TRƯỜNG ĐẠI HỌC BÁCH KHOA TP HCM
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

- - - - - - - - - - - -

ĐỒ ÁN KỸ THUẬT LẬP TRÌNH

Báo cáo

Xây dựng công cụ thu thập dữ liệu giá thành sản phẩm trên các trang Tiki, Lazada,...

Giáo viên hướng dẫn: Lưu Quang Huân
Sinh viên: Lộc Quốc Huy - 1812369
Nguyễn Việt Hưng - 1812499
Trần Quang Huy - 1812427

Tp. Hồ Chí Minh, Tháng 1/2021

Mục lục

1 Bảng phân công nội dung

Thành viên phụ trách	Nội dung	Phần trăm hoàn thành	Chú thích
Lộc Quốc Huy	Crawler Shopee,các tính năng công cụ	100	
Nguyễn Việt Hưng	Crawler Nguyễn Kim, Điện máy chợ lớn, Báo cáo	100	
Trần Quang Huy	Crawler Tiki, Lazada, Phân tích data	100	

2 Phân tích yêu cầu

3 Công cụ Crawl dữ liệu

3.1 Giải pháp

Nhóm đã sử dụng ngôn ngữ lập trình Python kết hợp với Selenium để crawl dữ liệu các sản phẩm điện tử (loại, tên, giá, link web) từ các trang thương mại điện tử như **Tiki,Shopee,Lazada,Nguyễn Kim, Điện máy chợ lớn**.

3.2 Kết quả

Nhóm em đã thu được lượng dữ liệu khoảng 10,000 item mỗi trang web nhằm để xây dựng một trang web dùng để tìm kiếm, phân loại, cũng như phân tích dữ liệu giá trên từng nền tảng thương mại điện tử.

4 Làm sạch chuẩn hóa dữ liệu

4.1 Format dữ liệu

Nhóm đã sử dụng Dataframe trong thư viện Pandas để format data theo các field và lưu trữ bộ Dataset (10,000 item) dưới dạng CSV nhằm hỗ trợ cho việc **Import** dữ liệu vào Database **MySQL**.

4.2 Làm sạch dữ liệu

Trong quá trình crawl không thể tránh khỏi có những kí tự đặc biệt, icon, các item trùng nhau, ... Do đó nhóm chúng em đã làm sạch toàn bộ dữ liệu cũng như update lại cơ sở dữ liệu đã hoàn thành

5 Phân tích dữ liệu

Với các file dữ liệu đã crawl ở các trang thương mại điện tử. Nhóm chúng em tiến hành phân tích dữ liệu nhờ công cụ Pyplot trong thư viện matplotlib để tạo nên các biểu đồ lượng sản

phẩm, giá trung bình của mỗi loại sản phẩm, giá trung bình của mỗi loại sản phẩm của mỗi trang,...

6 Xây dựng Webapp

Tiếp theo nhóm em đã bắt đầu xây dựng nên Web để hiển thị dữ liệu, cùng các biểu đồ phân tích được. Qua nền tảng Web nhóm chúng em đã tiến hành tạo thêm các công cụ như tìm kiếm sản phẩm theo tên, thể loại, giá của từng trang. Các công nghệ chúng em sử dụng để xây dựng nên trang web:

- Về Front-end : dùng HTML, CSS, JS kết hợp với JQuery
- Về phía Back-end : sử dụng Micro-Framework Flask của Python kết nối với Database MySQL

7 Source code

Link souce code github: https://github.com/tqhuy-bk/201CO10313_DA_KTLT

8 Tài liệu tham khảo

Tài liệu

[1] Web Scraping Using Selenium — Python - https://towardsdatascience.com/web-scraping-using-selenium-python-8a60f4cf40ab (Truy cập lần cuối ngày 26/11/2020)

[2] Format Data- https://pandas.pydata.org/pandas-docs/stable/user_guide/style.html (Truy cập lần cuối ngày 3/12/2020)

[3] Vẽ Biểu Đồ Với Thư Viện Matplotlib - https://codelearn.io/sharing/ve-bieu-do-voi-thu-vien-matplotlib-p1 (Truy cập ngày 18/12/2020)

[4] Make a Web Application Using Flask in Python 3 - https://www.digitalocean.com/community/tutorials/how-to-make-a-web-application-using-flask-in-python-3 (Truy cập lần cuối ngày 2/1/2021)