# CS4132
# Data Analytics

Name: _____ (  ) Date: _____

## Research Task – Pairwork

**Date issued: T3W4** **Date Due: T3W7 11 Aug, 2200hrs**

## Learning Web Scraping with ChatGPT (or equivalent LLM)

Web scraping is a valuable skill for data analysis, and different libraries offer unique advantages for web scraping. By learning a new library, you will expand your web scraping capabilities and learn how to solve specific data problems.

Advanced natural language processing models like ChatGPT (or Bard) can significantly improve the efficiency and effectiveness of web scraping processes by providing sample boiler code and examples.

In this research task, you will **work with a peer** to learn a new web scraping library in Python **with the assistance of LLM (i.e. ChatGPT 3.5 / Bard or equivalent)**. At the end of the task, you will create a step-by-step tutorial aimed at teaching a beginner basics of web scrapping with the assigned library.

*Note: Further references to ChatGPT will refer to ChatGPT or equivalent LLM such as Bard.*

## Task Outline

You are advice to use relevant prompt to generate a base using ChatGPT. However, note that you should not use text generated by ChatGPT verbatim. AI-generated output, which is based on the data that it has access to, may be inaccurate, irrelevant, illogical, or biased. You should **critically assess** AI-generated output for accuracy, objectivity, logic, and relevance and cross check with other sources.

Here are suggested steps for completing this research task:

1. Begin by using ChatGPT to generate a brief explanation of web scraping and the possible ethical and legal concerns associated with it. Obtain case studies of legal cases arising from web scraping activities.

   *To prompt ChatGPT, you can ask questions such as "What is web scraping?" or "What are the ethical and legal concerns in web scraping?" Take note of the accuracy of ChatGPT's responses and use them as a starting point for your research.*

2. Familiarize yourself with the basics of the assigned web scraping library, including its installation, syntax, code structure, and usage for extracting data from web pages. You may also need to familiarize yourself with the syntax of HTML.

   *To generate an explanation of the basics of your assigned library, prompt ChatGPT by asking questions like "What are the basics of [library name]?" or "How do I use [library name] for web scraping?" Evaluate the accuracy of ChatGPT's responses and use them as a foundation for your learning.*

© NUS High School of Math & Science

3. Deepen your understanding of the assigned library by conducting additional research using other tutorials and websites. In your own words, expand the explanation of the library, providing more details, examples, and use cases that showcase your knowledge. Incorporate code snippets and emphasize the library's features and advantages. You can demonstrate your understanding with relevant code snippets to scrap the website http://books.toscrape.com/

   *For examples and use cases, prompt ChatGPT by asking questions such as "What are some examples of using [library name] for web scraping?" or "What are some use cases for [library name] in web scraping?"*

4. You will have a good understanding of how to web scrape with the assigned library after going through (2) and (3). Now, apply the library to a specific use case by finding a suitable website that allows web scraping. Ensure the chosen website has a clear statement or provision in its terms of service or robots.txt file that permits web scraping activities.
   a. First, research and identify a website that has interesting data and explicitly allows web scraping. This can be done by examining the site's terms of service or robots.txt file to verify that web scraping is permitted.
   b. Next, use the assigned library to extract relevant information from the chosen website, such as product details, user reviews, or any other data of interest. Tailor the scraping process to suit the structure and content of the selected website. The website chosen should ideally allow you to demo different features of the library not used in (3).
   c. Finally, export the scraped data to an Excel file, ensuring that it is well-organized and easily accessible for further analysis or use.

5. Critique the outputs from ChatGPT throughout the process, evaluating their accuracy and relevance to the task. Reflect on how effectively you were able to use the AI-generated information in your learning process and tutorial development.

   **Ensure that both you and your partner generate your own ChatGPT prompts individually.** Compare and contrast the output from ChatGPT and work collaboratively to integrate the best elements of each response into your final report.

6. In your final report, include a list of references you used to verify and enhance the tutorial. Ensure that your report follows a clear structure, with a step-by-step tutorial that caters to beginners and incorporates your learnings from ChatGPT, independent research, and hands-on experience with the library.

© NUS High School of Math & Science

## Final Deliverables

1. A comprehensive report in Jupyter Notebook format. You should use the draft provided for you as a base. You may add in additional sections if appropriate.

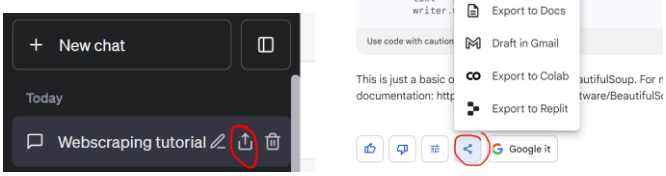
   The report should serve as a step-by-step tutorial that teaches beginners how to use the assigned web scraping library, including:

   a) A brief introduction to web scraping and its related ethical and legal concerns, including some interesting legal case studies.
   b) A brief overview of the assigned library, and detailed installation instructions.
   c) Clear and executable code snippets illustrating the basics of the library and its usage for web scraping using http://books.toscrape.com/ as an example.
   d) Select a suitable case study website of your choice to showcase additional web scraping features. Include relevant code examples and explanations.
   e) Finally, a short conclusion of what was covered in the tutorial, and pros and cons of the library assigned.

   Note that all code in the jupyter notebook should be executable end to end without any compilation errors. The excel files generated from scraping in steps (c) and (d) should be attached in the submission.

2. A **Reference** section citing any additional references / links you may have used for the task, including fact checking.

3. An **Appendix** in the report containing the following:
   a) A list of prompts used in your interactions with ChatGPT.
   b) Submit the sharing link of your full chat transcript.

   e.g. 


4. Finally, submit a 1-page **individual reflection** for this task. Use the following questions to guide your reflection:
   - How did ChatGPT help you understand the new web scraping library? Describe specific instances where the AI assistant clarified a concept, provided an example, or guided you to apply the knowledge in a practical way. How did your understanding evolve over the course of your learning journey?
   - What strategies did you use to interact with ChatGPT effectively to learn the library? Did you have to adjust your learning strategies while using the AI assistant? If so, describe what changes you made and why.
   - How did you feel throughout the learning process? Did using ChatGPT influence your motivation or emotional response (e.g., frustration, satisfaction, curiosity)? Please give specific examples.
   - Reflect on your metacognitive strategies during this process. How did you plan, monitor, and evaluate your learning progress? Did you encounter any difficulties and if so, how did you overcome them?

© NUS High School of Math & Science

Note: The reflection is not about the right or wrong answers but about your personal learning experience. Be honest and detailed about your thoughts, feelings, and actions during the process.

**Reflection MUST NOT be generated from ChatGPT or related AI tools.**

**Late work:** 5% of grade will be deducted per day you are late.

## Grading Rubrics

| Criteria | Weightage | Description |
|---|---|---|
| Overview of Web Scraping | 20% | • Clear and concise explanation of web scraping and its importance.<br>• Comprehensive discussion of ethical and legal concerns, supported by relevant examples or case studies. |
| Web Scraping with Library X | 30% | • Clear explanation of the library and its features, with proper and easy to follow installation instructions.<br>• Relevant, well-explained code snippets demonstrating the use of the library for scraping bookstoscrape, including setup.<br>• Provides a balanced discussion of the advantages and disadvantages of the assigned library. |
| Case Study | 30% | • Appropriate selection of a website to scrape, with a clear explanation of why the website is suitable.<br>• Clearly defines the specific problem and outlines the approach taken to solve it using the assigned library.<br>• Detailed description of the implementation, including code snippets and a discussion of the results obtained. |
| Clarity and Organization of Report | 15% | • Well-structured and logically organized content throughout the report.<br>• Clear language, proper grammar, and appropriate terminology used across all sections.<br>• Reflects on the learning process, identifies challenges faced, and discusses how those challenges were overcome with specific authentic examples. |
| References & Appendixes | 5% | • Proper citation of all sources and references used throughout the report, following a consistent citation style. |

© NUS High School of Math & Science