|  |  |  |
| --- | --- | --- |
|  | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM** |  |

**ĐỒ ÁN MÔN HỌC**

**TÌM HIỂU VỀ CÔNG CỤ MÃ NGUỒN SCRAPY VÀ THU THẬP DỮ LIỆU SẢN PHẨM TRÊN SÀN THƯƠNG MẠI ĐIỆN MÁY XANH**

Ngành: **KHOA HỌC DỮ LIỆU**

Môn học: **MÃ NGUỒN MỞ TRONG KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn : Ths.Lê Nhật Tùng

Sinh viên thực hiện :2286400001 Phạm Quốc An

2286400481 Nguyễn Thị Thu Ngân

TP. Hồ Chí Minh, 2024

|  |  |
| --- | --- |
|  | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM** |

**ĐỒ ÁN MÔN HỌC**

**TÌM HIỂU VỀ CÔNG CỤ MÃ NGUỒN SCRAPY VÀ THU THẬP DỮ LIỆU SẢN PHẨM TRÊN SÀN THƯƠNG MẠI ĐIỆN MÁY XANH**

Ngành: **KHOA HỌC DỮ LIỆU**

Môn học: **MÃ NGUỒN MỞ TRONG KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn : Ths.Lê Nhật Tùng

Sinh viên thực hiện :2286400001 Phạm Quốc An

2286400481 Nguyễn Thị Thu Ngân

TP. Hồ Chí Minh, 2024

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

TPHCM, Ngày… tháng 10 năm 2024

**Giáo viên hướng dẫn**

(Ký tên, đóng dấu)

**LỜI CAM ĐOAN**

Chúng tôi, Phạm Quốc An và Nguyễn Thị Thu Ngân, xin cam đoan rằng:

Toàn bộ nội dung của bài báo cáo này là kết quả từ quá trình nghiên cứu và làm việc chung của cả hai chúng tôi. Các thông tin được trình bày trong báo cáo đều được thu thập từ các nguồn đáng tin cậy và đã được xử lý cẩn thận.

Chúng tôi đảm bảo rằng không có bất kỳ hành vi sao chép hay sử dụng thông tin không đúng đắn nào từ các nguồn khác. Tất cả các tài liệu tham khảo đã được ghi nguồn rõ ràng và tuân thủ đúng các quy định về trích dẫn học thuật.

Bài báo cáo này là công trình nghiên cứu chung của chúng tôi và chưa từng được nộp hoặc công bố ở bất kỳ đâu trước đây. Chúng tôi cam kết hoàn toàn chịu trách nhiệm về tính trung thực và chính xác của nội dung báo cáo này.

Chúng tôi hy vọng rằng bài báo cáo này sẽ mang đến một góc nhìn tổng quát và chi tiết về chủ đề “Tìm hiểu về công cụ mã nguồn Scrapy và thu thập dữ liệu....” và đồng thời đóng góp phần nhỏ vào việc nghiên cứu trong lĩnh vực nghiên cứu này.

TPHCM, ngày… tháng 10 năm 2024

**Sinh viên**

Phạm Quốc An,

Nguyễn Thị Thu Ngân

**DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT VÀ TỪ KHÓA**

**MỤC LỤC**

**DANH SÁCH BẢNG**

**DANH SÁCH HÌNH VẼ**

**CHƯƠNG 1: TỔNG QUAN**

**1.1**

**1.2**

**1.3**

**1.4 Đối tượng và phạm vi**

*1.4.1 Đối tượng*

Đối tượng nghiên cứu của đề tài là các sản phẩm được bán trên sàn thương mại điện tử Điện máy xanh. Các sản phẩm này thuộc nhiều danh mục khác nhau như điện tử, gia dụng, thiết bị số và phụ kiện,..... Các sản phẩm này có sự đa dạng về loại, giá cả và lượng tiêu thụ. Bằng cách sử dụng mã nguồn mở Scrapy để thu nhập và phân tích dữ liệu từ các sản phẩm này, đề tài nhằm hiểu rõ hơn về xu hướng giá và các yếu tố khác ảnh hưởng đến hiệu suất kinh doanh trên sàn thương mại điện tử. Từ đó, nghiên cứu sẽ cung cấp một cái nhìn tổng quan giúp các doanh nghiệp có thể đưa ra các đề xuất nhằm cải thiện chiến lược kinh doanh.

*1.4.2 Phạm vi*

Đề tài tập trung vào việc thu thập và phân tích dữ liệu sản phẩm từ sàn thương mại điện tử Điện máy xanh bằng cách sử dụng Scrapy để tự động thu thập thông tin từ các trang sản phẩm như tên, hình ảnh, giá cả. Dựa trên kết quả thu thập và phân tích dữ liệu, đề tài sẽ đưa ra đề xuất các chiến lược kinh doanh phù hợp. Phạm vi nghiên cứu sẽ tập trung vào các sản phẩm thuộc một số danh mục tiêu biểu và trong một khoảng thời gian giới hạn nhằm đảm bảo dữ liệu được thu thập đầy đủ và chính xác.

**1.5 Phương pháp nghiên cứu**

*1.5.1 Phương pháp nghiên cứu sơ bộ*

Trước khi tiến hành thu thập dữ liệu, chúng tôi đã thực hiện nghiên cứu sơ bộ để hiểu rõ hơn về Scrapy, bao gồm cách thức hoạt động của nó và những lợi ích mà framework này mang lại trong việc thu thập dữ liệu từ các trang web. Thông qua việc nghiên cứu sơ bộ, chúng tôi sẽ xác định các vấn đề cụ thể cần giải quyết và đề xuất các phương pháp nghiên cứu phù hợp.

*1.5.2 Phương pháp nghiên cứu tài liệu*

Chúng tôi sẽ thực hiện nghiên cứu tài liệu để tìm hiểu về các nghiên cứu liên quan đến việc thu thập dữ liệu từ sàn thương mại điện tử. Qua việc đánh giá các nghiên cứu trước đây, chúng tôi sẽ rút ra kinh nghiệm về cách thu thập, xử lý và quản lý dữ liệu. Đồng thời, nghiên cứu các phương pháp phân tích dữ liệu liên quan để xác định cách tiếp cận phù hợp và áp dụng chúng vào đề tài.

*1.5.3 Phương pháp nghiên cứu thống kê*

Sau khi hoàn thành việc thu thập dữ liệu từ Điện máy xanh, chúng tôi sẽ tiến hành phân tích và đánh giá dữ liệu bằng cách sử dụng các phương pháp thống kê. Các phương pháp thống kê bao gồm việc phân tích đơn biến để mô tả các đặc tính của dữ liệu như giá sản phẩm, danh mục hoặc phân tích tương quan giữa các yếu tố khác nhau. Các phép tính này sẽ giúp tôi đưa ra nhận định về xu hướng thị trường và sự biến động của giá cả trong các danh mục sản phẩm khác nhau.

*1.5.4 Phương pháp thực nghiệm*

Chúng tôi sẽ sử dụng Scrapy để tiến hành thực nghiệm thu thập dữ liệu từ sàn thương mại điện tử Điện máy xanh. Các bước thực nghiệm bao gồm việc thu thập thông tin sản phẩm từ các danh mục cụ thể và đồng thời đảm bảo quá trình thu thập diễn ra một cách liên tục và ổn định. Chúng tôi cũng sẽ tiến hành đánh giá kết quả thu thập, so sánh dữ liệu thu được với dữ liệu thực tế và phân tích hiệu quả của quá trình này.

*1.5.5 Phương pháp đánh giá*

Chúng tôi sẽ thực hiện đánh giá hiệu quả của việc sử dụng Scrapy bằng cách so sánh độ chính xác của dữ liệu thu thập được với dữ liệu thực tế trên trang Điện máy xanh. Các chỉ số về thời gian thu thập, tần suất truy cập, và chất lượng dữ liệu sẽ được phân tích để đề xuất các hướng cải tiến.

**1.6 Những đóng góp nghiên cứu của đề tài**

*1.6.1 Trong lĩnh vực học thuật*

Đề tài đóng góp một phần đáng kể trong việc mở rộng hiểu biết về việc thu thập dữ liệu tự động trên các sàn thương mại điện tử bằng công cụ Scrapy. Bằng cách áp dụng Scrapy vào việc thu thập và xử lý dữ liệu từ Điện máy Xanh, nghiên cứu này cung cấp một ví dụ thực tiễn về khả năng của các công cụ mã nguồn mở trong lĩnh vực thu thập dữ liệu web. Kết quả của nghiên cứu cũng đóng góp vào việc phát triển một cơ sở dữ liệu phong phú về các sản phẩm điện tử, từ đó tạo nền tảng cho các nghiên cứu khác về thương mại điện tử, phân tích hành vi tiêu dùng và ứng dụng học máy. Đồng thời, nghiên cứu này cũng bổ sung thêm tài liệu về phương pháp và quy trình thu thập dữ liệu hiệu quả, phục vụ cho các nghiên cứu trong tương lai.

*1.6.2 Trong thực tiễn kinh doanh*

Đề tài không chỉ dừng lại ở nghiên cứu lý thuyết mà còn mang tính thực tiễn cao. Doanh nghiệp có thể sử dụng dữ liệu này để phân tích thị trường, xác định xu hướng tiêu dùng và từ đó phát triển các chiến lược kinh doanh hiệu quả hơn. Đặc biệt, thông tin về giá giúp các công ty tối ưu hóa quy trình định giá, cải thiện chiến lược tiếp thị và tăng khả năng cạnh tranh. Những kết quả này có thể dẫn đến tăng doanh thu, cải thiện hiệu quả kinh doanh và vị thế vững chắc hơn trên thị trường thương mại điện tử.

**Chương 2: CƠ SỞ LÝ THUYẾT**