|  |  |
| --- | --- |
| 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 TRANG WEB**

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 TRANG WEB"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**

Các sàn thương mai điện tử đang càng phát triển , cũng đang trở thành xu hướng mới . Với hình thức kinh doanh online dần trở nên phổ biến . Phân tích sản phẩm cảu các cửa hàng trở nên quan trọng trong việc phát triển của các doanh nghiệp . Thu thập dữ liệu có thể giúp ta thu được các thông tin về sản phẩm được bán sau đó đưa ra chiến lược để hiểu rõ về các mặt hàng tiêu dùng hiện nay . Dữ liệu đưa ra giúp ta xây dựng được mô hình kinh doanh phù hợp và tăng năng xuất hiệu quả . Đề tài "TÌM HIỂU CÔNG CỤ MÃ NGUỒN MỞ SCRAPY VÀ THU THẬP VÀ PHÂN TÍCH DỮ LIỆU CÁC SẢN PHẨM CỦA WEB (BOOK TO SCRAP) " là dự án nhằm cung cấp các dữ liệu trong một trang web kinh doanh online , thu thập các sản phẩm nhanh chóng từ công cụ mã nguồn mở .

* 1. **Nhiệm vụ của đề tài**

Nhiệm vụ của đề tài là áp dụng các công cụ mã nguồn mở để thu thập và phân tích dữ liệu nhằm tìm ra các sản phẩm phổ biến . Thu thập và phân tích dữ liệu là nhiệm vụ quan trọng đối với doanh nghiệp muốn phát triển trên thị trường có thể nắm bắt được thông tin quan trọng để tối ưu được cách hoạt động và đạt được lợi nhuận cao .

* + 1. *Tính cấp thiết của đề tài*

Thương mại điện tử ngày càng phát triển hơn , nhu cầu mua sắm trực truyến cũng tăng cao . Đi theo phát triển này theo sau đó là các doanh nghiệm phải luôn cạnh tranh nhau về chất lượng của sản phẩm , hiểu rõ xu hướng và nhu cầu của người tiêu dùng . Từ đó có thể nắm bắt được cách hoạt động để phát triển trên thị trường hiện nay .

Để hiểu rõ được thị trường , nhà quản lý phải lập được kế hoạch và tối ưu được nhiều chiến lược bao gồm như các chương trình khuyến mãi , quảng cáo hay giá của sản phẩm hợp lý hơn với người sử dụng . Công cụ mã nguồn mở Scrapy là công cụ để thu thập các dữ liệu một cách tiết kiệm nhanh nhất , không tốn nhiều thời gian và chi phí rẻ hơn thuê nhân sự . Các doanh nghiệp có thể thu thập được dữ liệu nhanh chóng và hiệu quả từ các trang web mà không cần tới nhiều nhân sự .

Kinh doanh hiện nay chủ yếu là từ internet , dữ liệu là yếu tố cần thiết và đưa ra quyết định đúng cho doanh nghiệp . Công cụ Scrapy được dùng để thu thập dữ liệu cho các doanh nghiệp và nhà quản lý có thể đưa ra được các chiến lược tối ưu hơn về cách hoạt động và còn nâng cao cạnh tranh trên thị trường . Dữ liệu thu được sẽ giúp doanh nghiệp hay nhà quản lý có quyết định chính xác hơn về sản phẩm hay giá cả so với thị trường .

Ngoài đó , dữ liệu vẫn có thể sử dụng để biết được nhu câu trong tương lai . Có thể phát triển các mô hình dự báo trên các dữ liệu đã thu thập được , nhằm giảm thiếu rủi ro trong hoạt động kinh doanh .

Với những lý do đã nêu , việc nghiên cứu và sử dụng Scrapy để thu thập dữ liệu mang tính cấp thiết khi thương mại điện tử đang phát triển mạnh , còn là tiềm năng trong việc phân tích và dự báo giúp doanh nghiệp đưa ra được các chiến lược tối ưu hơn trong cách hoạt động .

* + 1. Ý nghĩa khoa học và thực tiễn của đề tài

Ý nghĩa khoa học : Scrapy được viết bằng Python , thực hiện đề tài cũng sẽ được xem là có cơ hội để cải thiện kỹ năng lập trình ngôn ngữ Python. Scrapy mạnh trong việc xây dựng các chương trình tự động hóa thu thập dữ liệu. Hiểu rõ về scrapy giúp người học hiểu rõ hơn về quá trình thu thập dữ liệu của trang web.

Ý nghĩa thực tiễn :Thu thập thông tin bình thường sẽ tốn nhiều thời gian dường như không hiệu quả , công cụ Scrapy giúp tự động hóa quá trình này giúp tiết giận tơig gian và giảm nhân lực. Ngoài ra , ứng dụng còn giúp xử lý các dữ liệu lớn nhanh ,dữ liệu thu thập được sẽ giúp doanh nghiệp hiểu rõ hơn về nhu cầu của người tiêu dùng . Đưa ra các chiến lược tối ưu hiệu quả cho các nhà quản lý , giúp các doanh nghiệp dễ dàng điều chỉnh và tối ưu sản phẩm . Dữ liệu thu thập được từ các sản phẩm , giúp doanh nghiệp phân tích thông qua các từ khóa mua sắm nhiều nhằm mục tiêu tới cảm nhận của người tiêu dùng

* 1. **Mục tiêu**

1.3.1 Mục tiêu tổng quát

Đề tài cung cấp một phương pháp thu thập và phân tích dữ liệu thu được từ sản phẩm phổ biến giúp khách hàng có thể nắm bắt được những thông tin cần thiết trước khi mua sản phẩm. Sau đó, đề xuất các chiến lược phù hợp để tối ưu hóa cách hoạt động cho doanh nghiệp và đưa ra kế hoạch kinh doanh hiệu quả .

* + 1. *Mục tiêu cụ thể*

Vớibài nghiên cứu này, công cụ Scrapy là 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, 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 web , đảm bảo dữ liệu thu thập được đầy đủ 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 trang web 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 trang web 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 trang web . 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 sản phẩm được yêu thích nhiều và đang xu hướng. Trang web bán hàng 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 sản phẩm được bán ra.

* + 1. *Phạm vi*

Phạm vi tập trung vào thu thập và phân tích dữ liệu từ các trang web bán hàng. Đề tài sẽ áp dụng các phương pháp thu thập và phân tích dữ liệu để nhận diện các sản phẩm đang được yêu thích hoặc được mua nhiều nhất. Dựa trên kết quả này, sẽ đề xuất các chiến lược tối ưu cách hoạt động cho các nghiệp các kinh doanh phù hợp và đạt hiệu quả cao hơn . Điều này giúp mang lại lợi ích cho các doanh nghiệp đưa ra được các kinh doanh phù hợp .

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

Trước khi tiến hành thu thập dữ liệu , để đảm bảo được việc thu thập và phân tích nhóm phải xác định được phạm vi của đề tài bao gồm việc tìm hiểu về thương mại đại tử , yếu tố ảnh hướng đến kinh doanh của cửa hàng trực tuyến và phương pháp phân tích phổ biến . Qua đó , nhóm có thể đưa ra được phương pháp phù hợp với đề tài để có thể giải quyết được các vấn đề nghiên cứu bằng cách thu thập dữ liệu sao cho hiệu quả nhất .

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

**Nhóm sẽ tiến hành nghiên cứu tài liệu thông qua các bài báo hay các nghiên cứu đã có trước đó 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, nhóm sẽ xác định các phương pháp thu thập dữ liệu và phân tích xu hướng sản phẩm phù hợp 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 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 trang web . 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 trang web .

* + 1. ***Phương pháp thực nghiệm***

**Nhóm** sẽ tiến hành thực nghiệm trên dữ liệu thu thập được từ trang thương mại điện tử. 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, nhóm 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 trang web.

* + 1. ***Phương pháp đánh giá***

Sau khi đã thu thập và phân tích , nhóm sẽ thực hiện phương pháp đánh giá để đo được hiệu quả các dữ liệu đã được phân tích . Quá trình này bao gồm việc so sách doanh số hay độ phổ biến của sản phẩm đối với khách hàng để đưa ra chiến lược kinh doanh hiệu quả . Nhóm sẽ đánh giá để kiểm tra dữ liệu so với thực tế từ đó đưa ra được các giải pháp cụ thể để cải thiện .

**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.
  + 1. *. Ứng dụng*

Công cụ Scrapy ứng dụng được trong nhiều lĩnh vực , là một mã nguồn mở mạnh mẽ thường dùng để thu thập dữ liệu từ các trang web . Scrapy dùng để thu thập dữ liệu và có thể ứng dụng được nhiều lĩnh vực giúp cho việc thu thập và phân tích dữ liệu đạt hiệu quả . Một số ứng dụng của Scrapy :

* Ứng dụng vào thương mại điện tử : Scrapy sử dụng để thu thập cái thông tin về các web thương mại . Từ đó phân tích ra chiến lược cho nhà quản lý để đưa ra cách hoạt động phù hợp .
* Phân tích xu hướng của sản phẩn : Thu thập các dữ liệu đang xu hướng trên thị trường sau đó dự đoán đưa ra chiến lược kinh doanh hợp lý và hiểu hơn về những sản phẩm đang được người tiêu dùng ưa chuộng.
* Thu thập các đánh giá từ người tiêu dùng : Scrapy được dùng để thu thập dữ liệu , từ các dữ liệu thu được sẽ giúp nhà quản lý phân tích được mức độ hài lòng của khách hơn , sau đó xem xét và điều chỉnh .
* Tiết kiệm chi phí : Quét web giúp thu được dữ liệu lớn và có giá trị theo cách rẻ hơn và nhanh hơn nhiều so với việc người ta phải thực hiện thủ công. [1]
* Tối ưu hóa giá động : những người hoạt động trong lĩnh vực thương mại điện tử và du lịch có thể theo dõi giá của đối thủ bất cứ lúc nào và điều chỉnh giá của họ liên quan đến điều này. Điều này là do những thứ công nghệ thú vị đó, việc định giá linh hoạt thông qua việc thu thập dữ liệu giúp các công ty có thể đưa ra mức giá hợp lý, đồng thời tăng doanh thu của họ.[1]
  1. **PYTHON .**
     1. **. Giới thiệu về ngôn ngữ Python .**

Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (ML). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Phần mềm Python được tải xuống miễn phí, tích hợp tốt với tất cả các loại hệ thống và tăng tốc độ phát triển.[2]

Python **được thiết kế với tư tưởng giúp người học dễ đọc, dễ hiểu và dễ nhớ**; vì thế ngôn ngữ Python có hình thức rất clear, cấu trúc rõ ràng, thuận tiện cho người mới học. Cấu trúc của Python cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu, nói cách khác thì so với các ngôn ngữ lập trình khác, chúng ta có thể sử dụng ít dòng code hơn để viết ra một chương trình trong Python.[3]

* + 1. *. Ưu điểm và hạn chế*

Ưu điểm :

* Cú pháp đơn giản :

[1] Schmidt A., “Tại sao doanh nghiệp có thể sử dụng tính năng Quét web để thu thập dữ liệu? - ProxyLa bàn.” Accessed: Oct. 20, 2024. [Online]. Available: https://proxycompass.com/vi/why-might-a-business-use-web-scraping-to-collect-data/

[2] “Python là gì? - Giải thích về ngôn ngữ Python - AWS,” Amazon Web Services, Inc. Accessed: Oct. 20, 2024. [Online]. Available: https://aws.amazon.com/vi/what-is/python/

[3] TopDev, “Python là gì? Tổng hợp kiến thức cho người mới bắt đầu,” TopDev. Accessed: Oct. 20, 2024. [Online]. Available: https://topdev.vn/blog/python-la-gi/