CSIS 4260 – Assignment 2

Learning web scraping and text analysis

Part 1: Web scraping

In this part you will research and compare two of the popular web scraping libraries and recommend your choice based on benchmarks and ease of use.

Three common and popular libraries are Scrapy, BeautifulSoup / MechanicalSoup, and Playwright (browser automation) but you are not limited to these. You can use any other library as well.

Once you select a library, please use it to scrape any public site, e.g. a particular subreddit or a news website and collect at least 100 articles / pages **on a topic** of your choice.

Part 2: Text analysis

In this part you will apply at least 2 algorithms to do text analysis on the data you collected in part 1. You can choose either a library, algorithm or api (including LLM apis such as Llama, chatgpt, deepseek etc). If using commercial API, please be aware that they might be paid and, in that case, limit your expenses to under $5.

Please come up with a summary of each page along with an importance score which you will drive by analysing the article/page text from the context of your topic. The importance score should also have direction (positive or negative).

Please present the outcome of this analysis in a tabulated form along with your code.