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

**ĐỒ ÁN MÔN HỌC**

**<TÌM HIỂU VỀ CÔNG CỤ SELENIUM VÀ MONGODB>**

Ngành: **<KHOA HỌC DỮ LIỆU >**

Chuyên ngành: **<KHOA HỌC DỮ LIỆU >**

Giảng viên hướng dẫn :ThS. LÊ NHẬT TÙNG

Sinh viên thực hiện :

2286400003-NGUYỄN HOÀNG ANH

2286400041-BẠCH QUANG TÙNG

2286400004-NGUYỄN TUẤN ANH

Lớp: 22DKHA1

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**

Chúng tôi, Nguyễn Tuấn Anh, Bạch Quang Tùng và Nguyễn Hoàng Anh, xin cam đoan rằng:

Tất cả thông tin và nghiên cứu được trình bày trong bài báo cáo này là chính xác, trung thực và khách quan. Dữ liệu đã được thu thập và phân tích một cách cẩn thận từ các nguồn tin cậy và chính thống.

Mọi thông tin hoặc quan điểm được trích dẫn từ nguồn khác đều được ghi rõ nguồn gốc và tuân thủ quy định về trích dẫn. Chúng tôi cam kết rằng không có bất kỳ hành vi sao chép hoặc sử dụng thông tin không hợp lệ nào từ các nguồn bên ngoài.

Bài báo cáo này là kết quả nghiên cứu độc lập của chúng tôi và chưa từng được công bố ở bất kỳ đâu trước đây. Chúng tôi đã thực hiện đầy đủ các quy tắc và quy định của môn học, bao gồm cả việc tham khảo và sử dụng công cụ nghiên cứu.

Chúng tôi mong rằng bài báo cáo này sẽ mang lại cái nhìn tổng quan rõ ràng và toàn diện về chủ đề “Dạy học lập trình trên trang web titv”, góp phần vào sự phát triển của lĩnh vực nghiên cứu này.

TPHCM, ngày 09 tháng 06 năm 2024  
Sinh viên  
Nguyễn Tuấn Anh  
Bạch Quang Tùng  
Nguyễn Hoàng Anh

**Mục Lục**

**Trang phụ bìa**

**Lời cam đoan**

**Chương 1: Tổng Quan**  
1.1.Giới thiệu  
1.2. Mục tiêu của báo cáo  
1.3. Phương pháp nghiên cứu

1.4. Nhiệm vụ đồ án

1.5. Cấu trúc đồ án

**Chương 2: Cơ sở lý thuyết**

2.1. Giới thiệu khái quát về công cụ Selenium và MongoDB  
2.2. Định nghĩa và lịch sử phát triển  
2.3. Các thành phần của Selenium  
• 2.3.1. Selenium WebDriver  
• 2.3.2. Selenium IDE  
• 2.3.3. Selenium Grid  
2.4. Tính năng nổi bật  
2.5. Ứng dụng của Selenium trong thu thập dữ liệu  
2.6. Giới thiệu về MongoDB và ứng dụng trong lưu trữ dữ liệu

2.7. Tổng quan về website TITV  
2.8.. Lịch sử hình thành và phát triển

2.9. Mô hình kinh doanh và dịch vụ  
2.10. Thị trường và đối thủ cạnh tranh

**Chương 3 : Phương pháp thu thập dữ liệu**  
3.1. Xác định mục tiêu thu thập dữ liệu

3.2. Thiết kế quy trình thu thập dữ liệu  
3.3. Sử dụng Selenium trong thu thập dữ liệu từ TITV  
• 3.3.1. Cài đặt và cấu hình Selenium  
• 3.3.2. Viết mã tự động hóa  
• 3.3.3. Xử lý và lưu trữ dữ liệu bằng MongoDB

**Chương 4 :Kết quả thực nghiệm**  
4.1. Phân tích dữ liệu thu thập được

4.2. Đánh giá chất lượng dữ liệu  
4.3. So sánh với dữ liệu nguồn khác

4.4. Phân tích dữ liệu bằng MongoDB

4.5 Truy vấn dữ liệu từ dữ liệu MongoDB

**Chương 5: Kết luận và kiến nghị**

5.1Thảo luận  
• 5.1.1 Những thách thức trong quá trình thu thập dữ liệu

• 5.1.2 Đề xuất cải tiến quy trình thu thập dữ liệu

5.2. Kết luận  
• 5.2.1. Tóm tắt kết quả nghiên cứu (tích hợp video cào dữ liệu)

• 5.2.2. Định hướng nghiên cứu tiếp theo

**Tài liệu tham khảo**

**Phụ lục**  
Mã nguồn mẫu sử dụng Selenium  
Biểu đồ và bảng số liệu liên quan

**Chương 1: Tổng Quan**

**1.1.Giới thiệu**

Trong thời đại kỹ thuật số hiện nay, học tập trực tuyến đã trở thành một xu hướng ngày càng phổ biến trên toàn cầu, đặc biệt là sau ảnh hưởng mạnh mẽ của đại dịch COVID-19, khi hàng triệu người buộc phải chuyển sang hình thức học tập từ xa. Các nền tảng học tập trực tuyến đã mang đến một môi trường thuận lợi, linh hoạt và hiệu quả cho người học ở mọi độ tuổi, từ học sinh, sinh viên đến những người đã đi làm muốn mở rộng kỹ năng hoặc chuyển đổi nghề nghiệp. Tại Việt Nam, sự phát triển nhanh chóng của các nền tảng học trực tuyến như TITV không chỉ đáp ứng nhu cầu học tập của người dùng mà còn góp phần thúc đẩy quá trình chuyển đổi số trong giáo dục.

Bên cạnh các lợi ích mà nền tảng học trực tuyến mang lại, một thách thức lớn đối với các nhà phát triển là làm thế nào để hiểu được nhu cầu và sở thích học tập của người dùng. Việc nắm bắt các xu hướng học tập, sự tương tác với các khóa học, hoặc độ phổ biến của các chủ đề có thể giúp các nền tảng này cải tiến nội dung và gia tăng trải nghiệm người học. Trong bối cảnh đó, khai thác dữ liệu từ các nền tảng học tập trực tuyến đã trở thành một phương pháp vô cùng hữu ích, cho phép các nhà nghiên cứu và phát triển dựa vào dữ liệu thực tế để đưa ra những cải tiến chính xác và hiệu quả.

Vai trò của tự động hóa và thu thập dữ liệu trong việc tối ưu hóa nền tảng giáo dục:

-Với sự phát triển của công nghệ tự động hóa và các công cụ thu thập dữ liệu, các nền tảng học tập trực tuyến ngày nay không chỉ đóng vai trò là nơi cung cấp kiến thức mà còn là nguồn dữ liệu quý giá cho việc nghiên cứu và phân tích xu hướng học tập. Trong số các công cụ hỗ trợ tự động hóa, Selenium là một trong những lựa chọn phổ biến nhất, cho phép thực hiện các tác vụ như duyệt web, tương tác với giao diện người dùng và lấy dữ liệu tự động từ các trang web động. Công cụ này không chỉ giúp tiết kiệm thời gian mà còn đảm bảo tính chính xác và tính toàn vẹn của dữ liệu thu thập.

-Để quản lý và tổ chức lượng dữ liệu khổng lồ từ các nền tảng học trực tuyến, MongoDB là một lựa chọn lý tưởng vì khả năng lưu trữ linh hoạt và hỗ trợ quản lý dữ liệu phi cấu trúc. Với khả năng xử lý các loại dữ liệu không đồng nhất, MongoDB giúp các nhà phát triển tổ chức và phân tích dữ liệu một cách hiệu quả, tạo điều kiện thuận lợi cho việc phân tích hành vi người dùng và đưa ra những cải tiến chiến lược cho nền tảng học tập.

Tầm quan trọng của đề tài trong việc nâng cao chất lượng giáo dục trực tuyến:

-Mục tiêu của đề tài này là nghiên cứu quy trình tự động hóa thu thập dữ liệu từ nền tảng TITV, sử dụng Selenium để tự động hóa các thao tác truy cập và lấy dữ liệu, và ứng dụng MongoDB trong việc lưu trữ và quản lý dữ liệu thu thập được. Đề tài không chỉ mang lại lợi ích cho TITV mà còn có giá trị đối với các nhà nghiên cứu và nhà phát triển mong muốn khai thác dữ liệu từ các nền tảng trực tuyến để hiểu rõ hơn về hành vi học tập của người dùng.

-Thông qua quá trình thu thập và phân tích dữ liệu, báo cáo này hy vọng cung cấp các phát hiện quan trọng về thói quen học tập, các chủ đề được yêu thích, và các yếu tố ảnh hưởng đến quyết định tham gia khóa học của người dùng. Từ đó, đề tài có thể đưa ra những đề xuất thiết thực nhằm cải thiện chất lượng nội dung, tối ưu hóa giao diện và trải nghiệm người dùng, đồng thời góp phần nâng cao hiệu quả học tập của người dùng trên nền tảng TITV. Kết quả nghiên cứu không chỉ đóng góp giá trị thực tiễn cho nền tảng TITV mà còn góp phần thúc đẩy ứng dụng công nghệ tự động hóa và phân tích dữ liệu trong lĩnh vực giáo dục trực tuyến tại Việt Nam.

Sự kết hợp giữa tự động hóa thu thập dữ liệu bằng Selenium và quản lý dữ liệu bằng MongoDB là một giải pháp tối ưu để nâng cao hiệu quả và chất lượng giáo dục trực tuyến. Với đề tài này, không chỉ TITV mà các nền tảng học tập trực tuyến khác cũng có thể tận dụng để cải thiện nội dung và đáp ứng tốt hơn nhu cầu của người học, tạo ra những trải nghiệm học tập trực tuyến hấp dẫn, hiệu quả và bền vững hơn trong kỷ nguyên số.  
**1.2. Mục tiêu của báo cáo**

Báo cáo này tập trung vào việc xây dựng và triển khai một quy trình tự động hóa thu thập dữ liệu từ nền tảng học tập trực tuyến TITV, sử dụng Selenium - công cụ hỗ trợ tự động hóa kiểm thử phần mềm và thu thập dữ liệu web. Mục tiêu cụ thể của báo cáo được chia thành các khía cạnh như sau:

1. Khám phá và nghiên cứu các phương pháp sử dụng Selenium cho việc tự động hóa thu thập dữ liệu: Với sự phát triển của công nghệ hiện nay, việc thu thập dữ liệu không chỉ dừng lại ở việc thu thập thủ công, mà còn hướng tới các giải pháp tự động hóa và tối ưu. Selenium, với khả năng thực hiện các tác vụ tự động hóa trên nền web, đóng vai trò quan trọng giúp tiết kiệm thời gian và đảm bảo tính chính xác của dữ liệu. Báo cáo sẽ đi sâu vào tìm hiểu cách Selenium có thể thao tác và thu thập dữ liệu từ các trang có giao diện động, yêu cầu tương tác người dùng.
2. Ứng dụng MongoDB trong quản lý dữ liệu phi cấu trúc: MongoDB là một trong những cơ sở dữ liệu NoSQL phổ biến, đặc biệt hiệu quả trong việc lưu trữ và truy xuất dữ liệu không cấu trúc. Báo cáo sẽ trình bày chi tiết cách lưu trữ và tổ chức dữ liệu từ TITV bằng MongoDB, nhằm giúp người đọc thấy rõ vai trò của hệ quản trị cơ sở dữ liệu này trong các dự án khai thác và phân tích dữ liệu từ các trang web. Cơ sở dữ liệu được xây dựng sẽ được dùng để lưu trữ các thông tin khóa học và người học trên TITV, tạo nên nền tảng phân tích xu hướng học tập.
3. Phân tích dữ liệu và đánh giá chất lượng dữ liệu thu thập: Sau khi thu thập, dữ liệu sẽ được kiểm tra và đánh giá dựa trên các tiêu chí như tính đầy đủ, tính chính xác, tính liên quan và tính nhất quán. Thông qua MongoDB, dữ liệu sẽ được truy xuất và phân tích để rút ra các xu hướng học tập phổ biến, từ đó đưa ra những phân tích chuyên sâu về nội dung khóa học, loại khóa học nào được ưa chuộng, cũng như các đặc điểm học viên như độ tuổi, trình độ. Việc này không chỉ giúp cải thiện nội dung học tập mà còn là cơ sở để nâng cao trải nghiệm người dùng.
4. Đề xuất các phương án cải tiến nền tảng TITV từ dữ liệu phân tích: Cuối cùng, báo cáo sẽ đưa ra một số gợi ý nhằm tối ưu hóa nền tảng học tập TITV dựa trên các kết quả phân tích được từ dữ liệu thu thập. Cụ thể, các khóa học, chủ đề, và phương pháp giảng dạy sẽ được xem xét và đề xuất cải tiến nhằm đáp ứng đúng nhu cầu của người học và tăng cường sự hài lòng của người dùng trên nền tảng TITV.

Với các mục tiêu trên, báo cáo sẽ đóng góp giá trị cả về lý thuyết lẫn thực tiễn cho các dự án khai thác dữ liệu từ website giáo dục, đặc biệt là các nền tảng giáo dục trực tuyến.

**1.3. Phương pháp nghiên cứu**

Báo cáo này áp dụng phương pháp nghiên cứu kết hợp giữa nghiên cứu định tính và định lượng, với các bước cụ thể như sau:

1. Nghiên cứu tài liệu  
 Trong bước này, báo cáo sẽ phân tích và tổng hợp các tài liệu về công cụ Selenium, cơ sở dữ liệu MongoDB, và các kỹ thuật thu thập dữ liệu web. Mục đích của nghiên cứu tài liệu là để nắm vững những khái niệm lý thuyết và ứng dụng thực tế của Selenium trong việc thao tác trên các trang web động, đồng thời tìm hiểu cơ chế lưu trữ và quản lý dữ liệu phi cấu trúc của MongoDB. Tài liệu được sử dụng bao gồm sách, các bài báo khoa học, tài liệu hướng dẫn từ nhà phát triển, và các nghiên cứu điển hình (case studies) về sử dụng Selenium và MongoDB trong thu thập và phân tích dữ liệu web.

2. Thu thập dữ liệu từ thực nghiệm  
 Phương pháp thu thập dữ liệu thực nghiệm sẽ được tiến hành bằng cách thiết lập và triển khai một hệ thống tự động hóa dựa trên Selenium để lấy thông tin từ nền tảng TITV. Các thao tác này bao gồm truy cập trang web, lấy dữ liệu từ các khóa học, và lưu trữ dữ liệu vào MongoDB. Quy trình này được lặp đi lặp lại nhiều lần để đảm bảo tính chính xác và đầy đủ của dữ liệu thu thập được.

3. Phân tích và xử lý dữ liệu  
 Dữ liệu sau khi được thu thập sẽ được kiểm tra để loại bỏ các yếu tố dư thừa hoặc lỗi, sau đó tiến hành phân tích dựa trên các câu hỏi nghiên cứu đặt ra từ đầu báo cáo. Phân tích dữ liệu trong MongoDB cho phép thực hiện các truy vấn phức tạp, giúp khám phá những điểm nổi bật của dữ liệu, như sự phân bố các loại khóa học, tỉ lệ người tham gia, và mức độ hoàn thành các khóa học trên nền tảng TITV. Từ các kết quả phân tích, báo cáo có thể xây dựng các biểu đồ và bảng số liệu để minh họa cho các xu hướng học tập.

4. Đề xuất và đánh giá  
 Sau khi có kết quả phân tích, báo cáo sẽ tiến hành đánh giá và đề xuất các phương án cải tiến cho nền tảng TITV, nhằm mục tiêu tối ưu hóa trải nghiệm người dùng và cải thiện nội dung khóa học. Các đề xuất này được xây dựng dựa trên các phát hiện từ dữ liệu phân tích, chẳng hạn như việc bổ sung các khóa học phổ biến hoặc cải thiện các chủ đề ít người quan tâm.

**1.4. Nhiệm vụ đồ án**

Đồ án này nhằm giải quyết các yêu cầu trong việc thu thập và phân tích dữ liệu từ các nền tảng giáo dục trực tuyến, cụ thể là nền tảng TITV tại Việt Nam. Nhiệm vụ cụ thể của đồ án bao gồm:

1. Nghiên cứu công cụ Selenium và MongoDB: Để thiết kế một quy trình thu thập và lưu trữ dữ liệu hiệu quả, đầu tiên cần nghiên cứu và nắm vững các chức năng của Selenium và MongoDB. Selenium được sử dụng để tự động hóa việc lấy dữ liệu từ TITV, trong khi MongoDB đóng vai trò là nơi lưu trữ và quản lý dữ liệu thu thập được. Đồ án sẽ mô tả chi tiết cách Selenium truy cập, tương tác và thu thập dữ liệu từ trang web, đồng thời hướng dẫn cách thức tổ chức và lưu trữ dữ liệu này trong MongoDB.
2. Thiết lập hệ thống tự động hóa thu thập dữ liệu: Nhiệm vụ tiếp theo là thiết kế và triển khai một hệ thống tự động hóa thu thập dữ liệu. Để đáp ứng yêu cầu này, đồ án sẽ xây dựng một quy trình tự động hóa hoàn chỉnh từ bước truy cập, lấy dữ liệu đến lưu trữ vào MongoDB. Mỗi khóa học, lượt xem, và thông tin chi tiết về các khóa học trên TITV sẽ được lưu trữ để phục vụ cho việc phân tích sau này.
3. Phân tích và trình bày dữ liệu thu thập được: Với dữ liệu đã được lưu trữ, đồ án sẽ tiến hành phân tích nhằm rút ra những xu hướng học tập của người dùng trên TITV. Các phân tích bao gồm sự phổ biến của các khóa học, các chủ đề được yêu thích, và thời gian học tập trung bình của người dùng. Đồ án cũng sẽ trình bày kết quả này dưới dạng các biểu đồ và bảng dữ liệu, giúp minh họa rõ ràng và trực quan cho các phát hiện từ dữ liệu.
4. Đề xuất các giải pháp cải thiện nền tảng TITV: Dựa trên kết quả phân tích, đồ án sẽ đưa ra các đề xuất để cải thiện trải nghiệm người dùng trên TITV. Các giải pháp này bao gồm cải tiến nội dung khóa học, đề xuất bổ sung các chủ đề mới, và tối ưu hóa giao diện để tăng cường sự hài lòng của người học.

**1.5. Cấu trúc đồ án**

Đồ án được tổ chức thành bốn chương như sau:

* Chương 1: Tổng Quan

- Giới thiệu về bối cảnh nghiên cứu, các mục tiêu và phương pháp nghiên cứu, cùng với các nhiệm vụ của đồ án. Phần này đặt nền tảng cho toàn bộ nội dung báo cáo, giúp người đọc có cái nhìn tổng quan về mục tiêu và hướng tiếp cận của đồ án.

* Chương 2: Cơ Sở Lý Thuyết

- Trình bày chi tiết các khía cạnh lý thuyết liên quan đến Selenium và MongoDB. Chương này cũng giải thích cách thức hoạt động của Selenium trong việc thu thập dữ liệu từ các trang web và vai trò của MongoDB trong quản lý và lưu trữ dữ liệu phi cấu trúc.

* Chương 3: Phương Pháp Thu Thập Dữ Liệu:

- Chương này trình bày chi tiết các phương pháp được sử dụng để thu thập và xử lý dữ liệu từ nền tảng học trực tuyến TITV, trong đó có các công cụ và quy trình cụ thể để tự động hóa quy trình lấy dữ liệu. Các bước chính bao gồm sử dụng Selenium để truy cập và tương tác tự động với các trang web động của TITV, từ đó thu thập dữ liệu như tên khóa học, danh mục, số lượng người tham gia, lượt xem, đánh giá, và xếp hạng.

* Chương 4: Kết Quả Thực Nghiệm

- Tóm tắt quy trình thu thập dữ liệu từ TITV, kết quả phân tích dữ liệu, và các phát hiện quan trọng. Chương này bao gồm các biểu đồ và bảng số liệu để minh họa cho các xu hướng học tập từ nền tảng TITV.

* Chương 5: Kết Luận và Kiến Nghị - Đưa ra kết luận dựa trên các phát hiện từ quá trình thu thập và phân tích dữ liệu. Chương này cũng sẽ đề xuất các cải tiến cho nền tảng TITV nhằm tối ưu hóa trải nghiệm người dùng và nâng cao chất lượng khóa học.

**Chương 2 : Cơ Sở Lý Thuyết**

**2.1. Giới thiệu khái quát về công cụ Selenium và MongoDB**

Trong bối cảnh công nghệ số ngày nay, tự động hóa và quản lý dữ liệu là những yếu tố quan trọng quyết định thành công của các dự án thu thập dữ liệu trên quy mô lớn. Selenium và MongoDB đã nổi lên như những công cụ hàng đầu, hỗ trợ các kỹ sư phần mềm và nhà khoa học dữ liệu tối ưu hóa quá trình thu thập và quản lý dữ liệu từ các trang web động. Selenium là một công cụ mã nguồn mở mạnh mẽ cho phép tự động hóa các thao tác trên trình duyệt, trong khi MongoDB cung cấp một hệ quản trị cơ sở dữ liệu NoSQL linh hoạt, giúp lưu trữ dữ liệu không cấu trúc một cách hiệu quả.

Công cụ Selenium hỗ trợ các lập trình viên thu thập dữ liệu mà không cần thực hiện các thao tác thủ công trên trang web, giúp tiết kiệm thời gian và công sức. Điều này cực kỳ hữu ích đối với các trang web phức tạp như TITV, nơi thông tin về các khóa học liên tục được cập nhật và không thể lấy thông tin dễ dàng thông qua các phương pháp truyền thống. Cùng với đó, MongoDB giúp tối ưu hóa việc lưu trữ các thông tin này, cung cấp khả năng truy vấn mạnh mẽ và xử lý khối lượng dữ liệu lớn, tạo ra nền tảng cho việc phân tích và quản lý dữ liệu hiệu quả hơn. **2.2. Định nghĩa và lịch sử phát triển**

Selenium và MongoDB đều có hành trình phát triển riêng biệt, đáp ứng các nhu cầu khác nhau trong hệ sinh thái công nghệ thông tin. Selenium được phát triển vào năm 2004 bởi Jason Huggins tại ThoughtWorks, với mục tiêu ban đầu là tạo ra một công cụ để tự động hóa các bài kiểm thử trên trình duyệt web. Từ đó, Selenium đã mở rộng và cải tiến qua các giai đoạn phát triển khác nhau, với các phiên bản đáng chú ý như Selenium WebDriver, Selenium IDE và Selenium Grid. Mỗi phiên bản đều cải thiện các tính năng tự động hóa, hỗ trợ người dùng một cách linh hoạt trong việc kiểm thử và thu thập dữ liệu từ các trang web động.

MongoDB, ra đời vào năm 2007, là một hệ quản trị cơ sở dữ liệu NoSQL tiên phong trong việc xử lý dữ liệu phi cấu trúc. Được thiết kế bởi 10gen (nay là MongoDB Inc.), MongoDB ra đời nhằm đáp ứng nhu cầu lưu trữ dữ liệu ngày càng lớn và phức tạp trong các ứng dụng hiện đại. Với kiến trúc lưu trữ theo dạng tài liệu JSON, MongoDB không chỉ linh hoạt hơn so với các hệ quản trị cơ sở dữ liệu quan hệ truyền thống mà còn cho phép xử lý và mở rộng quy mô dễ dàng. Sự phát triển của MongoDB đã góp phần thay đổi cách thức lưu trữ và quản lý dữ liệu trong kỷ nguyên dữ liệu lớn (Big Data), nơi dữ liệu không còn chỉ là các bảng thông tin tĩnh mà còn là các tài liệu phức tạp với nhiều loại hình và cấu trúc khác nhau. **2.3. Các thành phần của Selenium**

Selenium bao gồm ba thành phần chính là Selenium WebDriver, Selenium IDE, và Selenium Grid. Mỗi thành phần đóng vai trò quan trọng trong việc cung cấp các phương pháp tiếp cận và giải quyết các thách thức riêng trong tự động hóa trình duyệt web, từ kiểm thử đến thu thập dữ liệu. Hiểu rõ cách hoạt động của từng thành phần giúp tối ưu hóa quy trình làm việc và tăng cường khả năng áp dụng trong các tình huống cụ thể.*• 2.3.1. Selenium WebDriver:*

Selenium WebDriver là một trong những thành phần mạnh mẽ nhất của Selenium, được thiết kế để tương tác trực tiếp với trình duyệt web thông qua API. Đặc điểm nổi bật của WebDriver là khả năng điều khiển trình duyệt thực sự, không phải qua một bản mô phỏng, cho phép người dùng tái tạo các hành vi của con người trên trang web. Selenium WebDriver hỗ trợ nhiều ngôn ngữ lập trình phổ biến như Python, Java, C#, Ruby và JavaScript, giúp các lập trình viên dễ dàng tích hợp vào các dự án của mình. Đặc biệt, WebDriver có khả năng tương thích với nhiều loại trình duyệt khác nhau, bao gồm Chrome, Firefox, Safari và Microsoft Edge. Khả năng linh hoạt của WebDriver cho phép tự động hóa các thao tác phức tạp như tìm kiếm và điền biểu mẫu, nhấp chuột và kiểm tra sự thay đổi của trang sau các thao tác này.

Khả năng hỗ trợ kiểm thử chéo trên nhiều trình duyệt giúp WebDriver trở thành công cụ tự động hóa không thể thiếu cho các doanh nghiệp và nhà phát triển, đặc biệt là khi cần kiểm thử ứng dụng web trên nhiều nền tảng khác nhau. Selenium WebDriver còn cho phép quản lý các cửa sổ trình duyệt, kiểm tra tương tác của người dùng với các phần tử trên trang và chạy các bài kiểm thử phức tạp, giúp người dùng tăng tốc quá trình phát triển phần mềm và thu thập dữ liệu từ các trang web động.

*• 2.3.2. Selenium IDE:*

Selenium IDE là một tiện ích mở rộng trình duyệt nhẹ và dễ sử dụng, được thiết kế dành cho những người dùng mới bắt đầu làm quen với Selenium. IDE có khả năng ghi lại các thao tác của người dùng trên trang web và phát lại chúng, nhờ đó giúp tự động hóa những tác vụ đơn giản mà không cần viết mã phức tạp. Đây là công cụ lý tưởng để ghi lại các kịch bản kiểm thử, đặc biệt hữu ích cho các dự án nhỏ hoặc các bài kiểm thử đơn giản, lặp đi lặp lại. Các thao tác được ghi lại dưới dạng mã lệnh, cho phép người dùng dễ dàng phát lại, chỉnh sửa hoặc sử dụng các đoạn mã đã ghi lại để phát triển các bài kiểm thử phức tạp hơn.

Khả năng ghi và phát lại thao tác giúp Selenium IDE trở thành công cụ hữu ích không chỉ cho kiểm thử tự động mà còn cho các nhu cầu thu thập dữ liệu đơn giản. Ngoài ra, IDE hỗ trợ chuyển đổi sang nhiều ngôn ngữ lập trình khác nhau, giúp dễ dàng mở rộng hoặc tích hợp với các công cụ và môi trường phát triển khác. *• 2.3.3. Selenium Grid:*

Selenium Grid là thành phần cho phép người dùng chạy các bài kiểm thử song song trên nhiều máy tính và trình duyệt khác nhau, nhờ đó tiết kiệm thời gian và tăng cường tính linh hoạt của quy trình kiểm thử. Grid hoạt động dựa trên kiến trúc chủ-tớ (hub-node), trong đó "hub" là máy chủ trung tâm phân phối các bài kiểm thử đến các "node" là các máy chủ hoặc trình duyệt khác nhau. Việc sử dụng Selenium Grid rất hữu ích cho các công ty và đội ngũ phát triển phần mềm khi muốn kiểm thử ứng dụng của mình trên nhiều nền tảng và trình duyệt khác nhau mà không cần thực hiện thủ công.  
 Tính năng song song hóa và phân phối này của Grid giúp tăng cường hiệu quả kiểm thử, đặc biệt là trong các hệ thống lớn và phức tạp. Khả năng tích hợp của Selenium Grid với các công cụ CI/CD (Continuous Integration/Continuous Deployment) còn giúp tối ưu hóa quy trình phát triển và kiểm thử liên tục, góp phần tăng cường độ tin cậy và chất lượng của phần mềm.

**2.4. Tính năng nổi bật**

Selenium nổi bật với nhiều tính năng vượt trội, giúp nó trở thành công cụ phổ biến cho việc tự động hóa và thu thập dữ liệu từ các trang web động. Các tính năng bao gồm:

1. Khả năng Tương Thích Với Nhiều Trình Duyệt: Selenium hỗ trợ nhiều trình duyệt như Chrome, Firefox, Safari và Edge, mang lại tính linh hoạt cao cho quá trình phát triển và kiểm thử ứng dụng. Khả năng kiểm thử đồng thời trên nhiều nền tảng giúp tiết kiệm thời gian và giảm thiểu rủi ro phát sinh lỗi khi ứng dụng được sử dụng trên các hệ điều hành và trình duyệt khác nhau.
2. Hỗ Trợ Nhiều Ngôn Ngữ Lập Trình: Selenium cho phép người dùng viết mã bằng nhiều ngôn ngữ lập trình như Python, Java, C#, JavaScript và Ruby, tạo điều kiện thuận lợi cho các lập trình viên với nền tảng khác nhau. Sự đa dạng này giúp dễ dàng tích hợp Selenium vào các dự án có sẵn, tăng cường hiệu quả phát triển và mở rộng quy mô.
3. Khả Năng Tự Động Hóa Linh Hoạt: Selenium có khả năng tự động hóa nhiều thao tác phức tạp, bao gồm cả thao tác với các trang web có nội dung động và các phần tử phức tạp như menu điều hướng, hộp thoại và các form nhập liệu. Đặc biệt, Selenium cho phép điều hướng và tương tác với các phần tử bị ẩn hoặc chỉ xuất hiện khi có hành động cụ thể.
4. Mở Rộng Với Các Công Cụ Kiểm Thử Khác: Selenium dễ dàng tích hợp với các công cụ kiểm thử và CI/CD khác như Jenkins, Maven và Docker, giúp tối ưu hóa quy trình phát triển và đảm bảo tính linh hoạt trong việc mở rộng hoặc thay đổi yêu cầu.

**2.5. Ứng dụng của Selenium trong thu thập dữ liệu**

Selenium là công cụ lý tưởng cho việc thu thập dữ liệu từ các trang web động, đặc biệt khi dữ liệu cần được cập nhật thường xuyên hoặc lấy từ các trang phức tạp. Với khả năng tự động hóa, Selenium có thể giúp tiết kiệm thời gian và công sức trong các tác vụ như cuộn trang, tìm kiếm, nhấp chuột, và lấy dữ liệu từ các trang yêu cầu tương tác người dùng. Ví dụ, khi cần thu thập dữ liệu từ các khóa học trực tuyến trên trang TITV, Selenium có thể giúp tự động cuộn trang, nhấn vào các khóa học, và lấy thông tin chi tiết về từng khóa học một cách nhanh chóng và chính xác.

Các lĩnh vực ứng dụng khác của Selenium trong thu thập dữ liệu bao gồm nghiên cứu thị trường, phân tích đối thủ cạnh tranh và khảo sát người dùng, đặc biệt đối với các trang thương mại điện tử, nơi thông tin sản phẩm thay đổi liên tục. Với khả năng tự động hóa các tác vụ phức tạp, Selenium cung cấp giải pháp hiệu quả cho việc thu thập dữ liệu từ các nguồn khác nhau, ngay cả khi dữ liệu được bảo vệ hoặc yêu cầu các thao tác xác thực phức tạp.

**2.6. Giới thiệu về MongoDB và ứng dụng trong lưu trữ dữ liệu**

MongoDB là hệ quản trị cơ sở dữ liệu NoSQL tiên tiến, cho phép lưu trữ và quản lý các loại dữ liệu phi cấu trúc và bán cấu trúc, đặc biệt phù hợp với các ứng dụng hiện đại yêu cầu tính linh hoạt cao. Khác với các hệ thống cơ sở dữ liệu quan hệ truyền thống, MongoDB lưu trữ dữ liệu theo dạng tài liệu JSON, cho phép lưu trữ các tài liệu phức tạp với nhiều thuộc tính và cấu trúc không đồng nhất. Điều này giúp MongoDB dễ dàng lưu trữ dữ liệu từ các trang web, nơi cấu trúc dữ liệu có thể thay đổi theo thời gian.

Trong bối cảnh nghiên cứu này, MongoDB đóng vai trò quan trọng trong việc lưu trữ các dữ liệu thu thập từ TITV, bao gồm thông tin về các khóa học, giảng viên, đánh giá và đánh giá xếp hạng. Với các tính năng như truy vấn linh hoạt, mở rộng quy mô dễ dàng và khả năng tích hợp với các công cụ phân tích dữ liệu, MongoDB trở thành một lựa chọn tối ưu cho việc lưu trữ và quản lý dữ liệu từ các trang web có lưu lượng truy cập cao và khối lượng dữ liệu lớn.

**2.7. Tổng quan về website TITV**

TITV là một nền tảng học trực tuyến cung cấp các khóa học về lập trình và công nghệ thông tin, đáp ứng nhu cầu học tập và phát triển kỹ năng của người dùng tại Việt Nam. TITV cung cấp nhiều khóa học với các cấp độ từ cơ bản đến nâng cao, với nội dung đa dạng bao gồm từ lập trình cơ bản đến các lĩnh vực chuyên sâu như trí tuệ nhân tạo và khoa học dữ liệu. Mỗi khóa học được thiết kế nhằm mang lại trải nghiệm học tập hiệu quả và toàn diện, phù hợp với người mới bắt đầu lẫn các chuyên gia trong ngành.2.8.. Lịch sử hình thành và phát triển

Ban đầu, TITV tập trung vào các khóa học lập trình cơ bản, nhằm tạo ra nền tảng kiến thức vững chắc cho người học. Với sự phát triển của công nghệ và nhu cầu ngày càng tăng của người học, TITV mở rộng nội dung, trở thành một hệ sinh thái học tập toàn diện với hàng loạt khóa học về công nghệ. TITV hiện là một trong những nền tảng giáo dục trực tuyến nổi bật ở Việt Nam, phục vụ cho đông đảo người dùng từ các bạn sinh viên đến các chuyên gia trong lĩnh vực công nghệ.

**2.9. Mô hình kinh doanh và dịch vụ**

TITV hoạt động theo mô hình freemium, cho phép người dùng tiếp cận một số tài liệu học tập miễn phí và trả phí cho các khóa học chuyên sâu hơn. Mô hình này mang đến cho người học nhiều lựa chọn, từ việc học miễn phí kiến thức nền tảng đến việc đầu tư vào các khóa học cao cấp. Bên cạnh đó, TITV còn cung cấp dịch vụ tư vấn nghề nghiệp và hỗ trợ học tập qua các dự án thực tế, giúp người học áp dụng kiến thức vào thực tiễn. **2.10. Thị trường và đối thủ cạnh tranh**

Trong thị trường giáo dục trực tuyến cạnh tranh khốc liệt tại Việt Nam, TITV đối mặt với sự cạnh tranh từ nhiều đối thủ lớn như Udemy, Coursera, và các nền tảng trong nước như Kyna, Edumall. Tuy nhiên, TITV có ưu thế với nội dung được cá nhân hóa và đáp ứng sát nhu cầu học tập của người Việt. Nhờ vào khả năng linh hoạt và am hiểu nhu cầu người học, TITV đã khẳng định vị thế của mình trên thị trường giáo dục trực tuyến, hướng đến mở rộng và phát triển thêm nhiều nội dung học tập phong phú hơn trong tương lai.

**Chương 3: Phương Pháp Thu Thập Dữ Liệu**

**3.1. Xác định mục tiêu thu thập dữ liệu**

**3.2. Thiết kế quy trình thu thập dữ liệu  
3.3. Sử dụng Selenium trong thu thập dữ liệu từ TITV***• 3.3.1. Cài đặt và cấu hình Selenium  
• 3.3.2. Viết mã tự động hóa  
• 3.3.3. Xử lý và lưu trữ dữ liệu bằng MongoDB*

**Chương 4: Kết Quả Thực Nghiệm**

**4.1. Phân tích dữ liệu thu thập được**

**4.2. Đánh giá chất lượng dữ liệu  
4.3. So sánh với dữ liệu nguồn khác**

**4.4. Phân tích dữ liệu bằng MongoDB**

**4.5 Truy vấn dữ liệu từ dữ liệu MongoDB**

**Chương 5: Kết Luận và Kiến Nghị**

**5.1 Thảo luận***• 5.1.1 Những thách thức trong quá trình thu thập dữ liệu*

*• 5.1.2 Đề xuất cải tiến quy trình thu thập dữ liệu*

**5.2. Kết luận***• 5.2.1. Tóm tắt kết quả nghiên cứu (tích hợp video cào dữ liệu)*

*• 5.2.2. Định hướng nghiên cứu tiếp theo*

**Tài Liệu Tham Khảo**

**Phụ Lục**

**Code cào dữ liệu TIVI**

**Code Truy Vấn**

**Hình ảnh lịch sử commit trên GitHub**