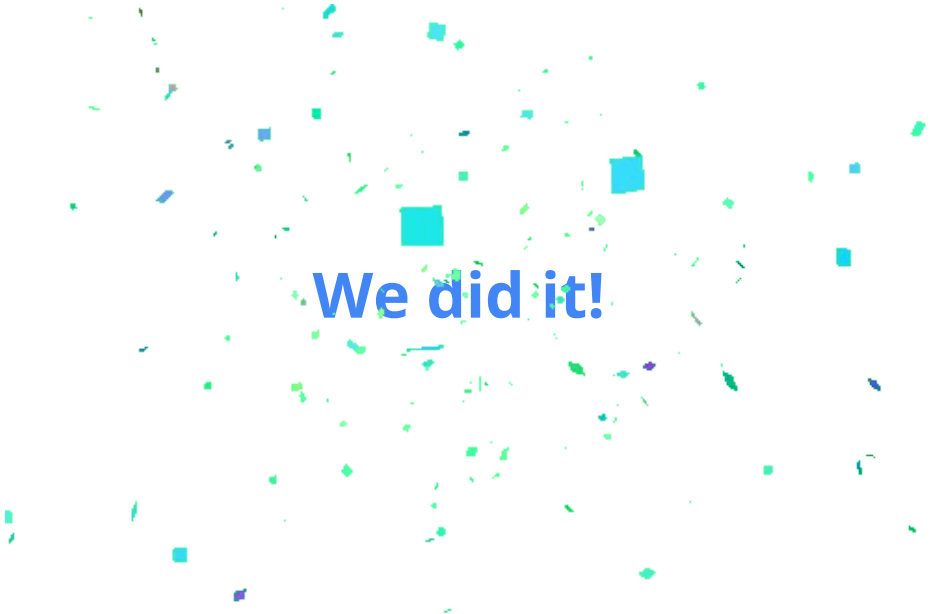
```
from selenium import webdriver

# Create a webdriver instance for the headless Chrome web browser
options = webdriver.ChromeOptions()
options.add_argument('--headless')
driver = webdriver.Chrome(options=options)
```

CODING SQUAD

Let's head to our Jupyter Notebook and explore Selenium!

CODING SQUAD



Request for you

Codecademy - Python & Machine Learning Squad

📝Thanks for attending An Introduction to Web Scraping! How did it go?

codecademy

Thank you for RSVPing for An Introduction to Web Scraping! We hope you had a chance to attend and connect with fellow members at the Python & Machine Learning Squad meetup.

If you were able to attend, please take a moment to provide your thoughts and feedback here. Your feedback helps us continue to create an excellent experience for you.

Share Feedback

codecademy