Lab 13: Beautiful Soup

Michael Avalos-Garcia, Jesus Andres Bernal and Paul Whipp

3/16/2019

Summary:

Because we had already completed Lab 12 entirely, this lab was easy to complete. We jumped to our virtual environments and called pip, then headed to the slides to work out the rest. The slides led us to the examples and documentation we needed.

Task 1:

Lab 12 was completed early and turned in.

Task 2:

Only two quick commands were needed to wrap up these tasks, and neither required any troubleshooting.

```
(cst205env) C:\Users\PaulW\Documents\School Spring 2019\CST 205\Python>pip install lxml
Requirement already satisfied: lxml in c:\users\paulw\documents\school spring 2019\cst 205\python\envs\cst205env\lib\sit
e-packages (4.3.2)

(cst205env) C:\Users\PaulW\Documents\School Spring 2019\CST 205\Python>pip install beautifulsoup4
Requirement already satisfied: beautifulsoup4 in c:\users\paulw\documents\school spring 2019\cst 205\python\envs\cst205e
nv\lib\site-packages (4.7.1)
Requirement already satisfied: soupsieve>=1.2 in c:\users\paulw\documents\school spring 2019\cst 205\python\envs\cst205e
nv\lib\site-packages (from beautifulsoup4) (1.8)

(cst205env) C:\Users\PaulW\Documents\School Spring 2019\CST 205\Python>
```

Task 3:

This task was a straight forward rewrite of the provided example code. The first step was to start a headless Mozilla browser, then to package up the right information into a Request object. Then we passed the Request object to urlopen. We then took the response and converted it to a BeautifulSoup object . Once we had the object, we looped through all links and found all the ‘img’ tags. We printed out the ‘src’ value, and the task was complete.

Code Used:

```
# 3/16/2019 -- CST 205 -- Lab 13 -- Beautiful Soup Web Scraping
# By Michael Avalos-Garcia, Jesus Andres Bernal, and Paul Whipp
# Description: This program creates a headless Mozilla browser, converts the
#              HTML DOM into a BeautifulSoup object, and then searches for all
#              images and prints their URLs

from bs4 import BeautifulSoup
from urllib.request import Request, urlopen

# My website from CST 336
my_site = 'http://pwhipp-cst336.herokuapp.com/labs/lab1/'
# package up request info
req = Request(
    my_site,
```

```
    headers={'User-Agent': 'Mozilla/5.0'}
)
# make request, convert to BeautifulSoup object
resp = urlopen(req)
bs_obj = BeautifulSoup(resp.read(), 'lxml')
# search bs_obj
for link in bs_obj.findAll("img"):
    if 'src' in link.attrs:
        print(link.attrs['src'])
```