DATA X

Web Scraping

Berkeley SCET

Created By: Fellipe Marcellino

Table of Content

- Motivation
- HTML Basics
- BeautifulSoup
- Additional Resources

Motivation

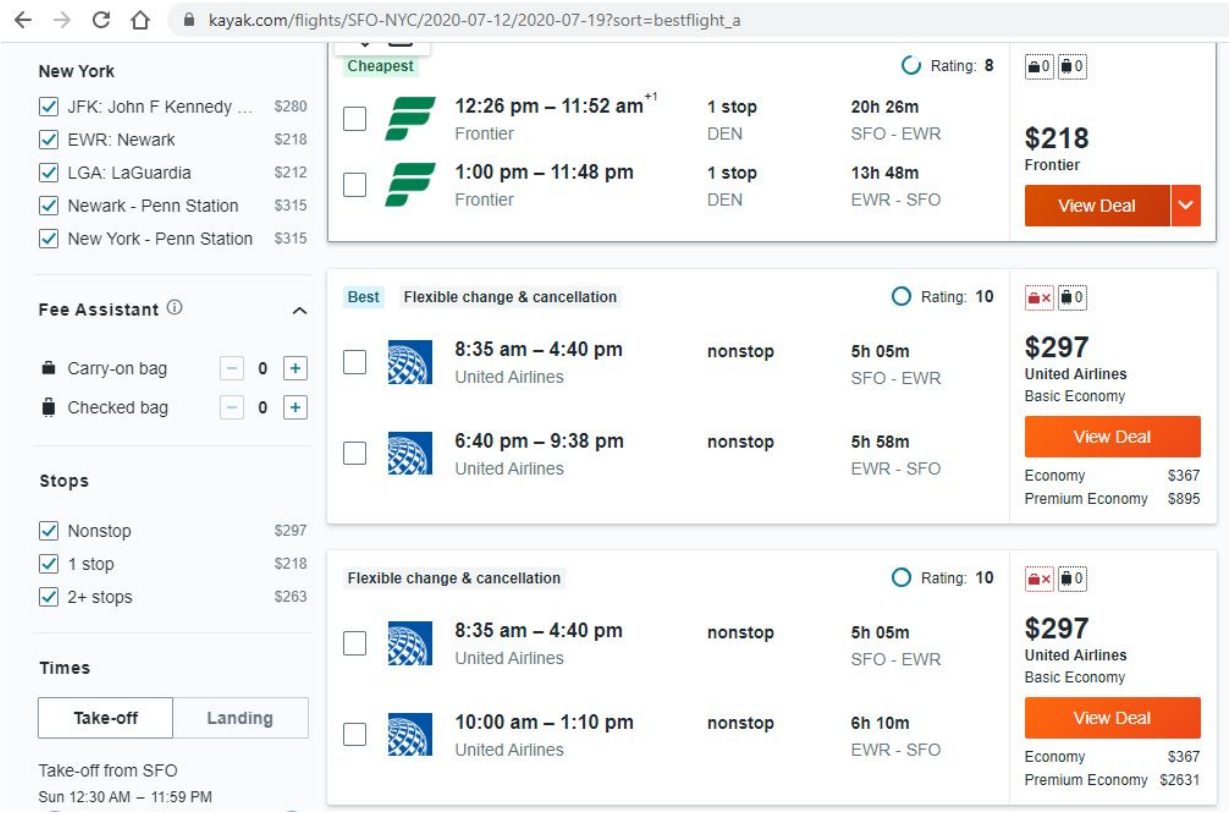
Why Web Scraping ?

"Web Scraping is the practice of gathering data through any means other than API.", Ryan Mitchell

- Data in real world is not always structured in data tables and offered via APIs
- There is a lot of valuable information available online to be extracted
- Web Scraping is a powerful skillset to have as a Data Scientist
- Always make sure to respect the law and Terms of Service of the targeted website!

Use case: Price comparison

Platforms like Kayak rely heavily on web scraping to run their businesses



Accessed on June 12, 2020

Berkeley SCET

Use case: Sentiment Analysis

We can do web scraping to collect reviews from websites like Amazon and then use sentiment analysis techniques

DMG41

Still prefer the QC35 II's

Reviewed in the United States on July 11, 2019

Color: Triple Black

...band that is causing this. Also the ear cups aren't as soft as the QC35 II's. Noise Cancelling - Its excellent, but not a noticeable difference over the QC35 II's. Phone Calls - These really shine during phone calls. Read more ›

573 people found this helpful

Helpful | Comment | Report abuse

Kurt L

Bose rushed this to market before it's ready - buggy app, inconsistent performance.

Reviewed in the United States on July 11, 2019

Color: Triple Black | Verified Purchase

...power off the headphones. You have to wait for them to auto power-off. The noise cancellation is excellent....when the rest of the bugs don't get in the way of it. Which is almost never. Bose clearly rushed this dumpster fire of a product to... Read more ›

328 people found this helpful

Helpful | Comment | Report abuse

Nima

Bose to retake the throne of noise canceling headphones!

Reviewed in the United States on July 2, 2019

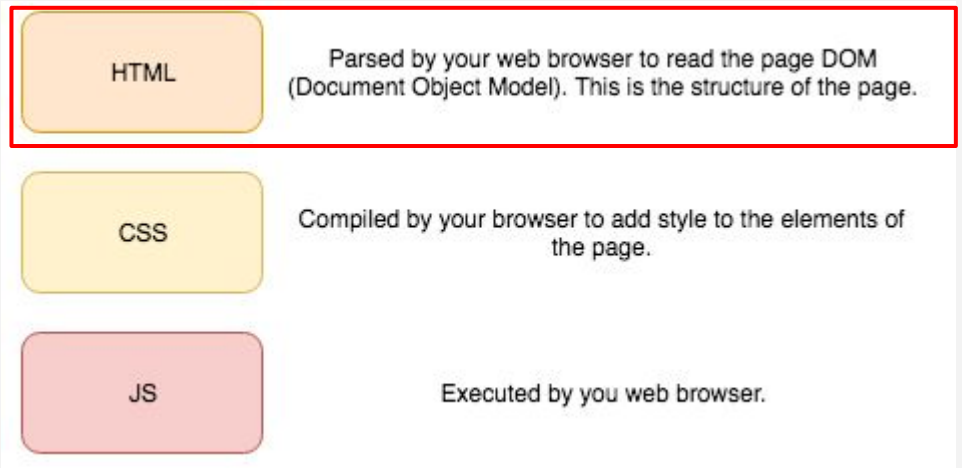
Color: Triple Black | Verified Purchase

Bose QC35 (and before that QC25) used to be the best noise canceling headphones on the market, until Sony WH-1000XM3 arrived last year. Bose couldn't sit idly by, so here we are. I received my NCH 700 pair today and have been using them for the past... Read more ›

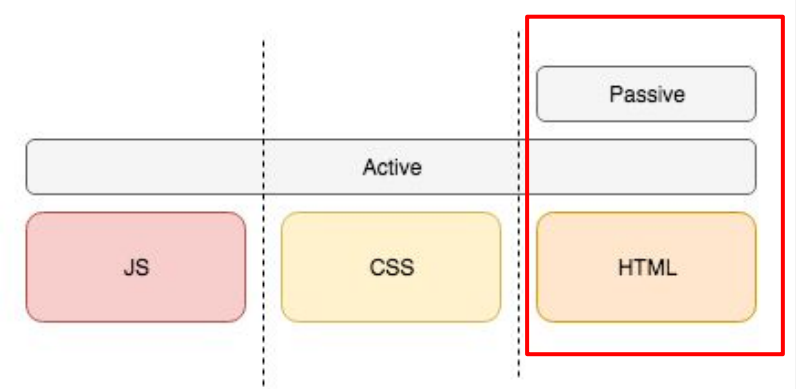
Extracted from Amazon.com on June 12, 2020

HTML Basics

Web page structure



The 3 main languages of a web page



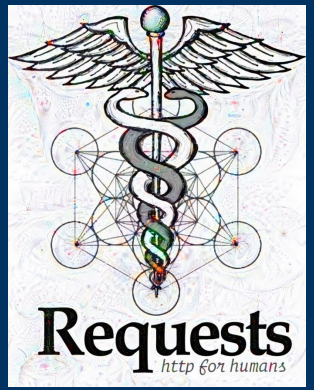
The 2 types of web scraping

We will focus on the HTML language, but we will provide reference to libraries that support CSS and JS as well.

Berkeley SCET

Source: https://www.sipios.com/blog-tech/concrete-example-of-web-scraping-with-financial-data
(Last access: June 18, 2020)

Requests



"Requests is an elegant and simple HTTP library for Python, built for human beings."

Requests **allows you to get HTML code from websites through HTTP/1.1 requests in an easy way**

```
>>> import requests

>>> r = requests.get('https://api.github.com/events')
>>> r.text
'[{"repository":{"open_issues":0,"url":"https://github.com/...
```

```
>>> r.content
b'[{"repository":{"open_issues":0,"url":"https://github.com/...
```

Documentation: https://requests.readthedocs.io/en/master/

HTML Tags

HTML tags are hidden keywords that determine how your web browser will format and display the content.

```
<!DOCTYPE html>
<html>
    <head>
        <title>Example Title</title>
    </head>

    <body>
        <h1>Example Text</>
        <p>Example paragraph</p>
    </body>
</html>
```

Example of HTML code structure

- Open a tag with <> and close with </>
- Nested structure (child, parent, sibling)
- Common tags: *head, body, p*, *div*, *table*

Berkeley SCET

HTML Attributes

"HTML attributes provide additional information about HTML elements."

```
<!DOCTYPE html>
<html>
    <head>
        <title>Example Title</title>
    </head>

    <body>
        <h1 id = "h1_tag">Example Text</>
        <p class = "example">Example paragraph</p>
    </body>
</html>
```

Example of HTML code structure with attributes

- <tag_name **attribute_name = *Value***>*Content*</tag name>
- **class:** used to identify <u>multiple elements</u> in the HTML code
- **id:** used to identify a <u>specific element</u> in the HTML code
- More info: https://www.w3schools.com/html/default.asp

Berkeley SCET

Web Scraping with BeautifulSoup

BeautifulSoup

"BeautifulSoup is a Python library for pulling data out of HTML and XML files. It commonly saves programmers hours or days of work."

```
html_doc = """
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>

<p class="story">Once upon a time there were three little sisters; and their names were
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.</p>

<p class="story">...</p>
"""

from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')
```

```
soup.find_all("a")
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

soup.find_all(id="link2")
# [<a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>]
```

```
soup.find_all("a", class_="sister")
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]
```

Documentation: https://www.crummy.com/software/BeautifulSoup/bs4/doc/#

Data-X website scraping

data-x.blog/syllabus/

Home | About Data-X | Resources | Data-X Online | Berkeley Syllabus | Projects | Posts | Labs | Contact

SCHEDULE OF TOPICS

- On a weekly basis, class sessions may start with a "meet a mentor" and/or "application model case study" section.*
- All slides and notebook samples will be updated at this site.

Topic 1:	Introduction **Theory**: Overview of Frameworks for obtaining insights from data (Slides). **Tools**: Python Review
Code	1. Introduction to GitHub 2. Setting up Anaconda Environment 3. Coding with Python Review
Homework	HW1 HW2
Project	Module 1: Project Introduction
Topic 2:	**Tools:** Linear Regression, Data as a Signal with Correlation
Code	—
Reading	
Project	Module 2: Team Formation 1
Topic 3:	**Theory**: Regression -ML
Code	Coding with Numpy

		Detailed Description
Week	**Part**	
Lecture1	**Topic 1:**	Introduction Theory: Overview of Frameworks for obtaining insights from data (Slides). Tools: Python Review
	Code	1. Introduction to GitHub 2. Setting up Anaconda Environment 3. Coding with Python Review
	Homework	HW1 HW2
	Project	Module 1: Project Introduction
Lecture2	**Topic 2:**	Tools: Linear Regression, Data as a Signal with Correlation
	Code	—
	Project	Module 2: Team Formation 1
Lecture3	**Topic 3:**	Theory: Regression -ML
	Code	Coding with Numpy
	Reading	DataCamp, tutorialpoint,
	Project Module 3	Module 3: Team Formation 2
Lecture4	**Topic 4:**	Theory: Classification and Logistic Regression

Berkeley SCET

Additional Resources

Other tools

Selenium

Active web scraping that is compatible with Javascript websites

https://pypi.org/project/selenium/

Scrapy

Very fast and robust. Good for large projects.

https://pypi.org/project/Scrapy/

Useful article: https://medium.com/analytics-vidhya/scrapy-vs-selenium-vs-beautiful-soup-for-web-scraping-24008b6c87b8