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 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

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 nhữngngười 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

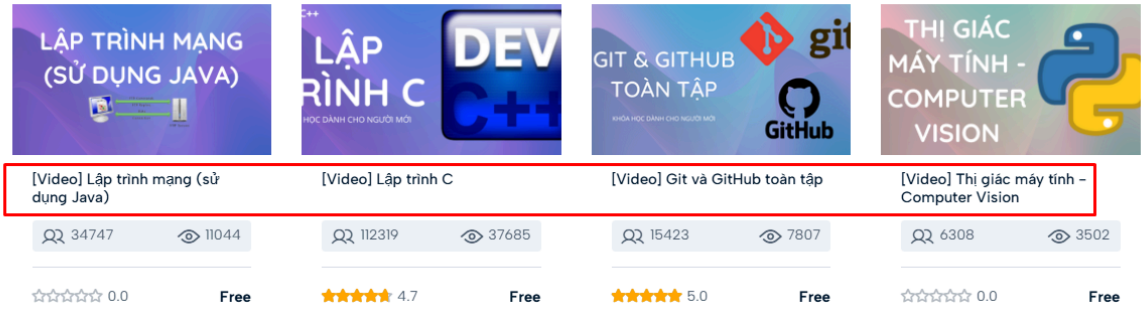
3.1. Xác định mục tiêu thu thập dữ liệu

Việc thu thập dữ liệu từ trang web dạy học TITV là một hoạt động quan trọng nhằm hiểu rõ hơn về các khóa học mà nền tảng này cung cấp, đồng thời phân tích chất lượng và mức độ tương tác của người dùng với các chương trình học. Mục tiêu của quá trình thu thập dữ liệu này sẽ được xác định cụ thể như sau:

1. Thu thập tên môn học

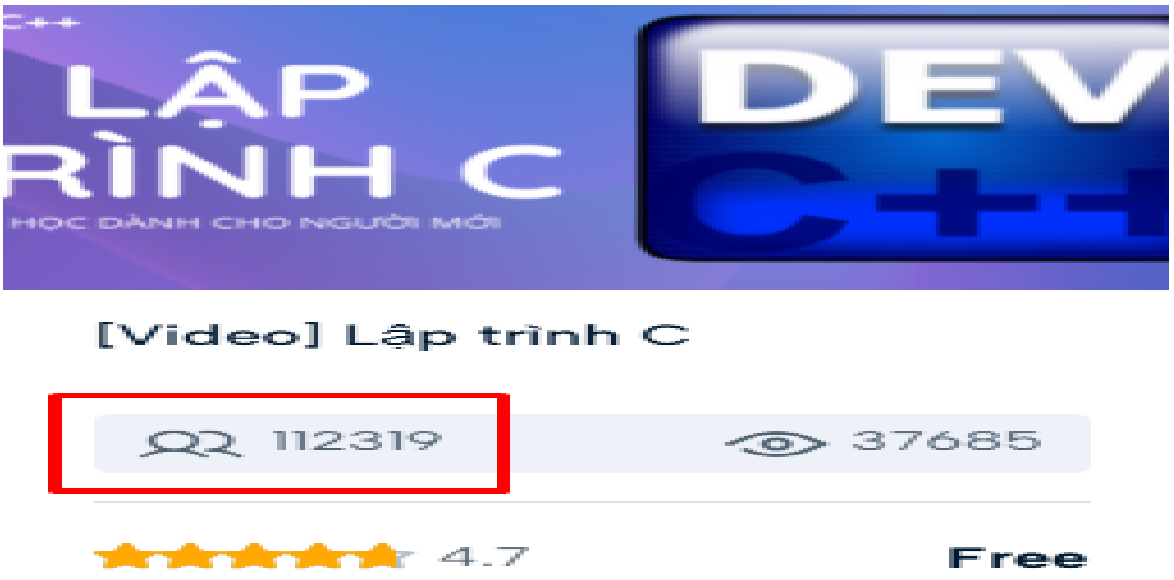
Một trong những mục tiêu đầu tiên là thu thập danh sách các khóa học hiện có trên TITV. Điều này bao gồm việc lấy thông tin về tên môn học, mô tả ngắn gọn về nội dung của từng khóa học, và các thông tin liên quan như thời gian học và mức độ khó. Thông tin này sẽ giúp người học có cái nhìn tổng quan về các lựa chọn học tập sẵn có và tạo cơ sở cho việc phân tích nội dung giảng dạy.

Hình minh họa tên một số môn học



2. Số người tham gia

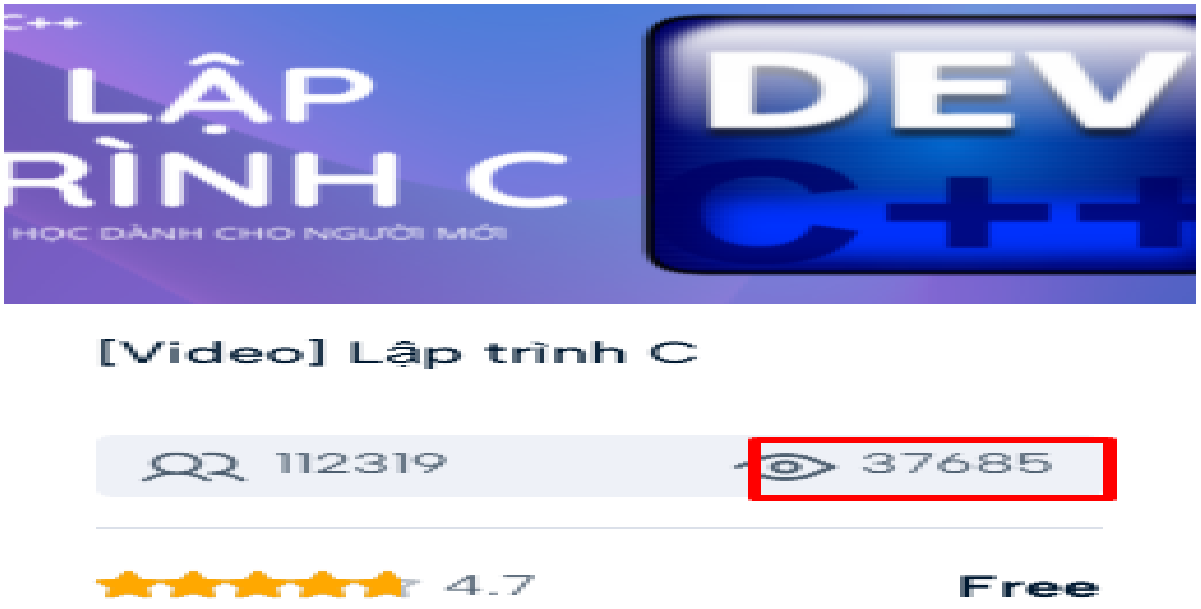
Một yếu tố quan trọng khác là thu thập dữ liệu về số lượng người tham gia vào mỗi khóa học. Thông tin này không chỉ phản ánh mức độ phổ biến của khóa học mà còn giúp phân tích xu hướng học tập trong cộng đồng người học. Số lượng người tham gia cũng có thể được xem như một chỉ số để đánh giá mức độ thành công của khóa học trên nền tảng.Hình ảnh minh họa cho số người tham gia :



3. Số người xem

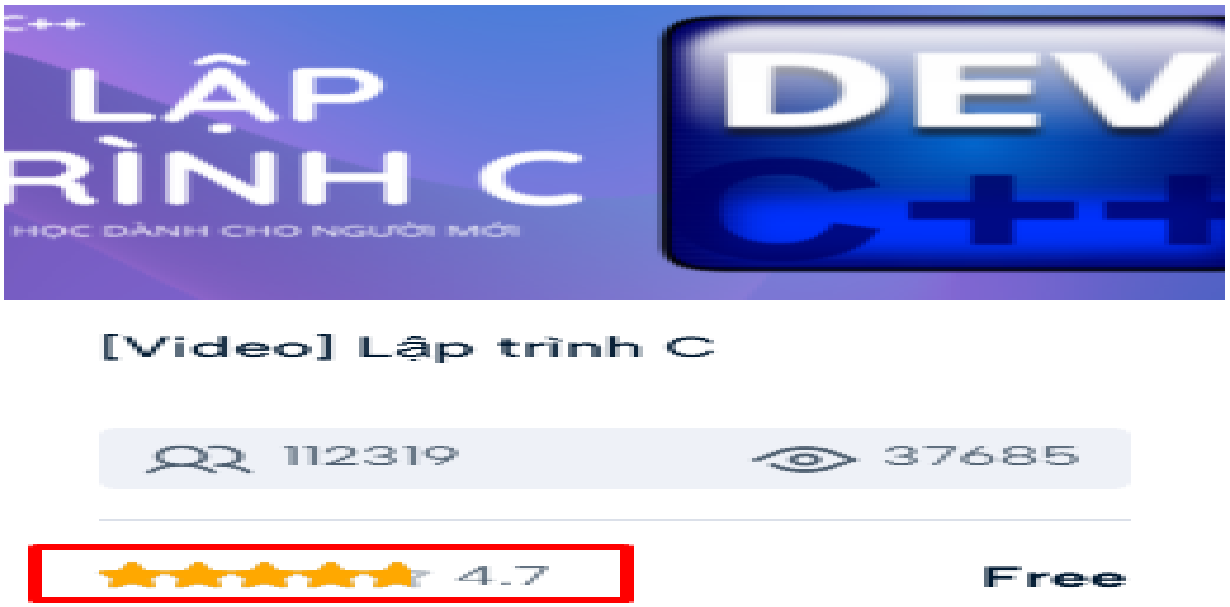
Việc theo dõi số người xem các video giảng dạy cũng là một mục tiêu chính trong quá trình thu thập dữ liệu. Số liệu này sẽ cung cấp thông tin về mức độ tiếp cận và sự quan tâm của người dùng đối với các khóa học. Nếu một khóa học có số người xem cao nhưng số người tham gia lại thấp, điều này có thể chỉ ra rằng mặc dù nội dung hấp dẫn nhưng có thể có rào cản trong việc tham gia học tập, như thời gian hay điều kiện học tập.

Hình ảnh minh họa cho người xem :



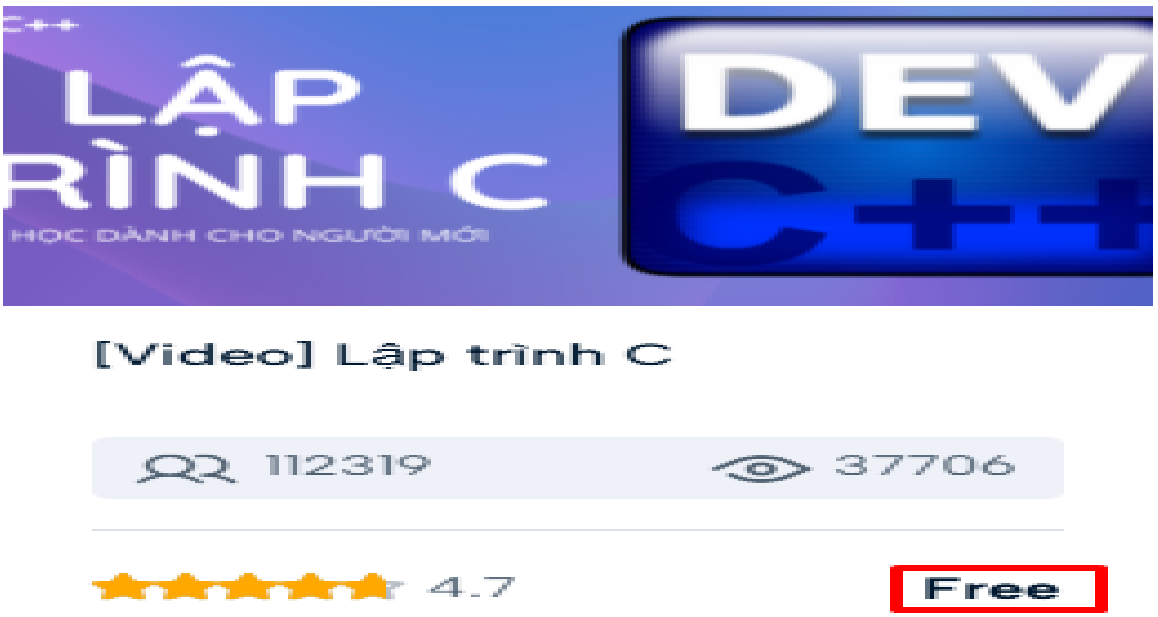
4. Thu thập thông tin về rating

Một mục tiêu quan trọng khác là thu thập thông tin về rating (điểm đánh giá) của các khóa học. Rating thường được thể hiện dưới dạng số sao hoặc điểm số từ người học. Thông tin này không chỉ cho biết chất lượng khóa học theo cảm nhận của người học mà còn có thể là yếu tố quyết định đối với những người dùng mới khi họ lựa chọn khóa học. Bằng cách phân tích rating, chúng ta có thể hiểu rõ hơn về cảm nhận của người học và tác động của nó đến quyết định tham gia học.Hình minh họa cho rating



5. Giá cả

Việc theo dõi giá cả của các khóa học là một yếu tố quan trọng trong quá trình thu thập dữ liệu. Số liệu này sẽ cung cấp thông tin về khả năng tiếp cận tài chính của người dùng đối với các khóa học lập trình. Nếu một khóa học có giá cao nhưng số lượng người đăng ký lại thấp, điều này có thể chỉ ra rằng mặc dù nội dung có chất lượng cao, nhưng chi phí có thể là một rào cản đối với người học. Ngược lại, nếu một khóa học có giá hợp lý và số lượng người tham gia đông, điều này cho thấy rằng mức giá đó được coi là phù hợp và hấp dẫn đối với đa số người học. Việc phân tích giá cả và xu hướng đăng ký sẽ giúp cải thiện chiến lược giá của trang web, nhằm tối ưu hóa số lượng học viên tham gia.Hình ảnh minh họa :



3.2. Thiết kế quy trình thu thập dữ liệu

Thiết kế quy trình thu thập dữ liệu là một yếu tố then chốt trong nghiên cứu, ảnh hưởng trực tiếp đến độ chính xác và độ tin cậy của kết quả phân tích. Đối với dự án nghiên cứu các khóa học trên trang web TITV, quy trình thu thập dữ liệu sẽ được thiết kế một cách chi tiết và có hệ thống nhằm đảm bảo dữ liệu thu thập được không chỉ đầy đủ mà còn có chất lượng cao. Dưới đây là các bước cụ thể trong quy trình này:

3.2.1. Xác định nguồn dữ liệu

Nguồn dữ liệu chính trong dự án này là trang web TITV, nơi cung cấp nhiều thông tin về các khóa học trực tuyến. Việc xác định rõ nguồn dữ liệu là bước đầu tiên và rất quan trọng, bởi vì nó sẽ quyết định phương pháp thu thập và các công cụ cần thiết. Các thông tin cần thu thập từ TITV bao gồm:

- Tên môn học: Tên gọi chính thức của các khóa học.
- Rating: Thông tin về điểm đánh giá mà người học dành cho mỗi khóa học, thường được thể hiện dưới dạng số sao hoặc điểm số.
- Ratting : Phản hồi từ người học, có thể bao gồm nhận xét, cảm nhận và kinh nghiệm của họ khi tham gia khóa học.
- Số người tham gia: Số lượng học viên đăng ký tham gia khóa học, điều này có thể phản ánh mức độ phổ biến của khóa học.
- Số người xem: Mức độ tiếp cận của khóa học, thể hiện qua số lượt xem video giảng dạy.

3.2.2. Lập kế hoạch thu thập dữ liệu

Sau khi xác định nguồn dữ liệu, bước tiếp theo là lập kế hoạch thu thập dữ liệu một cách chi tiết. Kế hoạch này sẽ bao gồm:

- Chọn công cụ thu thập dữ liệu: Quyết định sử dụng Selenium để tự động hóa quy trình thu thập. Selenium là một công cụ mạnh mẽ cho phép lập trình viên tương tác với trình duyệt web và thu thập dữ liệu từ các trang web động. Nó hỗ trợ nhiều ngôn ngữ lập trình như Python,và có thể được tích hợp vào các dự án phức tạp khác.
- Xác định các yếu tố cần thu thập: Liệt kê rõ ràng các trường dữ liệu cần thu thập, định dạng của chúng và cách thức truy cập vào từng phần thông tin trên trang web. Điều này sẽ giúp cho việc viết mã trở nên dễ dàng và hiệu quả hơn.
- Thiết lập lịch trình: Đưa ra thời gian cụ thể cho từng giai đoạn trong quy trình thu thập dữ liệu, từ việc chuẩn bị công cụ đến chạy các script thu thập và xử lý dữ liệu. Kế hoạch này sẽ giúp theo dõi tiến độ và đảm bảo rằng mọi bước đều được thực hiện đúng hạn.

3.2.3. Thực hiện thu thập dữ liệu

Quá trình thu thập dữ liệu sẽ được thực hiện qua các bước cụ thể như sau:

- Cài đặt và cấu hình Selenium: Bắt đầu bằng việc cài đặt Selenium trên máy tính. Điều này bao gồm việc tải về các thư viện cần thiết và thiết lập môi trường phát triển, như Python hoặc Java, tùy thuộc vào ngôn ngữ lập trình. Cần đảm bảo rằng các trình điều khiển (driver) cho trình duyệt được sử dụng (như ChromeDriver cho Google Chrome) hoặc fire fox cũng được cài đặt đúng cách.

Viết mã tự động hóa: Lập trình các script bằng ngôn ngữ đã chọn để tự động hóa quá trình thu thập dữ liệu. Mã sẽ bao gồm các lệnh để mở trang web TITV, điều hướng đến phần khóa học, và thu thập các thông tin cần thiết. Dưới đây là một ví dụ mô tả các bước trong mã:

```
from selenium import webdriver

from selenium.webdriver.firefox.service import Service

from selenium.webdriver.common.by import By

from selenium.webdriver.support.ui import WebDriverWait

from selenium.webdriver.support import expected_conditions as EC

import time

from pymongo import MongoClient


# Đường dẫn tới geckodriver

gecko_path =
r"C:\Users\HP\Downloads\geckodriver-v0.35.0-win64\geckodriver.exe"

service = Service(gecko_path)


# Kết nối tới MongoDB

mongo_client = MongoClient("mongodb://localhost:27017/")

db = mongo_client["course_database"]

courses_collection = db["courses"]
```

```
# Hàm thu thập thông tin khóa học
def collect_course_info():
    driver = webdriver.Firefox(service=service)
    driver.get("https://titv.vn/")
    time.sleep(10)

    all_courses = {}

    # Vòng lặp để thu thập các khóa học
    for data_id in range(75, 82):
        li_element = WebDriverWait(driver, 20).until(
            EC.element_to_be_clickable((By.CSS_SELECTOR,
f'span[data-id="{data_id}"]'))
        )
        li_name = li_element.text
        li_element.click()
        time.sleep(10)

        course_elements = WebDriverWait(driver, 20).until(
            EC.presence_of_all_elements_located((By.CSS_SELECTOR,
'div.ms_lms_courses_card_item'))
        )

        # Lặp qua từng khóa học
        for course in course_elements:
            try:
                course_name = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
```

```
                        'div.ms_lms_courses_card_item_wrapper >
div:nth-child(2) > a:nth-child(1) > h3'))
                ).text


                price_element = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_info_price >
div.ms_lms_courses_card_item_info_price_single > span'))
                ).text


                rating_element = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_info_rating_quantity >
span'))
                ).text


                members_count = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_meta_block >
i.stmlms-members + span'))
                ).text


                views_count = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_meta_block >
i.stmlms-views + span'))
                ).text


                # Chuyển đổi các trường sang dạng số
```

```
                rating_numeric = float(rating_element) if rating_element
else 0.0

                price_numeric = float(price_element.replace(',',
'').replace('đ', '').strip()) if price_element and price_element.lower() !=
"free" else 0.0

                members_count_numeric = int(members_count.replace(',',
'').strip()) if members_count else 0

                views_count_numeric = int(views_count.replace(',',
'').strip()) if views_count else 0


                course_info = {

                    "name": course_name,

                    "price": price_numeric,

                    "rating": rating_numeric,

                    "members_count": members_count_numeric,

                    "views_count": views_count_numeric,

                    "category": li_name,

                }

                all_courses[course_name] = course_info


            except Exception as e:

                print(f"Error collecting course info: {e}")

                continue


        driver.back()

        time.sleep(10)


    driver.quit()


    # Lưu thông tin khóa học vào MongoDB
```

```
    for course_name, course_info in all_courses.items():
        try:
            courses_collection.update_one(
                {"name": course_name},
                {"$set": course_info},
                upsert=True
            )
        except Exception as e:
            print(f"Error saving course {course_name}: {e}")

    return all_courses

# Chạy đoạn 1 để thu thập thông tin khóa học
collect_course_info()
```

- Chạy script: Sau khi hoàn tất viết mã, tiến hành chạy script để thu thập dữ liệu. Trong quá trình chạy, cần theo dõi các lỗi có thể xảy ra và đảm bảo rằng dữ liệu được thu thập đầy đủ và chính xác. Việc ghi log cũng rất quan trọng để theo dõi các vấn đề phát sinh trong quá trình thu thập dữ liệu.

4. Xử lý và lưu trữ dữ liệu

Sau khi thu thập dữ liệu, bước tiếp theo là xử lý và lưu trữ dữ liệu một cách hiệu quả:

- Xử lý dữ liệu: Dữ liệu thu thập được cần được làm sạch để loại bỏ các thông tin không cần thiết, trùng lặp hoặc sai sót. Việc xử lý này sẽ bao gồm việc chuẩn hóa định dạng dữ liệu, chẳng hạn như chuyển đổi các đánh giá thành

định dạng số, xử lý các ký tự đặc biệt, và xác định các giá trị null. Sử dụng các thư viện như Pandas trong Python có thể giúp đơn giản hóa quá trình này.

Lưu trữ dữ liệu: Sử dụng MongoDB để lưu trữ dữ liệu đã xử lý. MongoDB là một cơ sở dữ liệu NoSQL phù hợp cho việc lưu trữ dữ liệu phi cấu trúc. Cấu trúc cơ sở dữ liệu cần được thiết kế sao cho dễ dàng truy vấn và mở rộng trong tương lai. Dưới đây là một ví dụ về cách lưu trữ dữ liệu vào MongoDB:

```
# Kết nối tới MongoDB
mongo_client = MongoClient("mongodb://localhost:27017/")
db = mongo_client["course_database"]
courses_collection = db["courses"]
course_info = {
    "name": course_name,
    "price": price_numeric,
    "rating": rating_numeric,
    "members_count": members_count_numeric,
    "views_count": views_count_numeric,
    "category": li_name,
    "reviews": []
}
all_courses[course_name] = course_info
```

5. Kiểm tra và đánh giá

Cuối cùng, sau khi hoàn tất quy trình thu thập và lưu trữ, cần tiến hành kiểm tra và đánh giá dữ liệu để đảm bảo chất lượng:

- Đánh giá chất lượng dữ liệu: Kiểm tra tính chính xác và độ tin cậy của dữ liệu thu thập được. Việc này bao gồm việc so sánh dữ liệu thu thập với các nguồn dữ liệu khác (nếu có) để xác định tính chính xác và mức độ đầy đủ. Sử dụng

các phương pháp thống kê để phân tích sự phân bố của dữ liệu có thể giúp phát hiện ra các bất thường.

Chương 4: Kết Quả Thực Nghiệm

4.1. Phân tích dữ liệu thu thập được

Phân tích dữ liệu thu thập từ nền tảng TITV giúp đưa ra cái nhìn toàn diện và chi tiết về các khóa học, xu hướng học tập, và mức độ hài lòng của người học. Quá trình phân tích này sẽ được thực hiện dựa trên các biến quan trọng như tên khóa học, số người tham gia, số lượt xem, và đánh giá của người học (rating).

1. Phân tích tên khóa học và nội dung

Dữ liệu về tên khóa học và mô tả nội dung cung cấp cơ sở để phân loại các khóa học theo chủ đề, cấp độ, và thời gian học tập. Việc này giúp xác định các lĩnh vực học tập phổ biến, các lĩnh vực còn thiếu, và các khóa học thu hút nhiều người học. Thông tin từ đây có thể tạo điều kiện cho việc phân tích sâu hơn về xu hướng sở thích của người học, từ đó đưa ra các đề xuất nhằm cải thiện nội dung khóa học theo nhu cầu thực tế.

2. Phân tích số người tham gia

Số lượng người tham gia là một chỉ số quan trọng để đánh giá mức độ phổ biến của khóa học. Bằng cách phân tích dữ liệu này, chúng ta có thể phát hiện ra những khóa học được ưa chuộng, từ đó rút ra các yếu tố thành công của các khóa học này, chẳng hạn như nội dung, giảng viên, hoặc phương pháp giảng dạy. Bên cạnh đó, số liệu này còn giúp đánh giá tổng thể mức độ quan tâm của người học đến các khóa học trên nền tảng TITV, cung cấp thông tin cho việc phát triển và cải tiến sản phẩm học tập trong tương lai.

3. Phân tích số lượt xem

Số lượt xem các video giảng dạy phản ánh mức độ tiếp cận và sự quan tâm của người dùng đối với từng khóa học. Nếu một khóa học có số lượt xem cao nhưng số

người tham gia thấp, điều này có thể gợi ý về các rào cản trong việc tham gia, chẳng hạn như thời gian học kéo dài hoặc yêu cầu trình độ cao. Phân tích xu hướng xem còn có thể giúp xác định thời điểm cao điểm trong tuần hoặc trong ngày khi người học có xu hướng tham gia nhiều nhất, từ đó giúp tối ưu hóa thời gian tổ chức và phân phối nội dung.

4. Phân tích rating và phản hồi của người học

Dữ liệu rating, thường biểu thị qua số sao hoặc điểm số, cung cấp cái nhìn trực tiếp về mức độ hài lòng của người học với mỗi khóa học. Phân tích rating cho phép đánh giá chất lượng khóa học một cách khách quan và có thể định lượng được. Nếu có nhiều khóa học có rating cao, đó là tín hiệu tích cực về chất lượng của nền tảng. Ngược lại, nếu có nhiều khóa học có rating thấp, cần xem xét các yếu tố cần cải thiện. Kết hợp với các phản hồi chi tiết từ người học, việc phân tích này giúp xác định các điểm mạnh, điểm yếu của khóa học, từ đó nâng cao trải nghiệm học tập.

5. Phân tích tổng hợp

Tổng hợp các phân tích trên cho phép tạo ra bức tranh toàn cảnh về nền tảng TITV, giúp nhận diện các xu hướng học tập của người dùng, phân tích độ thành công của từng khóa học, và xác định các yếu tố quan trọng đóng góp vào sự hài lòng của người học. Từ đó, có thể đưa ra các đề xuất cải tiến, xây dựng các chiến lược phát triển nội dung mới, và tối ưu hóa trải nghiệm người dùng trên nền tảng.

4.2. Đánh giá chất lượng dữ liệu

Đánh giá chất lượng dữ liệu là một bước quan trọng nhằm đảm bảo dữ liệu thu thập từ nền tảng TITV có đủ độ tin cậy và tính chính xác để phục vụ cho quá trình phân tích. Để đạt được điều này, dữ liệu cần phải qua kiểm tra ở một số tiêu chí quan trọng như tính đầy đủ, tính chính xác, tính nhất quán, và khả năng tái sử dụng.

1. Tính đầy đủ

Tính đầy đủ của dữ liệu đề cập đến việc tất cả các yếu tố cần thiết (như tên môn học, số lượng người tham gia, số người xem, rating, và các mô tả chi tiết về khóa học) đã được thu thập một cách toàn diện và không có thiếu sót. Trong quá trình thu thập, cần đối chiếu dữ liệu với trang nguồn để đảm bảo các trường dữ liệu không bị bỏ sót hoặc bị lỗi do quá trình lấy dữ liệu. Ví dụ, một số khóa học có thể thiếu thông tin về đánh giá hoặc số lượng người tham gia; vì vậy, dữ liệu cần được kiểm tra để đảm bảo rằng các trường hợp này đã được xử lý và không ảnh hưởng đến phân tích.

2. Tính chính xác

Tính chính xác của dữ liệu được đảm bảo qua việc kiểm tra xem liệu các số liệu đã được ghi nhận một cách đúng đắn. Chẳng hạn, trong khi chuyển đổi dữ liệu từ dạng text về dạng số, như số lượng người tham gia hoặc rating, các lỗi định dạng có thể xảy ra. Điều này đòi hỏi phải kiểm tra và đối chiếu kết quả giữa các lần thu thập và so sánh với số liệu hiển thị trên nền tảng. Việc sử dụng các phương pháp thống kê, như phân tích phân bố của các số liệu, cũng giúp phát hiện ra các bất thường hoặc dữ liệu nhiễu.

3. Tính nhất quán

Tính nhất quán là một yếu tố then chốt để đảm bảo rằng tất cả dữ liệu thu thập đều tuân theo cùng một tiêu chuẩn và định dạng. Nếu có sự sai lệch trong định dạng của các trường dữ liệu, chẳng hạn như sự không nhất quán trong cách ghi số liệu người tham gia hoặc số lượt xem, quá trình phân tích có thể bị ảnh hưởng. Các công cụ như Python (Pandas) có thể được sử dụng để chuẩn hóa và đồng bộ hóa dữ liệu, đảm bảo rằng tất cả các giá trị đều ở cùng định dạng và dễ dàng so sánh.

4. Độ tin cậy

Độ tin cậy của dữ liệu cần được đảm bảo qua việc xác thực nguồn dữ liệu từ nền tảng TITV. Việc thu thập và lưu trữ thông tin trong MongoDB cũng cần được thiết kế sao cho việc truy vấn không gây ra các lỗi sai sót, đồng thời mã hóa và bảo mật dữ liệu để đảm bảo tính riêng tư của các thông tin nhạy cảm. Đánh giá độ tin cậy này

giúp đảm bảo rằng dữ liệu sau khi lưu trữ có thể được sử dụng trong các giai đoạn phân tích tiếp theo mà không bị sai lệch.

5. Khả năng tái sử dụng

Khả năng tái sử dụng là yếu tố giúp dữ liệu đã thu thập có thể được sử dụng lâu dài và phục vụ cho các nghiên cứu, phân tích trong tương lai. Việc lưu trữ dữ liệu theo định dạng JSON trong MongoDB không chỉ tối ưu hóa cho việc truy vấn, mà còn giúp mở rộng cơ sở dữ liệu dễ dàng khi có dữ liệu mới. Đồng thời, sử dụng các tiêu chuẩn trong lưu trữ và mã hóa dữ liệu sẽ giúp các dữ liệu này có thể dễ dàng tích hợp vào các mô hình phân tích khác, cho phép thực hiện phân tích và đánh giá chất lượng khóa học liên tục và có hệ thống.

4.3. So sánh với dữ liệu nguồn khác

Việc so sánh dữ liệu thu thập từ trang TITV với các nguồn dữ liệu khác là bước quan trọng nhằm đánh giá độ chính xác và độ nhất quán của dữ liệu. Đối chiếu dữ liệu với các nguồn tương tự hoặc từ các nền tảng học tập khác có thể giúp kiểm chứng thông tin và đảm bảo dữ liệu phản ánh chính xác xu hướng và hành vi người học. Các bước cụ thể trong việc so sánh với dữ liệu nguồn khác bao gồm:

1. Lựa chọn các nguồn dữ liệu so sánh: Các nguồn dữ liệu có thể bao gồm những nền tảng học trực tuyến phổ biến khác hoặc các báo cáo về xu hướng giáo dục trực tuyến. Những nền tảng lớn như Coursera, Udemy hoặc edX thường cung cấp số liệu về số lượng người học, đánh giá và lượt xem, giúp đối chiếu các chỉ số như mức độ phổ biến, chất lượng khóa học và xu hướng người học trong ngành.
2. So sánh số lượng người tham gia và người xem: Đối chiếu số lượng người tham gia và lượt xem các khóa học trên TITV với các nền tảng học tập tương tự nhằm đánh giá mức độ phổ biến của nền tảng và sự tương tác của người học. Nếu dữ liệu từ TITV cho thấy một số khóa học có lượt xem cao nhưng ít người đăng ký so với các nền tảng khác, điều này có thể phản ánh các yếu tố

như khả năng tiếp cận hoặc các yếu tố khác ảnh hưởng đến quyết định tham gia học tập.

3. Đánh giá rating và phản hồi người dùng: So sánh rating và các phản hồi người dùng trên TITV với các nền tảng khác để xác định mức độ hài lòng và chất lượng của các khóa học. Rating từ các nền tảng lớn có thể cung cấp cái nhìn tổng quan về những yếu tố người học ưu tiên, từ đó đánh giá xem các khóa học trên TITV có đáp ứng các yêu cầu của người học không, và những yếu tố nào cần cải thiện để tăng cường sự hài lòng của người dùng.
4. So sánh về nội dung khóa học: Nội dung các khóa học trên TITV có thể được đối chiếu với các nền tảng khác nhằm đánh giá tính cạnh tranh và sự đa dạng trong các chủ đề mà TITV cung cấp. Đối chiếu này giúp xác định xem liệu TITV có đáp ứng đúng các xu hướng học tập hiện tại hay cần điều chỉnh hoặc bổ sung thêm các chủ đề mới để nâng cao trải nghiệm học tập.
5. Phân tích các yếu tố đặc thù của TITV: So sánh dữ liệu từ TITV với các nền tảng khác giúp xác định những đặc điểm nổi bật mà TITV có thể cải thiện hoặc tận dụng để phát triển. Những yếu tố này có thể bao gồm các chương trình học có tính cá nhân hóa, thời lượng khóa học phù hợp, và các hỗ trợ học tập khác.

Thông qua các so sánh này, dữ liệu từ TITV có thể được xác thực và cải thiện về độ tin cậy. Điều này giúp đưa ra các khuyến nghị để nâng cao chất lượng nội dung và chiến lược phát triển nền tảng, từ đó đáp ứng tốt hơn nhu cầu người học và tạo ra lợi thế cạnh tranh.

4.4.Truy vấn dữ liệu từ dữ liệu MongoDB

1. Kết nối đến MongoDB

-Đầu tiên, chúng ta cần kết nối đến cơ sở dữ liệu MongoDB. Sử dụng thư viện Mongoose, đoạn mã dưới đây thực hiện việc kết nối:

```
from pymongo import MongoClient

# Kết nối tới MongoDB
mongo_client = MongoClient("mongodb://localhost:27017/")
db = mongo_client["course_database"]
courses_collection = db["courses"]
```

-Cấu trúc Dữ liệu

MongoDB lưu trữ dữ liệu dưới dạng tài liệu JSON, trong đó mỗi tài liệu có thể chứa các cặp key-value khác nhau. Cấu trúc này giúp người dùng dễ dàng quản lý và truy xuất thông tin theo nhu cầu.

Ví dụ về cấu trúc tài liệu:

```
course_info = {
    "name": course_name,
    "price": price_numeric,
    "rating": rating_numeric,
    "members_count": members_count_numeric,
    "views_count": views_count_numeric,
    "category": li_name,
}
```

Lưu Thông tin vào Collection

-Sau khi xác định dữ liệu, chúng ta có thể sử dụng phương thức insertOne() để lưu một tài liệu hoặc insertMany() để lưu nhiều tài liệu cùng lúc

```
# Lưu thông tin khóa học vào MongoDB
for course_name, course_info in all_courses.items():
    try:
        courses_collection.update_one(
            {"name": course_name},
            {"$set": course_info},
            upsert=True
        )
    except Exception as e:
        print(f"Error saving course {course_name}: {e}")

return all_courses
```

-Kết quả Tạo ra course_database và collection chứ thông tin thu thập được

Hình minh họa :

courses

	_id ObjectId	name String	category String	members_count Int32	price Double
1	ObjectId('6720da75c3a017…	"[Video] Lập trình C"	"Kiến thức nền tảng"	112320	0
2	ObjectId('6720da75c3a017…	"[Video] Lập trình Java …	"Kiến thức nền tảng"	271431	0
3	ObjectId('6720da75c3a017…	"[Video] Lập trình Pytho…	"Data Science"	26604	0
4	ObjectId('6720da75c3a017…	"Toán rời rạc"	"Kiến thức nền tảng"	23736	0
5	ObjectId('6720da75c3a017…	"[Video] Cấu trúc dữ liệ…	"Kiến thức nền tảng"	17908	0
6	ObjectId('6720da75c3a017…	"Quản trị Hệ điều hành L…	"Mới học lập trình"	16373	0
7	ObjectId('6720da75c3a017…	"Cấu trúc dữ liệu và giả…	"Kiến thức nền tảng"	6161	0
8	ObjectId('6720da75c3a017…	"[Video] JDBC - Lập trìn…	"Java Fullstack"	10477	0
9	ObjectId('6720da75c3a017…	"[Video] Tự học SQL với …	"Kiến thức nền tảng"	19676	0
10	ObjectId('6720da75c3a017…	"[Video] SQL Server - Cơ…	"Kiến thức nền tảng"	38798	0
11	ObjectId('6720da75c3a017…	"[Video] Lập trình viên …	"Java Fullstack"	1	399000
12	ObjectId('6720da75c3a017…	"[Video] Lập trình viên …	"Java Backend"	0	339000

members_count Int32	price Double	rating Double	views_count Int32
112320	0	4.7	37707
271431	0	5	51378
26604	0	5	11686
23736	0	0	8130
17908	0	0	12066
16373	0	0	4012
6161	0	0	3196
10477	0	0	10719
19676	0	5	9730
38798	0	0	13566
1	399000	0	13469
0	339000	0	13102

Chương 5: Kết Luận và Kiến Nghị

5.1 Thảo luận

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

Quá trình thu thập dữ liệu từ trang web TITV đã mang lại những trải nghiệm phong phú về cả mặt kỹ thuật và quản lý dữ liệu, nhưng đồng thời cũng đặt ra nhiều thách thức cần giải quyết để đảm bảo tính chính xác, đầy đủ và hiệu quả. Dù TITV là một trang web tĩnh không có cơ chế chống bot, việc thu thập dữ liệu bằng Selenium và MongoDB gặp một số khó khăn nổi bật sau đây:

Thứ nhất, vấn đề định dạng và làm sạch dữ liệu: TITV không được thiết kế với cấu trúc dữ liệu chuẩn cho việc truy xuất tự động, dẫn đến sự thiếu đồng nhất giữa các trường dữ liệu. Ví dụ, số lượt xem, số lượng học viên tham gia và điểm đánh giá không được chuẩn hóa và có thể chứa các ký tự không cần thiết, làm tăng khả năng sai lệch khi phân tích. Các khoảng trống trong dữ liệu hoặc các trường có thông tin không đầy đủ đòi hỏi phải có quy trình xử lý và làm sạch kỹ lưỡng, nhằm loại bỏ những phần dữ liệu không chính xác mà không gây ảnh hưởng đến tổng thể.

Thứ hai, sự phụ thuộc vào công cụ Selenium: Mặc dù Selenium là công cụ mạnh mẽ cho việc tự động hóa trình duyệt, nó yêu cầu thiết lập trình điều khiển (driver) và tối ưu mã lệnh để đảm bảo rằng quá trình thu thập diễn ra ổn định. Mỗi thao tác thu thập cần sự điều hướng cẩn trọng, đặc biệt là với các trang có nội dung tải động chậm. Điều này đôi khi làm chậm lại toàn bộ quy trình và ảnh hưởng đến hiệu quả tổng thể, đòi hỏi kỹ thuật lập trình tốt và các giải pháp tối ưu như bổ sung các lệnh chờ có điều kiện, điều này cũng yêu cầu phải có khả năng xử lý lỗi tốt nếu có vấn đề trong quá trình kết nối hoặc điều hướng.

Thứ ba, quản lý dữ liệu thu thập với khối lượng lớn: Khi dữ liệu được lưu trữ vào MongoDB, việc đảm bảo tính nhất quán và hiệu quả truy xuất thông tin là một bài toán không nhỏ. Nếu không có chiến lược lưu trữ hợp lý, dữ liệu có thể bị phân mảnh, gây khó khăn khi thực hiện các truy vấn phức tạp sau này. Việc chuẩn hóa

cấu trúc lưu trữ và kiểm tra định kỳ giúp hạn chế các vấn đề này nhưng đồng thời cũng đòi hỏi công sức và thời gian để giám sát.

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

Để cải thiện quy trình thu thập dữ liệu trên TITV và tăng cường hiệu quả cũng như chất lượng của dữ liệu thu thập được, một số đề xuất đã được đưa ra dựa trên các thách thức đã gặp phải. Những cải tiến này không chỉ giúp tiết kiệm thời gian mà còn nâng cao tính chính xác và khả năng tái sử dụng dữ liệu trong tương lai:

1. Tối ưu hóa công cụ thu thập bằng BeautifulSoup và Requests cho các trang tĩnh: Đối với những phần trang không có nội dung động, sử dụng BeautifulSoup kết hợp với Requests sẽ là lựa chọn tối ưu hơn so với Selenium, vì nó không cần tải toàn bộ trình duyệt và giúp giảm đáng kể thời gian xử lý. Công cụ này cũng cung cấp các phương pháp hiệu quả để xử lý HTML và trích xuất dữ liệu với cấu trúc phức tạp, giúp cải thiện đáng kể độ chính xác và tốc độ thu thập.
2. Xây dựng hàm tự động làm sạch và chuẩn hóa dữ liệu: Các hàm này sẽ thực hiện các thao tác xử lý định dạng tự động, chẳng hạn như chuyển đổi đánh giá từ chuỗi thành số thực, loại bỏ ký tự đặc biệt trong số lượng học viên và lượt xem, hoặc xử lý các giá trị trống. Sử dụng các thư viện mạnh mẽ như Pandas và NumPy trong Python sẽ giúp quá trình này trở nên hiệu quả hơn, đồng thời tăng cường tính nhất quán và tính chính xác của dữ liệu sau khi thu thập.
3. Thiết lập lịch trình thu thập dữ liệu hợp lý và theo dõi tiến trình: Việc xây dựng một lịch trình thu thập dữ liệu rõ ràng, bao gồm thời gian cụ thể để chạy script và tần suất kiểm tra lại dữ liệu, sẽ giúp giảm thiểu rủi ro khi truy cập liên tục vào hệ thống TITV. Cơ chế ghi log sẽ ghi nhận lại các lỗi phát sinh hoặc dữ liệu bất thường, cho phép kiểm tra và khắc phục kịp thời. Điều này giúp duy trì hiệu quả của toàn bộ quá trình thu thập trong dài hạn.

5.2. Kết luận

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

Dự án nghiên cứu thu thập dữ liệu từ TITV đã đạt được một số thành tựu quan trọng trong việc phân tích và đánh giá chất lượng của các khóa học trực tuyến. Các dữ liệu được thu thập bao gồm tên khóa học, mô tả ngắn gọn, số lượng học viên tham gia, số lượt xem và đánh giá của người học, cung cấp một bức tranh toàn diện về các khóa học trên nền tảng này. Qua việc xử lý và chuẩn hóa dữ liệu, dự án không chỉ cung cấp cái nhìn sâu sắc về các khóa học phổ biến, mà còn giúp phát hiện các vấn đề tiềm ẩn về tính hấp dẫn và khả năng tiếp cận của nội dung học tập.

Kết quả phân tích đã làm rõ một số xu hướng quan trọng. Ví dụ, các khóa học có điểm đánh giá cao thường được người học quan tâm hơn và có xu hướng giữ chân học viên lâu hơn. Đồng thời, những khóa học có số lượt xem cao nhưng ít học viên tham gia cho thấy có thể tồn tại các rào cản đối với người dùng, như yêu cầu thời gian học tập quá cao hoặc cấu trúc khóa học phức tạp. Từ đó, những thông tin này có thể là cơ sở để TITV điều chỉnh các khóa học sao cho phù hợp hơn với nhu cầu và mong muốn của người học.

Video cào dữ liệu

• *5.2.2. Định hướng nghiên cứu tiếp theo*

Dựa trên những phát hiện và kinh nghiệm thu thập được trong quá trình nghiên cứu, có một số hướng phát triển có thể được khai thác thêm để mở rộng và làm phong phú kết quả nghiên cứu. Các hướng nghiên cứu tiếp theo bao gồm:

1. Phân tích tương tác người dùng: Khám phá sâu hơn về hành vi của người học trên các khóa học khác nhau, chẳng hạn như thời gian trung bình học viên dành cho mỗi video, tỷ lệ hoàn thành khóa học và mức độ hài lòng tổng thể.

Kết quả từ những phân tích này có thể giúp TITV tùy chỉnh trải nghiệm học tập cho người dùng một cách chính xác hơn.

2. Ứng dụng học máy trong dự đoán xu hướng học tập: Sử dụng các mô hình học máy để xác định mối quan hệ giữa các yếu tố như độ dài khóa học, số lượng bài giảng, đánh giá và tỷ lệ tham gia học tập. Mô hình này có thể giúp dự báo các yếu tố ảnh hưởng đến việc học viên lựa chọn khóa học và mức độ hoàn thành, từ đó giúp TITV phát triển những khóa học hấp dẫn hơn.
3. Phân tích nội dung khóa học và cải thiện chất lượng giảng dạy: Việc tiến hành phân tích nội dung chi tiết của các khóa học như tài liệu, bài tập và các phương pháp giảng dạy có thể giúp đánh giá chính xác hơn chất lượng của từng khóa học. Điều này không chỉ hỗ trợ cho quá trình đánh giá mà còn góp phần nâng cao tiêu chuẩn chất lượng cho các khóa học hiện tại và trong tương lai.

Phụ Lục

Code cào dữ liệu TIVI

```
from selenium import webdriver

from selenium.webdriver.firefox.service import Service

from selenium.webdriver.common.by import By

from selenium.webdriver.support.ui import WebDriverWait

from selenium.webdriver.support import expected_conditions as EC

import time

from pymongo import MongoClient



# Đường dẫn tới geckodriver

gecko_path =
r"C:\Users\HP\Downloads\geckodriver-v0.35.0-win64\geckodriver.exe"

service = Service(gecko_path)
```

```
# Kết nối tới MongoDB
mongo_client = MongoClient("mongodb://localhost:27017/")
db = mongo_client["course_database"]
courses_collection = db["courses"]


# Hàm thu thập thông tin khóa học
def collect_course_info():
    driver = webdriver.Firefox(service=service)
    driver.get("https://titv.vn/")
    time.sleep(10)


    all_courses = {}


    # Vòng lặp để thu thập các khóa học
    for data_id in range(75, 82):
        li_element = WebDriverWait(driver, 20).until(
            EC.element_to_be_clickable((By.CSS_SELECTOR,
f'span[data-id="{data_id}"]'))
        )
        li_name = li_element.text
        li_element.click()
        time.sleep(10)


        course_elements = WebDriverWait(driver, 20).until(
            EC.presence_of_all_elements_located((By.CSS_SELECTOR,
'div.ms_lms_courses_card_item'))
        )


        # Lặp qua từng khóa học
```

```
        for course in course_elements:
            try:
                course_name = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_wrapper >
div:nth-child(2) > a:nth-child(1) > h3'))
                ).text

                price_element = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_info_price >
div.ms_lms_courses_card_item_info_price_single > span'))
                ).text

                rating_element = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_info_rating_quantity >
span'))
                ).text

                members_count = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_meta_block >
i.stmlms-members + span'))
                ).text

                views_count = WebDriverWait(course, 20).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR,
                        'div.ms_lms_courses_card_item_meta_block >
i.stmlms-views + span'))
```

```
                ).text


                # Chuyển đổi các trường sang dạng số
                rating_numeric = float(rating_element) if rating_element
else 0.0
                price_numeric = float(price_element.replace(',',
'').replace('đ', '').strip()) if price_element and price_element.lower() !=
"free" else 0.0
                members_count_numeric = int(members_count.replace(',',
'').strip()) if members_count else 0
                views_count_numeric = int(views_count.replace(',',
'').strip()) if views_count else 0


                course_info = {
                    "name": course_name,
                    "price": price_numeric,
                    "rating": rating_numeric,
                    "members_count": members_count_numeric,
                    "views_count": views_count_numeric,
                    "category": li_name,
                }
                all_courses[course_name] = course_info


            except Exception as e:
                print(f"Error collecting course info: {e}")
                continue


        driver.back()
        time.sleep(10)
```

```
    driver.quit()


    # Lưu thông tin khóa học vào MongoDB
    for course_name, course_info in all_courses.items():
        try:
            courses_collection.update_one(
                {"name": course_name},
                {"$set": course_info},
                upsert=True
            )
        except Exception as e:
            print(f"Error saving course {course_name}: {e}")


    return all_courses


# Chạy đoạn 1 để thu thập thông tin khóa học
collect_course_info()
```

Code Truy Vấn

```
from pymongo import MongoClient


# Kết nối tới MongoDB
mongo_client = MongoClient("mongodb://localhost:27017/")
db = mongo_client["course_database"]
courses_collection = db["courses"]


# Nhóm 1: Tìm kiếm và liệt kê các khóa học
print("=== Tìm kiếm và liệt kê các khóa học ===")
```

```
# 1. Tìm tất cả các khóa học
all_courses = list(courses_collection.find())
print("Tất cả các khóa học:", all_courses)


# 2. Tìm khóa học theo tên
course_by_name = courses_collection.find_one({"name": "[Video] Microsoft Excel cơ bản"})
print("Khóa học theo tên:", course_by_name)


# 3. Tìm khóa học trong danh mục "Kiến thức nền tảng"
foundation_courses = list(courses_collection.find({"category": "Kiến thức nền tảng"}))
print("Khóa học Kiến thức nền tảng:", foundation_courses)


# 4. Tìm khóa học có số lượng thành viên lớn hơn 20,000
popular_courses = list(courses_collection.find({"members_count": {"$gt": 20000}}))
print("Khóa học có số lượng thành viên lớn hơn 20,000:", popular_courses)


# 5. Tìm khóa học có giá là 0
free_courses = list(courses_collection.find({"price": 0}))
print("Khóa học miễn phí:", free_courses)


# Nhóm 2: Đếm và cập nhật dữ liệu
print("\n=== Đếm và cập nhật dữ liệu ===")


# 6. Đếm số lượng khóa học trong danh mục "Data Science"
data_science_count = courses_collection.count_documents({"category": "Data Science"})
print("Số lượng khóa học Data Science:", data_science_count)
```

```
# 7. Tìm khóa học có rating cao nhất
highest_rated_course = courses_collection.find_one(sort=[("rating", -1)])
print("Khóa học có rating cao nhất:", highest_rated_course)

# 8. Tìm khóa học có số lượt xem lớn hơn 10,000
popular_viewed_courses = list(courses_collection.find({"views_count":
{"$gt": 10000}}))
print("Khóa học có lượt xem lớn hơn 10,000:", popular_viewed_courses)

# 9. Cập nhật số lượng thành viên cho khóa học "Git và GitHub toàn tập"
courses_collection.update_one(
   {"name": "[Video] Git và GitHub toàn tập"},
   {"$set": {"members_count": 16000}}
)
updated_course = courses_collection.find_one({"name": "[Video] Git và
GitHub toàn tập"})
print("Khóa học sau khi cập nhật số lượng thành viên:", updated_course)

# 10. Xóa khóa học có tên "[Video] Nguyên lý Hệ điều hành"
courses_collection.delete_one({"name": "[Video] Nguyên lý Hệ điều hành"})
print("Đã xóa khóa học '[Video] Nguyên lý Hệ điều hành'.")

# Nhóm 3: Tìm khóa học theo điều kiện khác
print("\n=== Tìm khóa học theo điều kiện khác ===")

# 11. Tìm tất cả khóa học có rating = 0
zero_rating_courses = list(courses_collection.find({"rating": 0}))
print("Khóa học có rating bằng 0:", zero_rating_courses)
```

```
# 12. Tìm khóa học theo ID
course_by_id = courses_collection.find_one({"_id":
"6720b69a572a75111ce409bb"})
print("Khóa học theo ID:", course_by_id)


# 13. Tìm tất cả khóa học có số lượt xem từ 5,000 đến 10,000
views_range_courses = list(courses_collection.find({"views_count": {"$gte":
5000, "$lte": 10000}}))
print("Khóa học có số lượt xem từ 5,000 đến 10,000:", views_range_courses)


# Nhóm 4: Cập nhật và nhóm khóa học
print("\n=== Cập nhật và nhóm khóa học ===")


# 14. Cập nhật rating cho tất cả khóa học có rating = 0 thành 1
courses_collection.update_many(
    {"rating": 0},
    {"$set": {"rating": 1}}
)
print("Đã cập nhật rating cho tất cả khóa học có rating bằng 0.")


# 15. Tìm khóa học có số lượng thành viên lớn hơn 5,000 và thuộc danh mục
"Data Science"
data_science_popular = list(courses_collection.find({"members_count":
{"$gt": 5000}, "category": "Data Science"}))
print("Khóa học Data Science có số lượng thành viên lớn hơn 5,000:",
data_science_popular)


# 16. Tìm khóa học có giá khác 0
paid_courses = list(courses_collection.find({"price": {"$ne": 0}}))
```

```
print("Khóa học có giá khác 0:", paid_courses)


# Nhóm 5: Tóm tắt dữ liệu
print("\n=== Tóm tắt dữ liệu ===")


# 17. Đếm số lượng khóa học trong từng danh mục
category_count = courses_collection.aggregate([
    {"$group": {"_id": "$category", "count": {"$sum": 1}}}
])
print("Số lượng khóa học trong từng danh mục:")
for category in category_count:
    print(category)


# 18. Tìm khóa học có số lượng thành viên tối thiểu
min_members_course = courses_collection.find_one(sort=[("members_count",
1)])
print("Khóa học có số lượng thành viên tối thiểu:", min_members_course)


# 19. Tìm khóa học có số lượt xem nhiều nhất
most_viewed_course = courses_collection.find_one(sort=[("views_count",
-1)])
print("Khóa học có số lượt xem nhiều nhất:", most_viewed_course)


# 20. Lấy danh sách tên khóa học
course_names = [course["name"] for course in courses_collection.find()]
print("Danh sách tên khóa học:", course_names)


# Nhóm 6: Thông tin và tìm kiếm nâng cao
print("\n=== Thông tin và tìm kiếm nâng cao ===")
```

```
# 21. Tìm tất cả khóa học có tên chứa "Video"
video_courses = list(courses_collection.find({"name": {"$regex": "Video"}}))
print("Khóa học có tên chứa 'Video':", video_courses)

# 22. Tìm khóa học có đánh giá nhỏ nhất
lowest_rating_course = courses_collection.find_one(sort=[("rating", 1)])
print("Khóa học có đánh giá thấp nhất:", lowest_rating_course)

# 23. Tìm các khóa học có số lượng thành viên ít hơn 1,000
small_member_courses = list(courses_collection.find({"members_count": {"$lt": 1000}}))
print("Khóa học có số lượng thành viên < 1,000:", small_member_courses)

# 24. Tìm khóa học nào có số lượt xem nhiều hơn khóa học "Lập trình mạng (sử dụng Java)"
reference_course = courses_collection.find_one({"name": "[Video] Lập trình mạng (sử dụng Java)"})
if reference_course:
    more_views_than_reference = list(courses_collection.find({"views_count": {"$gt": reference_course["views_count"]}}))
    print("Khóa học có số lượt xem nhiều hơn khóa học 'Lập trình mạng (sử dụng Java)':", more_views_than_reference)

# 25. Đếm tổng số lượt xem của tất cả các khóa học
total_views = courses_collection.aggregate([{"$group": {"_id": None, "total_views": {"$sum": "$views_count"}}}])
for total in total_views:
    print("Tổng số lượt xem của tất cả các khóa học:", total['total_views'])
```

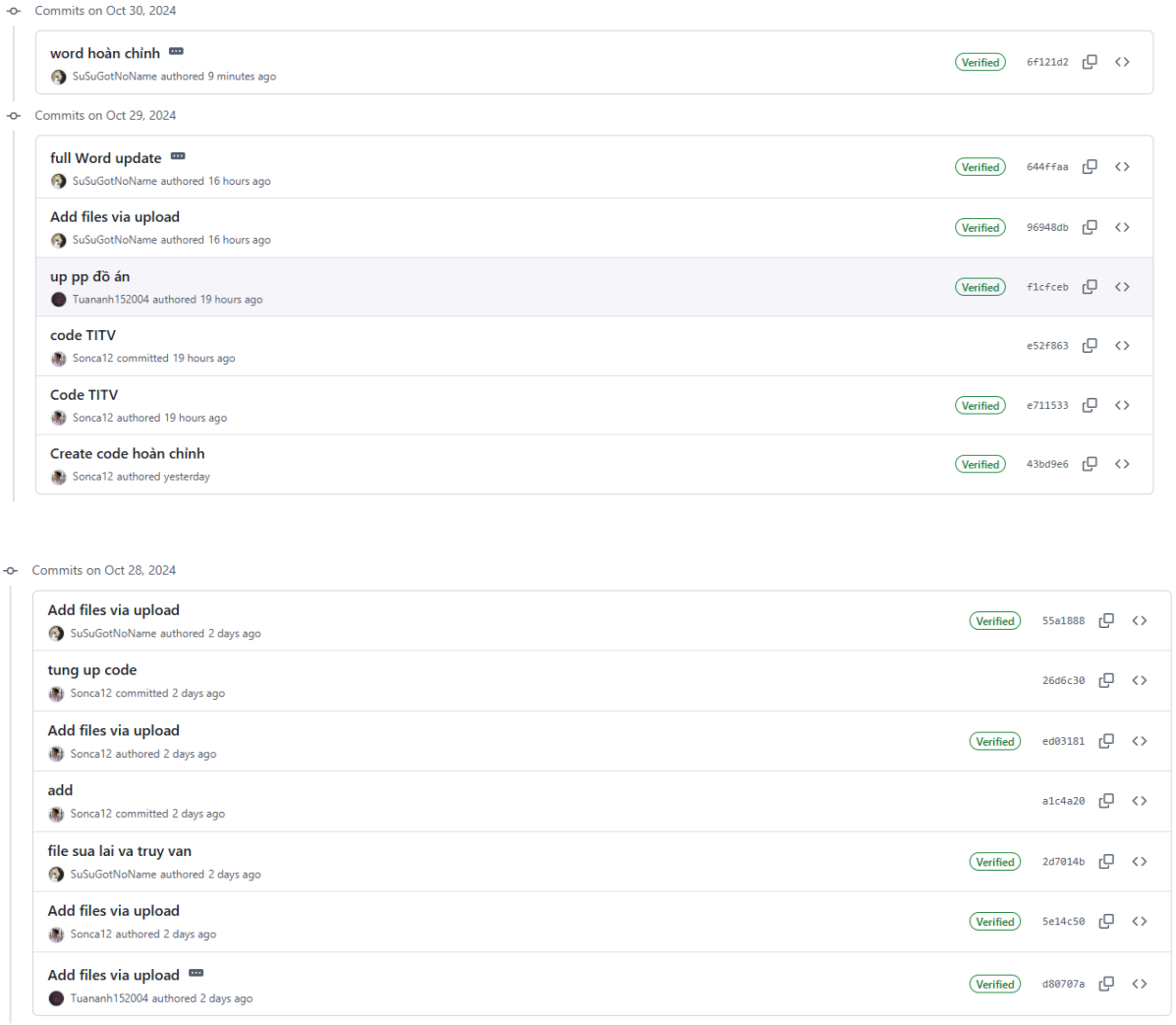
```
# 26. Tìm khóa học miễn phí và có số lượng thành viên lớn hơn 10,000
free_popular_courses = list(courses_collection.find({"price": 0, "members_count": {"$gt": 10000}}))
print("Khóa học miễn phí có số lượng thành viên > 10,000:", free_popular_courses)

# 27. Tìm khóa học có giá cao nhất
highest_price_course = courses_collection.find_one(sort=[("price", -1)])
print("Khóa học có giá cao nhất:", highest_price_course)

# 28. Tìm các khóa học có đánh giá là 5
top_rated_courses = list(courses_collection.find({"rating": 5}))
print("Khóa học có đánh giá 5:", top_rated_courses)

# 29. Tìm khóa học nào có số lượt xem thấp hơn khóa học có số lượt xem cao nhất
most_viewed_course = courses_collection.find_one(sort=[("views_count", -1)])
if most_viewed_course:
    less_views_than_most_viewed = list(courses_collection.find({"views_count": {"$lt": most_viewed_course["views_count"]}}))
    print("Khóa học có số lượt xem thấp hơn khóa học có số lượt xem cao nhất:", less_views_than_most_viewed)

# 30. Tìm các khóa học thuộc danh mục "Kiến thức nền tảng" và có đánh giá lớn hơn 1
foundation_high_rating_courses = list(courses_collection.find({"category": "Kiến thức nền tảng", "rating": {"$gt": 1}}))
```

```
print("Khóa học trong 'Kiến thức nền tảng' có đánh giá > 1:",
foundation_high_rating_courses)


# 31. Tìm tất cả khóa học có thành viên từ 1,000 đến 20,000

member_range_courses = list(courses_collection.find({"members_count":
{"$gte": 1000, "$lte": 20000}}))

print("Khóa học có số lượng thành viên từ 1,000 đến 20,000:",
member_range_courses)


# 32. Tìm khóa học nào có nhiều hơn 3,000 lượt xem và giá là 0

views_free_courses = list(courses_collection.find({"views_count": {"$gt":
3000}, "price": 0}))

print("Khóa học miễn phí có > 3,000 lượt xem:", views_free_courses)


# 33. Cập nhật số lượt xem cho khóa học theo ID

courses_collection.update_one({"_id": id("6720b69a572a75111ce409bc")},
{"$inc": {"views_count": 1}})

updated_view_course = courses_collection.find_one({"_id":
id("6720b69a572a75111ce409bc")})

print("Khóa học sau khi cập nhật số lượt xem:", updated_view_course)


# 34. Tìm khóa học đầu tiên trong danh mục "Data Science"

first_data_science_course = courses_collection.find_one({"category": "Data
Science"})

print("Khóa học đầu tiên trong danh mục 'Data Science':",
first_data_science_course)


# 35. Tìm khóa học nào có thành viên giữa 5,000 và 15,000 và có giá là 0

free_member_range_courses = list(courses_collection.find({"members_count":
{"$gte": 5000, "$lte": 15000}, "price": 0}))
```

```
print("Khóa học miễn phí có thành viên từ 5,000 đến 15,000:",
free_member_range_courses)


# 36. Liệt kê tất cả các danh mục có trong khóa học
distinct_categories = courses_collection.distinct("category")
print("Danh mục khóa học:", distinct_categories)


# 37. Tìm khóa học nào có số lượng thành viên tối đa
max_members_course = courses_collection.find_one(sort=[("members_count",
-1)])
print("Khóa học có số lượng thành viên tối đa:", max_members_course)


# 38. Tìm tất cả khóa học có rating < 3
low_rating_courses = list(courses_collection.find({"rating": {"$lt": 3}}))
print("Khóa học có rating < 3:", low_rating_courses)


# 39. Tìm tất cả khóa học có tên bắt đầu bằng "[Video]"
video_starting_courses = list(courses_collection.find({"name": {"$regex":
r"^\[Video\]"}}))
print("Khóa học có tên bắt đầu bằng '[Video]':", video_starting_courses)


# 40. Tính trung bình số lượt xem của tất cả các khóa học
average_views = courses_collection.aggregate([{"$group": {"_id": None,
"average_views": {"$avg": "$views_count"}}}])
for avg in average_views:
   print("Trung bình số lượt xem của tất cả các khóa học:",
avg['average_views'])



# Đóng kết nối
```

```
mongo_client.close()
```

Hình ảnh lịch sử commit trên GitHub



Commits on Oct 27, 2024

Added SuSuGotNoName committed 3 days ago		a07d035
code hoàn chỉnh lưu vào mongodb Sonca12 authored 3 days ago	Verified	4cd4119
lụm và làm tiếp nha các bé yêu Sonca12 authored 3 days ago	Verified	4f740db

Commits on Oct 20, 2024

phần 1 và 2 của đồ án Sonca12 authored last week	Verified	4d4fc40
Tùng code nha các bbi <3 Sonca12 authored last week	Verified	f73a87d

Commits on Oct 16, 2024

phần bìa , mục 1 , cam kết Sonca12 committed 2 weeks ago		49f45e7

Commits on Oct 1, 2024

Add SuSuGotNoName committed 29 days ago		3281354
Merge branch 'main' of https://github.com/Tuananh152004/nhom-3-nhoc SuSuGotNoName committed 29 days ago		663e7ce

Commits on Sep 30, 2024

phan tong quan Sonca12 committed last month		001a9f8
Merge branch 'main' of https://github.com/Tuananh152004/nhom-3-nhoc SuSuGotNoName committed on Sep 30		7268c2a
Merge branch 'main' of https://github.com/Tuananh152004/nhom-3-nhoc Sonca12 committed on Sep 30		80ca10e
Add Sonca12 committed on Sep 30		88d916a
Add report Sonca12 committed on Sep 30		3bf0cc4
Add Trang bia DACS.doc Sonca12 committed on Sep 30		87d221f

added Tuananh152004 committed on Sep 30		7f36b04
Added SuSuGotNoName committed on Sep 30		34ed997
add Tuananh152004 committed on Sep 30		e3a1c0d
Merge branch 'main' of https://github.com/Tuananh152004/nhom-3-nhoc SuSuGotNoName committed on Sep 30		4bf68ee
Added SuSuGotNoName committed on Sep 30		e1382a1
Merge branch 'main' of https://github.com/Tuananh152004/nhom-3-nhoc Tuananh152004 committed on Sep 30		7f06b28
added Tuananh152004 committed on Sep 30		2b80fb5

Liên kết trang wed :

Link:TITV – Giữ cho mọi thứ đơn giản

Link Github nhóm 3 nhóc : Tuananh152004/nhom-3-nhoc