|  |  |
| --- | --- |
| A blue and red text on a white background  Description automatically generated | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM** |

**ĐỒ ÁN MÔN HỌC**

**TÌM HIỂU CÔNG CỤ MÃ NGUỒN MỞ SCRAPY VÀ THU THẬP DỮ LIỆU CÁC SẢN PHẨM XU HƯỚNG CỦA FAHASHA**

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 : 2286400028\_Nguyễn Thị Thanh Tâm

2286400009\_Bùi Minh Huy

2286400007\_Trần Tuấn Đạt

TP. Hồ Chí Minh, 2024

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . .

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

TPHCM, Ngày......tháng......năm 2024

**Giáo viên hướng dẫn**

(Ký tên, đóng dấu)

LỜI CAM ĐOAN

Tôi, Bùi Minh Huy, Trần Tuấn Đạt, Nguyễn Thị Thanh Tâm xin cam đoan rằng:

Mọi thông tin và nghiên cứu được trình bày trong bài báo cáo này là trung thực và khách quan được thu thập và phân tích một cách cẩn thận dựa trên các nguồn chính thống và đáng tin cậy.

Bất kỳ thông tin hoặc ý kiến nào được trích dẫn từ các nguồn khác đều được nêu rõ nguồn gốc và được trích dẫn theo đúng quy định. Tôi cam đoan rằng không có bất kỳ sự sao chép hoặc sử dụng thông tin không đúng đắn nào từ các nguồn khác.

Bài báo cáo này là công trình nghiên cứu độc lập của tôi chưa từng được công bố ở bất kỳ nơi nào khác. Tôi cam đoan rằng đã tuân thủ đầy đủ các quy tắc và quy định của môn học bao gồm cả việc tham khảo và sử dụng công cụ nghiên cứu.

Tôi hy vọng rằng bài báo cáo này sẽ cung cấp một cái nhìn tổng quan rõ ràng và toàn diện về chủ đề “TÌM HIỂU CÔNG CỤ MÃ NGUỒN MỞ SCRAPY VÀ THU THẬP DỮ LIỆU CÁC SẢN PHẨM XU HƯỚNG CỦA FAHASHA"và sẽ đóng góp một phần nhỏ vào lĩnh vực nghiên cứu này.

TPHCM, ngày 09 tháng 06 năm 2024

Sinh viên

Bùi Minh Huy

Trần Tuấn Đạt

Nguyễn Thị Thanh Tâm

**CHƯƠNG 1: TỔNG QUAN**

* 1. **Giới thiệu đề tài**

Trong bối cảnh thương mại điện tử ngày càng phát triển, việc tối ưu hóa hiệu quả kinh doanh của các cửa hàng trực tuyến trở nên vô cùng quan trọng. Các doanh nghiệp không chỉ cần hiểu rõ về khách hàng mà còn phải nhận diện được các yếu tố ảnh hưởng đến thành công hoặc thất bại của các cửa hàng. Việc phân tích và phân cụm dữ liệu cửa hàng trực tuyến giúp xác định các nhóm cửa hàng có đặc điểm và hiệu suất kinh doanh tương đồng, từ đó đề xuất các chiến lược phát triển phù hợp. Đề tài "TÌM HIỂU CÔNG CỤ MÃ NGUỒN MỞ SCRAPY VÀ THU THẬP DỮ LIỆU CÁC SẢN PHẨM XU HƯỚNG CỦA FAHASHA" được chọn nhằm mục tiêu cung cấp một cách tiếp cận khoa học và có hệ thống để hỗ trợ doanh nghiệp trong việc cải thiện hiệu quả hoạt động và tăng cường cạnh tranh trên thị trường.

* 1. **Nhiệm vụ của đề tài**

Nhiệm vụ của đề tài "TÌM HIỂU CÔNG CỤ MÃ NGUỒN MỞ SCRAPY VÀ THU THẬP DỮ LIỆU CÁC SẢN PHẨM XU HƯỚNG CỦA FAHASA" là quá trình áp dụng các công cụ mã nguồn mở để thu thập và xử lý dữ liệu nhằm tìm ra các sản phẩm của nhà sách. Từ đó, phân tích và tìm ra các đặc điểm chung của từng nhóm. Điều này giúp những khách hàng có thể biết được những thông tin cần thiết trước khi mua sản phẩm, qua đó có thể tối ưu hóa chiến lược marketing và bán hàng, cải thiện dịch vụ khách hàng và trải nghiệm người dùng, và tối ưu hóa hoạt động kinh doanh.

* + 1. *Tính cấp thiết của đề tài*

Thương mại điện tử đang ngày càng phát triển một cách mạnh mẽ do nhu cầu mua hàng online của người Việt đang ngày càng gia tăng. Do đó các doanh nghiệp không chỉ cạnh tranh về sản phẩm mà còn phải phân tích xu hướng nhu cầu tiêu dùng của khách hàng thông qua dữ liệu lớn(Big Data). Vì vậy, việc thu thập dữ liệu sản phẩm sẽ giúp doanh nghiệp nhận diện và đáp ứng kịp thời nhu cầu của người tiêu dùng.

Việc nắm bắt các loại sách đang được yêu thích cho phép các nhà bán lẻ và người tiêu dùng đưa ra quyết định quan trọng về quản lý hàng, lên kế hoạch mua sắm hợp lý, và tối ưu hoá các chiến lược tiếp thị, bao gồm khuyến mãi, quảng cáo, và định giá sản phảm. Công cụ mã nguồn mở Scrapy giúp tự động hoá quy trình thu thập dữ liệu, tiết kiệm thời gian. Chúng ta có thể nhanh chóng thu thập thông tin hàng hoá từ nhiều trang web khác nhau mà không tốn kém chi phí thuê nhân sự để thực hiện việc này. Trong môi trường kinh doanh hiện đại, quyết định dựa trên dữ liệu là yếu tố sống còn. Việc sử dụng Scrapy để thu thập dữ liệu các sản phẩm xu hướng giúp doanh nghiệp, nhà quản lý, và các nhà nghiên cứu dựa vào dữ liệu để đưa ra các quyết định chiến lược chính xác hơn về sản phẩm, giá cả và thị trường.

Dữ liệu từ các sản phẩm xu hướng có thể được phân tích để dự báo nhu cầu trong tương lai. Các nhà nghiên cứu hoặc doanh nghiệp có thể sử dụng dữ liệu này để phát triển các mô hình dự báo nhằm tối ưu hóa sản xuất, nhập khẩu và phân phối sản phẩm một cách hiệu quả hơn.

Với những lý do trên, đề tài nghiên cứu về việc sử dụng Scrapy để thu thập dữ liệu các sản phẩm xu hướng của Fahasa không chỉ mang tính cấp thiết đối với ngành thương mại điện tử mà còn mở ra nhiều tiềm năng trong việc phân tích và dự báo thị trường.

* 1. **Mục tiêu** 
     1. *Mục tiêu tổng quát*

Đề tài nhằm cung cấp một phương pháp thu thập và phân tích dữ liệu thu được từ các loại sách được yêu thích của nhà sách Fahasa, giúp khách hàng có thể nhắm bắt được những thông tin cần thiết trước khi mua sản phẩm. Từ đó, đề xuất các chiến lược phát triển phù hợp để tối ưu hóa hiệu quả kinh doanh và nâng cao năng lực cạnh tranh cho các doanh nghiệp.

* + 1. *Mục tiêu cụ thể*

Vớibài nghiên cứu này, chúng tôi sử dụng công cụ Scrapy, một framework thu thập dữ liệu mã nguồn mở, để tự động thu thập thông tin về các sản phẩm xu hướng trên trang web Fahasa, bao gồm tên sản phẩm, giá cả, đánh giá của khách hàng và các thông tin liên quan khác. Thu thập và xử lý dữ liệu từ trang Fahasa, đảm bảo dữ liệu thu thập được đầy đủ và nhất quán để phục vụ cho các bước phân tích và nghiên cứu tiếp theo. Sử dụng Scrapy để thiết lập spider thu thập dữ liệu từ các trang sản phẩm bán chạy, đồng thời xây dựng các pipeline để lưu trữ và quản lý dữ liệu thu thập được. Thực hiện các quá trình tự động hóa việc thu thập dữ liệu từ Fahasa thông qua việc lập trình spider để lấy thông tin sản phẩm từ các trang có sản phẩm xu hướng. Phân tích dữ liệu sản phẩm thu thập được từ Fahasa để nhận diện các xu hướng nổi bật dựa trên thông tin về sản phẩm, giá cả và đánh giá của người tiêu dùng. Đồng thời, đánh giá hiệu quả của Scrapy trong việc thu thập dữ liệu. Sau khi hoàn tất thu thập và xử lý dữ liệu, tôi sẽ tiến hành phân tích kết quả để đưa ra các nhận định về xu hướng thị trường của Fahasa. Từ đó, đề xuất các chiến lược kinh doanh dựa trên dữ liệu thực tế và xu hướng thị trường.

* 1. **Đối tượng và phạm vi**
     1. *Đối tượng*

Đối tượng nghiên cứu của đề tài này là các loại sách được yêu thích nhiều nhất và đang xu hướng. Trang web bán hàng của nhà sách Fahasa bao gồm nhiều thông tin như: tên sản phẩm, giá cả, đánh giá của khách hàng và các thông tin liên quan khác. Bằng cách thu thập và phân tích dữ liệu từ trang web bán hàng của nhà sách, đề tài giúp các khách hàng và doanh nghiệp có thể nhắm bắt rõ hơn về các thông tin của nhiều loại sách được bán ra.

* + 1. *Phạm vi*

Đề tài tập trung vào thu thập và phân tích dữ liệu từ các trang web bán hàng của hiệu sách Fahasa. Nhóm sẽ áp dụng các phương pháp thu thập và phân tích dữ liệu để nhận diện các loại sách đang được yêu thích hoặc được mua nhiều nhất. Dựa trên kết quả này, chúng tôi sẽ đề xuất các chiến lược kinh doanh phù hợp và đánh giá hiệu quả của chúng. Điều này giúp mang lại giá trị lý thuyết và thực tiễn cho các doanh nghiệp trên thị trường thương mại điện tử.

* 1. **Phương pháp nghiên cứu**
     1. *Phương pháp nghiên cứu sơ bộ*

Trước khi tiến hành thu thập dữ liệu, nhóm sẽ tiến hành một nghiên cứu sơ bộ để hiểu rõ hơn về lĩnh vực nghiên cứu và các yếu tố quan trọng liên quan. Nghiên cứu này bao gồm việc tìm hiểu về thương mại điện tử, các yếu tố ảnh hưởng đến hiệu suất kinh doanh của cửa hàng trực tuyến, và các phương pháp phân tích dữ liệu phổ biến. Thông qua việc nghiên cứu sơ bộ, nhóm sẽ xác định các vấn đề cụ thể cần giải quyết và đề xuất các phương pháp nghiên cứu phù hợp.

* + 1. *Phương pháp nghiên cứu tài liệu*

**Nhóm sẽ tiến hành nghiên cứu tài liệu để thu thập thông tin về các phương pháp và công cụ phân tích dữ liệu trong lĩnh vực thương mại điện tử và mã nguồn mở. Qua việc đánh giá các nghiên cứu trước đây và các công trình khoa học liên quan, chúng em sẽ xác định các phương pháp thu thập dữ liệu và phân tích xu hướng sản phẩm hiệu quả nhất, đặc biệt là các phương pháp sử dụng công cụ Scrapy. Sau đó, nhóm sẽ áp dụng các phương pháp này vào việc thu thập và phân tích dữ liệu sản phẩm xu hướng từ trang web Fahasa để đưa ra các đề xuất chiến lược kinh doanh.**

* + 1. ***Phương pháp nghiên cứu thống kê***

Trong quá trình phân tích dữ liệu, nhóm sẽ sử dụng các phương pháp thống kê để mô tả và phân tích các biến số quan trọng liên quan đến các sản phẩm xu hướng của Fahasa. Các phương pháp thống kê bao gồm phân tích đơn biến, phân tích đa biến, phân tích phương sai và kiểm tra độ tương quan giữa các biến. Thông qua việc áp dụng các phương pháp này, nhóm sẽ đánh giá mối quan hệ giữa các biến số như giá cả, đánh giá của khách hàng và mức độ phổ biến của sản phẩm. Từ đó, nhóm có thể xác định các yếu tố ảnh hưởng đến xu hướng mua hàng của người tiêu dùng và hiệu quả kinh doanh của nhà sách trực tuyến Fahasa.

* + 1. ***Phương pháp thực nghiệm***

**Nhóm** sẽ tiến hành thực nghiệm trên dữ liệu thu thập được từ trang thương mại điện tử Fahasa. Quá trình này bao gồm việc tiền xử lý dữ liệu, áp dụng các phương pháp phân tích để nhận diện các xu hướng sản phẩm, và đánh giá hiệu quả của các chiến lược kinh doanh được đề xuất. Thông qua việc thực nghiệm thực tế, chúng tôi sẽ kiểm tra và đảm bảo tính khả thi cũng như hiệu quả của phương pháp nghiên cứu trong việc tối ưu hóa hiệu quả kinh doanh và nâng cao khả năng cạnh tranh của nhà sách trực tuyến Fahasa.

* + 1. ***Phương pháp đánh giá***

Cuối cùng, chúng tôi sẽ thực hiện phương pháp đánh giá để đo lường hiệu quả của các phương pháp phân tích dữ liệu đã áp dụng. Quá trình này bao gồm việc so sánh các chỉ số và thước đo về hiệu quả kinh doanh giữa các nhóm sản phẩm xu hướng từ Fahasa. Chúng tôi sẽ đánh giá các chỉ số như mức độ phổ biến của sản phẩm, giá cả, và phản hồi từ khách hàng để xác định mức độ hiệu quả của các chiến lược kinh doanh được đề xuất.

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

* 1. **SCARPY.**
     1. ***Giới thiệu về trích xuất dữ liệu từ Srcapy***

Scrapy là một framework mạnh mẽ và linh hoạt được phát triển bằng Python, chuyên dùng để trích xuất dữ liệu từ các trang web. Scrapy hỗ trợ người dùng xây dựng các spider, là các chương trình có khả năng thu thập dữ liệu một cách tự động và có tổ chức từ nhiều nguồn web khác nhau. Công cụ này đặc biệt hữu ích trong việc thu thập dữ liệu để phục vụ cho các mục đích như nghiên cứu, phân tích dữ liệu, xây dựng các hệ thống dự đoán và nhiều ứng dụng khác liên quan đến việc thu thập thông tin trên mạng.

Quy trình hoạt động của Scrapy dựa trên mô hình lặp đi lặp lại, nơi các spider sẽ gửi yêu cầu (HTTP requests) đến một URL và sau đó xử lý phản hồi (HTTP responses) để trích xuất thông tin mong muốn từ HTML hoặc XML. Người dùng có thể dễ dàng xác định những phần dữ liệu nào cần được lấy bằng cách sử dụng các quy tắc truy vấn như XPath hoặc CSS selectors.

* + 1. *Ưu điểm và hạn chế*

Ưu điểm:

* **Hiệu suất cao và nhanh chóng**: Scrapy được thiết kế để xử lý nhiều yêu cầu đồng thời (concurrent requests) giúp tiết kiệm thời gian khi thu thập dữ liệu từ nhiều trang web cùng một lúc.
* **Tích hợp sẵn các công cụ mạnh mẽ**: Scrapy cung cấp một loạt các công cụ hỗ trợ như việc xử lý yêu cầu, điều hướng trang web, và trích xuất dữ liệu, giúp người dùng không phải cài đặt thêm nhiều thư viện bên ngoài.
* **Dễ dàng mở rộng và tùy chỉnh**: Với Scrapy, người dùng có thể dễ dàng tùy chỉnh spider hoặc các tính năng khác để đáp ứng yêu cầu cụ thể của dự án. Framework này cũng hỗ trợ việc tạo ra các middleware để quản lý yêu cầu và phản hồi một cách linh hoạt.
* **Quản lý tốt các lỗi và ngoại lệ**: Scrapy có cơ chế tự động xử lý các lỗi thường gặp như lỗi kết nối hoặc phản hồi không mong muốn, đảm bảo quá trình thu thập dữ liệu ít bị gián đoạn.
* **Hỗ trợ việc lưu trữ dữ liệu**: Scrapy cung cấp các pipelines để xử lý dữ liệu trước khi lưu trữ. Dữ liệu có thể được xuất ra nhiều định dạng khác nhau như JSON, CSV, XML, hoặc lưu trực tiếp vào cơ sở dữ liệu.
* **Mã nguồn mở và có cộng đồng lớn**: Là một dự án mã nguồn mở, Scrapy được cộng đồng lập trình viên hỗ trợ nhiệt tình. Người dùng có thể tìm thấy nhiều tài liệu, các gói mở rộng, và sự hỗ trợ kỹ thuật từ cộng đồng.

Hạn chế:

* **Cài đặt và cấu hình ban đầu phức tạp**: Đối với người mới, việc cài đặt và cấu hình Scrapy có thể hơi phức tạp, đặc biệt là khi thiết lập môi trường phát triển hoặc quản lý các yêu cầu đồng thời.
* **Chỉ hoạt động với Python**: Scrapy chỉ hỗ trợ Python, do đó những ai quen với các ngôn ngữ lập trình khác sẽ phải học Python trước khi sử dụng.
* **Không phải lúc nào cũng thân thiện với các trang web động**: Scrapy hoạt động tốt với các trang tĩnh, nhưng khi gặp các trang web động sử dụng JavaScript để tải nội dung, Scrapy có thể gặp khó khăn trong việc trích xuất dữ liệu. Trong trường hợp này, người dùng thường phải tích hợp thêm Selenium hoặc Splash để xử lý.
* **Quản lý trạng thái phức tạp**: Đối với các trang web yêu cầu duy trì trạng thái như phiên đăng nhập (session) hoặc cookie, Scrapy yêu cầu phải cấu hình kỹ lưỡng, và đôi khi không thuận tiện bằng các công cụ khác như Selenium.
* **Không phù hợp cho các dự án nhỏ**: Đối với những dự án thu thập dữ liệu nhỏ hoặc ngắn hạn, việc sử dụng Scrapy có thể là quá phức tạp và nặng nề so với các công cụ nhẹ hơn như Beautiful Soup.