Crawl Data

1. Thế nào là Crawl Data
2. Tìm hiểu các kỹ thuật: Beauty Soup, Scrapy và Selenium
3. Mục đích sử dụng của các công cụ
4. Nêu sự khác biệt giữa các công cụ này trong Web Crawl Data
5. Nêu ra giải pháp chọn lựa các công cụ này trong từng tính huống cụ thể
6. Xây dựng ứng dụng cào dữ liệu web theo hướng dẫn: https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/ . Lưu dữ liệu cào được xuống file raw_data.csv
7. Với dữ liệu đã cào được, hãy thực hiện các xử lý cơ bản
    a. Đọc dữ liệu lên data frame
    b. Dữ liệu rỗng
    c. Dữ liệu trùng
    d. Dữ liệu sai định dạng
    e. Dữ liệu lỗi Unicode
    f. Dữ liệu chứa nhiều thông tin cần tách ra
    g. Dữ liệu thiếu
    h. ……
8. Lưu file dữ liệu đã xử lý với tên clean_data.csv

- Sinh viên nộp bài
    - File word trình bày nội dung các câu hỏi
    - Source code cào dữ liệu
    - Kết quả dữ liệu cào được raw_data.csv
    - Source code xử lý dữ liệu
    - Kết quả sau khi xử lý clean_data.csv