**Bài tập cá nhân Tuần 01**

**I. Yêu cầu:**

Xây dựng hệ thống thu thập dữ liệu tự động từ các trang thương mại điện tử sử dụng Docker và Scrapy, BeautifuSoup

- Crawl dữ liệu sản phẩm từ 2 trang trở lên

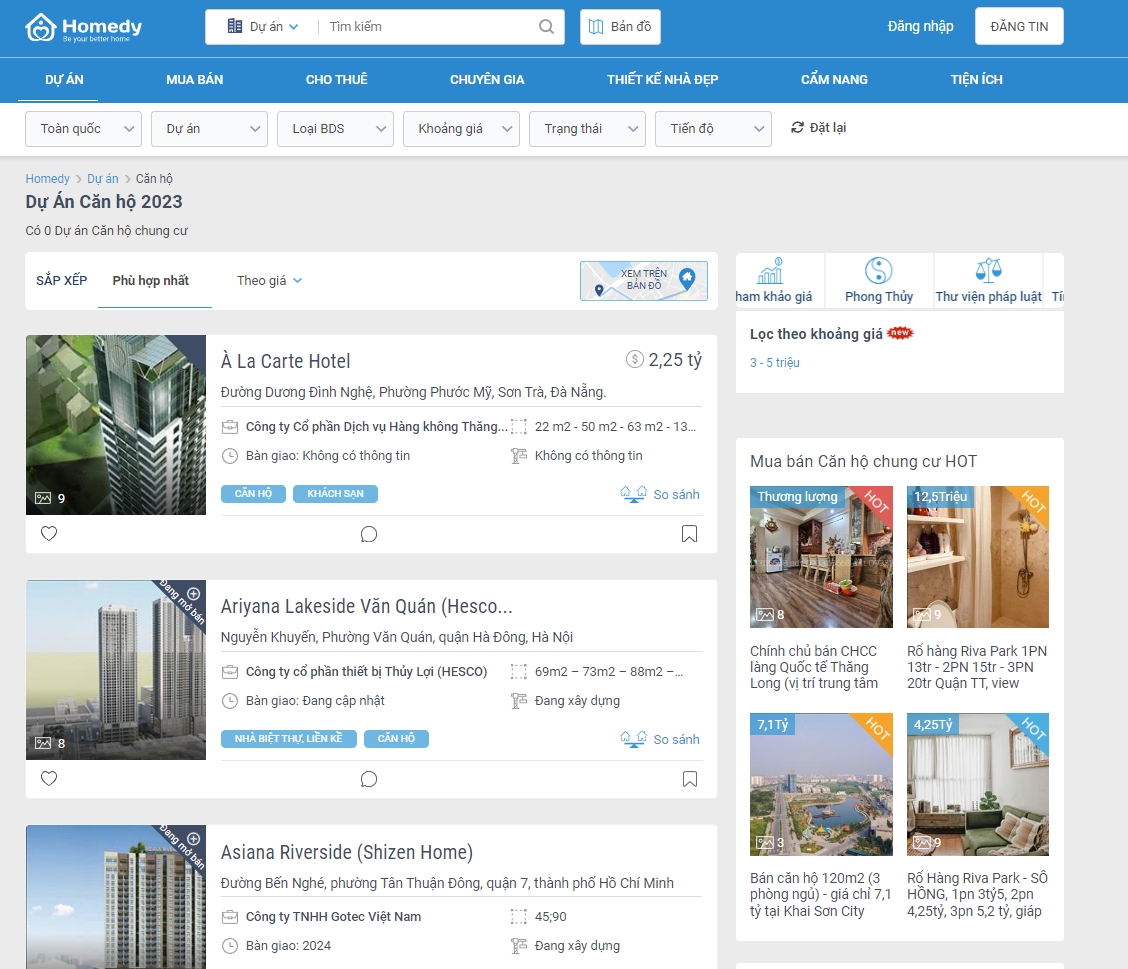
- Thu thập các thông tin cần thiết của trang

- Lưu trữ vào mysql, postgre, mongodb, csv, txt

- Sử dụng docker để đóng gói và triển khai

Link project: <https://github.com/HieuTrungMc/BigData-T01>

Đề tài: Xây dựng hệ thống thu thập dữ liệu tự đồng từ trang web bất động sản “homedy.com”, khu vực Dự án căn hộ



Trang web cụ thể để crawl: <https://homedy.com/du-an-can-ho>

Công nghệ sử dụng: Python, Scrapy, Docker

Các dữ liệu mẫu thu thập được:

A close up of text

Description automatically generated

Title price address area

À La Carte Hotel," 2,25 tỷ","Đường Dương Đình Nghệ, Phường Phước Mỹ, Sơn Trà, Đà Nẵng.",22 m2 - 50 m2 - 63 m2 - 135 m2

Ariyana Lakeside Văn Quán (Hesco Văn Quán),N/A,"Nguyễn Khuyến, Phường Văn Quán, quận Hà Đông, Hà Nội",69m2 – 73m2 – 88m2 – 89m2 – 90m2 – 95m2 -111m2

Asiana Riverside (Shizen Home),N/A,"Đường Bến Nghé, phường Tân Thuận Đông, quận 7, thành phố Hồ Chí Minh",45;90

Astral City, 295 triệu,"Mặt tiền Quốc lộ 13, TP Thuận An, Tỉnh Bình Dương",45;90

Athena Monopoly," 3,3 - 3,7 tỷ","Phường Vinh Tân, Thành phố Vinh, Nghệ An",Không có thông tin

Chung cư Huỳnh Tấn Phát," 1,7 tỷ","Đường Huỳnh Tấn Phát, Phường Phú Thuận, Quận 7, Hồ Chí Minh",50m2 – 90 m2

Green Center (Làng sinh thái golf)," 2,95 - 6,6 tỷ","xã Tân Mỹ, huyện Đức Hòa, tỉnh Long An.",Không có thông tin

Khu dân cư Bảo Thịnh Residence, 6 tỷ,"Hà Huy Giáp, Phường Thạnh Xuân, Quận 12, TP.Hồ Chí Minh",250m2 - 280m2

Khu đô thị Phương Đông Vân Đồn, 2 tỷ,"Xã Đông Xá, đặc khu kinh tế Vân Đồn, Quảng Ninh",87;120;140;200;300;500

Các loại handler để lưu dữ liệu: CSV, MariaDB, MongoDB, TXT plaintext

A screenshot of a computer

Description automatically generated

Triển khai Docker

A screen shot of a computer

Description automatically generated

**A black screen with blue lines

Description automatically generated**