COUPANG WEB AUTOMATION ASSIGNMENT REPORT

Name: Nelton
Email: neltontan@outlook.com
Date: 20 Jul 2025
GitHub: https://github.com/neonewton/PUBLIC_Nelton_assignment_coupang

1. Overview

Objective: To build a robust web scraper that can extract product data from any e-commerce website, supporting dynamic content, pagination, and anti-bot protections.

2. Approach & Tools

Programming Language: Python 3.11

Libraries & Tools:

- Selenium (Zalora only)
- Playwright (shein and decathlon)
- BeautifulSoup
- CSV

Scraper Features:

- JavaScript-rendered content support
- Pagination via "View More" button clicks
- Image, title, price, and product URL extraction
- CAPTCHA detection and graceful handling
- Retry logic and basic anti-bot delay wait simulation

3. Site Attempts & Challenges

1. Zalora

URL: https://www.zalora.sg/c/men/shoes/c-27
Time Spent: ~45 minutes

- Initial attempts using Selenium failed due to aggressive anti-bot mechanisms.
- Issues encountered:
    - Scraping too quickly without delays or backoff strategies triggered immediate blocking.
    - Browser automation fingerprints (e.g., navigator.webdriver = true) were detected.
    - Static IP requests from the same address flagged scraping activity.

    - Manual browsing in Chrome worked fine, but automation was blocked due to:
        - Missing browser artifacts (e.g., chrome.runtime, extension traces)
        - Unusual input timing, lack of user interactions of mouse & keyboard
        - Non-standard screen dimensions or headless behavior
- Playwright, despite being more stealth-capable, was blocked immediately likely due to Zalora's advanced bot detection systems.

2. SHEIN

URL: https://sg.shein.com/pdsearch/Shoes%20For%20Men/
Time Spent: ~15 minutes

- Used Playwright to initiate scraping.
- Encountered visual CAPTCHAs and anti-bot overlays shortly after loading the page.
- Promotions and popups further complicated DOM extraction.
- Similar to Zalora, scraping attempts were blocked almost immediately.

3. Decathlon

URL: https://www.decathlon.sg/c/men/shoes.html
Time Spent: ~20 minutes

- JavaScript-heavy site, but manageable with Playwright and strategic delays.
- "View More" button required dynamic clicking and waiting for AJAX content.
- No CAPTCHA or serious anti-bot triggers encountered.
- Successfully scraped 123 product items after iterative improvements.
  (Fields extracted:Title, Price, Image URL, Product URL)
  Final CSV output file: decathlon_output.csv

8. Conclusion

This project evaluated three e-commerce sites for web scraping feasibility. Zalora and SHEIN employed strong anti-bot protections, including CAPTCHAs and fingerprint detection, which blocked automation tools like Selenium and Playwright. Decathlon, while JavaScript-heavy, allowed successful scraping using Playwright with intelligent delays and retry logic. The scraper ultimately extracted 123 products. This task highlighted the need for adaptive tooling, respect for site constraints, and awareness of bot detection techniques in real-world scenarios. Decathlon was chosen as the final site due to its balance of accessibility and dynamic complexity, aligning well with the assignment's goals on data extraction and automation.