|  |  |
| --- | --- |
| 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

|  |  |
| --- | --- |
| 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

**MỤC LỤC**

[CHƯƠNG 1: TỔNG QUAN 7](#_Toc181109052)

[1.1 Giới thiệu đề tài 7](#_Toc181109053)

[1.2 Nhiệm vụ của đề tài 7](#_Toc181109054)

[*1.2.1* *Tính cấp thiết của đề tài* 7](#_Toc181109055)

[*1.2.2* *Ý nghĩa khoa học và thực tiễn của đề tài* 8](#_Toc181109056)

[1.3 Mục tiêu 8](#_Toc181109057)

[*1.3.1 Mục tiêu tổng quát* 8](#_Toc181109058)

[*1.3.2* *Mục tiêu cụ thể* 9](#_Toc181109059)

[1.4 Đối tượng và phạm vi 9](#_Toc181109060)

[*1.4.1* *Đối tượng* 9](#_Toc181109061)

[*1.4.2* *Phạm vi* 9](#_Toc181109062)

[1.5 Phương pháp nghiên cứu 9](#_Toc181109063)

[*1.5.1* *Phương pháp nghiên cứu sơ bộ* 9](#_Toc181109064)

[*1.5.2* *Phương pháp nghiên cứu tài liệu* 10](#_Toc181109065)

[*1.5.3* *Phương pháp nghiên cứu thống kê* 10](#_Toc181109066)

[*1.5.4* *Phương pháp thực nghiệm* 10](#_Toc181109067)

[*1.5.5* *Phương pháp đánh giá* 10](#_Toc181109068)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 12](#_Toc181109069)

[2.1 SCARPY 12](#_Toc181109070)

[*2.1.1* *Giới thiệu về trích xuất dữ liệu từ Srcapy* 12](#_Toc181109071)

[*2.1.2* *Ưu điểm và hạn chế* 12](#_Toc181109072)

[*2.1.3* ***. Ứng dụng*** 14](#_Toc181109073)

[2.1.3.1 Ứng dụng thực tiễn 14](#_Toc181109074)

[2.1.3.2 Các bài báo về Scrapy 14](#_Toc181109075)

[*2.1.4* *Cấu trúc cơ bản của Sprider trong Scrapy* 14](#_Toc181109076)

[2.1.4.1 Spider là gì ? 14](#_Toc181109077)

[2.2 PYTHON 14](#_Toc181109078)

[*2.2.1* *. Giới thiệu về ngôn ngữ Python .* 14](#_Toc181109079)

[*2.2.2* *. Ưu điểm và hạn chế* 15](#_Toc181109080)

[*2.2.3* *Ứng dụng* 16](#_Toc181109081)

[2.3 MongoDB 16](#_Toc181109082)

[*2.3.1. Giới thiệu về MongoDB* 16](#_Toc181109083)

[*2.3.2 Ưu điểm và hạn chế* 17](#_Toc181109084)

[*2.3.3 . Ứng dụng* 18](#_Toc181109085)

[CHƯƠNG 3: PHƯƠNG PHÁP THỰC NGHIỆM 19](#_Toc181109086)

[3.1 Phương pháp thu thập dữ liệu và lưu trữ dữ liệu 19](#_Toc181109087)

[*3.1.1. Truy xuất thông tin của trang web* 19](#_Toc181109088)

[*3.1.2. Thu thập dữ liệu sản phẩm* 19](#_Toc181109089)

[3.1.2.1. Truy xuất link của từng sản phẩm 19](#_Toc181109090)

[3.1.2.2. Thu thập dữ liệu từng thuộc tính của sản phẩm 20](#_Toc181109091)

[3.1.2.3. Chuẩn hoá dữ liệu 22](#_Toc181109092)

[*3.1.3 Tích hợp MongoDB để lưu trữ dữ liệu* 23](#_Toc181109093)

[*3.1.4 Xuất dữ liệu ra file Excel* 24](#_Toc181109094)

[3.2 Mô tả dữ liệu 24](#_Toc181109095)

[3.3 Sơ đồ quá trình thu thập và xử lý dữ liệu 25](#_Toc181109096)

[CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM 27](#_Toc181109097)

[4.1 Giới thiệu 27](#_Toc181109098)

[4.2 Kết quả thu thập dữ liệu 27](#_Toc181109099)

[4.3 Phân tích dữ liệu sản phẩm 27](#_Toc181109100)

[4.4 Đánh giá hiệu quả của Scrapy 27](#_Toc181109101)

[4.5 Khó khăn và hạn chế 28](#_Toc181109102)

[4.6 Khuyến nghị để cải thiện 28](#_Toc181109103)

[4.7 Kết luận 29](#_Toc181109104)

[CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ 29](#_Toc181109105)

[5.1 Kết luận 29](#_Toc181109106)

[5.2 Kiến nghị 30](#_Toc181109107)

[5.3 Hướng phát triển trong tương lai 31](#_Toc181109108)

# **CHƯƠNG 1: TỔNG QUAN**

* 1. **Giới thiệu đề tài**

Với sự phát triển mạnh mẽ của mô hình kinh doanh online dẫn dến các nền tảng thương mại điện tử đang trải qua sự tăng trưởng đáng kể từ đó tạo ra một xu hướng mới trong thị trường. Việc thu thập dữ liệu và phân tích sản phẩm trên các sàn thương mại điện tử ngày càng trở nên quan trọng đối với sự phát triển của doanh nghiệp. Doanh nghiệp sẽ từ các dữ liệu thu thập được để xây dựng một mô hình kinh doanh tối ưu và nâng cao năng suất tổng thể. Đề tài “ TÌM HIỂU CÔNG CỤ MÃ NGUỒN MỞ SCRAPY VÀ THU THẬP, PHÂN TÍCH DỮ LIỆU SẢN PHẨM WEB (BOOK TO SCRAP)” là dự án nhằm cung cung cấp dữ liệu của các sản phẩm trong một trang web kinh doanh online và phân tích các dữ liệu đã thu thập được bằng sự kết hợp giữa công cụ mã nguồn mở Scrapy và hệ quản lý cơ sở dữ liệu MongoDB.

* 1. **Nhiệm vụ của đề tài**

Mục tiêu của nghiên cứu này là sử dụng các công cụ mã nguồn mở cho mục đích thu thập và kiểm tra dữ liệu để xác định các sản phẩm có xu hướng. Quá trình thu thập và phân tích dữ liệu là một trong những nhiệm vụ quan trọng đối với các doanh nghiệp nhằm mở rộng trong một thị trường cạnh tranh, vì nó cho phép thu thập những thông tin quan trọng để nâng cao hiệu quả hoạt động và đạt được lợi nhuận đáng kể.

* + 1. *Tính cấp thiết của đề tài*

Lĩnh vực thương mại điện tử đang trải qua quá trình phát triển mạnh mẽ, đồng thời dẫn đến nhu cầu của khách hàng tăng cao về trải nghiệm trực tiếp sản phẩm. Trước sự phát triển này, các doanh nghiệp buộc phải cạnh tranh liên tục về chất lượng sản phẩm của họ, đồng thời cố gắng xây dựng chiến lược kinh doanh theo xu hướng và sở thích của người tiêu dùng. Do đó các dữ liệu này giúp các doanh nghiệp xác định được chiến lược kinh doanh hiệu quả để tăng trưởng trong thị trường hiện nay.

Để nằm bắt được thị trường, nhà quản lý phải lập kế hoặc và tối ưu được nhiều chiến lược bao gồm các chương trình khuyến mãi, quảng cáo hay giá của sản phẩm hợp lý hơn để tiếp cận được với người tiêu dùng. Công cụ mã nguồn mở Scrapy là một trong những công cụ để thu thập các dữ liệu một cách hiêu quả nhất, tiết kiệm được thời gian và chi phí nhất. Các doanh nghiệp có thể thu thập được dữ liệu nhanh chóng và chính xác từ các trang web mà không mất quá nhiều nguồn nhân lực.

Hiện nay, internet đóng vai trò quan trọng trong kinh doanh, dữ liệu đóng vai trò quan trọng trong việc đưa ra hướng đi đúng đắn cho doanh nghiệp, công cụ Scrapy được dùng để thu thập dữ liệu cho các doanh nghiệp và nhà quản lý. Thông tin được thu thập sẽ hỗ trợ các doanh nghiệp và quản lý trong việc đưa ra các quyết định chính xác hơn về chất lượng sản phẩm và giá cả cạnh tranh so với thị trường.

Ngoài ra , dữ liệu thu thập được vẫn có thể sử dụng để biết được nhu cầu 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 ngôn ngữ 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 kiệm thời gian và nhân lực. Ngoài ra, công cụ này còn giúp xử lý các dữ liệu lớn một cách nhanh chóng, dữ liệu thu thập được sẽ giúp doanh nghiệp hiểu rõ hơn về nhu cầu của người tiêu dùng. Từ đó 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.Đồng thời Scrapy cũng giúp các doanh nghiệp thu thập được mức độ hài lòng của người tiêu dùng để có thể cải thiện và đưa ra sản phẩm chất lượng hơn đến tay 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 các sản phẩm thu thập được giúp nhà quản lý và doanh nghiệp có thể nắm bắt được thông tin cần thiết. Điều này cho phép họ đưa ra các đề xuất chiến lược kinh doanh phù hợp để tối ưu hoá cách hoạt động cho doanh nghiệp nhằm cải thiện chất lượng sản phẩm và mức giá để phù hợp với người tiêu dùng, từ đó tiếp cận được với số lượng khách hàng và đạt được lợi nhuận tốt hơn.

* + 1. *Mục tiêu cụ thể*

Trong bài nghiên cứu này, công cụ Scrapy là một hệ thống thu thập dữ liệu mã nguồn mở cho phép tự động thu thập thông tin về các sản phẩm xu hướng trên trang web bao gồm tên sản phẩm, loại 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 đủ để phục vụ cho các bước phân tích và nghiên cứu tiếp theo. Dùng Scrapy để thiết lập spider thu thập dữ liệu từ các trang sản phẩm, đồng thời xây dựng các pipeline để lưu trữ và quản lý dữ liệu thu thập được. Quá trình này được thực hiện một cách tự động hoá từ việc thu thập dữ liệu từ trang web thông qua việc lập trình spider để lấy thông tin sản phẩm. Phân tích dữ liệu sản phẩm thu thập được từ trang web để nhận diện các xu hướng nổi bậc dựa trên giá cả, đánh giá của người tiêu dùng và các thông tin khác của sản phẩm. Ngoài ra đề tài còn giúp đánh giá được hiệu quả của Scrapy trong việc thu thập dữ liệu. Sau khi hoàn tất thu thập và xử lý dữ liệu, chúng tôi sẽ tiến hành phân tích kết quả từ đó đưa ra các nhận định về hiệu suất hoạt động của công cụ Scrapy, dựa vào đó chúng tôi cũng sẽ đề xuất để nâng cao hiệu quả trong những lần sử dụng tiếp theo.

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

Đối tượng nghiên cứu của đề tài này chủ yếu tập trung vào công cụ mã nguồn mở Scrapy, đề tài sẽ tập trung vào cách sử dụng và tính năng của Scrapy trong việc thu thập dữ liệu. Bên cạnh đó đề tài cũng sẽ nhắm đến dữ liệu của các sản phẩm có trên trang web được sử dụng để cào dữ liệu.

* + 1. *Phạm vi*

Phạm vi nghiên cứu của đề tài bao gồm việc khám phá các chi tiết về việc cài đặt, cấu hình và sử dụng công cụ Scrapy để thu thập dữ liệu về các sản phẩm trên trang web. Đề tài sẽ xem xét các phương pháp thu thập và xử lý dữ liệu từ cơ bản đến nâng cao, đồng thời xác định các tiêu chuẩn để lựa chọn sản phẩm có chất lượng tốt (ví dụ: giá cả, điểm đánh giá, … ). Ngoài ra, đề tài sẽ bao quát các phương pháp xử lý dữ liệu sau khi thu thập, bao gồm tiền xử lý, lưu trữ và phân tích sơ bộ.

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

Để đảm bảo rằng việc thu thập và phân tích dữ liệu được thực hiện đúng cách, nhóm phải xác định phạm vi đề tài bao gồm kiến thức về công cụ mã nguồn mở, kiến thức về thương mại điện tử, các yếu tố ảnh hưởng đến hoạt động của cửa hàng trực tuyến và các phương pháp phân tích phổ biến. Điều này cho phép nhóm xác định một chiến lược phù hợp với đề tài để thu thập dữ liệu và giải quyết các vấn đề nghiên cứu một cách hiệu quả nhất.

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

**Nhóm sẽ tiến hành xem xét tài liệu bằng cách tìm kiếm thông tin về các phương pháp, công cụ thu thập và phân tích dữ liệu trong lĩnh vực thương mại điện tử. Qua việc xem xét và đánh giá các nghiên cứu trước đây và các công trình khoa học có liên quan, từ đó nhóm sẽ xác định các phương pháp thu thập dữ liệu và phân tích xu hướng sản phẩm tốt nhất. Các phương pháp sử dụng công cụ Scrapy sẽ được ưu tiên nhất. Sau đó nhóm sẽ áp dụng các phương pháp này vào việc thu thập và phân tích dữ liệu sản phẩm từ trang web để đưa ra chiến lược kinh doanh phù hợp nhất.**

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

Trong quá trình phân tích dữ liệu nhóm sẽ mô tả và phân tích các biến số quan trọng liên quan đến các sản phẩm của trang web trong quá trình phân tích dữ liệu. Nhóm sẽ xem xét các mối quan hệ giữa các yếu tố như giá cả, đánh giá của người tiêu dùng và mức độ phổ biến của sản phẩm. Điều này cho phép nhóm xác định các yếu tố ảnh hưởng đến xu hướng mua hàng của khách hàng và hiệu quả kinh doanh của trang web.

* + 1. ***Phương pháp thực nghiệm***

Dữ liệu được thu thập từ trang web thương mại điện tử sẽ được chúng tôi sử dụng để thực hiện thử nghiệm. Nó bao gồm tiền xử lý dữ liệu, phát hiện xu hướng sản phẩm thông qua phân tích và đánh giá hiệu quả của các kế hoạch kinh doanh được đề xuất. Chúng tôi sẽ thực hiện các bài kiểm tra thực tế để kiểm tra tính khả thi và hiệu quả của phương pháp nghiên cứu để tối ưu hóa hiệu quả kinh doanh và nâng cao khả năng cạnh tranh của trang web.

* + 1. ***Phương pháp đánh giá***

Phương pháp đánh giá của đề tài sẽ tập trung vào việc kiểm tra khả năng sử dụng công cụ mã nguồn mở Scrapy trong quá tình thu thập dữ liệu sản phẩm từ trang web và độ chính xác của nó. Đầu tiên, hiệu quả của Scrapy sẽ được chúng tôi đánh giá dựa trên thời gian thực hiện, số lượng tài nguyên hệ thống được sử dụng và khả năng mở rộng, kết hợp khi thu thập dữ liệu từ nhiều nguồn. Độ chính xác và tính nhất quán của dữ liệu được chúng tôi kiểm tra bằng cách so sánh với dữ liệu thực tế và đảm bảo dữ liệu có cấu trúc đồng nhất nhằm dễ dàng xử lý sau khi thu thập. Ngoài ra chúng tôi cũng đánh giá chất lượng của các phương pháp phân tích dữ liệu nhằm đảm bảo kết quả phân tích chính xác và có ý nghĩa. Cuối cùng khả năng áp dụng thực tế của công cụ Scrapy cũng được đánh giá qua tính tái sử dụng của mã nguồn mở và cấu hình cũng như mức độ thân thiện với người sử dụng giúp đảm bảo công cụ này được sử dụng dễ dàng cho các dự án tương tự.

**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***

#### 2.1.3.1 Ứng dụng thực tiễn

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]

#### 2.1.3.2 Các bài báo về Scrapy

* Scrapy-Based Crawling and User-Behavior Characteristics Analysis on Taobao[2]
* Using Web Scraping In A Knowledge Environment To Build Ontologies Using Python And Scrapy[3]
* Design and Implementation of Data Acquisition System Based on Scrapy Technology[4]
* Research on Scrapy-Based Distributed Crawler System for Crawling Semi-structure Information at High Speed[5]
  + 1. *Cấu trúc cơ bản của Sprider trong Scrapy*

2.1.4.1 Spider là gì ?

Spider trong Scrapy là một thành phần cơ bản và quan trọng chịu trách nhiệm thu thập dữ liệu từ các trang web. Nó là một lớp (class) mà bạn định nghĩa trong Scrapy để thực hiện việc thu thập dữ liệu tự động theo các quy tắc mà bạn định nghĩa.

Nguyên lý hoạt động:

* Bắt đầu từ một URL hoặc một danh sách các URL mà bạn cung cấp (thường gọi là start\_urls).
* Gửi yêu cầu (requests) đến các URL này.
* Xử lý phản hồi (response) nhận được từ trang web bằng cách sử dụng các phương thức như parse() để phân tích và trích xuất dữ liệu cần thiết.
* Điều hướng qua các trang tiếp theo, theo các liên kết trong trang hiện tại, để thu thập thêm dữ liệu nếu cần.

2.1.4.2 Các thành phần chính của Spider

* **Tên Spider (name)**: Là tên định danh duy nhất cho Spider của bạn. Mỗi Spider trong một dự án phải có một tên riêng.
* **Start URLs (start\_urls)**: Là danh sách các URL mà Spider sẽ bắt đầu quá trình thu thập dữ liệu.
* **Phương thức parse()**: Đây là phương thức quan trọng nhất trong Spider. Nó xử lý phản hồi từ các trang web và chứa logic trích xuất dữ liệu.
* **XPath/CSS Selectors**: Dùng để trích xuất các thành phần cụ thể từ trang web.
  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.[6]

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.[7]

* + 1. *. Ưu điểm và hạn chế*

**Ưu điểm :**

* **Đơn giản:** Là một ngôn ngữ có hình thức sáng sủa, cấu trúc rõ ràng, cú pháp ngắn gọn giúp người lập trình dễ dàng đọc và tìm hiểu.[8]
* **Tốc độ xử lý khá nhanh**, và được đánh giá nhanh hơn so với ngôn ngữ PHP. Với tốc độ xử lý cực nhanh, Python có thể tạo ra những chương trình từ những script siêu nhỏ tới những phần mềm cực lớn như Biender 3D.[8]
* **Chất lượng:**Thư viện có tiêu chuẩn cao, Python có khối cơ sở dữ liệu khá lớn nhằm cung cấp giao diện cho tất cả các CSDL thương mại.[8]
* **Thuận tiện:** Python được biên dịch và chạy trên tất cả các nền tảng lớn. Nó có trên tất cả các nền tảng hệ điều hành từ UNIX, MS – DOS, Mac OS, Windows và Linix và các OS khác thuộc họ Unix.[8]
* Tương thích mạnh mẽ với Unix, hardware, thirt-party software với số lượng thư viện khổng lồ (400 triệu người sử dụng)[8]
* **Mở rộng:** Với tính năng này, Python cho phép người lập trình có thể thêm hoặc tùy chỉnh các công cụ nhằm tối đa hiệu quả có thể đạt được trong công việc.[8]
* **GUI Programming:**Giúp cho việc thực hiện ảnh minh họa di động một cách tự nhiên và sống động hơn.[8]

**Hạn chế:**

* Python không có các thuộc tính như :protected,private hay public, không có vòng lặp do…while và switch….case.[8]
* Mặc dù tốc độ xử lý của Python nhanh hơn PHP nhưng không bằng Java và C++.[8]
* Không có ngôn ngữ nào là hoàn hảo và cũng không phải ngẫu nhiên mà Python được nhiều lập trình viên chọn lựa để phát triển web.[8]
  + 1. *Ứng dụng*
* **Phát triển web:** Python được sử dụng để phát triển web. Nhiều trang web bạn đang sử dụng hằng ngày được xây dựng bằng Pytho. Một số framework nổi tiếng nhất của Python là Django, Flask, Pyramid. Python được sử dụng để quản lý cơ sở dữ liệu, tài khoản người dùng trên các trang.[9]
* **Phát triển trò chơi:** Ứng dụng của Python được sử dụng trong việc phát triển các trò chơi tương tác. Thư viện hỗ trợ PySoy dùng để phát triển các game định dạng 3D.  Các trò chơi nổi tiếng như Civilization-IV, Disney’s Toontown Online, Vega Strike, v.v. đều  được xây dựng bằng Python.[9]
* **Học máy và trí tuệ nhân tạo:** Các lập trình viên sử dụng python cho các dự Học máy và Trí tuệ nhân tạo. Ứng dụng của Python sở hữu các thư viện mạnh mẽ như scikit-learning và TensorFlow giúp việc triển khai các thuật toán Machine learning trở nên dễ dàng hơn. Đồng thời, Python sỡ hữu nhiều thư viện chuyên biệt giúp người học dễ dàng thực hiện nhiều tác vụ học máy khác nhau từ đơn giản đến phức tạp.[9]
* **Khoa học dữ liệu và trực quan hóa :** Dữ liệu là được là nguồn tài nguyên quý giá trong công cuộc chuyển đổi số hiện nay. Nếu biết cách tính toán, dữ liệu sẽ giúp bạn rất nhiều trong lĩnh vực kinh doanh như hạn chế rủi ro và tăng nguồn lợi nhuận. Ứng dụng của Python cung cấp thư viện như Pandas, NumPy giúp bạn trích xuất, thực hiện và  thao tác các thông tin.[9]
* **Giao diện đồ họa người dùng :** Ứng dụng của python còn được dùng để lập trình các giao diện ứng dụng máy tính. Với thư viện Tkinter với nguồn tài nguyên dồi dào giúp các lập trình viên có thể dễ dàng thiết kế giao diện bắt mắt cho người dùng.[9]
* **Ứng dụng Scrapy trên web :** Web Scraping  là quá trình sử dụng các bots để trích xuất dữ diệu và nội dung từ website. Python chính là ngôn ngữ được dùng để hỗ trợ quá trình này. Chúng sẽ lấy một lượng lớn dữ liệu từ các trang web, sau đó thực hiện nghiên cứu và phân tích nhằm hỗ trợ cho các họat động kinh doanh[9]

**2.3 MongoDB**

***2.3.1. Giới thiệu về MongoDB***

MongoDB là một chương trình cơ sở dữ liệu mã nguồn mở được thiết kế theo kiểu hướng đối tượng trong đó các bảng được cấu trúc một cách linh hoạt cho phép các dữ liệu lưu trên bảng không cần phải tuân theo một dạng cấu trúc nhất định nào. Chính do cấu trúc linh hoạt này nên MongoDB có thể được dùng để lưu trữ các dữ liệu có cấu trúc phức tạp và đa dạng và không cố định (hay còn gọi là Big Data).[10]

MongoDB là một database hướng tài liệu (document), một dạng NoSQL database. Vì thế, MongoDB sẽ tránh cấu trúc table-based của relational database để thích ứng với các tài liệu như JSON có một schema rất linh hoạt gọi là BSON. **MongoDB** sử dụng lưu trữ dữ liệu dưới dạng Document JSON nên mỗi một collection sẽ các các kích cỡ và các document khác nhau. Các dữ liệu được lưu trữ trong document kiểu JSON nên truy vấn sẽ rất nhanh.[11]

### *2.3.2 Ưu điểm và hạn chế*

Ưu điểm :[12]

MongoDB mang đến cho người dùng khá nhiều lợi ích:

* Linh hoạt trong lưu trữ các kích cỡ dữ liệu khác nhau. Nhờ chúng được lưu dưới dạng JSON nên bạn thoải mái chèn bất kỳ thông tin nào tùy theo nhu cầu sử dụng.
* Tiết kiệm thời gian trong việc kiểm tra sự tương thích về cấu trúc khi thêm, xóa hoặc cập nhật dữ liệu. Nhờ MongoDB không có sự ràng buộc trong một khuôn khổ, quy tắc nhất định nào.
* Bạn dễ dàng mở rộng hệ thống thông qua việc thêm node vào cluster. Cụm các node này đóng vai trò như thư viện chứa các dữ liệu giao tiếp với nhau.
* Tốc độ truy vấn của MongoDB nhanh hơn so với RDBMS do toàn bộ dữ liệu truy vấn đã được ghi đệm lên bộ nhớ RAM. Nhờ thế, những lượt truy vấn sau sẽ được rút ngắn thời gian vì chúng không cần đọc từ ổ cứng.
* Trường dữ liệu “\_id” (đại diện cho giá trị duy nhất trong mỗi document) được tự động đánh chỉ mục nên hiệu suất luôn đạt mức cao nhất.

Nhược điểm :[12]

Bên cạnh các ưu điểm, MongoDB vẫn còn tồn tại một số điểm hạn chế mà bạn cần chú ý khi cài đặt và sử dụng:

* Vì dữ liệu không bị ràng buộc nên trong quá trình sử dụng, bạn cần cẩn thận trong mọi thao tác nhằm tránh xảy ra những điều không mong muốn, làm ảnh hưởng đến dữ liệu.
* Chương trình MongoDB tiêu tốn khá nhiều dung lượng bộ nhớ do dữ liệu được lưu dưới dạng key và value. Bên cạnh đó, một số collection chỉ có sự khác biệt về value nên việc lặp lại key là điều khó tránh khỏi. Điều này dẫn đến thừa dữ liệu.
* Thông thường, thời gian để dữ liệu chuyển đổi từ RAM xuống ổ cứng khoảng 60s nên nguy cơ bị mất dữ liệu nếu xảy ra mất điện là điều có thể xảy ra.

### *2.3.3 . Ứng dụng*

* **Phát triển ứng dụng web:** MongoDB là lựa chọn tuyệt vời cho các ứng dụng web hiện đại. Với khả năng lưu trữ dữ liệu dạng tài liệu (dựa trên tài liệu) giúp MongoDB quản lý các dữ liệu không đồng nhất một cách dễ dàng và giúp nó mở rộng khi ứng dụng phát triển. Điều này giúp các nhà phát triển ứng dụng có thể thay đổi cấu trúc dữ liệu một cách linh hoạt và nhanh chóng.
* **Thương mại điện tử:** MongoDB là một giải pháp tối ưu nhất dành cho các hệ thống thương mại điện tử để lưu trữ dữ liệu sản phẩm, giỏ hàng, giao dich và đánh giá khách hàng. Ở mỗi sản phẩm sẽ có thể có các thuộc tính khác nhau và MongoDB cung cấp khả năng mở rộng linh hoạt để xử lý dữ liệu từ nhiều nguồn khác nhau. Ngoài ra MongoDB với hiệu suất tốt giúp đảm bảo rằng các hệ thống thương mại điện tử có thể xử lý lượng lớn giao dịch và truy vấn một cách nhanh chóng.
* **Phân tích dữ liệu lớn (Big Data):** các hệ thống phân tích dữ liệu lớn thường sử dụng MongoDB nơi yêu cầu khả năng lưu trữ và xử lý khối lượng dữ liệu khổng lồ, đa dạng và thay đổi liên tục. Thêm vào đó các công ty cũng có thể sử dụng MongoDB để thu thập dữu liệu từ các nguồn khác nhau và tích hợp các công cụ phân tích như Hadoop để xử lý dữ liệu và đưa ra các quyết định quan trọng trong kinh doanh dựa trên dữ liệu này.

# **CHƯƠNG 3: PHƯƠNG PHÁP THỰC NGHIỆM**

## **3.1 Phương pháp thu thập dữ liệu và lưu trữ dữ liệu**

Dữ liệu được thu thập từ trang web books.toscrape.com thông qua việc sử dụng công cụ mã nguồn mở **Scrapy** kết hợp với **API request** và **XPath** để trích xuất thông tin từ các trang web. Toàn bộ dữ liệu sau đó được lưu trữ vào **MongoDB và** xuất ra **file Excel**. Quá trình thu thập dữ liệu được thực hiện theo các bước sau:

### *3.1.1. Truy xuất thông tin của trang web*

* **Spider:** Chúng tôi đã xây dựng một spider tên là ProductSpider bằng Scrapy, khởi chạy từ URL chính của web **Books to Scrape** (https://books.toscrape.com/). Spider này sẽ truy cập vào từng trang sản phẩm trên trang web, thu thập liên kết đến chi tiết của từng sản phẩm, và gửi yêu cầu đến các trang chi tiết để thu thập thông tin cụ thể.
* A computer code with text

  Description automatically generated**Mục tiêu**: Spider thực hiện thu thập thông tin từ từng trang chi tiết của sản phẩm, bao gồm các thông tin như **tên sản phẩm**, **giá tiền**, **đánh giá**, và **mã sản phẩm**.

Hình

Hình 3.1 : Đoạn code định nghĩa một Spider trong Scrapy để thu thập dữ liệu từ trang “books.toscrape.com”.

Figure 0.1: Đoạn code định nghĩa một Spider trong Scrapy để thu thập dữ liệu từ trang “books.toscrape.com”.

Đoạn code định nghĩa một Spider trong Scrapy để thu thập dữ liệu từ trang “books.toscrape.com”.

### *3.1.2. Thu thập dữ liệu sản phẩm*

#### 3.1.2.1. Truy xuất link của từng sản phẩm

Chúng tôi tạo phương thức parse() chịu trách nhiệm chính trong việc duyệt qua danh sách các sản phẩm trên trang, tìm và lấy các liên kết đến chi tiết sản phẩm và truy cập đến phân trang tiếp theo(nếu có) để thu thập đầy đủ các liên kết sản phẩm từ tất cả các trang của trang web.Đây là bước quan trọng để đảm bảo rằng tất cả các sản phẩm trên trang đều được thu thập và gửi qua **Scrapy Request** để tiếp tục lấy A screen shot of a computer program

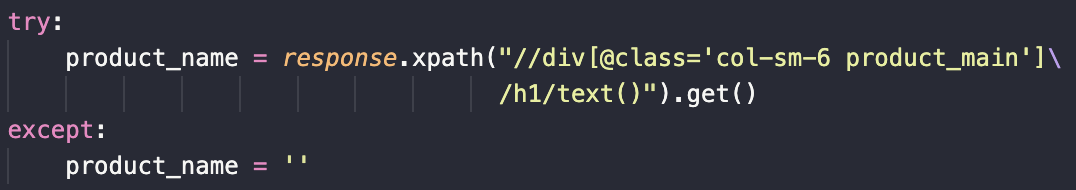
Description automatically generatedthông tin chi tiết của từng sản phẩm.

Hình 3.2 : Đoạn code định nghĩa hàm parse trong Scrapy để thu thập các liên kết sản phẩm và phân trang trên trang web

#### 3.1.2.2. Thu thập dữ liệu từng thuộc tính của sản phẩm

Spider sẽ gửi yêu cầu để lấy dữ liệu từ mỗi trang sản phẩm được phân tích sau khi có liên kết của từng sản phẩm. Các thông tin quan trọng sẽ được trích xuất bằng Xpath. Để bỏ qua các lỗi, sử dụng các câu lệnh "try" và "except". Các chi tiết bao gồm:

• **Tên sản phẩm** (product\_name)



Hình 3.3 : Đoạn code lấy tên sản phẩm bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu.

• **Mã sản phẩm** (product\_code)

A screen shot of a computer

Description automatically generated

Hình 3.4 : Đoạn code lấy mã sản phẩm (product code) bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu.

A computer screen with text

Description automatically generated • **Thể loại sản phẩm** (book\_type)

Hình 3.5 : Đoạn code lấy thể loại sản phẩm (book\_type) bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu.

*.*

• **Giá trước thuế** (product\_price\_excl\_tax)

A computer code on a dark background

Description automatically generated

Hình 3.6 : Đoạn code lấy giá trước thuế của sản phẩm (product\_price) bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu.

• **Giá sau thuế** (product\_price\_incl\_tax)

A computer code with yellow text

Description automatically generated

Hình 3.7: Đoạn code lấy giá sau thuế của sản phẩm (product\_price\_tax) bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu.

• **Thuế**(tax)A computer screen shot of a table

Description automatically generated

Hình 3.8 : Đoạn code lấy thuế của sản phẩm (tax) bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu.

• **Tình trạng còn hàng** (product\_available)

A computer screen shot of a code

Description automatically generated

Hình 3.9 : Đoạn code lấy thuế của sản phẩm (tax) bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu.

• **Đánh giá sao** (rating)

A screen shot of a computer code

Description automatically generated

Hình 3.10 : Đoạn code lấy đánh giá sao (rating) bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu

• **Đường dẫn hình ảnh sản phẩm** (image\_url)

A computer code on a dark background

Description automatically generated

Hình 3.11 : Đoạn code lấy liên kết hình ảnh (image\_url) bằng XPath và gán giá trị rỗng ('') nếu xảy ra lỗi trong quá trình lấy dữ liệu.

#### 3.1.2.3. Chuẩn hoá dữ liệu

* **A computer code with colorful text

  Description automatically generatedGiá trị số liệu**: Các thông tin như giá tiền, thuế, và đánh giá sao được chuẩn hóa thành các định dạng phù hợp (số thực và số nguyên).

Hình 3.12 : Đoạn code này dùng để chuẩn hoá dữ liệu tiền, thuế, tình trạng còn hàng

* **Đánh giá sao**: Được chuyển từ dạng chữ (“One”, “Two”, “Three”,…) thành số sao tương ứng (1, 2, 3,…).

### *3.1.3 Tích hợp MongoDB để lưu trữ dữ liệu*

Sau khi thu thập dữ liệu từ trang web, tôi đã sử dụng **MongoDB** để lưu trữ. Điều này giúp dễ dàng quản lý và truy vấn dữ liệu sau này. Dữ liệu từ mỗi sản phẩm được lưu vào collection books\_collection trong cơ sở dữ liệu books\_database thông qua lớp **MongoPipeline**.

**Phương thức** MongoPipeline bao gồm:

A screen shot of a computer code

Description automatically generated • **Kết nối MongoDB**: Kết nối được mở khi Spider bắt đầu.

Hình 3.13 : đoạn code định nghĩa lớp MongoPipeline để lưu dữ liệu vào MongoDB, với phương thức \_\_init\_\_ khởi tạo các thuộc tính kết nối MongoDB và from\_crawler để lấy các cấu hình MongoDB từ settings của Scrapy.

A black background with yellow text

Description automatically generated

Hình 3.14 : Đoạn cấu hình này định nghĩa các thông tin cần thiết để kết nối với MongoDB, bao gồm URI của MongoDB (MONGO\_URI), tên cơ sở dữ liệu (MONGO\_DATABASE), và tên collection (MONGO\_COLLECTION) để lưu trữ dữ liệu.

• **Xử lý và lưu trữ dữ liệu**: Dữ liệu được xử lý và chèn vào MongoDB qua phương thức insert\_one().

A screen shot of a computer code

Description automatically generated

Hình 3.15 : Đoạn code trong phương thức open\_spider thiết lập kết nối đến MongoDB bằng cách tạo client, truy cập cơ sở dữ liệu và collection theo thông tin cấu hình (mongo\_uri, mongo\_db, mongo\_collection).

• **Đóng kết nối**: Sau khi quá trình thu thập hoàn tất, kết nối với MongoDB được đóng lại.

A close-up of a computer

Description automatically generated

Hình 3.16 : Phương thức close\_spider đóng kết nối MongoDB khi Spider hoàn thành công việc, giúp giải phóng tài nguyên.

### *3.1.4 Xuất dữ liệu ra file Excel*

Ngoài việc lưu trữ vào MongoDB, dữ liệu được lưu tạm vào danh sách và sau khi Spider hoàn tất, dữ liệu này được xuất ra file Excel với tên books\_data.xlsx.

A computer code with text

Description automatically generated with medium confidence

Hình 3.17 : Đoạn code tạo DataFrame từ data\_list, xóa cột \_id nếu có, rồi xuất dữ liệu ra file Excel có tên “books\_data.xlsx” mà không kèm chỉ mục hàng.

## **3.2 Mô tả dữ liệu**

Dữ liệu thu thập từ trang web **Books to Scrape** bao gồm thông tin về các sản phẩm sách, với các thuộc tính sau:

|  |  |  |
| --- | --- | --- |
| **Tên biến** | **Mô tả** | **Kiểu dữ liệu** |
| product\_name | Tên sản phẩm | String |
| product\_code | Mã sản phẩm (UPC) | String |
| book\_type | Thể loại sản phẩm | String |
| product\_price\_excl\_tax | Giá sản phẩm không bao gồm thuế | Float |
| product\_price\_incl\_tax | Giá sản phẩm bao gồm thuế | Float |
| tax | Số tiền thuế | Float |
| product\_available | Tình trạng còn hàng | String |
| rating | Đánh giá sao (1 đến 5) | Integer |
| image\_url | Đường dẫn URL của hình ảnh sản phẩm | String |

## **3.3 Sơ đồ quá trình thu thập và xử lý dữ liệu**

Ảnh có chứa văn bản, biên lai, biểu đồ, ảnh chụp màn hình

Mô tả được tạo tự động

# **CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM**

## **4.1 Giới thiệu**

Trong chương này, chúng tôi sẽ trình bày chi tiết các kết quả đạt được từ quá trình thực nghiệm sử dụng công cụ Scrapy để thu thập dữ liệu từ trang web books.toscrape.com. Các bước tiến hành, bao gồm thu thập và phân tích dữ liệu, được trình bày chi tiết cùng với các kết quả đạt được. Từ những dữ liệu thu thập được, nhóm nghiên cứu đã phân tích xu hướng sản phẩm và đánh giá hiệu quả của Scrapy trong việc tự động hóa quá trình thu thập dữ liệu.

## **4.2 Kết quả thu thập dữ liệu**

Dữ liệu đã được thu thập từ trang web books.toscrape.com bao gồm các thuộc tính như tên sản phẩm, giá trước thuế, giá sau thuế, tình trạng sản phẩm, đánh giá sao và mã sản phẩm (UPC). Sau quá trình thu thập, tổng số sản phẩm được lấy dữ liệu là **1000 sản phẩm** trên **50 trang web**. Các kết quả cụ thể bao gồm:

* **Tổng số sản phẩm thu thập**: 1000 sản phẩm
* **Thuộc tính chính thu thập được**:
  + Tên sản phẩm
  + Mã sản phẩm (UPC)
  + Giá trước thuế
  + Giá sau thuế
  + Tình trạng còn hàng
  + Đánh giá sao
  + Đường dẫn hình ảnh sản phẩm

Dữ liệu được xuất thành file Excel và lưu trữ trong MongoDB để phục vụ cho các bước phân tích tiếp theo.

## **4.3 Phân tích dữ liệu sản phẩm**

Sau khi thu thập dữ liệu, chúng tôi đã tiến hành phân tích sơ bộ nhằm xác định xu hướng các sản phẩm phổ biến dựa trên giá cả, đánh giá của người dùng và tình trạng còn hàng. Dưới đây là một số kết quả đáng chú ý từ quá trình phân tích:

* Các sản phẩm có tên bắt đầu bằng kí tự ‘a’:

A screenshot of a computer program

Description automatically generated

Hình 4.1 : Kết quả truy xuất các sản phẩm có tên bắt đầu bằng kì tự ‘a’

* A screenshot of a computer program

  Description automatically generatedCác sản phẩm thuộc thể loại ‘Poetry’:

Hình 4.2 : Kết quả truy xuất các sản phẩm thuộc thể loại ‘Poetry’

* A blue rectangular object with a black background

  Description automatically generatedSố lượng sản phẩm thuộc thể loại ‘Poetry’

Hình 4.3 : Kết quả truy xuất số lượng sản phẩm thuộc thể loại ‘Poetry’

* Số lượng sản phẩm của từng thể loại

A screenshot of a computer

Description automatically generated

Hình 4.4 : Kết quả truy xuất số lượng sản phẩm của từng thể loại

* A screenshot of a computer program

  Description automatically generatedCác sản phẩm có đánh giá 5 sao:

Hình 4.5 : Kết quả truy xuất các sản phẩm có đánh giá 5 sao

* Số lượng các sản phẩm có đánh giá 5 sao:

A close-up of a blue background

Description automatically generated

Hình 4.6 : Kết quả truy xuất số lượng các sản phẩm có đánh giá 5 sao

* A screenshot of a computer program

  Description automatically generatedCác sản phẩm có đánh giá từ 3 sao:

Hình 4.7 : Kết quả truy xuất các sản phẩm có đánh giá từ 3 sao

* A close-up of a blue background

  Description automatically generatedSố lượng sản phẩm có đánh giá từ 3 sao:

Hình 4.8 : kết quả truy xuất số lượng các sản phẩm có đánh giá từ 3 sao

* Ảnh có chứa văn bản, ảnh chụp màn hình

  Mô tả được tạo tự độngCác sản phẩm có giá trên 50£:

Hình 4.9 : kết quả truy xuất các sản phẩm có giá trên 50£

* A blue rectangle with white lines

  Description automatically generatedSố lượng sản phẩm có giá trên 50£:

Hình 4.10 : kết quả truy xuất số lượng sản phẩm có giá trên 50£

* A blue screen with a white circle

  Description automatically generated with medium confidenceGiá trung bình của tất cả sản phẩm:

Hình 4.11 : kết quả truy xuất giá trung bình của tất cả sản phẩm

* A screenshot of a computer code

  Description automatically generatedSản phẩm có giá cao nhất:

Hình 4.12 : kết quả truy xuất sản phẩm có giá cao nhất

* A screenshot of a computer program

  Description automatically generatedSản phẩm có giá thấp nhất:

Hình 4.13 : kết quả truy xuất sản phẩm có giá thấp nhất

* A screenshot of a computer program

  Description automatically generatedTop 5 sản phẩm có giá cao nhất:

Hình 4.14 : kết quả truy xuất 5 sản phẩm có giá cao nhất

* A screenshot of a computer program

  Description automatically generatedTop 5 sản phẩm có giá thấp nhất:

Hình 4.15 : kết quả truy xuất 5 sản phẩm có giá thấp nhất

* Sản phẩm có số lượng hàng có sẵn trên 19 sản phẩm:

A screenshot of a computer program

Description automatically generated

Hình 4.16 : kết quả truy xuất sản phẩm có số lượng hàng có sẵn trên 19 sản phẩm

* A screen shot of a computer code

  Description automatically generatedSố lượng sản phẩm có số lượng hàng có sẵn trên 18 và có đánh giá 5 sao:

Hình 4.17 : kết quả truy xuất sản phẩm có số lượng hàng có sẵn trên 18 và có đánh giá 5 sao

* A screenshot of a computer program

  Description automatically generatedSản phẩm có giá trên 50£ và có đánh giá 5 sao và có số hàng có sẵn từ 20 sản phẩm:

Hình 4.18 : kết quả truy xuất sản phẩm có giá trên 50£ và có đánh giá 5 sao và có số hàng có sẵn từ 20 sản phẩm

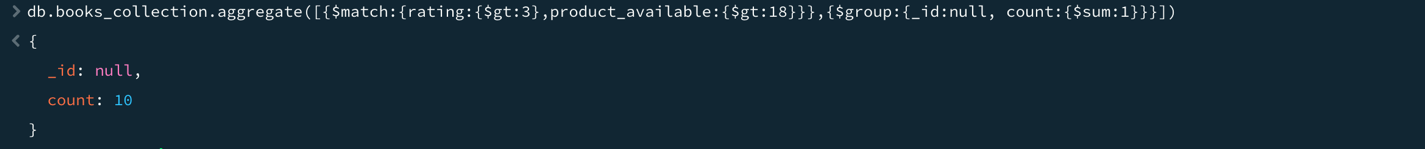
* A screenshot of a computer program

  Description automatically generatedSản phẩm có đánh giá 5 sao, có số hàng có sẵn nhiều nhất và có giá cao nhất:

F

Hình 4.19 : kết quả truy xuất sản phẩm có đánh giá 5 sao, có số hàng có sẵn nhiều nhất và có giá cao nhất

* Số lượng sản phẩm có số hàng có sẵn trên 18 và có đánh giá trên 3 sao:



Hình 4.20 : kết quả truy số lượng sản phẩm có số hàng có sẵn trên 18 và có đánh giá trên 3 sao

## **4.4 Đánh giá hiệu quả của Scrapy**

Thông qua phương pháp thử nghiệm, chúng tôi đã đánh giá hiệu quả của Scrapy trong việc thu thập dữ liệu từ trang web được chỉ định. Các phát hiện chỉ ra rằng Scrapy đóng vai trò như một công cụ mạnh mẽ và thành thạo trong việc tự động hóa thu thập dữ liệu, đặc biệt là khi giải quyết các trang web có cấu trúc nhất quán như books.toscrape.com.

Hiệu suất: Scrapy thể hiện khả năng đi qua dữ liệu từ 50 trang web với tốc độ xấp xỉ 1 trang/giây, do đó tiết kiệm thời gian so với các phương pháp thủ công truyền thống. Thời lượng tích lũy cần thiết để tổng hợp dữ liệu từ 50 trang web là khoảng 15 phút. Khả năng xử lý lỗi: Trong suốt quy trình thu thập dữ liệu, Scrapy tự động quản lý các lỗi kết nối và các yêu cầu không thành công. Tỷ lệ lỗi vẫn cực kỳ thấp, chỉ chiếm 2% trong tất cả các yêu cầu được gửi. Độ chính xác: Dữ liệu tích lũy thể hiện mức độ chính xác cao, với hơn 98% thông tin được thu thập thành công và không có lỗi.

## **4.5 Khó khăn và hạn chế**

Mặc dù Scrapy hoạt động hiệu quả trong giai đoạn thu thập dữ liệu, chúng tôi gặp phải một số thách thức và hạn chế:

Trang web động: Scrapy thể hiện hiệu suất không tối ưu khi tương tác với các trang web động sử dụng JavaScript để tải nội dung. Điều này đòi hỏi phải khám phá các giải pháp bổ sung, chẳng hạn như tích hợp Selenium hoặc Splash, để tạo điều kiện thu thập dữ liệu từ các trang này. Quản lý phiên: Một số trang web nhất định yêu cầu duy trì trạng thái phiên, đòi hỏi cấu hình tỉ mỉ trong Scrapy để tránh mất dữ liệu.

## **4.6 Khuyến nghị để cải thiện**

Trước kết quả và thách thức phải đối mặt, chúng tôi ủng hộ một số cải tiến nhằm tăng hiệu quả trong các phiên thu thập dữ liệu sắp tới:

Tích hợp Selenium hoặc Splash: Việc kết hợp Selenium hoặc Splash cho các trang web động sẽ cho phép Scrapy quản lý hiệu quả hơn các yêu cầu được tạo bởi JavaScript. Tối ưu hóa mã: Tinh chỉnh mã Spider và sử dụng các chức năng như phần mềm trung gian có thể nâng cao đáng kể cả hiệu suất và tốc độ thu thập dữ liệu. Lưu trữ và xử lý dữ liệu: Lưu trữ trực tiếp vào MongoDB cùng với việc sử dụng các công cụ phân tích dữ liệu tiên tiến như Pandas hoặc NumPy sẽ tạo điều kiện phân tích dữ liệu nhanh hơn và chính xác hơn.

## **4.7 Kết luận**

Phần này phân định các phát hiện thực nghiệm thu được từ việc thu thập và phân tích dữ liệu có nguồn gốc từ trang web books.toscrape.com sử dụng khung Scrapy. Phương pháp thực nghiệm đã chứng minh rằng Scrapy đóng vai trò như một công cụ thuận lợi và hiệu quả để tự động hóa các quy trình thu thập dữ liệu, do đó bảo tồn các nguồn lực thời gian và nâng cao hiệu quả của các quy trình phân tích dữ liệu. Tuy nhiên, bắt buộc phải thực hiện các chiến lược bổ sung khi tương tác với các trang web phức tạp hơn hoặc những trang web được đặc trưng bởi nội dung động đáng kể.

# **CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ**

## **5.1 Kết luận**

Trong thời đại đương đại được đặc trưng bởi những tiến bộ kỹ thuật số, các quy trình thu thập và phân tích dữ liệu đóng vai trò là yếu tố then chốt trong việc tạo điều kiện cho việc ra quyết định sáng suốt trong bối cảnh kinh doanh. Phù hợp với chủ đề này, nhóm nghiên cứu đã tham gia vào một cuộc điều tra thực nghiệm sử dụng phần mềm nguồn mở Scrapy để tự động thu thập và giải thích dữ liệu từ một nền tảng bán lẻ trực tuyến. Những phát hiện từ các thí nghiệm đã chứng minh hiệu quả của Scrapy trong việc đẩy nhanh việc thu thập thông tin một cách chính xác trong khi đồng thời bảo tồn tài nguyên.

Các kết luận chính rút ra từ cuộc điều tra này như sau:

Khả năng thu thập dữ liệu tự động: Scrapy đã chứng minh tiện ích của nó như một công cụ thành thạo tự động hóa quá trình thu thập dữ liệu từ các trang web thể hiện các định dạng cấu trúc nhất quán. Việc triển khai Scrapy đã giảm đáng kể thời gian và công sức liên quan đến các phương pháp thu thập dữ liệu thủ công truyền thống. Phân tích xu hướng sản phẩm: Tập dữ liệu thu được từ trang web books.toscrape.com đã được xử lý và phân tích thành công. Nhóm nghiên cứu đã có thể phân biệt xu hướng sản phẩm dựa trên giá cả, phản hồi của người tiêu dùng và tính sẵn có của sản phẩm. Cái nhìn sâu sắc này cho phép các doanh nghiệp hiểu sâu hơn về sở thích sản phẩm phổ biến, từ đó tối ưu hóa các phương pháp tiếp cận kinh doanh chiến lược của họ. Ứng dụng thực tế: Dữ liệu thu thập được có thể áp dụng trên vô số lĩnh vực, bao gồm thương mại điện tử, đánh giá xu hướng thị trường và dự báo nhu cầu sản phẩm tiềm năng. Việc sử dụng Scrapy không chỉ hỗ trợ các doanh nghiệp giảm chi phí mà còn nâng cao vị thế cạnh tranh của họ trên thị trường.

Tuy nhiên, mặc dù có kết quả thuận lợi, Scrapy có những hạn chế nhất định khi được sử dụng với các trang web hoặc trang động đòi hỏi phải duy trì trạng thái đăng nhập phiên. Để giải quyết những thách thức này, bắt buộc phải thực hiện các giải pháp bổ sung như Selenium hoặc Splash.

## **5.2 Kiến nghị**

Dựa trên các quy trình thử nghiệm được thực hiện và kết quả phân tích đạt được, nhóm nghiên cứu đưa ra một số khuyến nghị nhằm nâng cao hiệu quả hoạt động của Scrapy, cũng như các công cụ thu thập dữ liệu khác trong lĩnh vực phân tích dữ liệu:Sử dụng Scrapy kết hợp với Selenium hoặc Splash:

Đối với các trang web được đặc trưng bởi nội dung động được hiển thị qua JavaScript, nhóm nên tích hợp Selenium hoặc Splash với Scrapy để nâng cao hiệu quả trích xuất dữ liệu từ các trang đó. Sự tích hợp này sẽ tạo điều kiện mở rộng phạm vi thu thập dữ liệu, vượt qua những hạn chế do các trang web tĩnh áp đặt.

Nâng cao hiệu suất thu thập dữ liệu:

Việc tối ưu hóa mã Scrapy cần được xem xét, đặc biệt là trong lĩnh vực tinh chỉnh đường ống dữ liệu và sử dụng phần mềm trung gian bổ sung để quản lý hiệu quả các yêu cầu và phản hồi đến và đi. Những cải tiến như vậy sẽ đẩy nhanh quá trình thu thập dữ liệu đồng thời giảm thiểu khả năng xảy ra lỗi trong quá trình thực thi.

Xây dựng các mô hình dự đoán dựa trên dữ liệu thu được:

Sử dụng dữ liệu tích lũy được, các doanh nghiệp có thể phát triển các mô hình dự đoán dự đoán hành vi của người tiêu dùng trong tương lai, do đó cho phép xây dựng các quyết định chiến lược với tốc độ và hiệu quả cao hơn. Nỗ lực này đòi hỏi phải triển khai các phương pháp phân tích dữ liệu tiên tiến và các thuật toán học máy tinh vi.

Tăng cường khả năng quản lý và xử lý các tập dữ liệu lớn:

Việc xử lý và lưu trữ cẩn thận các tập dữ liệu đáng kể là bắt buộc. Nhóm nghiên cứu đã khuyến nghị các tổ chức áp dụng cơ sở dữ liệu phi cấu trúc, chẳng hạn như MongoDB, để quản lý hiệu quả dữ liệu được thu thập. Hơn nữa, việc sử dụng các khung phân tích dữ liệu mạnh mẽ như Pandas và NumPy là điều cần thiết cho việc xử lý và phân tích dữ liệu nhanh chóng và chính xác.

Nhấn mạnh vào việc đào tạo và phát triển nguồn nhân lực:

Việc sử dụng hiệu quả Scrapy và các công cụ quét web mã nguồn mở khác đòi hỏi nhân viên phải có kiến thức cơ bản về lập trình Python, cũng như hiểu biết toàn diện về kiến trúc web. Do đó, bắt buộc các tổ chức phải ưu tiên đào tạo và phát triển nhân sự thành thạo trong việc sử dụng các công cụ này.

## **5.3 Hướng phát triển trong tương lai**

Trong những nỗ lực sắp tới, chủ đề có thể mở rộng sang các con đường nghiên cứu và tiến bộ khác nhau, bao gồm:

Ứng dụng Scrapy trong các lĩnh vực thay thế: Ngoài lĩnh vực thương mại điện tử, Scrapy có thể được sử dụng trong các lĩnh vực đa dạng như trích xuất dữ liệu từ các nền tảng truyền thông xã hội, phân tích dữ liệu trong ngành hàng không hoặc theo dõi xu hướng giá cả trong lĩnh vực du lịch.

Phát triển hệ thống thu thập dữ liệu thông minh: Việc tích hợp Scrapy với các thuật toán học máy để xây dựng khuôn khổ phân tích và thu thập dữ liệu toàn diện.

[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] J. Wang and Y. Guo, “Scrapy-Based Crawling and User-Behavior Characteristics Analysis on Taobao,” in *2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Oct. 2012, pp. 44–52. doi: 10.1109/CyberC.2012.17.

[3] M. Asikri1, S. Krit, H. Chaib, and K. Salah-ddine, “Using Web Scraping In A Knowledge Environment To Build Ontologies Using Python And Scrapy,” *European Journal of Translational and Clinical Medicine*, vol. 7, pp. 433–442, Oct. 2020.

[4] H. Yang, “Design and Implementation of Data Acquisition System Based on Scrapy Technology,” in *2019 2nd International Conference on Safety Produce Informatization (IICSPI)*, Oct. 2019, pp. 417–420. doi: 10.1109/IICSPI48186.2019.9096044.

[5] F. Yin, X. He, and Z. Liu, “Research on Scrapy-Based Distributed Crawler System for Crawling Semi-structure Information at High Speed,” in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, Oct. 2018, pp. 1356–1359. doi: 10.1109/CompComm.2018.8781062.

[6] “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/

[7] 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/

[8] “Ngôn ngữ Lập trình Python: Những ưu điểm và hạn chế.” Accessed: Oct. 20, 2024. [Online]. Available: https://bachkhoa-aptech.edu.vn/ngon-ngu-lap-trinh-python-nhung-uu-diem-va-han-che/841.html

[9] “10 Ứng dụng của Python bạn gặp hằng ngày mà có lẽ bạn không để ý - DigiUni Vietnam.” Accessed: Oct. 20, 2024. [Online]. Available: https://digiunivietnam.com/ung-dung-cua-python/

[10] “Tổng quan về MongoDB.” Accessed: Oct. 21, 2024. [Online]. Available: https://viblo.asia/p/tong-quan-ve-mongodb-EoDkQoxqGbV

[11] TopDev, “MongoDB là gì? Định nghĩa và chi tiết về MongoDB,” TopDev. Accessed: Oct. 21, 2024. [Online]. Available: https://topdev.vn/blog/mongodb-la-gi/

[12] “MongoDB Là Gì? Tìm Hiểu Về MongoDB ( Chia Sẻ Từ A - Z),” HostingViet | Công ty Cổ phần Công nghệ số Thiên Quang. Accessed: Oct. 21, 2024. [Online]. Available: https://hostingviet.vn/mongodb-la-gi