|  |  |
| --- | --- |
| logo (CMYK)-01 | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM** |

**ĐỒ ÁN MÔN HỌC**

**TÌM HIỂU SELENIUM VÀ MONGODB, THU THẬP DỮ LIỆU SẢN PHẨM NHÀ THUỐC PHARMACITY**

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 :

Hà Thế Anh MSSV: 2286400002

Nguyễn Nhật Nam MSSV: 2286400019

Hoàng Quang Minh MSSV: 2286400017

Lớp: 22DKHA1

TP. Hồ Chí Minh, 2024

|  |  |
| --- | --- |
| logo (CMYK)-01 | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM** |

**ĐỒ ÁN MÔN HỌC**

**TÌM HIỂU SELENIUM VÀ MONGODB, THU THẬP DỮ LIỆU SẢN PHẨM NHÀ THUỐC PHARMACITY**

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 :

Hà Thế Anh MSSV: 2286400002

Nguyễn Nhật Nam MSSV: 2286400019

Hoàng Quang Minh MSSV: 2286400017

Lớp: 22DKHA1

TP. Hồ Chí Minh, 2024

# LỜI CAM ĐOAN

Chúng tôi, Hà Thế Anh, Hoàng Quang Minh và Nguyễn Nhật Nam 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. Chúng tôi xin 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 chúng 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 Selenium và MongoDB, thu thập dữ liệu sản phẩm nhà thuốc Pharmacity" và sẽ đóng góp một phần nhỏ vào lĩnh vực nghiên cứu này.

TP.HCM, Ngày.....tháng.....năm 2024

**Sinh viên**

Hà Thế Anh

Hoàng Quang Minh

Nguyễn Nhật Nam

# NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

*TP. Hồ Chí Minh, ngày tháng năm 2023*

**Giáo viên hướng dẫn**

**(**Ký tên,đóng dấu**)**

# MỤC LỤC

[LỜI CAM ĐOAN i](#_Toc180505271)

[NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN ii](#_Toc180505272)

[MỤC LỤC iii](#_Toc180505273)

[DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT VÀ TỪ KHÓA v](#_Toc180505274)

[CHƯƠNG 1. TỔNG QUAN 1](#_Toc180505275)

[1.1. Giới thiệu đề tài 1](#_Toc180505276)

[1.2. Nhiệm vụ của đồ án 1](#_Toc180505277)

[*1.2.1.* *Tính cấp thiết của đề tài* 1](#_Toc180505278)

[*1.2.1.* *Ý nghĩa khoa học và thực tiễn của đề tài* 2](#_Toc180505279)

[1.3. Mục tiêu 3](#_Toc180505280)

[*1.3.1.* *Mục tiêu tổng quan* 3](#_Toc180505281)

[*1.3.2.* *Mục tiêu cụ thể* 4](#_Toc180505282)

[1.4. Đối tượng và phạm vi 4](#_Toc180505283)

[*1.4.1.* *Đối tượng* 4](#_Toc180505284)

[*1.4.2.* *Phạm vi* 4](#_Toc180505285)

[1.5. Phương pháp nghiên cứu 5](#_Toc180505286)

[*1.5.1.* *Phương pháp nghiên cứu sơ bộ* 5](#_Toc180505287)

[*1.5.2.* *Phương pháp nghiên cứu tài liệu* 5](#_Toc180505288)

[*1.5.3.* *Phương pháp nghiên cứu thống kê* 5](#_Toc180505289)

[*1.5.4.* *Phương pháp thực nghiệm* 6](#_Toc180505290)

[*1.5.5.* *Phương pháp đánh giá* 6](#_Toc180505291)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 7](#_Toc180505292)

[2.1. Selenium 7](#_Toc180505293)

[*2.1.1.* Giới thiệu và trích xuất dữ liệu từ Selenium. 7](#_Toc180505294)

[CHƯƠNG 3. KẾT QUẢ THỰC NGHIỆM 8](#_Toc180505295)

[CHƯƠNG 4. KẾT LUẬN VÀ KIẾN NGHỊ 9](#_Toc180505296)

[4.1. Kết luận 9](#_Toc180505297)

[*4.1.1.* 9](#_Toc180505298)

[4.2. Kiến nghị 9](#_Toc180505299)

[TÀI LIỆU THAM KHẢO 10](#_Toc180505300)

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

# TỔNG QUAN

## Giới thiệu đề tài

Trong bối cảnh cộng nghệ thông tin ngày càng phát triển, việc thu thập và quản lý dữ liệu sản phẩm một cách hiệu quả là yếu tố then chốt đối với các doanh nghiệp, đặc biệt là trong lĩnh vực dược phẩm. Để tối ưu hóa được quy trình này, công cụ tự động như Selenium được sử dụng với công việc thu thập dữ liệu từ các website một cách tự động, trong khi MongoDB đóng vai trò là một kho lưu trữ và quản lý được khối lượng thông tin sản phẩm một cách linh hoạt. Áp dụng hai công cụ này lại với nhau giúp các doanh nghiệp và trong đó có nhà thuốc Pharmacity có thể dễ dàng theo dõi và quản lý sản phẩm, nhanh chóng cập nhật được thông tin, từ đó nâng cao hiệu quả kinh doanh và cải thiện khả năng cạnh tranh.

## Nhiệm vụ của đồ án

Nhiệm vụ của đề tài “Tìm hiểu Selenium và MongoDB, thu thập dữ liệu sản phẩm nhà thuốc Pharmacity” là áp dụng được Selenium để tự động quá trình thu thập dữ liệu sản phẩm từ website của nhà thuốc Pharmacity và lưu trữ, quản lý bằng MongoDB. Thông qua quá trình thu thập và quản lý dữ liệu này, doanh nghiệp sẽ có cái nhìn sâu sắc hơn về cái sản phẩm từ đó tối ưu hóa chiến lược bán hàng, nâng cao chất lượng dịch vụ và tăng khả năng cạnh tranh trên thị trường.

### *Tính cấp thiết của đề tài*

Trong thời đại số hóa, việc thu thập và quản lý dữ liệu hiểu quả là yếu tố then chốt để doanh nghiệp tồn tại và phát triển. Nhà thuốc Pharmacity là một chuỗi của hàng bán thuốc lẻ lớn đang phải đối mặt với sự cạnh tranh khốc liệt từ nhiều đối thủ. Để đáp ứng được nhu cầu của thị trường và duy trì cạnh tranh thì doanh nghiệp cần phải ứng dụng được công nghệ tự động hóa quy thu thập và quản lý dữ liệu sản phẩm.

Dự án “Tìm hiểu Selenium và MongoDB, thu thập dữ liệu sản phẩm nhà thuốc Pharmacity” được triển khai nhằm giải quyết được bài toán thu thập và quản lý dữ liệu sản phẩm. Cụ thể việc thu thập và quản lý dữ liệu mang lại cho nhà thuốc Pharmacitiy những lợi ích sau:

+ Tối ưu hóa được chiến lược kinh doanh: Việc thu thập dữ liệu sẽ giúp doanh nghiệp hiểu rõ hơn về thị trường, từ đó đưa ra được chiến lược marketing và bán hàng chính xác tối ưu hóa hiệu quả kinh doanh.

+ Cải thiện dịch vụ và trải nghiệm của người dùng: Với dữ liệu được cập nhật tự động hóa liên tục, doanh nghiệp có thể nhanh chóng bắt kịp xu hướng và đáp ứng nhu cầu của khách hàng, nâng cao sự trải nghiệm dịch vụ của người dùng.

+ Tối ưu hóa vận hành: Dữ liệu chính xác về sản phẩm giúp doanh nghiệp quản lý kho hàng hiệu quả hơn, giảm lãng phí thời gian và tối ưu hóa quá trình phân phối.

+ Đưa ra quyết định dựa trên lữ liệu: Doanh nghiệp có thể dựa trên dữ liệu chính xác để đưa ra quyết định thay đổi chiến lược một cách đúng đắn.

+ Nâng cao khả năng cạnh tranh: Ứng dụng công nghệ vào thu thập dữ liệu, doanh nghiệp nhanh chóng thích ứng được xự thay đổi của thị trường và thay đổi chiến lược cho phù hợp, tăng cường vị thế cạnh tranh.

Với những lợi ích trên dự án không chỉ giúp doanh nghiệp nâng cao hiểu quả mà còn tăng cường khả năng và phát triển bền vững trong lĩnh vực dược phẩm.

### *Ý nghĩa khoa học và thực tiễn của đề tài*

**Ý nghĩa khoa học:** Dự án của nhóm chúng tôi sẽ đóng góp vào lĩnh vực tự động hóa thu thập dữ liệu bằng cách sử dụng công cụ Selenium và MongoDB. Việc áp dụng hai công cụ này giúp đơn giản hóa quá trình thu thập dữ liệu từ website, đặc biệt là dữ liệu về sản phẩm từ hệ thống nhà thuốc. Dự án không chỉ mở rộng kiến thức về cách tự động hóa trong thu thập dữ liệu mà còn giúp tích hợp những phương pháp xử lý dữ liệu cơ bản để khai thác thông tin hữu ích từ các nguồn dữ liệu trực tuyến. Điều này giúp các nhà phân tích và chuyên gia những công cụ hiện đại hơn trong công việc xử lý dữ liệu, tiềm năng ứng dụng được trong các dự án lớn.

**Ý nghĩa thực tiễn:** Dự án này mang lại nhiều lợi ích cho nhà thuốc Pharmacity trong việc tự động hóa thư thập và quản lý dữ liệu sản phẩm. Nhờ đó, dữ liệu sản phẩm được cập nhật liên tục giúp doanh nghiệp hiểu rõ hơn về xu hướng tiêu thụ trên thị trường, đẩy nhanh được chiến lược marketing và cải thiện dịch vụ đối với khách hàng. Việc áp dụng công cụ này giúp doanh nghiệp giảm bớt được thời gian và nguồn nhân lực, nắm bắt nhanh chóng trước những thay đổi của thị trường. Điều này đặc biệt quan trọng khi doanh nghiệp cần phải duy trì được lợi thế cạnh tranh trong môi trường phát triển hiện nay. Phương pháp thu thập và quản lý dữ liệu từ dự án có thể được mở rộng ra nhiều ngành hàng khác . Giúp nâng cao hiệu quả quản lý sản phẩm trong kinh doanh.

Ngoài ra, kết quả của dự án mang lại giúp doanh nghiệp có thể đưa ra những chiến lược dựa trên dữ liệu một cách chính xác thay vì dựa trên cảm tính. Điều này góp phần hỗ trở các nhà quản lý trong việc phát triển bền vững trong hoạt động kinh doanh. Nhờ sự kết hợp của lý thuyết và tự động hóa vào trong ứng dụng thực tiễn, dự án không chỉ mang lại giá trị khoa học mà còn đóng góp quan trọng vào việc nâng cao hiệu và phát triển bền vững của doanh nghiệp.

## Mục tiêu

### *Mục tiêu tổng quan*

Dự án này nhằm phát triển phương pháp tự động thu thập dữ liệu và quản lý dữ liệu sản phẩm nhà thuốc Pharmacity bằng công cụ Selenium để thu thập tự động dữ liệu, lưu trữ và quản lý dữ bằng MongoDB. Mục tiêu giúp doanh nghiệp nắm bắt nhanh chóng và chính xác thông tin sản phẩm, từ đó đưa ra được chiến lược kinh doanh hiệu quả, cải thiện hoạt động và tăng cường khả năng cạnh tranh trên thị trường.

### *Mục tiêu cụ thể*

Trong dự án này, nhóm tôi sẽ sử dụng các dữ liệu dược phẩm đã thu thập được từ website trực tuyến của nhà thuốc Pharmacity. Bao gồm thông tin sản phẩm, giá cả và đánh giá của khách hàng. Quá trình thực hiện bao gồm: xác định dữ liệu cần được thu thập và thu thập bằng Selenium, lưu trữ dữ liệu trong MongoDB, áp dụng phương pháp phân tích dữ liệu và đánh giá hiệu quả của hệ thống quản lý dữ liệu.

Cụ thể hơn, nhóm tôi sẽ triển khai công cụ Selenium để thu thập dữ liệu tự động và dùng MongoDB để quản lý và lưu trữ dữ liệu đã thu thập được. Sau đó, xử lý dữ liệu và phân tích dữ liệu để hỗ trợ doanh nghiệp đưa ra được các chiến lược tối ưu hóa vận hành, nâng cao chất lượng dịch vụ. Kết quả mang lại mong đợi là thu thập và quản lý dữ liệu một cách hiệu quả, giúp doanh nghiệp dễ dàng theo dõi và quản lý thông tin sản phẩm từ đó cải thiện hiệu suất kinh doanh và tăng cường khả năng cạnh tranh.

## Đối tượng và phạm vi

### *Đối tượng*

Đối tượng nghiên cứu của dự án là các sản phẩm đang được bán trên hệ thống trực tuyến của nhà thuốc Pharmacity. Dữ liệu sẽ bao gồm thông tin về sản phẩm, giá cả đánh giá cảu khách hàng. Mục tiêu của dự án là thu thập và phân tích dữ liệu để hiểu rõ hơn về những yếu tố ảnh hưởng đến hoạt động kinh doanh của Pharmacity, từ đó để xuất các giả pháp giúp cái thiện hiệu suất và tối ưu hóa quá trình vẫn hành.

### *Phạm vi*

Phạm vi phân tích tập trung vào công việc thu thập dư liệu sản phẩm từ website của nhà thuốc Pharmacity thông qua công cụ Selenium và lưu trữ, quản lý bằng MongoDB. Dự án tập trung vào phân tích đánh giá thông tin sản phẩm, phản hồi và xu hướng tiêu thụ của khách hàng, tối ưu hóa chiến lược bán hàng và cải thiện trải nghiệm khách hàng.

## Phương pháp nghiên cứu

### *Phương pháp nghiên cứu sơ bộ*

Trước khi tiến hành thu thập dữ liệu, chúng tôi sẽ thực hiện nghiên cứu sơ bộ để hiểu rõ hơn về ngành dược phẩm trực tuyến và các yếu tố liên quan. Nghiên cứu sơ bộ sẽ bao gồm việc tìm hiểu về hoạt động kinh doanh trực tuyến của nhà thuốc Pharmacity, các yếu tố ảnh hưởng đến doanh thu và đánh giá sản phẩm của khách hàng, cũng như các phương pháp thu thập dữ diện tự động và phân tích được một cách hiệu quả. Thông qua quá trình này, chúng tôi sẽ xác định những vấn đề cần được gải quyết và lựa chọn các phương pháp nghiên cứu phù hợp cho dự án.

### *Phương pháp nghiên cứu tài liệu*

Chúng tôi sẽ tiến hành nghiên cứu tài liệu để thu thập thông tin về các công cụ và phương pháp tự động hóa thu thập dữ liệu và quán lý dữ liệu bằng công cụ Selenium và MongoDB. Qua việc đánh giá các nghiên cứu trước đây và các công trình khóa học liên quan, chúng tôi xác định kĩ thuật phù hợp nhất để triển khai việc thu thập dữ liệu dược phẩm từ website của Pharmacity và quản lý chúng một cách hiệu quả. Từ đó, những công cụ này sẽ xây dựng được hệ thống tự động hóa thu thập dữ liệu và phân tích dữ liệu cho dự án.

### *Phương pháp nghiên cứu thống kê*

Trong quá trình phân tích dữ liệu thu thập từ website của Pharmacity, chúng tôi sẽ sử dụng các phương pháp thống kê để mô tả và phân tích các yếu tố liên quan đến sản phẩm. Nhóm tôi sẽ đánh giá số lượng khách mua hàng của tưng sản phẩm và đánh giá của khách hàng. Từ đó, chúng tôi sẽ đánh giá được các yếu tố ảnh hưởng đến hoạt động kinh doanh của Pharmacity đồng thời đề xuất các giải pháp để nâng cao năng suất kinh doanh.

### *Phương pháp thực nghiệm*

Nhóm sẽ thực hiện quá trình thực nghiệm dựa trên dữ liệu dược phẩm được thu thập từ website Pharmacity. Bap gồm các bước xử lý dữ liệu sản phẩm, áp dụng các phương phân tích để đánh giá xu hướng tiêu thụ của khách hàng, cũng như đánh giá những chiến lược kinh doanh đã đề ra. Thông qua thực nghiệm này, nhóm sẽ đưa ra được tính khả thi và hiệu quả của phương pháp nghiên cứu, nhằm tối ưu hóa quy trình quản lý sản phẩm, nâng cao hiệu suất cạnh tranh của Pharmacity trên thị trường trực tuyến.

### *Phương pháp đánh giá*

Sau khi thu thập và phân tích dữ liệu, nhóm sẽ tiến hành đánh giá và đo lường hiệu quả của quá trình phân tích mang lại. Bước này sẽ bao gồm so sánh về thông tin của sản phẩm, giá cả, phản hồi của khách hàng và xu hướng tiêu thụ để đánh giá hiệu suất của kinh doanh. Nhóm đối chiếu kết quả này với thực tế hoạt động của Pharmacity để đưa ra giải pháp cụ thể giúp tối ưu hóa quá trình khinh doanh, nâng cao hiệu quả quán lý sản phẩm và cải thiện tổng thể hoạt động của doanh nghiệp.

# CƠ SỞ LÝ THUYẾT

## Selenium

### *Giới thiệu và trích xuất dữ liệu từ Selenium.*

Selenium là một công cụ mã nguồn mở tự động hóa trình duyệt web, cho phép người dùng điều khiển các trình duyệt một cách tự động. Được phát triển bởi Jason Huggins vào nằm 2004 và cho đến hiện nay trở thành một công cụ phổ biến nhất dùng để kiểm thử, tự động hóa và tương tác với trang web. Selenium được thực hiện thông qua các mã lệnh của người dùng, cho phép tự động các thao tác như nhấp chuột, nhập dữ liệu, cuộn trang và các hành động khác do người thực hiện các câu mã lệnh đưa ra.

### *Ưu điểm và nhược điểm*

### *Thành phần chính của selenium*

Selenium gồm có 4 phần chính:

* Selenium Integrated Development Enviroment (Selenium IDE): Đây là một công cụ cung cấp giao diện người dùng đơn giản để ghi, chỉnh sửa và chạy các kịch bản thử nghiệm trên trình duyệt web.IDE thích hợp cho các bước kiểm thử đơn giản, nhưng hạn chế khi tự động hóa phức tạp hoặc thu thập dữ liệu từ trang web động.
* Selenium Webdriver: Đây là thành phần chính của Selenium với khả năg API linh hoạt để tương tác với các trình duyệt như Chrome, Firefox, Edge, Safari mà không cần qua lớp trung gian, Webdriver sử dụng API đơn giản nhưng mạnh mẽ để điều khiển được trình duyệt thực hiện các thao tác nhấp chuột, cuộn trang đồng thời thu thập dữ liệu.
* Selenium Grid: Công cụ có nhiệm vụ phục vụ cho việc thực thi đồng thời các kịch bản thử nghiệm trên nhiều trình duyệt và hệ thống khác nhau. Selenium Grid cho phép phân phối các kịch bản thử nghiệm trên các thiết bị và môi trường riêng biệt. Từ đó tăng hiệu suất và giảm thời gian kiểm thử.[1]
* Selenium RC (Remote Control): Đây là một công cụ cũ của Selenium nhưng đã được loại bỏ và thay thế bởi Senium Webdriver do sự phức tạp và hiệu suất thấp hơn. Selenium RC điều khiển trình duyệt thông qua JavaScript và hoạt động như một lớp trung gian giữa mã kiểm thử và trình duyệt.

### *Cách thức Selenium hoạt động*

# KẾT QUẢ THỰC NGHIỆM

# KẾT LUẬN VÀ KIẾN NGHỊ

## Kết luận

## 

## Kiến nghị

# TÀI LIỆU THAM KHẢO

[1] FPT C. ty C. phần B. lẻ K., “Selenium là gì? Tìm hiểu các tính năng nổi bật của Selenium trong lĩnh vực phần mềm.” Accessed: Oct. 22, 2024. [Online]. Available: https://fptshop.com.vn/tin-tuc/danh-gia/selenium-la-gi-167783