|  |  |
| --- | --- |
|  | 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 một cách tự động từ trang cá nhân trên mạng xã hội X.

* 1. Cấu trúc đồ án
     1. Trình bày cấu trúc đồ án
     2. Tóm tắt từng chương

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Khái niệm và phương pháp giải quyết vấn đề

2.2. Mô tả các công nghệ, hệ thống, các ràng buộc

2.3. Các giải pháp mới và mô hình toán

2.4. Lý giải xây dựng mô hình

# CHƯƠNG 3: PHƯƠNG PHÁP THỰC NGHIỆM

3.1. Phương pháp thu thập dữ liệu:CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

4.1. Cài đặt Selenium