Web Scraping

Using Python

What is it?

- Web scraping is a general term for techniques involving automating the gathering of data from a website.
- Example - List of Items, Data, Images

When viewing a website, the browser doesn’t show you all the source code behind the website, instead it shows you the HTML and some CSS and JS that the website sends to your browser.

- HTML is used to create the basic structure and content of a webpage
- CSS is used for the design and style of a web page, where elements are placed and how it looks
- JavaScript is used to define the interactive elements of a webpage

- For effective basic web scraping we only need to have a basic understanding of HTML.
- Python can view these HTML elements programmatically, and then extract information from the website.

Brief Overview of Python

- Created in 1990 by Guido van Rossum
- Python 3 released in 2008
- Specifically designed as an easy to use language
- High focus on readability of code

Why Choose Python?

- Designed for clear, logical code that is easy to read and learn.
- Lots of existing libraries and frameworks written in Python allow users to apply Python to a wide variety of tasks.
- Focuses on optimizing developer time, rather than a computer's processing time.
- Great documentation online:
  - **docs.python.org/3**

Uses of Python

- Automate simple tasks
  - Searching for files and editing them
  - Scraping information from a website
  - Automate emails and text messages
  - Fill out forms
- Data Science and Machine Learning
  - Analyze large data files
  - Create visualizations
  - Perform machine learning tasks
- Create websites
  - Use web frameworks such as Django and Flask to handle the backend of a website and user data

Installation

- There are several Python distributions you can choose from.
- Most used are -
  - Default Python Distribution
  - Anaconda which contains more tools and packages
- You can use the default Python Distribution because it’s light-weight and sufficient for this course.
- We are using Python 3.8 for this course, although Python 3.7 will also work fine.
- Add Python to “PATH” when installing.

Data Types

- A data type is an attribute of data which tells the compiler or interpreter how the programmer intends to use the data.
- Numeric, non-numeric and Boolean (true/false) data are the most used data types.
- Each programming language has its own classification largely reflecting its programming philosophy.