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

**ĐỒ ÁN MÔN HỌC**

**TÌM HIỂU KỸ THUẬT WEB SCRAPING BẰNG SELENIUM, THU THẬP VÀ PHÂN TÍCH DỮ LIỆU TỪ TRANG CÁ NHÂN TRÊN MẠNG XÃ HỘI X**

Ngành: **KHOA HỌC DỮ LIỆU**

Môn học: **MÃ NGUỒN MỞ TRONG KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn: ThS.Lê Nhật Tùng

Sinh viên thực hiện: 2286400022 - Hồ Nguyễn Hoàng Phát

2286400042 - Trần Lê Vân

TP. Hồ Chí Minh, 2024

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

**ĐỒ ÁN MÔN HỌC**

**TÌM HIỂU KỸ THUẬT WEB SCRAPING BẰNG SELENIUM, THU THẬP VÀ PHÂN TÍCH DỮ LIỆU TỪ TRANG CÁ NHÂN TRÊN MẠNG XÃ HỘI X**

Ngành: **KHOA HỌC DỮ LIỆU**

Môn học: **MÃ NGUỒN MỞ TRONG KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn: ThS.Lê Nhật Tùng

Sinh viên thực hiện: 2286400022 - Hồ Nguyễn Hoàng Phát

2286400042 - Trần Lê Vân

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, Hồ Nguyễn Hoàng Phát, Trần Lê Vân xin cam đoan rằng:

Tất cả thông tin và phân tích trình bày trong báo cáo này được thực hiện một cách chính xác và trung thực.

Mọi dữ liệu, nhận định hoặc ý kiến được trích dẫn từ các nguồn khác đều đã được nêu rõ nguồn gốc và trích dẫn đúng quy định. Tôi cam đoan rằng không có bất kỳ hành vi sao chép hoặc sử dụng thông tin không hợp pháp nào từ các nguồn khác.

Bài báo cáo này là kết quả của công trình nghiên cứu độc lập của chúng tôi và chưa từng được công bố tại bất kỳ nơi nào khác. Tôi cam đoan đã tuân thủ nghiêm ngặt các quy tắc và quy định của môn học, bao gồm việc tham khảo và áp dụng các công cụ nghiên cứu một cách hợp lệ. Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung bài báo cáo của mình.

Tôi hy vọng rằng bài báo cáo này sẽ cung cấp những thông tin hữu ích cho các nhà nghiên cứu , doanh nghiệp. Góp phần vào việc hiểu rõ hơn về mạng xã hội ngày nay.

TPHCM, ngày……tháng 10 năm 2024

**Sinh viên**

**MỤC LỤC**

**DANH MỤC CÁC KÝ HIỆU , CÁC CHỮ VIẾT TẮT**

**DANH MỤC CÁC BẢNG**

**DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ**

# CHƯƠNG 1 TỔNG QUAN

* 1. Giới thiệu đề tài

Mạng xã hội đã trở thành một phần không thể thiếu trong cuộc sống hiện nay của chúng ta. Mỗi ngày có hàng triệu dữ liệu được tạo ra và chia sẻ trên các nền tảng trực tuyến. Dữ liệu từ mạng xã hội như một kho tàng vô giá, đóng vai trò quan trọng trong việc phân tích hành vi người dùng và dự đoán xu hướng. Đề tài “ ” tập trung vào việc nghiên cứu và áp dụng kỹ thuật thu thập dữ liệu (web scraping) bằng Selemium, một công cụ giúp tương tác và cho phép mô phỏng các hành động của người dùng trên trình duyệt, lấy dữ liệu từ các trang web một cách tự động. Từ đó, dữ liệu thu thập sẽ được phân tích nhằm đưa ra các thông tin quan trọng về hành vi người dùng, mức độ tương tác và các yếu tố khác.

* 1. Nhiệm vụ đồ án

Nhiệm vụ của đề tài là hướng tới nghiên cứu và áp dụng kĩ thuật web scraping để trích xuất thông tin từ các trang cá nhân trên mạng xã hội X.Từ đó, phục vụ cho việc phân tích và khám phá các xu hướng, hiểu rõ hành vi của người dùng trên mạng xã hội.

* + 1. *Tính cấp thiết và lý do hình thành đề tài*

Trong thời đại số, mạng xã hội X đã trở thành một diễn đàn sôi động, nơi hàng triệu người dùng chia sẻ thông tin, ý kiến và tương tác với nhau mỗi ngày. Những nội dung như bài viết, bình luận, lượt thích, và các hành vi trực tuyến khác không chỉ phản ánh quan điểm cá nhân mà còn tạo ra những xu hướng xã hội, sở thích tiêu dùng và cảm xúc công chúng. Những dữ liệu này ẩn chứa nhiều thông tin giá trị về hành vi, quan điểm và xu hướng của người dùng, cung cấp nguồn tài nguyên phong phú cho các nhà nghiên cứu và nhà hoạch định chính sách trong việc hiểu sâu hơn về thị hiếu, hành vi và sự phát triển của xã hội.

Tuy nhiên, việc khai thác và phân tích dữ liệu trên mạng xã hội X vẫn còn nhiều hạn chế do tốc độ thay đổi nhanh chóng của dữ liệu. Quá trình thu thập dữ liệu thủ công từ các trang cá nhân không chỉ tốn nhiều thời gian và công sức mà còn dễ gặp phải các rào cản kỹ thuật từ việc các trang web với nhiều lớp mã JavaScript phức tạp. Hơn nữa, các công cụ phân tích dữ liệu truyền thống khó có thể xử lý và khai thác hiệu quả lượng thông tin không có cấu trúc và thay đổi liên tục này.

Do đó, kỹ thuật web scraping bằng công cụ Selenium, một công cụ tự động hóa trình duyệt web , mang lại giải pháp hiệu quả để vượt qua những khó khăn này. Selenium cho phép mô phỏng toàn bộ các thao tác của người dùng trên trình duyệt, từ việc đăng nhập cho đến tương tác với các phần tử trên trang, nhờ đó có thể thu thập dữ liệu từ những trang web động mà các phương pháp thông thường không thể tiếp cận.

Nghiên cứu này được thực hiện nhằm đáp ứng nhu cầu cấp thiết về một phương pháp hiệu quả để thu thập và phân tích dữ liệu từ mạng xã hội X, từ đó khám phá những giá trị tiềm ẩn trong dữ liệu này. Thông qua việc ứng dụng Selenium, đề tài hướng tới khám phá những giá trị tiềm ẩn trong dữ liệu giúp các nhà nghiên cứu xã hội học nắm bắt được xu hướng cộng đồng, và các tổ chức có thể đưa ra các quyết định sáng suốt hơn dựa trên dữ liệu. Với sự phát triển của kỹ thuật thu thập liệu và khả năng tự động hóa quy trình, đề tài này mở ra tiềm năng lớn cho việc ứng dụng dữ liệu mạng xã hội trong nhiều lĩnh vực khác nhau, từ tiếp thị kỹ thuật số đến nghiên cứu khoa học và quản lý xã hội.

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

Ý nghĩa khoa học: Nghiên cứu này đóng góp vào việc phát triển và mở rộng kiến thức về kỹ thuật web scraping và cách sử dụng Selenium trong thu thập dữ liệu từ các trang web động, đặc biệt là trên mạng xã hội. Bằng cách nghiên cứu quy trình thu thập dữ liệu tự động, đề tài giúp làm rõ các phương pháp và công cụ hiện đại, đồng thời cung cấp cái nhìn sâu hơn về những thách thức và giải pháp trong việc xử lý dữ liệu lớn từ môi trường trực tuyến. Kết quả của nghiên cứu sẽ tạo nền tảng cho các công trình tiếp theo trong lĩnh vực khoa học dữ liệu, hỗ trợ việc ứng dụng công nghệ vào phân tích hành vi người dùng và phân tích dữ liệu mạng xã hội.

Ý nghĩa thực tiễn: Nghiên cứu này mang lại lợi ích thực tiễn rõ rệt cho nhiều lĩnh vực. Đối với các doanh nghiệp, kết quả thu thập và phân tích dữ liệu từ mạng xã hội cung cấp thông tin quan trọng về hành vi khách hàng, xu hướng tiêu dùng, và cảm xúc cộng đồng. Điều này giúp họ xây dựng chiến lược tiếp thị chính xác hơn, tối ưu hóa chiến dịch quảng cáo. Ngoài ra, nghiên cứu còn hỗ trợ các nhà nghiên cứu xã hội trong việc phân tích xu hướng, hành vi cộng đồng và các vấn đề xã hội, góp phần vào quá trình ra quyết định dựa trên dữ liệu chính xác.

* 1. Mục tiêu nghiên cứu
     1. *Mục tiêu tổng quát:*

Đề tài nhằm cung cấp một hệ thống web scraping tự động, linh hoạt và hiệu quả bằng Selenium để thu thập dữ liệu từ trang cá nhân trên mạng xã hội X. Từ đó, phân tích thông tin thu thập được để khám phá những xu hướng và hành vi của người dùng trên nền tảng này.

* + 1. *Mục tiêu cụ thể:*

Trong bài nghiên cứu này, chúng tôi xây dựng một kích bản tự động sử dụng Selenium để thu thập dữ liệu từ trang cá nhân trên mạng xã hội X , bao gồm các thông tin công khai như bài đăng , ngày đăng bài , các reaction(like , love, …), bình luận, lượt chia sẻ bài. Tổ chức dữ liệu lại theo cấu trúc có thể phân tích. Lưu trữ dữ liệu đã thu thập vào cơ sở dữ liệu NoSQL , nhằm tận dụng khả năng lưu trữ dữ liệu linh hoạt và hiệu quả của MôngDB . Sử dụng truy vấn NoSQL trên MongoDB để thực hiện phân tích dữ liệu, khám phá xu hướng chính, hành vi của người dùng và các yếu tố như tương tác và nội dụng. Đánh giá hiệu quả của hệ thống web scraping dựa trên tốc độ thu thập, tính chính xác của dữ liệu, và khả năng mở rộng hệ thống trong việc xử lý các tập dữ liệu lớn từ mạng xã hội X.

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

Đối tượng nghiên cứu của đề tài là các trang cá nhân trên mạng xã hội X , bao gồm các tài khoản cá nhân , doanh nghiệp. Các tài khoản được chọn phải có hoạt động thường xuyên, đặc biệt là các tương tác như bài đăng, bình luận và lượt thích phải công khai giúp cung cấp dữ liệu phong phú cho quá trình phân tích.

* + 1. *Phạm vi giới hạn*

Đề tài này tập trung vào việc thu thập dữ liệu tự động từ các trang cá nhân trên mạng xã hội X bằng cách sử dụng kỹ thuật web scraping bằng Selenium. Dữ liệu thu thập bao gồm các thông tin như bài đăng, bình luận, lượt thích, và các tương tác công khai khác. Sau khi thu thập, dữ liệu sẽ được lưu trữ và xử lý trong cơ sở dữ liệu NoSQL (MongoDB), nhằm phục vụ cho quá trình phân tích. Từ đó, có thể thấy các xu hướng, hành vi người dùng, và mối liên hệ giữa nọi dung và tương tác.

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

Chúng tôi tiến hành tìm hiểu tổng quan về các kỹ thuật web scraping và khả năng thu thập dữ liệu từ mạng xã hội X. Nghiên cứu sơ bộ sẽ giúp xác định các yếu tố chính cần xem xét, như loại dữ liệu cần thu thập, giới hạn kỹ thuật của Selenium và các vấn đề về quyền riêng tư. Điều này giúp định hình các hướng đi cụ thể cho quá trình nghiên cứu sâu hơn.

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

Phương pháp này liên quan đến việc cái tài liệu, báo cáo nghiên cứu liên quan đến web scraping , Selenium , phân tích dữ liệu và NoSQL. Việc nghiên cứu tài liệu sẽ cung cấp các kiến thức nền tảng và những ví dụ thực tiễn trong việc sử dụng các công cụ này. Thông qua việc nghiên cứu tài liệu, chúng tôi sẽ có cơ sở để phát triển công cụ phù hợp cho đề tài.

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

Phương pháp thống kê sẽ được sử dụng để phân tích dữ liệu sau khi đã thu thập được. Các phương pháp thống kê mô tả sẽ giúp tóm tắt các đặc điểm chính của dữ liệu, như tần suất xuất hiện của các tương tác, xu hướng nội dung, rút ra các kết luận về hành vi người dùng và tương tác trên mạng xã hội X dựa trên mẫu dữ liệu đã thu thập.

* + 1. *Phương pháp nghiên cứu thực nghiệm*

Phương pháp thực nghiệm sẽ được tiến hành thông qua quá trình xây dựng hệ thống web scraping bằng Selenium. Nhóm nghiên cứu sẽ thử nghiệm các trường hợp trích xuất dữ liệu tự động từ các trang cá nhân, đánh giá hiệu suất của hệ thống và điều chỉnh khi cần. Việc thu thập và xử lý dữ liệu trong cơ sở dữ liệu NoSQL cũng sẽ được thử nghiệm để đảm bảo khả năng lưu trữ và truy xuất hiệu quả.

* + 1. *Phương pháp đánh giá*

Cuối cùng, chúng tôi sẽ thực hiện phương pháp đánh giá để đo lường hiệu quả của quá trình tự động thu thập dữ liệu và phân tích dữ liệu. Quá trình này sẽ xem xét tốc độ thu thập , tính chính xác và đầy đủ của dữ liệu, cũng như hiệu quả của quá trình phân tích dữ liệu thông qua các tiêu chí về độ tin cậy và các giá trị khác của các kết quả phân tích từ hệ thống.

* 1. Những đóng góp nghiên cứu của đề tài
     1. *Trong lĩnh vực học thuật*

Đề tài này đóng góp bằng cách mở rộng kiến thức về kỹ thuật web scraping và tự động thu thập dữ liệu từ các nền tảng mạng xã hội, đặc biệt là sử dụng Selenium. Điều này giúp bổ sung cho các tài liệu về phân tích dữ liệu phi cấu trúc và ứng dụng NoSQL (MongoDB) trong việc lưu trữ và quản lý dữ liệu. Ngoài ra, còn góp phần vào nghiên cứu hành vi người dùng trên mạng xã hội, mở ra tiềm năng cho những nghiên cứu sâu hơn về lĩnh vực này.

* + 1. *Trong thực tiễn kinh doanh*

Cung cấp một hệ thống có thể giúp doanh nghiệp thu thập dữ liệu về hành vi và xu hướng người dùng trên mạng xã hội, từ đó xây dựng các chiến lược phát triển trong tương lai. Ngoài ra, các doanh nghiệp có thể tận dụng hệ thống để hiểu rõ hơn về sự tương tác của khách hàng và cải thiện trải nghiệm người dùng.

* 1. Cấu trúc đồ án
     1. *Trình bày cấu trúc đồ án*

Chương 1: Tổng quan

Chương 2: Cơ sở lý thuyết

Chương 3: Phương pháp nghiên cứu

Chương 4: Kết quả thực nghiệm

* + 1. *Tóm tắt từng chương*

Chương 1: Tổng quan

Trình bày về lý do chọn đề tài , tính cấp thiết, mục tiêu nghiên cứu, phạm vi và giới hạn nghiên cứu, cũng như đóng góp khoa học và thực tiễn của đề tài. Từ đó, đưa ta các nền tảng lý thuyết cần thiết để giải thích về tầm quan trọng của kỹ thuật web scraping và phương pháp phân tích dữ liệu từ mạng xã hội.

Chương 2: Cơ sở lý thuyết

Cung cấp các kiến thức liên quan đến web scraping, Selenium, NoSQL và MongoDB, đồng thời trình bày các khái niệm về( dữ liệu phi cấu trúc) và phân tích dữ liệu từ mạng xã hội. Các lý thuyết và nghiên cứu trước đây cũng sẽ được tham khảo để làm cơ sở cho việc phát triển hệ thống.

Chương 3: Phương pháp nghiên cứu

Chương này mô tả chi tiết về các phương pháp thu thập dữ liệu và xử lý dữ liệu, bao gồm việc xây dựng hệ thống web scraping, lưu trữ và phân tích dữ liệu , cũng như cách thức thực nghiệm và đánh giá hiệu quả của hệ thống.

Chương 4: Kết quả thực nghiệm

Chương này tập trung vào việc phân tích kết quả thu thập được, đưa ra các nhận xét về xu hướng và hành vi người dùng trên mạng xã hội. Các kết quả phân tích sẽ được trình bày dưới dạng số liệu, và những kết luận liên quan.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Tổng quan về web scraping

*2.1.1. Khái niệm*

Web scraping là quá trình tự động thu thập dữ liệu từ các trang web . Hầu hết dữ liệu từ trang web là dữ liệu phi cấu trúc ở định dạng HTML, sau đó sử dụng các công cụ thể truy cập vào các trang web, trích xuất nội dung cần thiết và lưu trữ dưới dạng có cấu trúc để dễ dàng phân tích. Web scraping giúp truy cập nhanh chóng và hiệu quả vào dữ liệu công khai trên Internet, thường được sử dụng trong các lĩnh vực như phân tích dữ liệu, nghiên cứu thị trường, và theo dõi thông tin trực tuyến.

<https://duyanhweb.com.vn/web-scraping.html>

2.1.2. Cách thức hoạt động:

Quá trình thu thập dữ liệu bao gồm các bước sau:



Hình : Quy trình hoạt động của Web scraping

- Lập kế hoạch và xác định các dữ liệu cần thu thập.

- Tải trang web

- Phân tích HTML

- Trích xuất dữ liệu

- Lưu trữ dữ liệu

*2.1.2. Ưu điểm*

Tự động hóa quy trình thu thập dữ liệu: Giúp tiết kiệm thời gian và công sức so với việc thu thập dữ liệu thủ công

Thu thập lượng lớn dữ liệu: Có khả năng thu thập dữ liệu từ nhiều trang wed khác nhau trong một khoảng thời gian ngắn

Cập nhập liên tục: Có thể thiết lập công cụ Web Scraping để tự động cập nhập dữ liệu theo thời gian thực, giúp thong tin luôn chính xác và mới nhất.

Tăng hiệu quả phân tích: Cung cấp lượng thông tin luôn chính lớn cho các dự án phân tích , nghiên cứu thị trường và dự đoán xu hướng.

*2.1.3. Nhược điểm*

Phụ thuộc vào cấu trúc HTML: Nếu trang thay đổi cấu trúc HTML*,* mã scraping cần được cập nhập để phù hợp với cấu trúc mới.

Rủi ro pháp lý: Một số trang web có thể có điều khoản cấm web scraping, và việc thu thập dữ liệu có thể gây vi phạm điểu khoản sử dụng của trang đó , gây ra tranh chấp pháp lý cho người thực hiện.

2.2. Giới thiệu về python

2.2.1. Khái niệm

2.2.2. Tính năng và lợi thế của python

2.2.3. Ứng dụng

2.3. Giới thiệu về Selenium

2.3.1. Khái niệm

2.3.2. Các thành phần của Selenium

2.3.3. Ứng dụng

2.4. Giới thiệu về NoSQL

2.4.1. Khái niệm về NoSQL

2.4.2.

2.4.3.

2.5. Mongodb

2.5.1.Khái niệm

MongoDB là một hệ quản trị cơ sở dữ liệu mã nguồn mở, thuộc loại NoSQL,

2.5.2. Đặc điểm quan trọng của MongoDB

2.5.3. Ưu điểm:

2.5.4. Nhược điểm:

2.5.5. Ứng dụng của MongoDB