# 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.
Use case: Sentiment Analysis

We can do web scraping to collect reviews from websites like Amazon and then use sentiment analysis techniques.
HTML Basics
We will focus on the HTML language, but we will provide reference to libraries that support CSS and JS as well.

(Last access: June 18, 2020)
Requests allows you to get HTML code from websites through HTTP/1.1 requests in an easy way.

```python
>>> 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/](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.

Example of HTML code structure

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Title</title>
  </head>
  <body>
    <h1>Example Text</h1>
    <p>Example paragraph</p>
  </body>
</html>
```

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

“HTML attributes provide additional information about HTML elements.”

Example of HTML code structure with attributes

- `<tag_name attribute_name = Value>Content</tag name>`
- **class**: used to identify multiple elements in the HTML code
- **id**: used to identify a specific element in the HTML code
- More info: [https://www.w3schools.com/html/default.asp](https://www.w3schools.com/html/default.asp)
Web Scraping with BeautifulSoup
```
from bs4 import BeautifulSoup

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

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>]

tollk2"
# [<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>]
```

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

### 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

**Project**

- Module 1: Project Introduction

#### Topic 2:

**Tools:** Linear Regression, Data as a Signal with Correlation

**Code**

- —

**Reading**

- Module 2: Team Formation 1

**Project**

- Module 2: Team Formation 1

#### Topic 3:

**Theory:** Regression - ML

**Code**

- Coding with Numpy

**Reading**

- DataCamp, tutorialpoint

**Project Module 3**

- Module 3: Team Formation 2

#### Lecture 1

- 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

- Project

  - Module 1: Project Introduction

#### Lecture 2

- Topic 2: Tools: Linear Regression, Data as a Signal with Correlation

- Code

- Reading

- Project Module 3

- Topic 3: Theory: Regression - ML

- Code

- Project Module 3

#### Lecture 4

- Topic 4: Theory: Classification and Logistic Regression
Additional Resources
Other tools

**Selenium**
Active web scraping that is compatible with Javascript websites

[https://pypi.org/project/selenium/](https://pypi.org/project/selenium/)

**Scrapy**
Very fast and robust. Good for large projects.

[https://pypi.org/project/Scrapy/](https://pypi.org/project/Scrapy/)

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