Week 4 Lab Assignment Goals

- Retrieve web pages using Python
- Use BeautifulSoup to parse web pages

Step 0 : Install BeautifulSoup

- Open a Terminal window or command prompt
- Run ‘pip3 install beautifulsoup4’
- Check that it installed successfully
    - Type ‘python3’ and press Enter.
    - Type ‘from bs4 import BeautifulSoup’
    - There should be no error.

Step 1 : Create a GitHub repository

- Go to https://classroom.github.com/assignment-invitations/db5e1af935911ed902337cc5700282da
- Accept the assignment invite

Step 2 : Get starter code onto your machine

- Like last week, clone the assignment repository onto your machine
- Open a Terminal window or command prompt and ‘cd’ to the cloned directory

Step 3 : Retrieving and Parsing Web Pages - Part 1

- Examine the code in urllinks.py and urllink2.py and understand what it does
- Run ‘python3 urllink2.py’
    - Enter http://www.dr-chuck.com/page1.htm
    - Examine the results
    - urllink2.py prints out the contents of the webpage at that URL
- Run ‘python3 urllinks.py’
    - Enter the same URL
    - Examine the results
    - Urllinks.py prints out the links contained within <a> tags in that webpage
- Run urllink2.py and urllinks.py with the URL http://www.dr-chuck.com/page2.htm

Step 4 : Retrieving and Parsing Web Pages - Part 2

- Run urllink2.py with other URLs, e.g. https://www.michigandaily.com/
- What happens? Why?
- Modify urllink2.py to fix this error

Step 5 : Parsing Web Pages

- Complete exercise 4 from https://books.trinket.io/pfe/12-network.html#exercises.
- Write a program that uses urllib to retrieve the document at http://www.data.pr4e.org/romeo.txt, display the first 100 characters, and count the overall number of characters in the document.

Step 6 : Commit changes to GitHub

- Commit and push all your changes to GitHub