# *Internet Development II (ITAS256)*

# *Assignment 1 – Web Scraping*

Date assigned: January 28, 2025

Date due: **February 14, 2025**

**Learning Objectives**

Upon successful completion of this assignment, the student will be able to:

1. Use Python libraries to access websites.
2. Use Python web scraping techniques to scrape web pages.
3. Assemble data scraped from websites using Beautiful Soup.
4. Convert list data to json.
5. Write data to files.

**To be uploaded to Github:**

1. Create a git repository on your own account called <<username>>ITAS256-A01 and make me a collaborator.
2. This repo should be your source for all the work on this project. I don’t only want to see the final product here, but I want to be able to look at the change logs and see that you worked at the project over time.
3. When the project is done, add a tag to your code branch in the repo with the name ‘<<yourusername>>-complete’. Also add a single line to the portal assignment indicating you are done.

**To Start:**

1. You can do this assignment with the standard python library if you wish. You may also (and I recommend doing so) use the requests library and the Beautiful Soup library for making the requests and extracting the information.

**To Do:**

1. You are going to scrape the information off two websites to see if there are jobs that you want to apply for. The first is <https://www.itjobs.ca/>. This is a site that posts for resources for tech jobs in Canada. The second site will be <https://jobs.techtalent.ca/> which is, again, a database of tech jobs.
2. ITJobs: go to the itjobs website at <https://www.itjobs.ca/en/search-jobs/?location=British+Columbia&location-id=BC&location-type=2&search=1&sort_order=1&page=1> Note, in particular, the query string (everything after the ?). I have narrowed the search to include only jobs in BC and remote as well as changed the sort order to sort by date added (sort\_order=1). This will be a multi-page return (as of me writing this there were 64 pages returned) and the page=1 shows that this is the first page.
3. Look at the source code for the web page (Ctrl-U or F12 Elements). The group of jobs is buried inside a div with the class result-ctn. And each individual job within a div with class offer-details and some of its children classes (you will need to look and see).
4. From the page returned extract (HTML Parser or Beautiful Soup) and create a list (or similar) of the job title, company name, offer description and location.
5. You will need to loop and perform this operation for multiple pages. I recommend that you extract all jobs posted in the last two weeks (FROM THE DAY THE PROGRAM IS RUN), but that is up to you (it could be a lot). You must do at least 5 pages in total however,
6. Once you have extracted all this date, write this information as a JSON file called joblist.json with each record looking as below:

A screenshot of a computer code

Description automatically generated

1. TechTalent: go to the TechTalent site and look at the job board. I entered Information Technology for the “What” field and British Columbia for the “Where” field on the form to end up with a URL like: <https://jobs.techtalent.ca/?k=information%20technology&l=British%20Columbia,%20Canada>. If you would like to change the criteria do so, but keep it for jobs in high tech.
2. Once again, look at the source code for the web page. Determine where the position name, company and location information are found.
3. You will need to extract the data from this page using BeautifulSoup (or standard library functions) and store it in a list again.
4. This time the description is found on the linked page. The descriptions on this site can be lengthy. I recommend that you take all the text about the job and send it to ChatGPT and have chat reduce the job description to less than 20 words.
5. Once again, create a list of the jobs with the same information as above. Append that information to the joblist.json file you create above.

**Assignment Marking Scheme**

The following marking scheme will be used in marking the assignment.

|  | Out of |
| --- | --- |
| Scraped itjobs.ca and extracted jobs   * Scraped correctly * Data extracted to list * Did multiple pages (at least 5) in a loop but better for last 2 weeks | 35 |
| Wrote extracted information to JSON file. | 15 |
| Scraped techtalent.ca and extracted jobs   * Scraped correctly * Data extracted to list | 20 |
| Used ChatGPT to summarize job description | 5 |
| Appended information to JSON file   * Description blank if ChatGPT is not used | 10 |
| Coding Style  - Good use of functions/classes  - Variable names/function names correct. PEP standard followed  - Comments where needed and docstring for each function and file | 10 |
| Organization  - assignment folder and project named according to standards  - assignment added to Github repository properly  - git history available for viewing showing work | 5 |
| Total | 100 |