\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**CSC221 ADVANCED Python Programming**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

LAB8 WEB SCRAPING WITH PYTHON

# Objectives

In this lab assignment, students will learn how to:

* Build a web scraping program using **BeautifulSoup**, **Urllib** and **Pandas**

**PRE-REQs:** BeautifulSoup4 (bs4), Urllib and Pandas must be installed.

See **Anaconda – BeautifulSoup4 & Urllib Installation** (below)

# Instructions

I recommend you create a (new) folder for this week’s lab, for example **Lab8**. This will help you keep your work organized throughout the semester and make it easy to find your work for this course.

This lab is made up of two (2) parts within in a single Jupyter Notebook file. The Lab8-Files.zip contains seven (6) files:

* **CSC221-Lab8-WebScrapingWithPython.docx** ~ these lab instructions
* **Lab8-WebScrapeWiki.ipynb** ~ Jupyter notebook file required for this lab.
* **HowToWebScrapeWikiTutorial** ~ PDF version of PART A tutorial webpage
* **WebScrapingWiki-image1.jpg** ~ Tutorial table image 1
* **WebScrapingWiki-image2.jpg** ~ Tutorial table image 2
* **NC-Demographics-image1.jpg** ~ NC Demographics wiki table image

You should unzip Lab8-Files.zip into your Lab8 folder. The image (.jpg) files must be in the same directory as the Jupyter notebook file for this lab.

#### PART A

Part A requires you to complete an online tutorial ***How to Web Scrape Wikipedia***. The tutorial was written by *Alan Hylands*, and can be found at the following location:

### URL: <https://alanhylands.com/how-to-web-scrape-wikipedia-python-urllib-beautiful-soup-pandas/>

**Note:** A PDF version of this tutorial has also been included in Lab8-Files.zip

The tutorial instructions include the code and brief explanation as you work through each step. You need to complete and understand Part A because it will be used as a model for doing the same thing on a different Wikipedia page for Part B.

**IMPORTANT**: One additional step has been added to the tutorial requiring you to save the DataFrame created to a CSV file (**wiki\_tutorial.csv**). This CSV file must be submitted with your lab.

#### PART B

Part B requires you to complete another Wiki web scraping program very similar to what was done in Part A. The Wikipedia page to be web scraped contains demographic information for the state of North Carolina and can be found at location:

### URL: <https://en.wikipedia.org/wiki/Demographics_of_North_Carolina>

Refer back to the code in Part A as needed, then read and complete the instructions (**in blue**) contained within the Jupyter notebook file. Place your code in the marked cells [# INSERT CODE FOR STEPS *X – X*].

**IMPORTANT**: Again, one additional step has been added requiring you to save the DataFrame created to a CSV file (**NC\_demographics.csv**). This CSV file must be submitted with your lab.

Be sure to re-run all the cells at the end of each assignment to check for syntax errors and correct them. **(Kernel-> Restart & Run All)**.Use this **Run All** command throughout your assignments to sync up the code and remove problems.

# Submitting Assignment

Once you have successfully completed the lab, upload (submit) three (3) files:

* **Lab8-WebScrapewiki.ipynb**
* **wiki\_tutorial.csv**
* **NC\_demographics.csv**

via Blackboard for grading.

# Grading Rubric

Total points = 100

10 points ~ Tutorial (PART A)

5 points ~ Create wiki\_tutorial.csv file (PART A+)

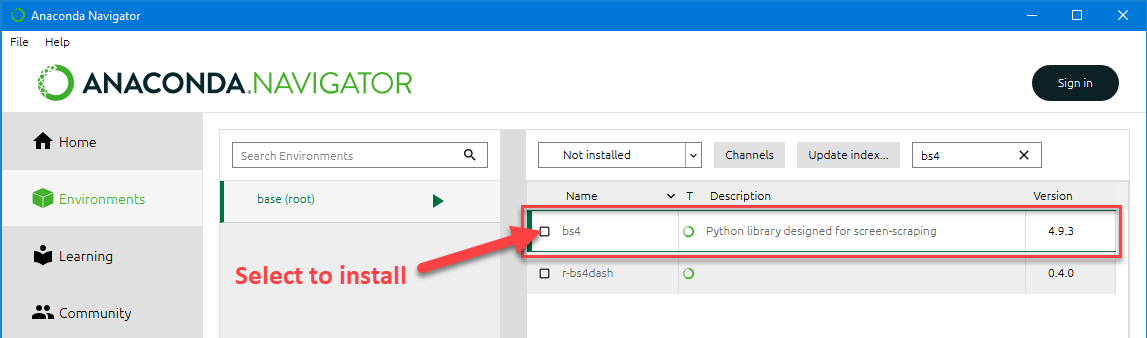
80 points ~ NC Demographics (PART B) (4 pts per step)

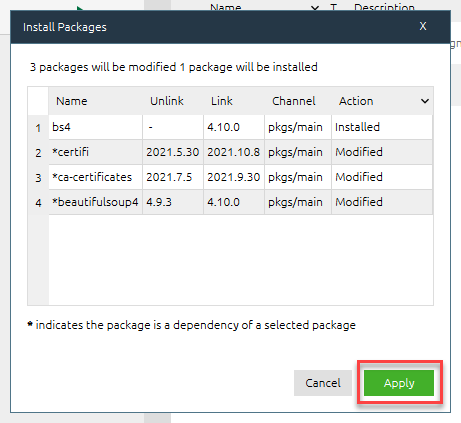
5 points ~ Create NC\_demographics.csv file (PART B+)

## **Anaconda – BeautifulSoup & Urllib Installation**

**BeautifulSoup (bs4)** may already be installed, however if not, install within Anaconda so the libraries are available in your Anaconda/Jupyter Notebook environment. To check to see if **BeautifulSoup** is installed in your Anaconda environment, **start** **Anaconda** to open the **Anaconda Navigator**. Click on the **Environments** link in the menu, make sure **base (root)** is selected, then change the drop-down to **Not installed**. Enter “**bs4**” in the search box, to search the uninstalled packages.

If the results of this search indicate **bs4** is not installed, then select it & click **Apply** to have it installed within **Anaconda**.





**Urllib** may already be installed, however if not, install within Anaconda so the libraries are available in your Anaconda/Jupyter Notebook environment. To check to see if **urllib** is installed in your Anaconda environment, **start** **Anaconda** to open the **Anaconda Navigator**. Click on the **Environments** link in the menu, make sure **base (root)** is selected, then change the drop-down to **Not installed**. Enter “**urllib**” in the search box, to search the uninstalled packages.

If the results of this search indicate **urllib** is not installed, then select it & click **Apply** to have it installed within **Anaconda**.

In my case, **urllib** is already installed as seen here:

