

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

**-----------------------------------**

**BÁO CÁO ĐỒ ÁN**

**MÔN : ĐỒ ÁN CHUYÊN NGÀNH**

**Đề tài: TÌM HIỂU SELENIUM WEBDRIVER VÀ VIẾT ỨNG DỤNG CRAWLER**

**Giảng viên hướng dẫn: TS. Huỳnh Ngọc Tín**

**Sinh viên thực hiện:**

Văn Minh Nguyên 14520616

Đỗ Đình Chiểu 14520087

**TP. Hồ Chí Minh, tháng 12 năm 2018**

Mục lục

[**I.** **Phát biểu bài toán** 4](#_Toc518373617)

[**1.** **Khảo sát hiện trạng** 4](#_Toc518373618)

[**1.1** **Nhu cầu thực tế** 4](#_Toc518373619)

[**1.2** **Hiện trạng đơn vị được khảo sát** 4](#_Toc518373620)

[**1.3** **Các phần mềm cùng chức năng có trên thị trường** 6](#_Toc518373621)

[**1.4** **Điểm hạn chế và vấn đề còn tồn tại của hệ thống**. 6](#_Toc518373622)

[**2.** **Danh sách các yêu cầu** 7](#_Toc518373623)

[a. **Biểu mẫu hóa đơn mua hàng**. 7](#_Toc518373624)

[**b.** **Biểu mẫu phiếu đặt hàng** 7](#_Toc518373625)

[**c.** **Biểu mẫu yêu cầu nhập hàng** 8](#_Toc518373626)

[**II.** **Mô hình use-case** 9](#_Toc518373627)

[**1.** **Sơ đồ use-case** 9](#_Toc518373628)

[**a.** **Use-case tổng quát** 9](#_Toc518373629)

[**b.** **Use-case quản trị viên** 9](#_Toc518373630)

[**c.** **Use-case khách hàng** 10](#_Toc518373631)

[**2.** **Danh sách các actor** 10](#_Toc518373632)

[**3.** **Danh sách các use-case** 10](#_Toc518373633)

[**4.** **Đặc tả use-case** 11](#_Toc518373634)

[**a.** **Use-case đăng nhập** 11](#_Toc518373635)

[**b.** **Use-case đăng ký** 11](#_Toc518373636)

[**c.** **Use-case tìm kiếm** 12](#_Toc518373637)

[**d.** **Use-case mua hàng** 12](#_Toc518373638)

[**e.** **Use-case cập nhập đơn hàng** 13](#_Toc518373639)

[**f.** **Use-case thêm sản phẩm** 14](#_Toc518373640)

[**g.** **Use-case sửa sản phẩm** 14](#_Toc518373641)

[**h.** **Use-case xóa sản phẩm** 15](#_Toc518373642)

[**i.** **Use-case thêm loại sản phẩm** 15](#_Toc518373643)

[**j.** **Use-case sửa loại sản phẩm** 16](#_Toc518373644)

[**k.** **Use-case xóa loại sản phẩm** 16](#_Toc518373645)

[**l.** **Use-case thêm khuyến mãi** 17](#_Toc518373646)

[**m.** **Use-case sửa khuyến mãi** 17](#_Toc518373647)

[**n.** **Use-case xóa khuyến mãi** 18](#_Toc518373648)

[**III.** **Phân tích** 19](#_Toc518373649)

[**1.** **Sơ đồ lớp** 19](#_Toc518373650)

[**a.** **Sơ đồ lớp** 19](#_Toc518373651)

[**b.** **Danh sách các đối tượng** 19](#_Toc518373652)

[**c.** **Mô tả chi tiết các đối tượng** 20](#_Toc518373653)

[**IV.** **Thiết kế dữ liệu** 23](#_Toc518373654)

[**1.** **Sơ đồ logic** 23](#_Toc518373655)

[**2.** **Mô tả chi tiết các thành phần trong sơ đồ logic** 26](#_Toc518373656)

[**V.** **Thiết kế kiến trúc:** 30](#_Toc518373657)

[**1.** **Kiến trúc hệ thống:** 30](#_Toc518373658)

[**2.** **Mô tả chi tiết thành phân bên trong:** 31](#_Toc518373659)

[**VI.** **Thiết kế giao diện** 31](#_Toc518373660)

[**1.** **Danh sách cá màn hình:** 31](#_Toc518373661)

[**2.** **Mô tả chi tiết các màn hình:** 32](#_Toc518373662)

[**VII.** **Kết luận** 62](#_Toc518373663)

[**1.** **Môi trường phát triển:** 62](#_Toc518373664)

[**2.** **Kết quả đạt được:** 62](#_Toc518373665)

[**3.** **Hướng phát triển:** 62](#_Toc518373666)

1. **Tìm hiểu về Selenium webdriver**
2. **Mô tả bài toán**
3. **Tầm quan trọng của việc thu thập dữ liệu**

Mỗi ngày con người gửi vào internet hàng tỉ dữ liệu. Chúng ta không hề biết rằng một lượt like của chúng ta trên Facebook cũng là một dữ liệu vô cùng quan trọng trong bài toán phân tích cảm xúc hay khi chúng ta tag tên người vào một bức ảnh cũng là một hành động cung cấp dữ liệu cho bài toán nhận diện khuôn mặt, vv… . Mỗi dữ liệu con người gửi vào internet dù ít hay nhiều đều có giá trị cho từng bài toán khác nhau. Hiện nay việc khai thác dữ liệu từ các trang mạng xã hội như Facebook, Instagram, twitter, vv… để phân tích thị trường, phục vụ y học, nghiên cứu tâm lý ngày càng phổ biến. Yêu cầu đầu tiên cho bài toán phân tích dữ liệu đó là thu thập dữ liệu. Có rất nhiều nguồn dữ liệu để thu thập, trong phạm vi đồ án xin phép được đề cập hai nguồn chính là Youtube và Facebook.

1. **Phân tích dữ liệu Youtube**

YouTube là một trang web chia sẻ video, là nơi người dùng có thể tải lên hoặc tải về máy tính hay điện thoại và chia sẻ các video clip. YouTube do 3 nhân viên cũ của PayPal tạo nên vào giữa tháng 2 năm 2005.

Dữ liệu từ Youtube bao gồm: Video, các bình luận, các tài liệu người dùng, các lượt yêu thích, vv.

Để sử dụng Youtube người dùng có thể tạo hoặc không tạo một tài khoản. Nếu người dùng muốn truy cập vào Youtube chỉ để xem thì không nhất thiết phải có một tài khoản Youtube tuy nhiên nếu muốn đăng tải video, bình luận, hay nhận thông báo từ Youtube thì người dùng cần phải có một tài khoản Youtube.

Dưới đây là bản mô tả các thuật ngữ Youtube.

|  |  |  |
| --- | --- | --- |
| STT | Thuật ngữ | Mô tả |
| 1 | Title | Là tiêu đề của một video được người dùng đặt lúc đăng tải hoặc live stream. |
| 2 | Channel | Mỗi người dùng được đăng ký một kênh gọi là channel. Channel hiển thị cùng với video ở phần title. |
| 3 | Like | Khi nội video phù hợp với người dùng và người dùng thấy nó hay người dùng có thể chọn like. |
| 4 | Dislike | Một dislike tương đương với một lượt bình chọn không thích nội dung video. |
| 5 | Description | Nội dung mô tả video được người dùng them vào khi đăng tải video hoặc live stream. |
| 6 | Comment | Phần bình luận văn bản của người dùng, thể hiện quan điểm của người dùng về video. |

1. **Phân tích dữ liệu Facebook**

Facebook là một website dịch vụ mạng xã hội và truyền thông xã hội do công ty Facebook, Inc điều hành với trụ sở tại Menlo Park, California.

Người tham gia Facebook cần đăng ký một tài khoản để liên kết vào mạng.

Dữ liệu từ Facebook bao gồm: các bài post, các profile, các video, trạng thái, bình luận, cảm xúc, vv…

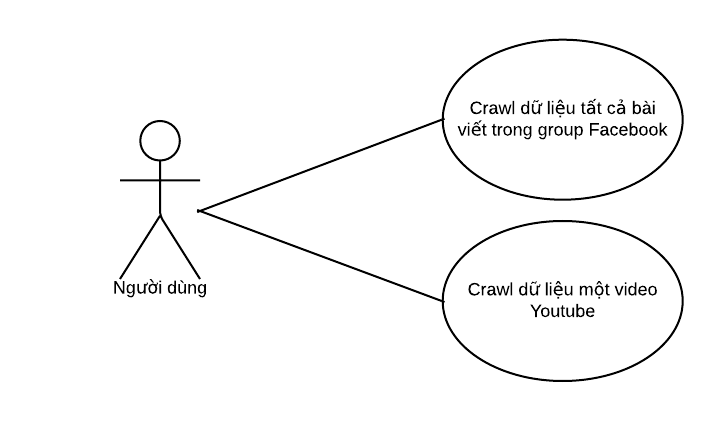
Dưới đây là bảng mô tả chi tiết các thành phần một bài post Facebook:

|  |  |  |
| --- | --- | --- |
| STT | Thuật ngữ | Mô tả |
| 1 | User name | Người dùng Facebook thường được mô tả ở phần đầu bài viết với một ảnh đại diện và một tên người dùng. |
| 2 | Caption | Chú thích kèm them để làm rỏ trạng thái. |
| 3 | Status (content) | Là phần nội dung bài viết hay dòng tâm trạng. |
| 4 | Like | Facebook chia độ quan tâm bài viết thành nhiều loại mức độ như: yêu thích, thích, ngạc nhiên, giận dữ , vv . |
| 5 | Comment | Phần bình luận văn bản của người dùng, thể hiện quan điểm của người dùng về bài viết. |

1. **Phát biểu bài toán**

Hiện nay các chương trình truyền hình, game show ngày càng nhiều với đa dạng các thể loại từ tìm kiếm tài năng, hẹn hò, hài kịch, đến kiến thức xã hội, giao lưu người nỗi tiếng, vv… Số lượng lớn cần phải đi kèm chất lượng tuy nhiên không phải game show nào cũng đạt được điều này. Vì vậy một bài toán đặt ra là làm sao phân tích được mức độ quan tâm của mọi người đối với từng chương trình truyền hình. Để làm được điều nay bước đầu tiên là phải đảm bảo thu thập một lượng dữ liệu đủ lớn về các chương trình truyền hình kể trên. Phạm vi bài toán gói gọn trong việc tạo một ứng dụng cho phép crawl dữ liệu từ Youtube và Facebook về một máy chủ MongoDB lưu trữ để làm đầu vào cho bước tiếp theo là phân tích dữ liệu.

1. **Phân tích**
2. **Lượt đồ usecase**

****

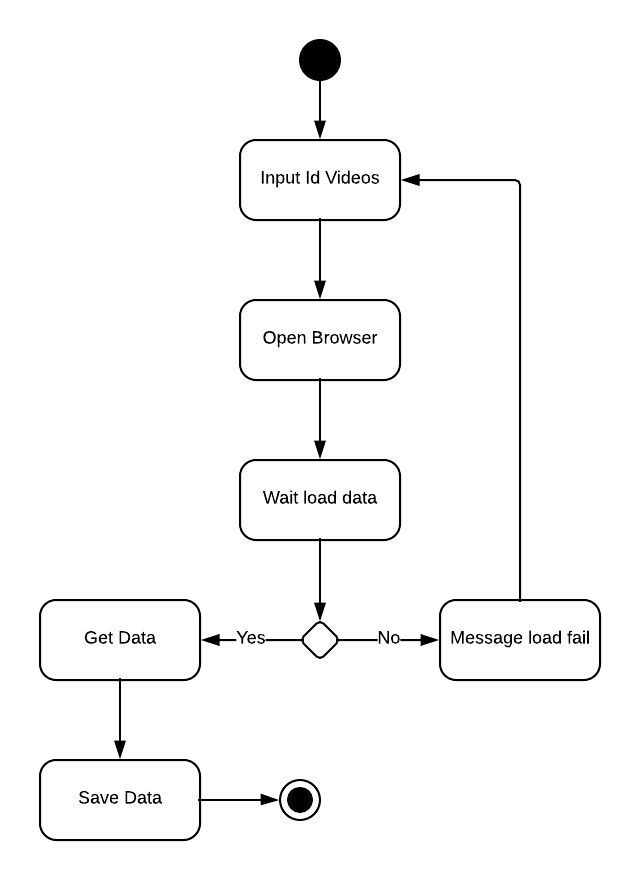
1. **Đặc tả usecase**
   1. **Crawl dữ liệu một video Youtube**

|  |  |
| --- | --- |
| **Tên usecase** | Crawl dữ liệu một video Youtube. |
| **Mô tả usecase** | Cho phép người dùng nhập vào id video và crawl dữ liệu của video đó lưu trên MongoDB server. |
| **Điều kiện bắt đầu** | * Id video tồn tại. * Trình duyệt Chrome hoạt động trên máy client. |
| **Luồn thực thi** | F1: Người dùng nhập vào id video  F2: Bật trình duyệt  F3: tải trang chứ video  F4: Tìm kiếm tất cả các thành phần cần lấy dữ liệu  F5: Lấy toàn bộ dữ liệu về  F6: Lưu vào MongoDB |
| **Ngoại lệ** | E1: Load trang timeout  E2: Không tìm thấy thành phần cần crawl. |
| **Sau thực thi** | Trả về một đối tượng Json chứa thông tin vửa crawl. |

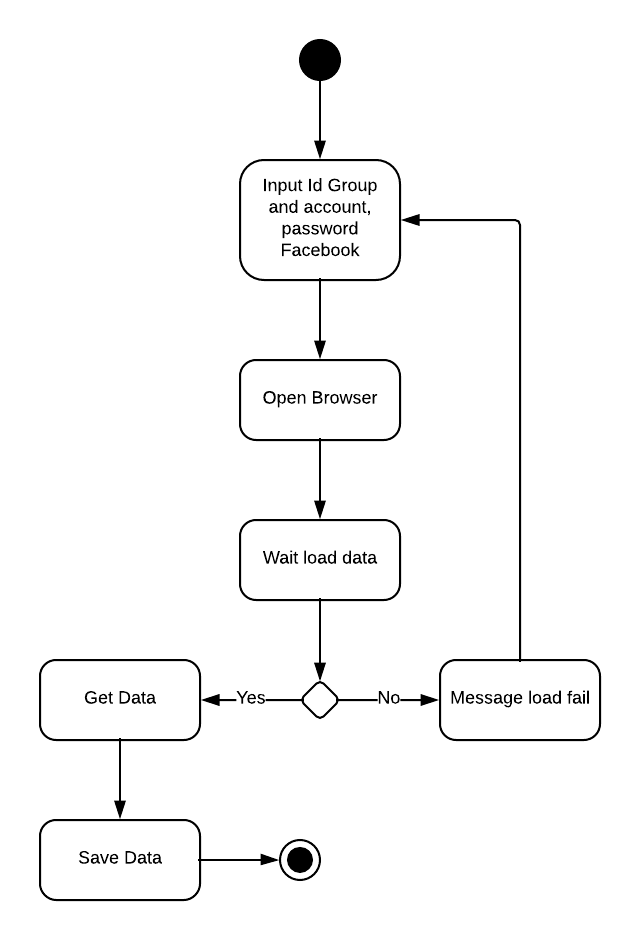
* 1. **Crawl dữ liệu tất cả các bài viết trong group Facebook**

|  |  |
| --- | --- |
| **Tên usecase** | Crawl dữ liệu tất cả các bài viết trong group Facebook. |
| **Mô tả usecase** | Cho phép người dùng nhập vào group id và crawl dữ liệu của tất cả các bài viết có trong group Facebook đó lưu trên MongoDB server. |
| **Điều kiện bắt đầu** | * Group Id tồn tại. * Trình duyệt Chrome hoạt động trên máy client. |
| **Luồn thực thi** | F1: Người dùng nhập vào group id.  F2: Bật trình duyệt và đăng nhập Facebook  F3: Tải trang  F4: Tìm kiếm tất cả các thành phần cần lấy dữ liệu  F5: Lấy toàn bộ dữ liệu về  F6: Lưu vào MongoDB |
| **Ngoại lệ** | E1: Load trang timeout  E2: Không tìm thấy thành phần cần crawl.  E3: Tài khoản facebook không hợp lệ. |
| **Sau thực thi** | Trả về một đối tượng Json chứa thông tin vửa crawl. |

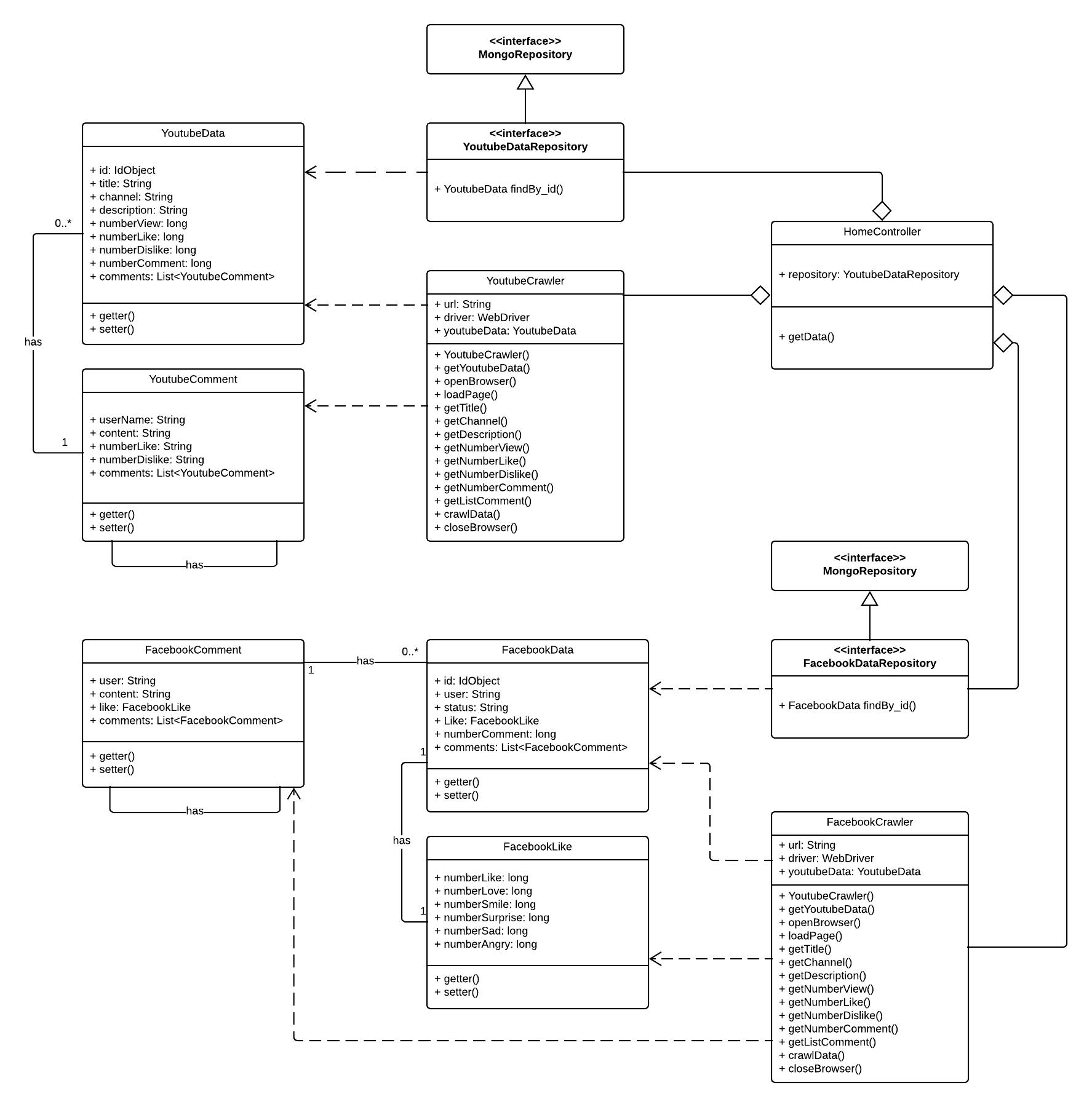
1. **Lượt đồ activity**
   1. **Crawl dữ liệu một video Youtube**

****

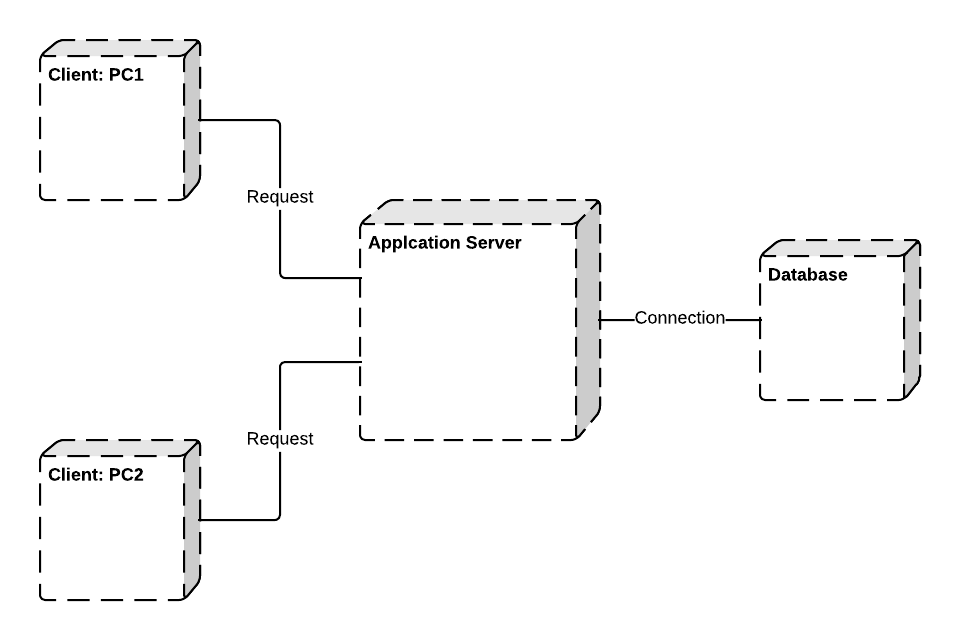
* 1. **Crawl dữ liệu tất cả các bài viết trong group Facebook**

****

1. **Lượt đồ lớp**

****

1. **Lượt đồ triển khai**

****

1. **Thiết kế dữ liệu**
   1. **YoutubeData**

Lớp lưu trữ dữ liệu của một video Youtube với các thành phần được mô tả như bên dưới:

|  |  |
| --- | --- |
| **Thuộc tính** | **Kiểu dữ liệu** |
| id | IdObject |
| title | String |
| channel | String |
| description | String |
| numberView | Long |
| numberLike | Long |
| numberDislike | Long |
| numberComment | Long |
| comments | List<YoutubeComment> |

* 1. **FacebookData**

Lớp lưu trữ dữ liệu của một bài post trên Facebook với các thành phần được mô tả như bên dưới:

|  |  |
| --- | --- |
| **Thuộc tính** | **Kiểu dữ liệu** |
| id | IdObject |
| user | String |
| status | String |
| like | FacebookLike |
| numberComment | Long |
| comments | List<FacebookComment> |

* 1. **YoutubeComment**

Lớp lưu trữ dữ liệu của một comment trên Youtube.

|  |  |
| --- | --- |
| **Thuộc tính** | **Kiểu dữ liệu** |
| userName | String |
| content | String |
| numberLike | long |
| numberDislike | long |
| comments | List<YoutubeComment> |

* 1. **FacebookComment**

Lớp lưu trữ dữ liệu của một comment trên Facebook.

|  |  |
| --- | --- |
| **Thuộc tính** | **Kiểu dữ liệu** |
| userName | String |
| content | String |
| like | FacebookLike |
| comments | List<FacebookComment> |

* 1. **FacebookLike**

Lớp lưu trữ dữ liệu của Like của một bài viết hoặc một comment Facebook.

|  |  |
| --- | --- |
| **Thuộc tính** | **Kiểu dữ liệu** |
| numberLike | Long |
| numberLove | Long |
| numberSmile | Long |
| numberSurprise | Long |
| numberSad | Long |
| numberAngry | Long |

1. **Mô tả chi tiết thành phần**

|  |  |  |
| --- | --- | --- |
| **Lớp đối tượng** | **Thuộc tính** | **Mô tả** |
| **YoutubeData** | id | Mã dùng để lưu trữ thông tin trên mongoDB |
| title | Tiêu đề video |
| channel | Kênh phát video |
| description | Mô tả video |
| numberView | Số lượt view video |
| numberLike | Số lượt yêu thích |
| numberDislike | Số lượt không thích |
| numberComment | Số lượt comment |
| comments | Chứ danh sách các comment về video. |
| **FacebookData** | id | Mã dùng để lưu trữ thông tin trên mongoDB |
| user | Tên người đăng bài |
| status | Nội dung bài đăng |
| like | Thông tin về các lượt like bài viết |
| numberComment | Số lượt comment |
| comments | Danh sách các comment |
| **YoutubeComment** | userName | Tên người dùng comment |
| content | Nội dung comment |
| numberLike | Số lượng like |
| numberDislike | Số lượng dislike |
| comments | Danh sách các reply |
| **FacebookComment** | userName | Tên người dùng comment |
| content | Nội dung comment |
| like | Thông tin chi tiết về số lượt thích |
| comments | Danh sách các reply |
| **FacebookLike** | numberLike | Số lượt thích |
| numberLove | Số lượt yêu thích |
| numberSmile | Số lượt haha |
| numberSurprise | Số lượt ngạc nhiên |
| numberSad | Số lượt buồn |
| numberAngry | Số lượt phẩn nộ |

1. **Thiết kế ứng dụng**
2. **Thiết kế kiến trúc phần mềm**

**Kiến trúc Rest API:** REST là viết tắt của REpresentational State Transfer. REST là tập hợp các hướng dẫn và kiến trúc sử dụng cho việc truyền dữ liệu. REST áp dụng phổ biến cho các Web app, nhưng cũng hoàn toàn có thể sử dụng cho các phần mềm nói chung.**Model:** Đây là thành phần chứa tất cả các nghiệp vụ logic, phương thức xử lý, truy xuất database, đối tượng mô tả dữ liệu như các Class, hàm xử lý…

**Kiến trúc REST có 1 số đặc điểm như sau**

* Nhất quán xuyên suốt các API. Ví dụ: luôn luôn sử dụng danh từ số nhiều thay vì khi số nhiều, khi số ít
* Tồn tại mà không lưu trạng thái (Stateless existence), ví dụ như không sử dụng session của server. Đây chính là lý do vì sao lại có cụm State Transfer (ST) trong REST. Tính chất này đảm bảo rằng mỗi khi client submit 1 request, request ấy sẽ bao gồm toàn bộ các thông tin mà server cần để xử lý request. Đây là điều cần thiết để build 1 ứng dụng phục vụ lên tới hàng triệu user.
* Sử dụng HTTP status code khi có thể
* Sử dụng URL Endpoint có phân tầng logic
* Đánh version trong URL thay vì trong HTTP Headers

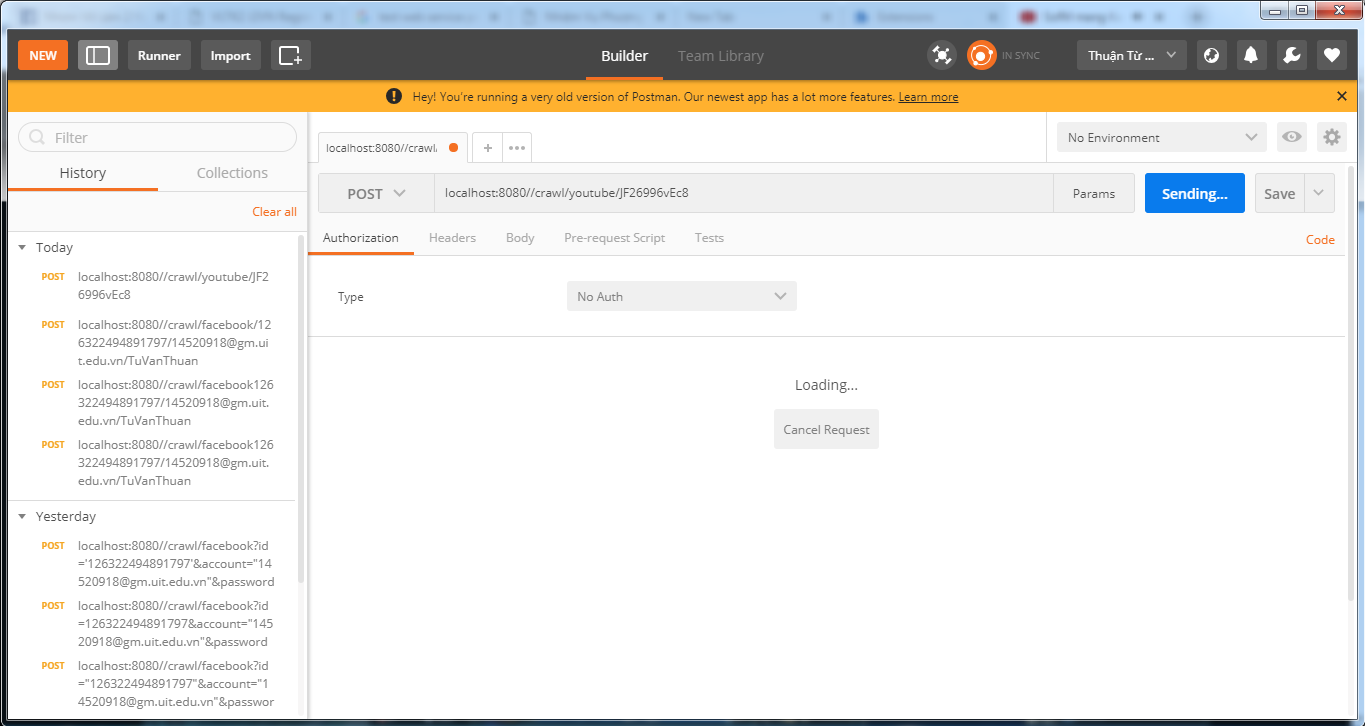


1. **Triển khai ứng dụng**

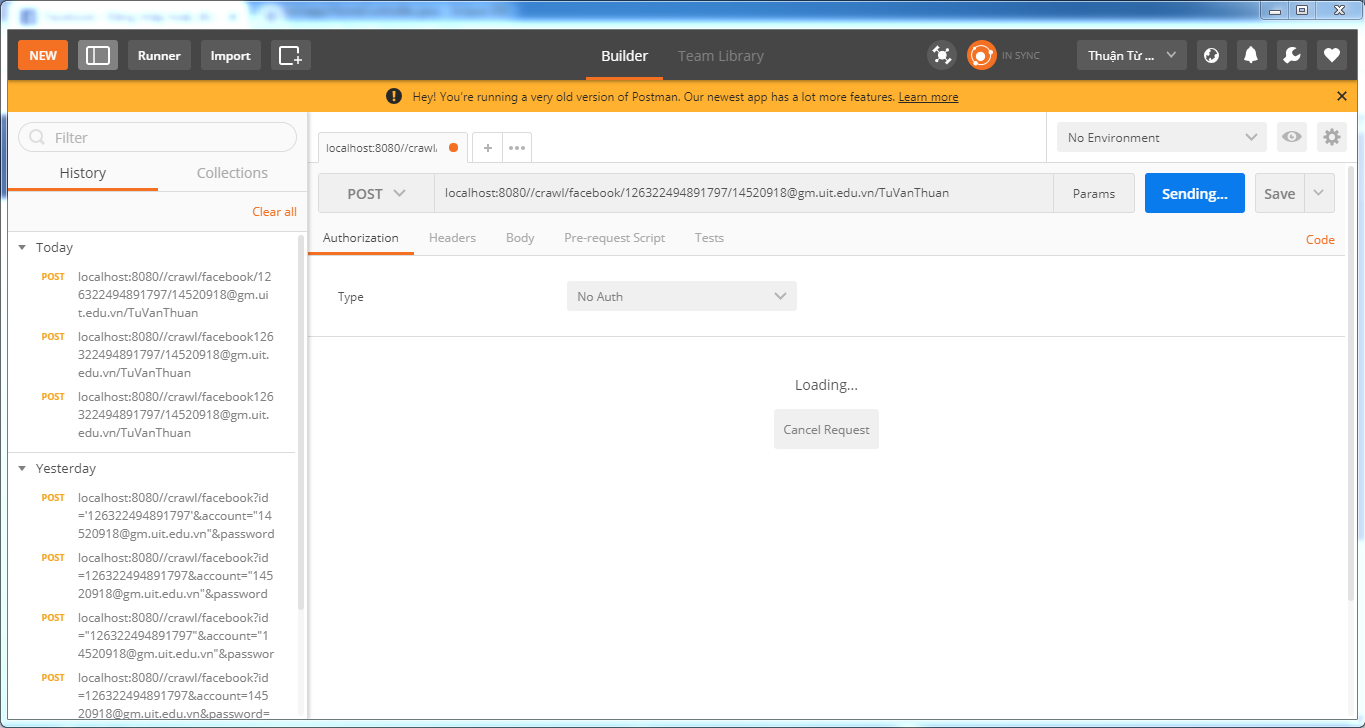
Ứng dụng được triển khai dưới dạng web service người dùng có thể truy cập theo phương thức POST để sữ dụng các chức năng cawl như sau:

Truy cập vào ứng dụng Postman và thực hiện như bên dưới.

* **Sử dụng crawl dữ liệu một video Youtube**



* **Sử dụng crawl dữ liệu một group Facebook**



1. **Kết luận**
2. **Môi trường phát triển:**

* Ứng dụng là một web service sử dụng cấu trúc rest trên nền tảng Java 8.
* Ứng dụng sử dụng Selenium WebDriver

1. **Kết quả đạt được:**

* Hiểu được cách thức sữ dụng các phương thức mà selenium hổ trợ.
* Nắm bắt được cách triển khai ứng dụng web service với Spring Boot.
* Triển khai được một web service cho phép thu thập dữ liệu từ Facebook và Youtube mặc dù chưa được đầy đủ các chức năng cần có.

1. **Khó khăn:**

* Selenium phụ thuộc hoàn toàn vào cấu trúc web nên việc thu thập dữ liệu từ Youtube là dễ tuy nhiên Facebook thì rất khó chạm đến các thành phần dữ liệu.
* Nếu cấu trúc website thay đổi thì yêu cầu người phát triển phái update source mới để có thể hoạt động.
* Facebook chia ra làm nhiều thành phần như người dùng đơn, nhóm, trang, vv … mỗi loại như vậy có một cấu trúc web khác nhau nên khó tận dụng lại mã nguồn.

1. **Hướng phát triển:**

Tiếp tục hoàn thiện các chức năng để có thể lấy được đầy đủ dữ liệu từ hai nguồn dữ liệu trên. Tiến hành phân tích dữ liệu, học máy và xữ lý ngôn ngữ tự nhiên để gia tang độ chính xác kết quả. Hướng đến ứng dụng phân tích dữ liệu lớn.