Objective:

Evaluate technical abilities in Python regarding web scraping.

Statement:

Below you will find a list of scraping tasks, in increasing difficulty, in order to evaluate and or build your web scraping and Python skills. If you have little or no experience web scraping, that is ok, as the idea is that you will learn along the way. It may be helpful to plan out how you are going to do each task before you hop right in. If you want to check your plan with Matt or Karthik, that is welcome. Also, feel free to ask any questions if you have hit a roadblock (after you've given it your best effort).

Some tools and recommended libraries to get you started (not a comprehensive list):

Visual Studio Code or PyCharm for your development environment

Python – version 3.8 or above

Pyenv – To manage virtual python environments (optional but helpful to contain all libraries you need for one project)

Beautiful Soup – For parsing html

Requests – For making web requests

Concurrent.futures - For implementing concurrency

Asyncio + httpx or aiohttp – For implementing asynchronous requests

Pandas – for data processing and analysis

* YouTube and Stack overflow will be very helpful in getting up to speed, reach out if you want help finding some resources to look at

Tasks:

Level 1:

Scrape http://books.toscrape.com/ and return a list of all the books on the first page. You can print to the terminal or to a file.

Level 2:

Starting from the homepage of http://books.toscrape.com/ , scrape all products on all pages and return the results in a more complex data format. You could choose JSON, CSV, or any model you find suitable for the task. Scrape more than one attribute this time, for example you could scrape name, price, in stock availability, URLs, etc. Print the data to the terminal or save to a file.

Level 3:

Do all in level 2, but also save your results to the cloud storage. Look up how to save to Google Cloud Storage or BigQuery. You will need Matt or Karthik’s help to give you a place to save to and the credentials. Bonus points if you can generate and send an email with a link to the file or a message saying it has finished.

Level 4:

Choose one or two of the following to implement into your current project.

1) Concurrency
2) Refactoring code to be object oriented
3) Asynchronous web requests
4) Web requests through a proxy service (ask Matt to help get this setup if you choose this one)