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

**ĐỒ ÁN MÔN HỌC**

**THU THẬP VÀ PHÂN TÍCH DỮ LIỆU TỪ TRANG CÁ NHÂN TRÊN MẠNG XÃ HỘI X.COM (TWITTER.COM) SỬ DỤNG CÔNG CỤ SELENIUM VÀ MONGODB**

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

Lớp: 22DKHA1

TP. Hồ Chí Minh, 2024

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

**ĐỒ ÁN MÔN HỌC**

**THU THẬP VÀ PHÂN TÍCH DỮ LIỆU TỪ**

**TRANG CÁ NHÂN TRÊN MẠNG XÃ HỘI X.COM (TWITTER.COM) SỬ DỤNG CÔNG CỤ SELENIUM VÀ MONGODB**

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

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, 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**

[NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN 3](#_Toc180956850)

[LỜI CAM ĐOAN 4](#_Toc180956851)

[MỤC LỤC 5](#_Toc180956852)

[DANH MỤC CÁC KÝ HIỆU , CÁC CHỮ VIẾT TẮT 8](#_Toc180956853)

[DANH MỤC CÁC BẢNG 9](#_Toc180956854)

[DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ 10](#_Toc180956855)

[CHƯƠNG 1. TỔNG QUAN 11](#_Toc180956856)

[1.1. Giới thiệu đề tài 11](#_Toc180956857)

[1.2. Nhiệm vụ đồ án 11](#_Toc180956858)

[*1.2.1.* *Tính cấp thiết và lý do hình thành đề tài* 11](#_Toc180956859)

[*1.2.2. Ý nghĩa khoa học và thực tiễn của đề tài* 12](#_Toc180956860)

[1.3. Mục tiêu nghiên cứu 13](#_Toc180956861)

[*1.3.1. Mục tiêu tổng quát* 13](#_Toc180956862)

[*1.3.2. Mục tiêu cụ thể* 13](#_Toc180956863)

[1.4. Đối tượng và phạm vi giới hạn 13](#_Toc180956864)

[*1.4.1. Đối tượng* 13](#_Toc180956865)

[*1.4.2.* *Phạm vi giới hạn* 14](#_Toc180956866)

[1.5. Phương pháp nghiên cứu 14](#_Toc180956867)

[*1.5.1.* *Phương pháp nghiên cứu sơ bộ* 14](#_Toc180956868)

[*1.5.2.* *Phương pháp nghiên cứu tài liệu* 14](#_Toc180956869)

[*1.5.3.* *Phương pháp nghiên cứu thống kê* 14](#_Toc180956870)

[*1.5.4.* *Phương pháp nghiên cứu thực nghiệm* 15](#_Toc180956871)

[*1.5.5.* *Phương pháp đánh giá* 15](#_Toc180956872)

[1.6. Những đóng góp nghiên cứu của đề tài 15](#_Toc180956873)

[*1.6.1.* *Trong lĩnh vực học thuật* 15](#_Toc180956874)

[*1.6.2.* *Trong thực tiễn kinh doanh* 15](#_Toc180956875)

[1.7. Cấu trúc đồ án 16](#_Toc180956876)

[*1.7.1.* *Trình bày cấu trúc đồ án* 16](#_Toc180956877)

[*1.7.2.* *Tóm tắt từng chương* 16](#_Toc180956878)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 17](#_Toc180956879)

[2.1. Web Scraping 17](#_Toc180956880)

[*2.1.1. Khái niệm* 17](#_Toc180956881)

[*2.1.2. Cách thức hoạt động* 17](#_Toc180956882)

[*2.1.3. Ưu điểm* 18](#_Toc180956883)

[*2.1.4. Nhược điểm* 19](#_Toc180956884)

[2.2. Python 19](#_Toc180956885)

[*2.2.1. Khái niệm* 19](#_Toc180956886)

[*2.2.2. Tính năng và lợi thế của python* 19](#_Toc180956887)

[*2.2.3. Nhược điểm* 20](#_Toc180956888)

[*2.2.4. Ứng dụng* 20](#_Toc180956889)

[2.3. Selenium 21](#_Toc180956890)

[*2.3.1. Khái niệm* 21](#_Toc180956891)

[*2.3.2. Các thành phần của Selenium* 21](#_Toc180956892)

[*2.3.3. Ứng dụng* 21](#_Toc180956893)

[*2.3.4. Các công trình công nghệ áp dụng Selenium* 21](#_Toc180956894)

[2.4. Giới thiệu về NoSQL 22](#_Toc180956895)

[*2.4.1. Khái niệm về NoSQL* 22](#_Toc180956896)

[*2.4.2. Tính năng nổi bật của NoSQL* 22](#_Toc180956897)

[*2.4.3. Các loại dữ liệu NoSQL* 22](#_Toc180956898)

[*2.4.4.Ứng dụng của NoSQL* 23](#_Toc180956899)

[2.5. MongoDB 24](#_Toc180956900)

[*2.5.1.Khái niệm* 24](#_Toc180956901)

[*2.5.2. Đặc điểm quan trọng của MongoDB* 24](#_Toc180956902)

[*2.5.3. Ưu điểm* 24](#_Toc180956903)

[*2.5.4. Nhược điểm* 25](#_Toc180956904)

[*2.5.5. Ứng dụng của MongoDB* 25](#_Toc180956905)

[2.6. Trình duyệt FireFox 25](#_Toc180956906)

[*2.6.1. Giới thiệu trình duyệt* 25](#_Toc180956907)

[*2.6.2. Vì sao chọn FireFox* 25](#_Toc180956908)

[CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM 26](#_Toc180956909)

[3.1. Trình bày công việc thực nghiệm 26](#_Toc180956910)

[*3.1.1. Mục tiêu* 26](#_Toc180956911)

[*3.1.2. Quy trình thực nghiệm* 26](#_Toc180956912)

[3.2. Kết quả thực nghiệm 26](#_Toc180956913)

[*3.2.1. Chi tiết quá trình cào dữ liệu* 26](#_Toc180956914)

[*3.2.2. Kết quả thực nghiệm* 29](#_Toc180956915)

[*3.4.1. So Sánh với mục tiêu ban đầu* 29](#_Toc180956916)

[*3.4.2. Đánh giá hiệu suất và độ chính xác* 30](#_Toc180956917)

[**3.5 Hạn chế và hướng phát triển** 30](#_Toc180956918)

[*3.5.1. Hạn chế của chương trình* 30](#_Toc180956919)

[*3.5.2. Hướng phát triển* 30](#_Toc180956920)

# DANH MỤC CÁC KÝ HIỆU , CÁC CHỮ VIẾT TẮT

Web Scraping

NoSQL

GDPR: Quy định chung về bảo vệ dữ liệu  
ToS: Điều khoản dịch vụ

# 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 Selenium, 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.

*1.2.2. 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.

1.3. Mục tiêu nghiên cứu

*1.3.1. 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.

*1.3.2. 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, bình luận, lượt chia sẻ. 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 MongoDB. 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.

1.4. Đối tượng và phạm vi giới hạn

*1.4.1. 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. Web Scraping

*2.1.1. Khái niệm*

Web scraping là quá trình tự động thu thập dữ liệu từ các trang web.

Có một cách phổ biến mà chúng ta vẫn hay dùng đó là sao chép và dán (Ctrl + C, Ctrl + V), nhưng đây là cách thủ công và khá rườm rà. Chúng ta chỉ có thể sử dụng nếu dữ liệu ta thu thập được ít, nhưng nếu cần thu thập nhiều dữ liệu dung cho mục đích khai thác và phân tích, ta không thể dùng theo cách phổ thông như vậy vì rất tốn thời gian.

Nhưng ngày nay, dữ liệu đang ngày càng trở nên quan trọng và phần lớn dữ liệu này đều có sẵn trên các trang web. Vậy làm thế nào để có thể thu thập số lượng lớn dữ liệu nhanh chóng? Khái niệm web scraping bắt đầu xuất hiện từ đây. Quá trình này được thực hiện với sự trợ giúp của các phần mềm trích xuất dữ liệu. Và nó tự động tải và thu thập dữ liệu từ các trang web dựa trên yêu cầu của người dung.[1]

*2.1.2. Cách thức hoạt động*

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

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



Hình 1: Quy trình hoạt động của Web scraping

- Gửi một yêu cầu HTTP đến server.

- Nhận HTML từ server.

- Phân tích HTML và trích xuất dữ liệu.

- Lưu trữ dữ liệu.

*2.1.2. Công dụng*

- Nghiên cứu thị trường: Web Scraping được sử dụng để lấy phản hồi từ khách hàng về dịch vụ, hoặc về sản phẩm thể xin ý kiến từ khách hàng. Hoặc nó có thể so sánh giá cả từ các đối thủ cạnh tranh.

- Phân tích dữ liệu lớn: Web Scraping có khả năng thu thập số lượng lớn dữ liệu và nhờ vào đó, có thể dự đoán xu hướng kinh doanh trong tương lai.

- Học máy: Máy được phép học và cải thiện khả năng của nó thay vì được lập trình sẵn. Mà để làm được điều đó, cần có một lượng dữ liệu cực kỳ lớn được thu thập và trích xuất từ hàng triệu trang web khác nhau thông qua Web Scraping.

- Phân tích dữ liệu tài chính: Web Scraping được sử dụng để tổng hợp thông tin từ các thị trường chứng khoán, tình hình biến động về kinh tế thế giới hoặc cụ thể hơn là khu vực, để có thể hổ trợ phân tích và nghiên cứu đầu tư.

- Tối ưu hóa công cụ tìm kiếm (SEO): Web Scraping giúp thu thập dữ liệu theo thời gian thực từ các trang mạng xã hội, nhật kí trực tuyến. Doanh nghiệp dựa vào các dữ liệu thu thập được để tìm ra xu hướng mới để tiếp cận khách hàng, nâng cao lượt hiển thị và thứ hạng của một trang web.

*2.1.3. 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.

*2.1.4. 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.

- Tính hợp pháp của Web Scraping: Đây không phải là vấn đề bất hợp pháp, nhưng nếu ta lạm dụng, có thể làm giảm hiệu suất máy chủ của trang web. Còn nhiều yếu tố liên quan đến đạo đức hoặc thậm chí là pháp luật. Một số quy định về bảo vệ dữ liệu (GDPR) ở châu Âu về thu thập dữ liệu cá nhân. Hoặc nhiều trang web có ToS (điều khoản dịch vụ) dẫn tới vi phạm pháp lý dù thực tế nó không phải là hành động bất hợp pháp.[2]

2.2. Python

*2.2.1. Khái niệm*

Python là một ngôn ngữ lập trình bậc cao, được thông dịch. Được Guido van Rossum tạo ra và phát hành vào năm 1991. Ngôn ngữ này ngày càng trở nên phổ biến và phát triển cực kỳ mạnh mẽ bởi cấu trúc cú pháp đơn giản, dễ đọc. Và đặc biệt được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau.

*2.2.2. Tính năng và lợi thế của python*

- Đơn giản và dễ tiếp cận: Đây là một ngôn ngữ có cú pháp rất đơn giản, dễ học và cũng dễ đọc. Là ngôn ngữ phù hợp cho người mới học lập trình.

- Được thông dịch (interpreted language): Python là ngôn ngữ được thông dịch, nghĩa là mã được thực thi từng dòng một mà không cần biên dịch thành mã máy. Điều này giúp dễ dàng kiểm tra và sữa lỗi, không quá phức tạp. [3]

- Sự linh hoạt: Bạn không cần khai báo kiểu dữ liệu của biến một cách rõ ràng. Python có thể tự động suy ra kiểu dữ liệu dựa trên giá trị được gán cho biến. Kiểu dữ liệu của biến được áp dụng khi chạy chương trình. Giúp làm tăng tốc độ và giúp ngăn ngừa lỗi làm cho mã mạnh mẽ hơn.

- Ngôn ngữ đa nền tảng: Python có thể chạy trên nhiều hệ điều hành khác nhau mà không cần thay đổi. Giúp dễ dàng phát triển ứng dụng Python trên nhiều nền tảng khác nhau mà không gặp bất kì rắc rối nào.

- Đa dạng thư viện: Gồm các thư viện cho khoa học dữ liệu (Numpy, Pandas), học máy (Scikit-learn), web (Django, Flask) cung cấp các công cụ và mô-đun cho nhiều tác vụ khác nhau giúp xây dựng các ứng dụng mà không cần viết lại mọi thứ từ đầu.

- Đa năng: Đây là ngôn ngữ có thể sử dụng cho nhiều mục đích khác nhau ví dụ như phát triển web, trí tuệ nhân tạo, trò chơi, khoa học dữ liệu, học máy, học sâu,...

- Hệ sinh thái: Python có một cộng đồng lớn cực kỳ đông đảo và năng động, nhiều diễn đàn hỗ trợ cho người mới là điều khiến ngôn ngữ này phát triển rất nhanh.

*2.2.3. Nhược điểm*

- Hiệu suất: Tốc độ chạy chương trình chậm hơn so với những ngôn ngữ phổ biến khác.

- Global Interpreter Lock (GIL): là một cơ chế trong Python ngăn nhiều luồng thực thi mã Python cùng một lúc. Điều này hạn chế tính song song và đồng thời của một số ứng dụng.

- Tiêu thụ bộ nhớ: Python có thể tiêu tốn rất nhiều bộ nhớ, đặc biệt khi làm việc với các tập dữ liệu lớn hoặc các thuật toán phức tạp. [4]

*2.2.4. Ứng dụng*

- Web Scraping: Python được sử dụng để thu thập dữ liệu từ các trang web dựa vào phân tích cú pháp HTML.

- Phát triển web: Các framework phổ biến như Django và Flask để phát triển các ứng dụng web.

- Tính toán khoa học (khoa học dữ liệu và máy học): SciPy là một thư viện mã nguồn mở dùng để tính toán khoa học. Numpy là thư viện toán học chuyên xử lý dữ liệu ma trận và mảng. Pandas là thư viện phân tích và mô hình hóa dữ liệu.

- Giao diện đồ họa người dùng (GUI): Python có thể dùng để phát triển giao diện thông qua sự hỗ trợ của các thư viện như Tkinter, PyQt, Kivy, wxWidgets.

- Quản lý cơ sở dữ liệu: Python cho phép truy cập các phần mềm quản lý cơ sở dữ liệu như MySQL, MongoDB, PostgreSQL.

- Giáo dục: Các câu lệnh khá đơn giản nên rất thích hợp để trở thành một ngôn ngữ lập trình cho người mới bắt đầu học. Python ngày càng được nhiều giáo viên giảng dạy ở các cấp bậc trung học, đại học.[4], [5]

2.3. Selenium

*2.3.1. Khái niệm*

Selenium là một automation testing framework miễn phí (mã nguồn mở).  
Nó được sử dụng để kiểm thử các ứng dụng web trên các trình duyệt (chrome, firefox, ms edge, …) và nền tảng khác nhau (Windows, Mac, Linux, …)*.*[6] Cho phép người dùng mô phỏng các hoạt động của người dùng trên trình duyệt như nhấp chuột, điền biểu mẫu và điều hướng giữa các trang web.

*2.3.2. Các thành phần của Selenium*

*Selenium IDE:* Là một công cụ ghi lại các thao tác trên trình duyệt của người dùng và cho phép tạo các kịch bản kiểm thử mà không cần lập trình.

*Selenium WebDriver:*  Là thành phần chính, cho phép điều khiển và tương tác trực tiếp với các thành phần của trình duyệt web một cách tự động.

*Selenium Grid:* Cho phép chạy các kịch bản kiểm thử song song trên nhiều máy tính và trình duyệt cùng một lúc, giúp tiết kiệm thời gian và tăng hiệu suất

*2.3.3. Ứng dụng*

Kiểm thử ứng dụng web: là công cụ phổ biến nhất để thực hiện kiểm thử tự động cho các ứng dụng web , nhằm đảm bảo chúng hoạt động đúng trên các trình duyệt và hệ điều hành khác nhau.

Tự động hóa tác vụ web: Tự động hóa các tác vụ lặp đi lặp lại trên web, như thu thập dữ liệu, gửi email tự động, …

Phát triển phần mềm: Sử dụng Selenium để kiểm tra các tính năng mới của ứng dụng trước khi phát hành, giúp phát hiện lỗi sớm trong quá trình phát triển.

*2.3.4. Các công trình công nghệ áp dụng Selenium*

"Web Scraping Using Python and Selenium to Build E-commerce Price Comparison System"

Tác giả: J. K. Sharma, A. Kumar

Tạp chí: International Journal of Innovative Research in Technology (2020)

Mô tả: Bài báo mô tả cách Selenium được sử dụng để thu thập dữ liệu từ các trang web thương mại điện tử nhằm xây dựng hệ thống so sánh giá trực tuyến.

<https://ieeexplore.ieee.org/abstract/document/9915892>

<https://www.browserstack.com/selenium>

<https://www.selenium.dev/>

<https://www.geeksforgeeks.org/browser-automation-using-selenium/>

<https://tech.cybozu.vn/tong-quan-ve-selenium-va-vai-tro-cua-cac-thanh-phan-74a12/#:~:text=Selenium%20l%C3%A0%20m%E1%BB%99t%20automation%20testing,Mac%2C%20Linux%2C%20%E2%80%A6>).

2.4. Giới thiệu về NoSQL

*2.4.1. Khái niệm về NoSQL*

NoSQL ( Not Only SQL) là một loại cơ sở dữ liệu không quan hệ , được thiết kế để xử lý và lưu trữ dữ liệu lớn theo cách linh hoạt hơn so với cơ sở dữ liệu quan hệ truyền thống (SQL) và được ứng dụng rộng rãi. NoSQL cho phép lưu trữ dữ liệu có cấu trúc , bán cấu trúc, phi cấu trúc mà không cần một schema cố định.

*2.4.2. Tính năng nổi bật của NoSQL*

Xử lý dữ liệu lớn: được thiết kế để xử lý các tập dữ liệu khổng lồ.

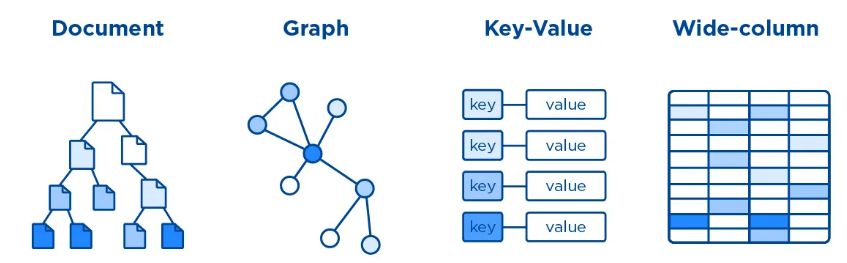
Khả năng mở rộng và phân phối: Hệ thống NoSQL có thể mở rộng dễ dàng, có khả năng chạy trên hàng nghìn máy chủ.

Truy vấn nhanh: được tối ưu hóa để trả về kết quả truy vấn một cách nhanh chóng, do không bị giới hạn bởi các mối quan hệ phức tạp như trong SQL.

Không cần cấu trúc(Schema-less): NoSQL cho phép lưu trữ dữ kiệu mà không yêu cầu cấu trúc cố định, mang lại sự linh hoạt trong việc quản lý dữ liệu.

Phát triển mã nguồn mở: nhiều giải pháp NoSQL được phát triển dưới dạng mã nguồn mở, cho phép cộng đồng tham gia vào việc cải tiến và phát triển công nghệ.

*2.4.3. Các loại dữ liệu NoSQL*



Hình 2: 4 loại cơ sở dữ liệu NoSQL[7]

Cơ sở dữ liệu document: Lưu trữ dữ liệu dưới dạng tài liệu (documents) , thường là JSON hoặc XML, cho phép linh hoạt trong cấu trúc dữ liệu. Mỗi tài liệu có thể chứa các cặp khóa-giá trị và hỗ trợ cấu trúc lồng nhau.

Cơ sở dữ liệu Wide-column: Lưu trữ dữ liệu theo cột , tối ưu hóa các truy vấn đề có tính toán trên các cột cụ thể.

Cơ sở dữ liệu Key-Value: Lưu trữ dưới dạng cặp khóa giá trị, mỗi khóa là duy nhất và phù hợp cho các ứng dụng cần truy cập nhanh

Cơ sở dữ liệu Graph: Dữ liệu được lưu trữ dưới dạng các nút và cạnh, tập trung vào việc lưu trữ và truy vấn dữ liệu lên quan đến mối quan hệ, lý tưởng cho các ứng dụng yêu cầu phân tích mối quan hệ phức tạp.

<https://docs.aws.amazon.com/whitepapers/latest/choosing-an-aws-nosql-database/types-of-nosql-databases.html>

<https://blazeclan.com/blog/dive-deep-types-nosql-databases/>

*2.4.4.Ứng dụng của NoSQL*

Mạng xã hội: Lưu trữ các bài đăng, bình luận và tương tác của người dùng trên các nền tảng như Facebook, Instagram.

Thương mại điện tử: Quản lý dữ liệu khách hàng, đơn hàng và tương tác người dùng trên các trang web như Amazon, eBay.

Phân tích dữ liệu lớn (Big Data): Thu thập và phân tích dữ liệu phi cấu trúc từ các nguồn khác nhau nhằm hỗ trợ các quyết định kinh doanh.

Phân tích thời gian thực: Theo dõi và phân tích dữ liệu theo thời gian thực từ cảm biến IoT hoặc ứng dụng trực tuyến.

2.5. MongoDB

*2.5.1.Khái niệm*

MongoDB là một hệ quản trị cơ sở dữ liệu (DBMS) thuộc loại NoSQL, được thiết kế để lưu trữ và truy xuất dữ liệu theo cách linh hoạt và không yêu cầu sự chuẩn bị cấu trúc cố định trước. Nó lưu trữ dữ liệu dưới dạng tài liệu (document) JSON[8]. Là môt hệ quản trị cơ sở dữ liệu phổ biến cho các ứng dụng web và các dự án phát triển nhanh, đặc biệt là khi có sự thay đổi thường xuyên trong cấu trúc dữ liệu.

*2.5.2. Đặc điểm quan trọng của MongoDB*

Các ad hoc query: MongoDB hỗ trợ các truy vấn linh hoạt, cho phép tìm kiếm theo trường dữ liệu, thực hiện các truy vấn thông thường, tìm kiếm theo biểu thức chính quy (regular expression), và truy vấn theo khoảng giá trị.

Indexing: Bất kỳ trường nào trong tài liệu BSON cũng có thể được tạo chỉ mục, giúp tăng tốc quá trình tìm kiếm và truy xuất dữ liệu.

Replication (Nhân bản): MongoDB hỗ trợ chức năng nhân bản, tạo ra một bản sao đồng nhất với phiên bản đang hoạt động. Điều này giúp bảo vệ dữ liệu khỏi mất mát và đảm bảo tính toàn vẹn của cơ sở dữ liệu trong trường hợp sự cố.

Aggregation: Các phép toán tập hợp trong MongoDB xử lý và trả về kết quả được tính toán. Chúng có thể nhóm các giá trị từ nhiều tài liệu lại với nhau và thực hiện nhiều phép toán đa dạng để trả về kết quả duy nhất. Điều này giống với GROUP BY và các hàm tổng hợp trong SQL.

Lưu trữ file: MongoDB có thể được sử dụng như một hệ thống lưu trữ file, tận dụng các chức năng của nó và hoạt động như một cách phân phối thông qua sharding, giúp quản lý và truy xuất dữ liệu lớn.

*2.5.3. Ưu điểm*

Dữ liệu lưu trữ phi cấu trúc, không có tính ràng buộc, toàn vẹn nên tính sẵn sàng cao, hiệu suất lớn và dễ dàng mở rộng lưu trữ.

Dữ liệu được caching (ghi đệm) lên RAM, hạn chế truy cập vào ổ cứng nên tốc độ đọc và ghi cao.

*2.5.4. Nhược điểm*

Không ứng dụng được cho các mô hình giao dịch nào có yêu cầu độ chính xác cao do không có ràng buộc.

Không có cơ chế transaction (giao dịch) để phục vụ các ứng dụng ngân hàng.

Dữ liệu lấy RAM làm trọng tâm hoạt động vì vậy khi hoạt động yêu cầu một bộ nhớ RAM lớn.

Mọi thay đổi về dữ liệu mặc định đều chưa được ghi xuống ổ cứng ngay lập tức vì vậy khả năng bị mất dữ liệu từ nguyên nhân mất điện đột xuất là rất cao.

*2.5.5. Ứng dụng của MongoDB*

Phát triển ứng dụng web: Nhiều công ty như Facebook, eBay, và Adobe sử dụng MongoDB để lưu trữ và quản lý lượng lớn dữ liệu người dùng.

Khoa học dữ liệu: MongoDB hỗ trợ phân tích dữ liệu lớn và có thể được sử dụng trong các ứng dụng học máy để lưu trữ và xử lý khối lượng lớn thông tin.

Hệ thống quản lý nội dung: MongoDB rất phù hợp cho việc quản lý nội dung động như blog hoặc bài viết trên trang web.

Ứng dụng IoT: Với khả năng xử lý dữ liệu thời gian thực, MongoDB là lựa chọn tốt cho các ứng dụng IoT cần lưu trữ và phân tích dữ liệu từ cảm biến*.*

2.6. Trình duyệt FireFox

*2.6.1. Giới thiệu trình duyệt*

*2.6.2. Vì sao chọn FireFox*

# CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM

## 3.1. Trình bày công việc thực nghiệm

### *3.1.1. Mục tiêu*

Truy cập trang cá nhân của một người nổi tiếng (có sức ảnh hưởng trên mạng xã hội) để thu thập dữ liệu các bài đăng, lượt tương tác, lượt xem,… Dựa vào đó để xác định được họ đăng những cái gì? Vấn đề họ thường bàn luận thuộc chủ đề gì? Và mọi người có quan tâm đến vấn đề đó hay không?

### *3.1.2. Quy trình thực nghiệm*

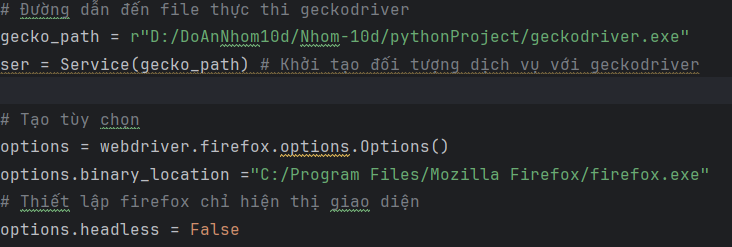
## 3.2. Kết quả thực nghiệm

### *3.2.1. Chi tiết quá trình cào dữ liệu*

- **Bước 1**: Kết nối đến cơ sở dữ liệu MongoDB



- **Bước 2**: Đường dẫn đến geckodriver



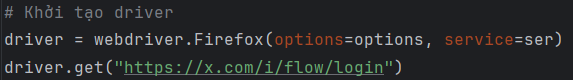
-- Khởi tạo đối tượng “Service” giúp tương tác với driver trong quá trình tự động hóa điều khiển trình duyệt.

-- Cấu hình trình duyệt: Khởi tạo đối tượng Options để thiết lập cấu hình của trình duyệt trước khi khởi động:

+ options.binary\_location: Chỉ định vị trí của file thực thi chương trình.

+ options.headless: Chạy chương trình ở chế độ hiển thị giao diện (False) giúp cho phép bạn quan sát quá trình hoạt động của chương trình, đặt (True) thì trình duyệt chạy ngầm giúp bạn tiết kiệm tài nguyên.

- **Bước 3**: Khởi tạo trình duyệt theo đối tượng cài đặt từ trước và truy cập trang:



- **Bước 4**: Đăng nhập tài khoản:

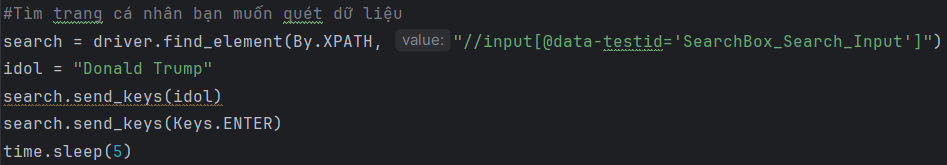
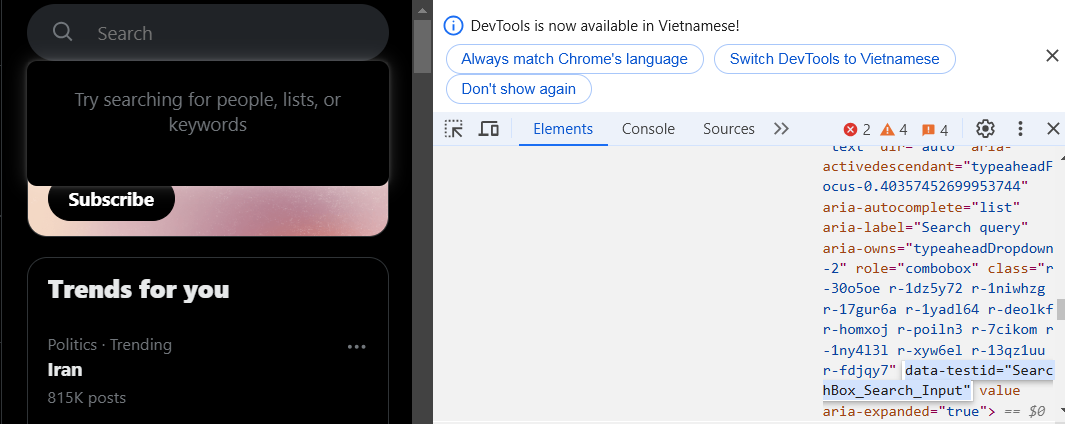


-- Tài khoản: Sử dụng thẻ XPATH để tìm và gửi thông tin đăng nhập.

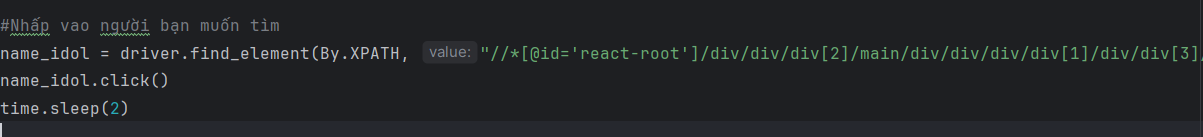
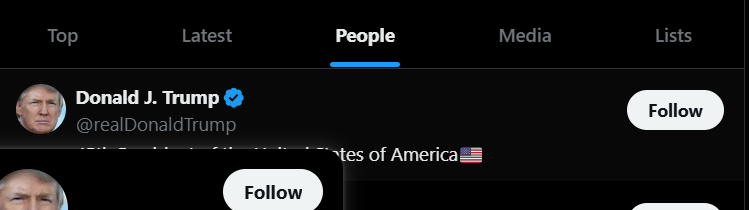
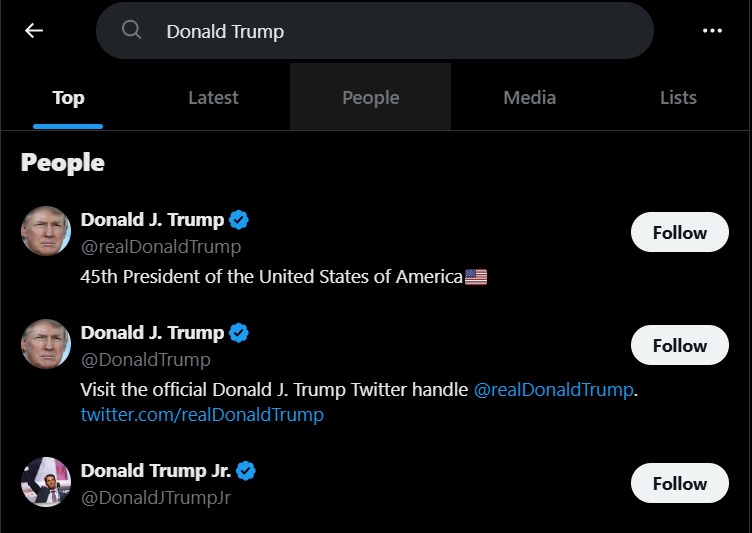
-- Xác nhận thông tin: Thỉnh thoảng, trang này sẽ hỏi xem gmail hoặc số điện thoại của bạn là gì để xác thực xem đó có phải là bạn không. Nhưng rất ít gặp, nếu không có thì bỏ qua.

-- Mật khẩu: Tương tự với bước nhập tài khoản.

- **Bước 5**: Tìm thanh tìm kiếm để nhập trang cá nhân bạn muốn truy cập:



- **Bước 6**: Nhấn vào tab People để tìm người:



-- Dựa vào thẻ XPATH để tìm và nhấn chuột để truy cập trang cá nhân.

- **Bước 7**:

### *3.2.2. Kết quả thực nghiệm*

Thu thập được file dữ liệu gồm….

#### 3.4 Đánh giá kết quả thực nghiệm

*3.4.1. So Sánh với mục tiêu ban đầu*

So sánh kết quả thực tế đạt được với mục tiêu ban đầu đã đề ra trong chương trình thực nghiệm

*3.4.2. Đánh giá hiệu suất và độ chính xác*

Đánh giá hiệu suất của chương trình cào dữ liệu (tốc độ, độ chính xác) và hiệu quả trong việc xử lý dữ liệu lớn.

**3.5 Hạn chế và hướng phát triển**

*3.5.1. Hạn chế của chương trình*

Nêu rõ những hạn chế còn tồn tại trong phần mềm (ví dụ: độ chính xác dữ liệu không cao, tốc độ cào còn chậm, cần cải thiện thêm tính năng).

*3.5.2. Hướng phát triển*

Đề xuất các hướng phát triển trong tương lai, như mở rộng phạm vi cào dữ liệu, tối ưu hiệu suất chương trình, cải thiện giao diện.

[1] “Introduction to Web Scraping,” GeeksforGeeks. Accessed: Oct. 19, 2024. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-web-scraping/

[2] “Is Web Scraping Legal?” Accessed: Oct. 20, 2024. [Online]. Available: https://oxylabs.io/blog/is-web-scraping-legal

[3] “Getting Started with Python Programming,” GeeksforGeeks. Accessed: Oct. 20, 2024. [Online]. Available: https://www.geeksforgeeks.org/getting-started-with-python-programming/

[4] “Python Language advantages and applications,” GeeksforGeeks. Accessed: Oct. 23, 2024. [Online]. Available: https://www.geeksforgeeks.org/python-language-advantages-applications/

[5] “Applications for Python,” Python.org. Accessed: Oct. 23, 2024. [Online]. Available: https://www.python.org/about/apps/

[6] “Tổng quan về Selenium và vai trò của các thành phần.” Accessed: Oct. 22, 2024. [Online]. Available: https://tech.cybozu.vn/tong-quan-ve-selenium-va-vai-tro-cua-cac-thanh-phan-74a12/

[7] Tiến C. L. V., “NoSQL là gì? Các thông tin về cơ sở dữ liệu NoSQL.” Accessed: Oct. 25, 2024. [Online]. Available: https://vietnix.vn/nosql-la-gi/

[8] “MONGODB LÀ GÌ? TÍNH NĂNG NỔI BẬT CỦA MONGODB MÀ BẠN CẦN BIẾT,” 200Lab Blog. Accessed: Oct. 26, 2024. [Online]. Available: https://200lab.io/blog/mongodb-la-gi/