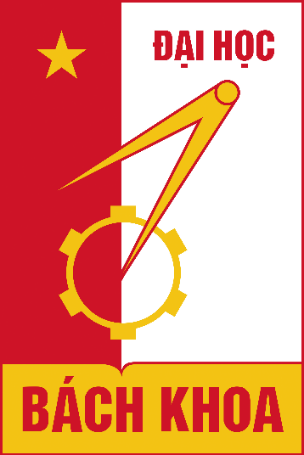
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

**VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

======&&======



**BÀI TẬP LỚN**

Xử lý ngôn ngữ tự nhiên

Đề tài: Phân tích cú pháp theo tiếp cận thống kê

Giáo viên hướng dẫn: PGS.TS Nguyễn Thị Thanh Hương

Thành viên nhóm 06:

|  |  |
| --- | --- |
| Nguyễn Văn Hoàn  Nguyễn Duy Đoàn  Nguyễn Đình Khánh  Đỗ Ngọc Tuyên | 20167973  20209546  20162145  20164442 |

**Hà Nội, tháng 12 năm 2020**

Nội dung

[I. Giới thiệu 2](#_Toc42808671)

[1. Đặt vấn đề 2](#_Toc42808672)

[2. Mô tả ứng dụng 2](#_Toc42808673)

[II. Các công cụ sử dụng 3](#_Toc42808674)

[1. Scrapy 3](#_Toc42808675)

[2. Elasticsearch 4](#_Toc42808676)

[III. Xây dựng hệ thống 5](#_Toc42808677)

[1. Thu thập dữ liệu 5](#_Toc42808678)

[2. Xử lý dữ liệu 5](#_Toc42808679)

[3. Các tính năng của hệ thống và các kỹ thuật được sử dụng tương ứng 6](#_Toc42808680)

[3.1 Tìm kiếm dựa trên title của cuốn sách, lọc theo tên tác giả 6](#_Toc42808681)

[3.2 Sắp xếp kết quả trả về 6](#_Toc42808682)

[3.3 Tự động gợi ý kết quả tìm kiếm (auto complete) 6](#_Toc42808683)

[4. Giao diện hệ thống 7](#_Toc42808684)

[5. Hướng phát triển hệ thống 9](#_Toc42808685)

[IV. Kết luận 10](#_Toc42808686)

[V. Tài liệu tham khảo 10](#_Toc42808687)

# Giới thiệu

## Đặt vấn đề

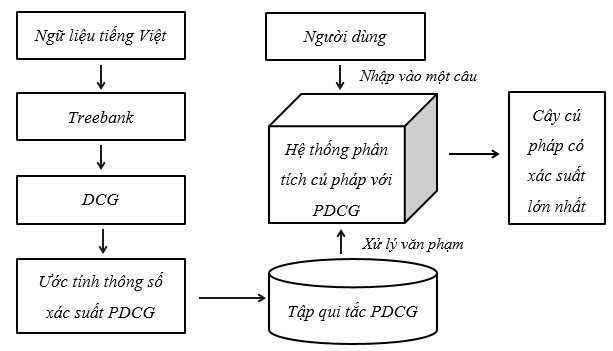
Máy tính rất tuyệt khi làm việc với các dữ liệu có cấu trúc như bảng tính và bảng cơ sở dữ liệu. Nhưng con người chúng ta lại thường giao tiếp bằng lời nói chứ không phải trong bảng. Điều đó thật đáng tiếc cho máy tính. *Thật không may chúng ta không sống trong phiên bản lịch sử này, nơi mà việc giao tiếp được thực hiện bằng bảng biểu, tức là dữ liệu có cấu trúc!*

Rất nhiều thông tin trên thế giới không có cấu trúc - văn bản thô bằng tiếng Anh, hoặc các ngôn ngữ khác trên thế giới. Làm thế nào chúng ta có thể làm cho máy tính có thể hiểu được văn bản phi cấu trúc và trích xuất dữ liệu từ chúng?

Xử lý ngôn ngữ tư nhiên hay còn được gọi là NLP là một lĩnh vực con của AI, trong đó, nó sẽ tập trung vào việc cho phép máy tính hiểu và xử lý được ngôn ngữ của con người. Xử Lý Ngôn Ngữ Tự Nhiên có vai trò hết sức quan trọng trong ngành Khoa Học Máy Tính. Nó có vô vàn ứng dụng hữu ích trong cuộc sống cũng như nghiên cứu. Chúng ta có thể điểm qua một vài ứng dụng của xử lý ngôn ngữ tự nhiên như:

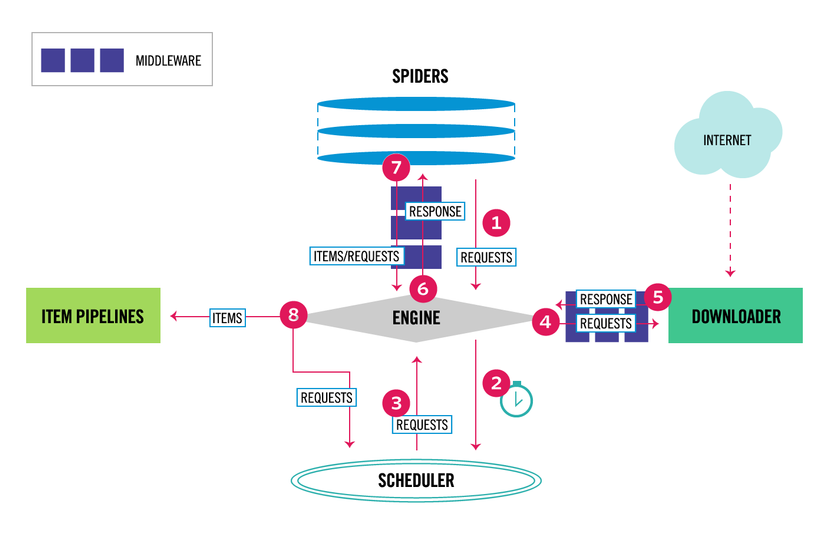
* Nhận dạng chữ viết
* Nhận dạng tiếng nói
* Tổng hợp tiếng nói
* Dịch tự động (Machine translate)
* Tìm kiếm thông tin (Information retrieval)
* Tóm tắt văn bản
* Khai phá dữ liệu (Data mining) và phát hiện tri thức

## Mô tả đề tài



# Các công cụ sử dụng

## Scrapy



 Scrapy là một framework được viết bằng Python, nó cấp sẵn 1 cấu trúc tương đối hoàn chỉnh để thực hiện việc thu thập thông tin và trích xuất dữ liệu từ trang web một cách nhanh chóng và dễ dàng.

Scrapy cho phép thu thập thông tin trang web đồng thời mà không phải đối mặt với vấn đề về luồng, tiến trình, đồng bộ hóa,… Nó xử lý các yêu cầu không đồng bộ một cách nhanh chóng. Scrapy đã phát triển được nhiều năm dựa trên kinh nghiệm của những người thực hiện thu thập thông tin web trên quy mô lớn, vì vậy nó giải quyết được rất nhiều thách thức mà các nhà phát triển đang phải đối mặt hàng ngày như là:

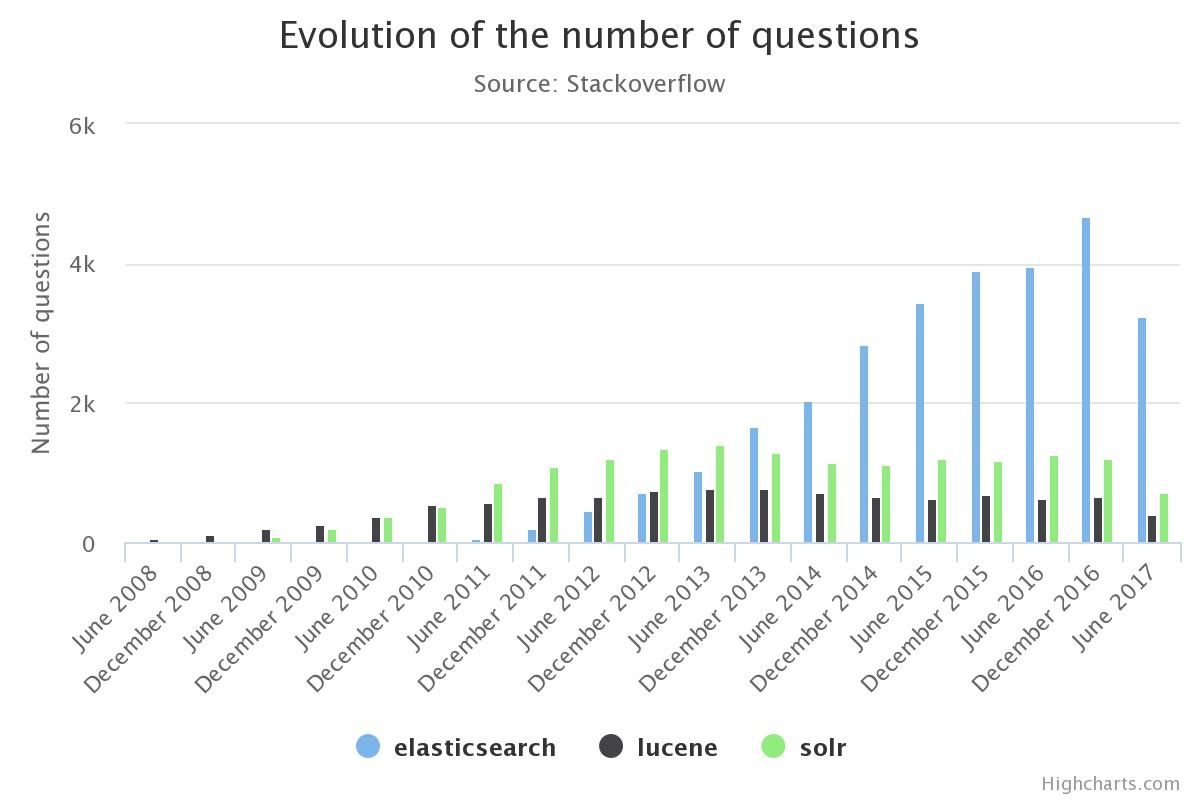
* Cung cấp cơ chế auto-throttling tự động điều chỉnh tốc độ thu thập dữ liệu dựa trên cả máy chủ web và máy tính người dùng.
* Tự động giữ lại các phiên làm việc. Nó xử lý cookies, đi qua nó một cách dễ dàng thông qua các yêu cầu (request). Xác thực cũng không phải là trở ngại ngay cả khi mẫu đăng nhập có CSRF token
* Nó có thể tránh các bẫy đổi hướng.
* Lọc các yêu cầu trùng lặp và cho phép tùy chỉnh hành vi lọc.

Do hộ trợ rất nhiều tính năng như kể trên nhóm em quyết định sử dụng scrapy để crawl dữ liệu sách phục vụ cho hệ thống tìm kiếm.

## Elasticsearch

Elasticsearch là một công cụ tìm kiếm dựa trên phần mềm lucene. Nó cung cấp một bộ máy tìm kiếm dạng phân tán, có đầy đủ công cụ với một giao diện web HTTP có hỗ trợ dữ liệu JSON. Elasticsearch được phát triển bằng Java và được phát hành dạng nguồn mở theo giấy phép Apache. ( wikipedia).

Elasticsearch là một server riêng biệt, không liên quan đến server web gốc và giao tiếp thông qua Restfull nên rất tiện dụng và nhanh chóng. Nên dễ dàng tích hợp vào hệ thống, bạn chỉ cần gửi request http lên là nó trả về kết quả. Ngoài ra nó là hệ thống phân tán hộ trợ việc mở rộng tìm kiếm dữ liệu từ nhiều nơi. Do đó nó là một trong những search engine phổ biển nhất hiện nay



# Xây dựng hệ thống

## Thu thập dữ liệu

#### 1.1 Nguồn gốc dữ liệu

Dữ liệu được thu thập từ trang vinabook.com

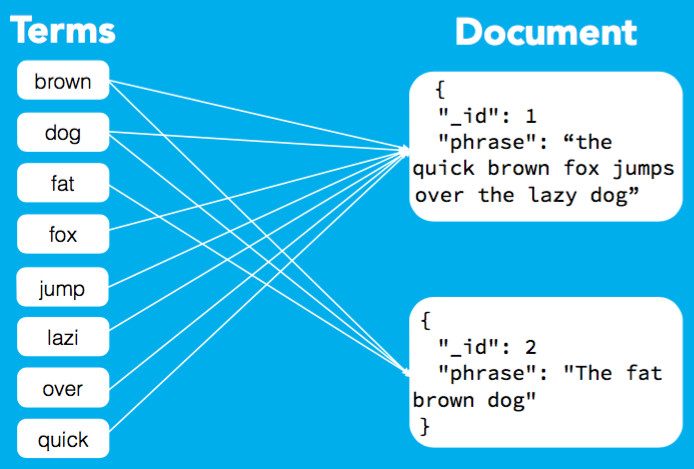
#### 1.2 Cấu trúc dữ liệu

Dữ liệu được lưu dưới dạng file book.csv với các thuộc tính của cuốn sách như dưới :

1. Author : tên tác giả
2. Category : thể loại
3. datePublished : ngày xuất bản
4. description : mô tả cuốn sách
5. image\_urls : link ảnh cuốn sách
6. images : link ảnh full
7. numberOfPages : số trang của cuốn sách
8. price : giá cuốn sách
9. sub\_category : danh mục phụ
10. title : tên cuốn sách

## Xử lý dữ liệu

Sử dụng plugin phân tích từ Tiếng Việt để tích hợp vào trong elasticsearch với khả năng tách từ tiếng Việt (tokenizer) và khả năng loại bỏ từ dừng.



Sử dụng các bộ Analysis với khả năng nhận dữ liệu vào và trả về các token, hỗ trợ cả tiếng Việt. Nó giúp việc đánh index dữ liệu trở lên đơn giản, nhanh chóng và hiệu quả.

## Các tính năng của hệ thống và các kỹ thuật được sử dụng tương ứng

### Tìm kiếm dựa trên title của cuốn sách, lọc theo tên tác giả

* Đánh index title cuốn sách và tên tác giả,...
* Sử dụng dụng tính năng edge\_ngram trong elasticsearch để thực hiện auto complete, một số bộ analysis như vi\_analyzer để tokenize tiếng Việt, icu\_standard để chia câu thành từng từ,...

### Sắp xếp kết quả trả về

Sử dụng config sort với kết quả trả về.

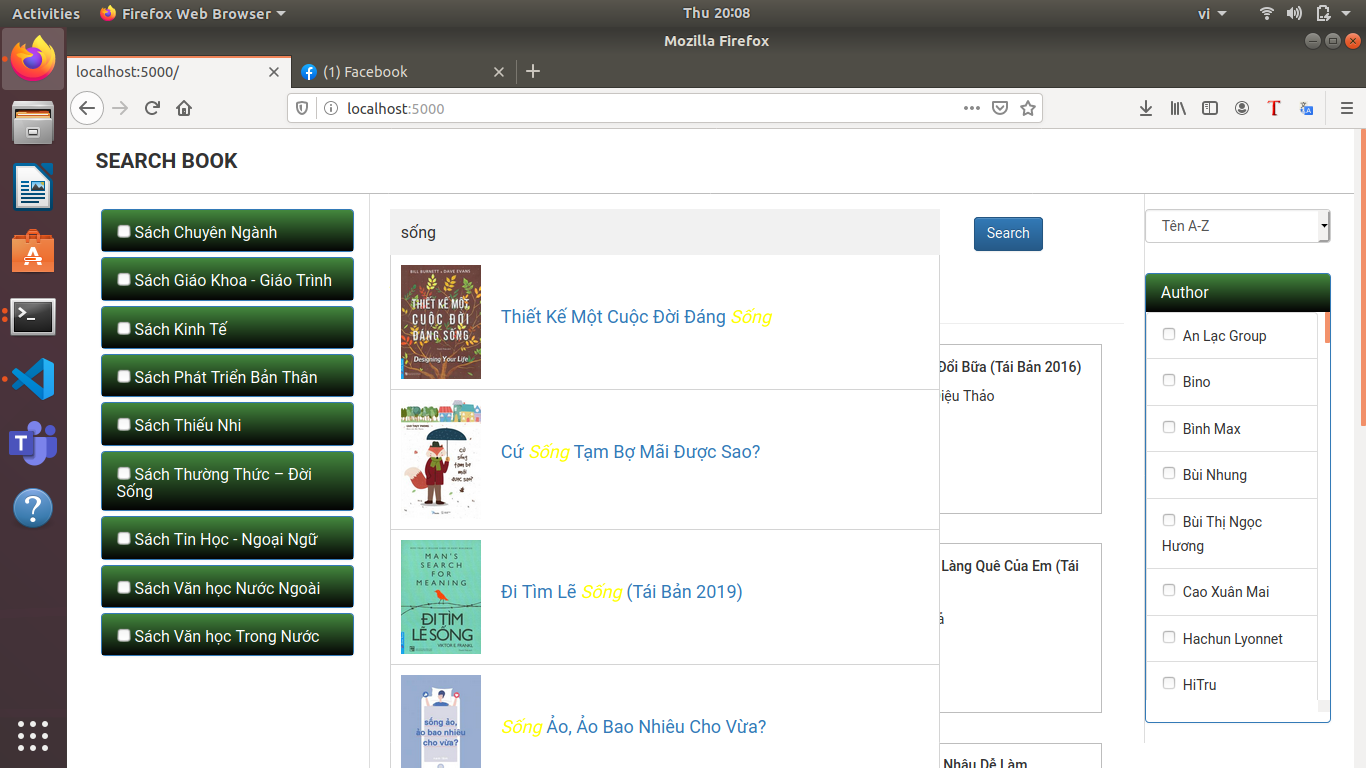
### Tự động gợi ý kết quả tìm kiếm (auto complete)

Sử dụng edge\_ngram trong elasticsearch

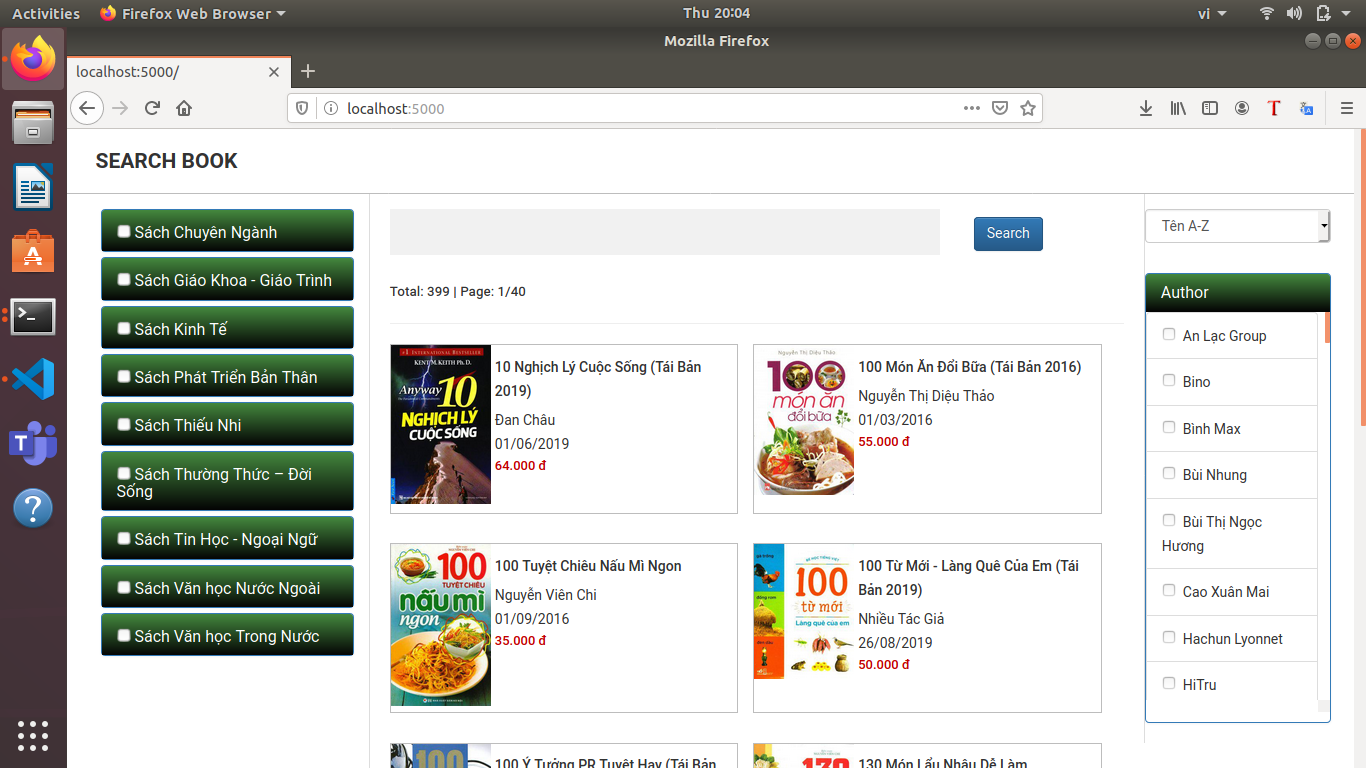
Nó sẽ ghép ký tự đầu tiên của từ với lần lượt các ký tự tiếp theo cho đến hết từ

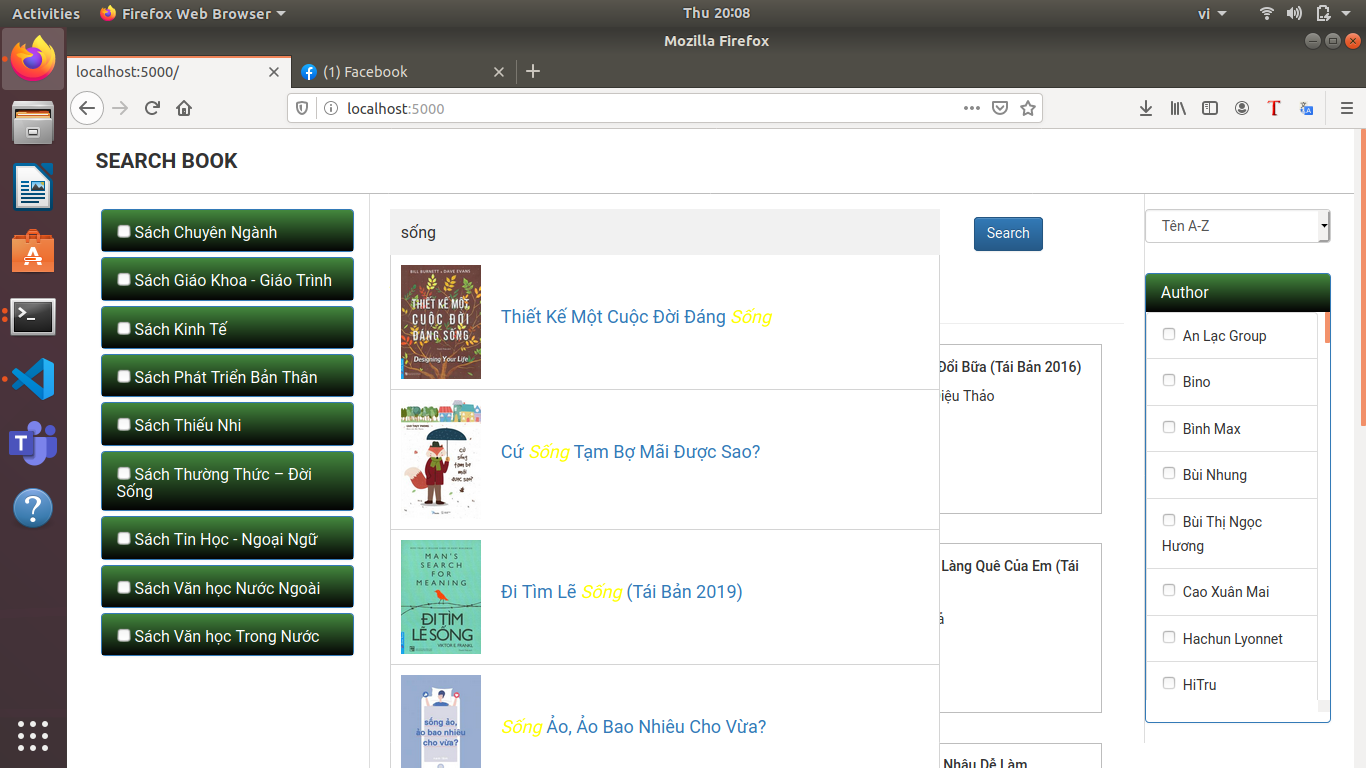
Ví dụ từ harry sẽ được đánh index những từ [h,ha,har,harr,harry]

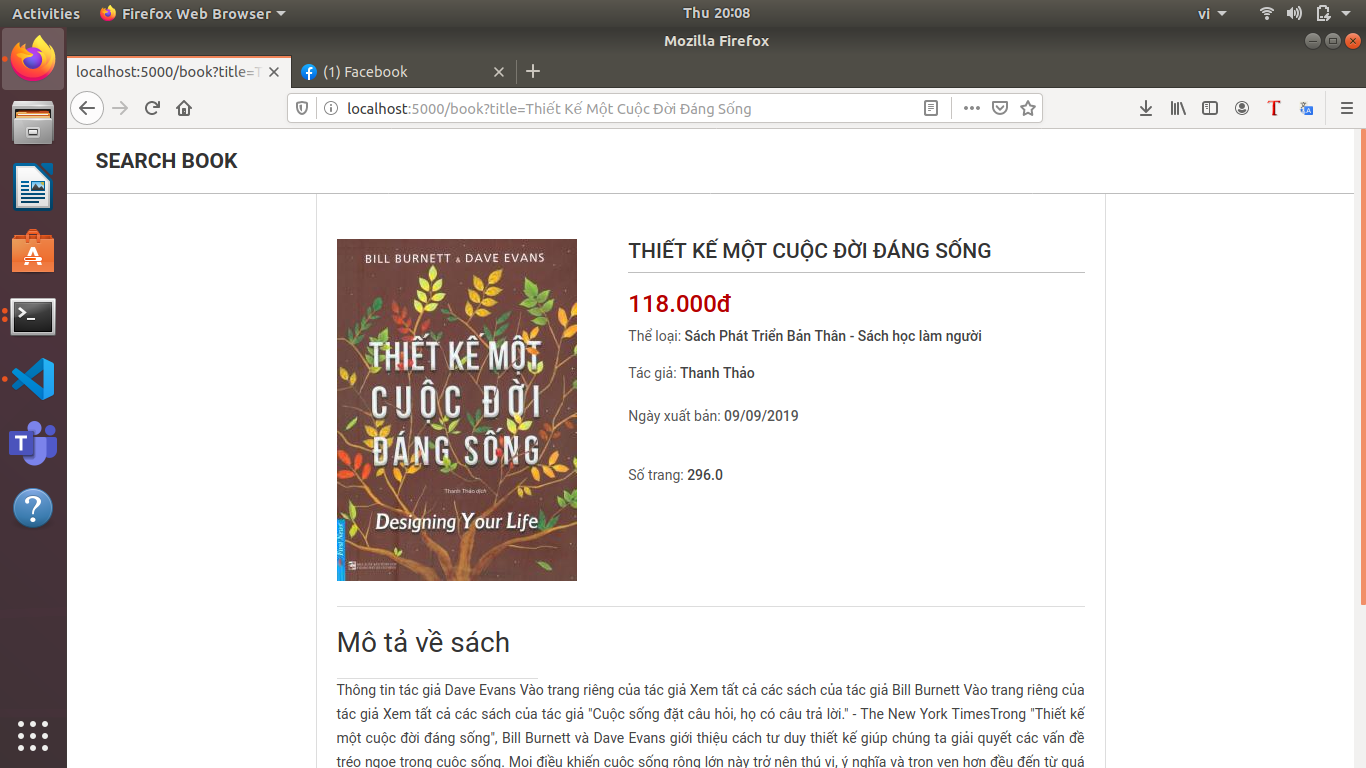
Việc này giúp hệ thống có thể tự động gợi ý kết quả ngay cả khi người dùng chưa gõ hết.



## Giao diện hệ thống







## Hướng phát triển hệ thống

- Cải thiện giao diện người dùng

- Cải thiện độ chính xác và tốc độ tìm kiếm

- Phát triển thêm một số chức năng

* Hiển thị sách theo các mục là giá thành, bán chạy, hot….
* Danh sách tìm kiếm có thể phân loại hiển thị như danh sách ban đầu
* Phát triển thêm các chức năng phụ khi vào xem trang chi tiết

# Kết luận

Sau thời gian làm việc nhóm đã hoàn thành cơ bản các yêu cầu đặt ra là xây dựng một hệ thống tìm kiếm sách tương đối hoàn chỉnh, tuy còn gặp một số khó khăn trong quá trình phát triển hệ thống.

Chúng em xin cảm ơn thầy Nguyễn Bá Ngọc đã hướng dẫn để chúng em hoàn thành đề tài này.

# Tài liệu tham khảo

Wikipedia.com

Slide tìm kiếm và trình diễn thông tin – Ts. Nguyễn Bá Ngọc

Vietnamese Analysis Plugin for Elasticsearch -

<https://github.com/duydo/elasticsearch-analysis-vietnamese>