Web Scraping using BeautifulSoup, Selenium, and Wireshark

Due: 2023-10-08

1 Objectives

The main aim of this assignment is to familiarize you with web scraping techniques using BeautifulSoup and Selenium and to understand the underlying network communications using Wireshark. You'll not only retrieve data from a website but also analyze the HTTP packets when communicating with that website.

2 Instructions

1. **Website Selection**: Choose a website that allows web scraping (always check the robots.txt file). The site should contain structured data, whether it be items from an e-commerce site, articles from a news portal, or posts from a forum.
2. **Packet Capture using Wireshark**
   - Installation: Install Wireshark (https://www.wireshark.org/) on your laptop. Ensure that you understand the basic operations of this tool.
   - Capture: Start capturing packets before you initiate your web scraping task. Filter the packets only to include HTTP traffic to and from your selected website.
   - Analysis: Once you've captured enough packets, stop the capture and analyze the packets. Look for details such as the request method (GET, POST), response codes (e.g., 200 OK, 404 Not Found), and any interesting headers or data.
3. **HTML Structure Analysis**: Analyze the HTML structure of the website to navigate and retrieve the desired data accurately.
4. **Data Retrieval**: Use BeautifulSoup for HTML parsing and Selenium for browser automation, especially if the website contains dynamic content.
5. **Determining the Quantity for Scraping**: Make sure you gather enough data without overloading or getting blocked by the website.
   - Page limit: Limit your scraping to a specific number of pages, for instance, up to 5 pages.
   - Data item limit: Restrict the data retrieval to a set number of items, such as 30 products or 15 news articles, etc.
   - Time constraint: Ensure your scraping process does not exceed a duration (15 minutes).
6. **Data Storage**: Store the retrieved data in a structured format, like a CSV or a database, which will be easier for further analysis or review.

3 Submission List

Submit your code and report via the KENTECH LMS System.

- **Experiment code**: Note that your submission code is in Python format (.py or .ipynb). Your code should be organized, and each major task should be in separate code cells for clarity.
- **A written report summarizing the results**: Ensure the entire assignment - from website selection to data storage - is documented in your report. Include code blocks, explanations, and observations.