

**TRƯỜNG ĐẠI HỌC QUY NHƠN**

**KHOA CÔNG NGHỆ THÔNG TIN**

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

**o0o**

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

**MÁY TÌM KIẾM – SEARCH ENGINE**

**Đ**

**Ề**

**TÀI:**

**TÌM KIẾM THÔNG TIN**

Quy Nhơn, ngày tháng năm 2022



Giáo viên hướng dẫn: TS.Lê Quang Hùng

Sinh viên: Ung Minh Hoài

Nguyễn Thị Thu Liễu

Thái Bá Tường

Lớp: CNTT 42A

**Mục lục**

[**I. Máy tìm kiếm – Search Engine** 2](#_Toc103029747)

[**II. Máy tìm kiếm thông tin** 5](#_Toc103029748)

[**1. Giới thiệu** 5](#_Toc103029749)

[**2. Bài toán** 8](#_Toc103029750)

[**III. Phương pháp** 8](#_Toc103029751)

[**1. Crawler** 8](#_Toc103029752)

[**2. Indexing** 10](#_Toc103029753)

[**3. Ranking** 10](#_Toc103029754)

[**4. Viewing** 11](#_Toc103029755)

[**5. Một số module khác** 12](#_Toc103029756)

[**IV. Các bước cài đặt** 13](#_Toc103029757)

[**1.** **Công cụ** 13](#_Toc103029758)

[**2.** **Các bước cài đặt** 13](#_Toc103029759)

[**V. Kết luận** 14](#_Toc103029760)

[**1.** **Kết quả đạt được** 14](#_Toc103029761)

[**2.** **Hạn chế** 14](#_Toc103029762)

[**3.** **Hướng phát triển** 14](#_Toc103029763)

1. **Máy tìm kiếm – Search Engine**

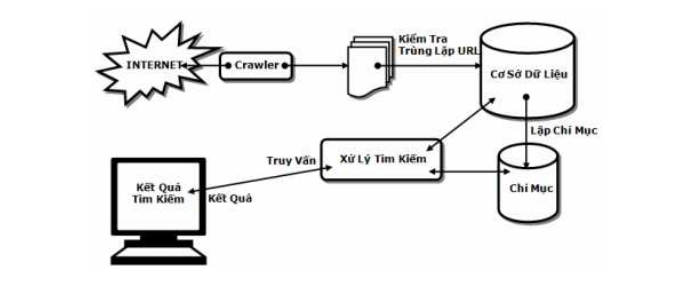
Máy tìm kiếm là một trong những công cụ được ứng dụng rộng rãi trong tất cả các lĩnh vực, nhất là đối với tìm kiếm thông tin trên Internet. Nhờ có nó mà con người có thể giải quyết được các vấn đề tra cứu thông tin một cách thuận tiện. Ngày nay, đã có rất nhiều ứng dụng và công cụ trợ giúp con người trong việc tìm kiếm thông tin trên Internet như Google hay Yahoo, nên cũng đã góp phần giúp giải quyết được vấn đề tìm kiếm thông tin trên Internet. Là một công cụ giúp chúng ta tìm kiếm thông tin dựa trên cơ sở các từ khóa, hình ảnh… liên quan đến vấn đề cần tìm trong một kho dữ liệu lớn.

Nguyên lý hoạt động của Search Engine:

Search engine điều khiển robots đi thu thập thông tin trên mạng thông qua các siêu liên kết (hyperlink). Khi robots phát hiện ra một website mới, nó gởi tài liệu (web page) về cho server chính để tạo cơ sở dữ liệu chỉ mục phục vụ cho nhu cầu tìm kiếm thông tin. Bởi vì thông tin trên mạng luôn thay đổi nên robots phải liên tục cập nhật các website cũ. Mật độ cập nhật phụ thuộc vào từng hệ thống search engine. Khi search engine nhận câu truy vấn từ user, nó sẽ tiến hành phân tích, tìm trong cơ sở dữ liệu chỉ mục và trả về những tài liệu thoả yêu cầu.

Các bước thực hiện của máy tìm kiếm:

* **Đầu vào:** Tập dữ liệu d = {d1, d2, …, dn} chứa n tài liệu và câu truy vấn q của người dùng.
* **Đầu ra:** Tập kết quả d = {d1, d2, …, dm} chứa m tài liệu (m <= n) được xếp hạng phù hợp với truy vấn người dùng.



**Hình 1**. Mô tả cơ chế hoạt động của máy tìm kiếm

Cơ chế hoạt động của máy tìm kiếm:

* Bước 1:
* Crawl là một thuật ngữ mô tả quá trình thu thập dữ liệu trên website của Googlebot. Bắt đầu từ một trang web hoặc sơ đồ trang web (site map) nào đó được người dùng submit trên công cụ Google Search Console hoặc từ danh sách các website từ lần thu thập dữ liệu trước đó, Google Spider sẽ tiến hành thu thập thông tin trên các trang này, và dò theo tất cả các liên kết trên trang đó giống như khi người dùng duyệt lần lượt tất cả các nội dung trên website.
* Googlebot sẽ lần lượt đi từ liên kết này đến các liên kết khác và thu thập tất cả các dữ liệu về trang web đầu tiên cùng tất cả các trang có liên quan đến trang đó. Thậm chí, quá trình này vẫn tiếp tục được tiến hành ở các trang có liên quan và chỉ kết thúc khi tất cả các liên kết có liên quan đến nhau được thu thập hết. Như vậy, chỉ từ một trang web ban đầu, dữ liệu thu về có thể lên đến hàng triệu trang khác.
* Các thông tin về những trang có liên quan này sẽ được thu thập về máy chủ Google phân tích và xem xét để đưa ra quyết định index và xác định chất lượng của website. Ngoài ra, thông qua việc crawl dữ liệu, Google cũng sẽ xác định xem website nào cần thu thập thông tin và tần suất cùng số lượng trang trên site đó cần tìm nạp.
* Xây dựng mô hình crawler đơn giản nhất:

1. Chọn URL khởi đầu.

2. Sử dụng HTML protocol để lấy trang web.

3. Trích xuất ra các link. Lưu lại trong queue.

4. Lặp đi lặp lại bước 2,3.

* Bước 2: Kiểm tra trùng lặp URL: là một bước xử lý trong hệ thống tìm kiếm, nhằm đảm bảo Crawler không bị xử lý hai tài liệu giống nhau trong quá trình duyệt.
* Bước 3: Cơ sở dữ liệu: dùng để chứa nội dung các tài liệu trên Internet. Hệ cơ sở dữ liệu phải hỗ trợ khả năng phân tán vì khối lượng dữ liệu lưu trong nó sẽ là rất lớn và có thể phân chia về mặt nội dung rất nhiều.
* Bước 4: Lập chỉ mục: là một quá trình xử lý trong hệ thống tìm kiếm, nhằm tạo ra bộ chỉ mục để hỗ trợ quá trình tìm kiếm cho tất cả các tài liệu được lưu trong cơ sở dữ liệu.
* Bước 5: Chỉ mục: : là cơ sở dữ liệu dùng để chứa bảng chỉ mục của hệ thống tìm kiếm.
* Bước 6: Xử lý tìm kiếm: là một module quan trọng trong hệ thống tìm kiếm. Xử lý tìm kiếm tiếp nhận và xử lý câu truy vấn của người dùng, sau đó tiến hành tìm kiếm trong cơ sở dữ liệu thông qua bảng chỉ mục. Kết quả trả về được ranking cung cấp các nội dung đáp ứng tốt nhất truy vấn của người tìm kiếm. Thông thường, xếp hạng này dựa vào: độ thân thiện của cấu trúc trang web đối với bot tìm kiếm, số lượng truy cập và độ tin cậy.

# **Máy tìm kiếm thông tin**

## **Giới thiệu**

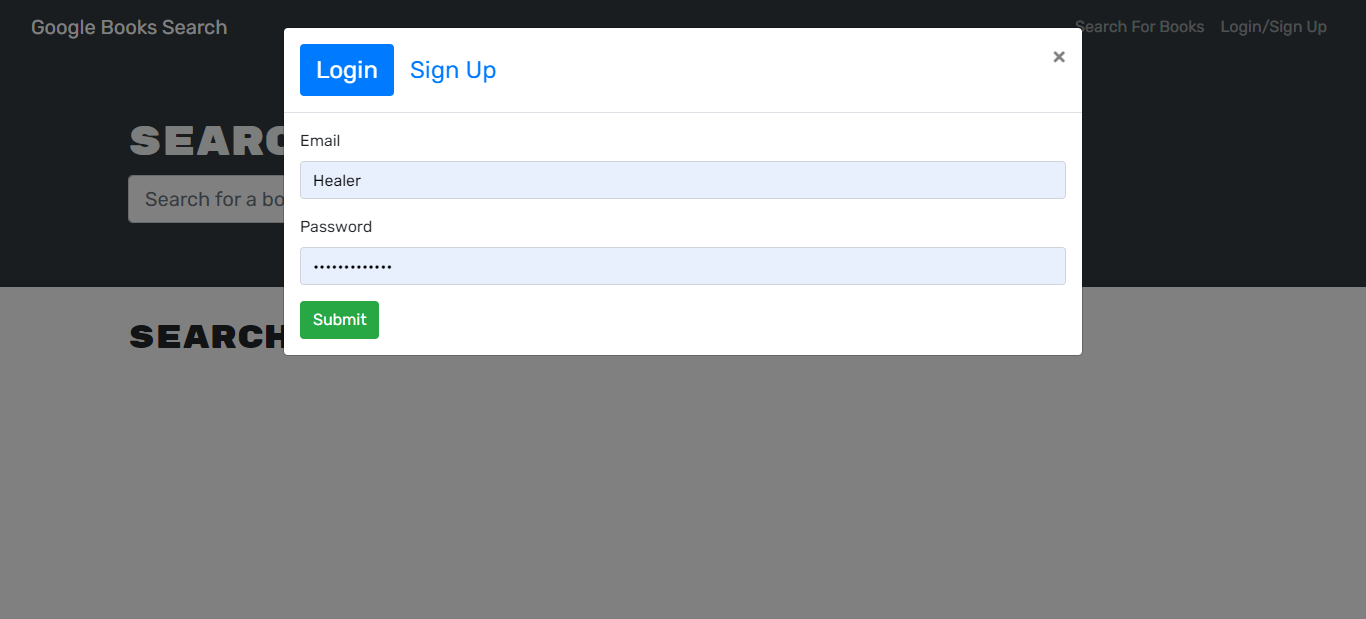
Lượng dữ liệu đang được tạo ra và lưu trữ ở toàn cầu gần như không thể tưởng tượng được và lượng dữ liệu ấy không dừng lại mà ngày một phát triển tăng lên nhanh chóng. Dẫn đến việc đặt ra một yêu cầu làm thế nào tìm kiếm thông tin nhanh, giảm được chi phí, tiết kiệm thời gian mà đem lại hiệu quả cao nhất.

Việc internet phát triển một cách mạnh mẽ làm thay đổi thói quen đọc sách của nhiều người. Thay vì đến trực tiếp thư viện để tìm kiếm sách mà mình muốn đọc mọi người có thể đọc trực tiếp trên các trang web bằng cách gõ vào tên hoặc nội dung có liên quan đến cuốn sách. Từ nhu cầu đó một trang web ra đời có tên Google Books Search giúp người dùng tìm kiếm sách một cách nhanh chóng và hiệu quả. Kết quả tìm kiếm sẽ được ranking theo độ tương tự với nội dung tìm kiếm.

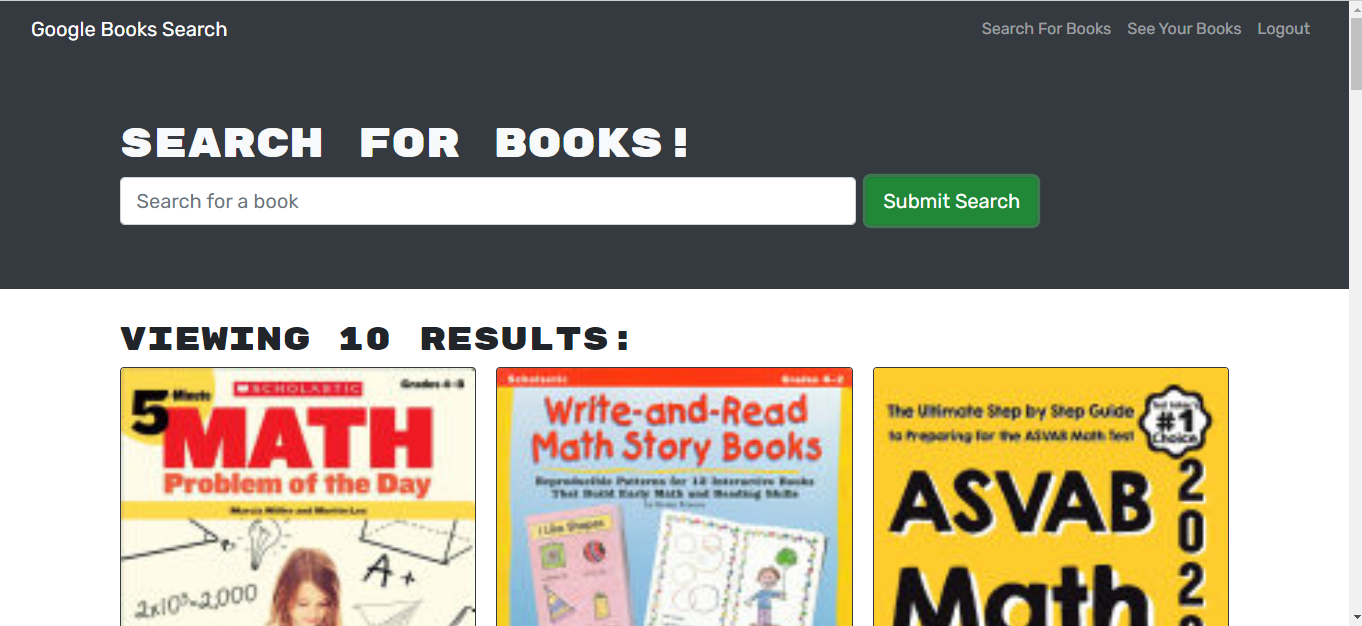
Ứng dụng này chuyển đổi công cụ tìm kiếm API GoogleBook hoạt động đầy đủ được xây dựng với API RESTful thành API GraphQL được xây dựng với Máy chủ Apollo. Sử dụng ngăn xếp MERN, với giao diện người dùng React, cơ sở dữ liệu MongoDB và server Node.js / Express.js và API.

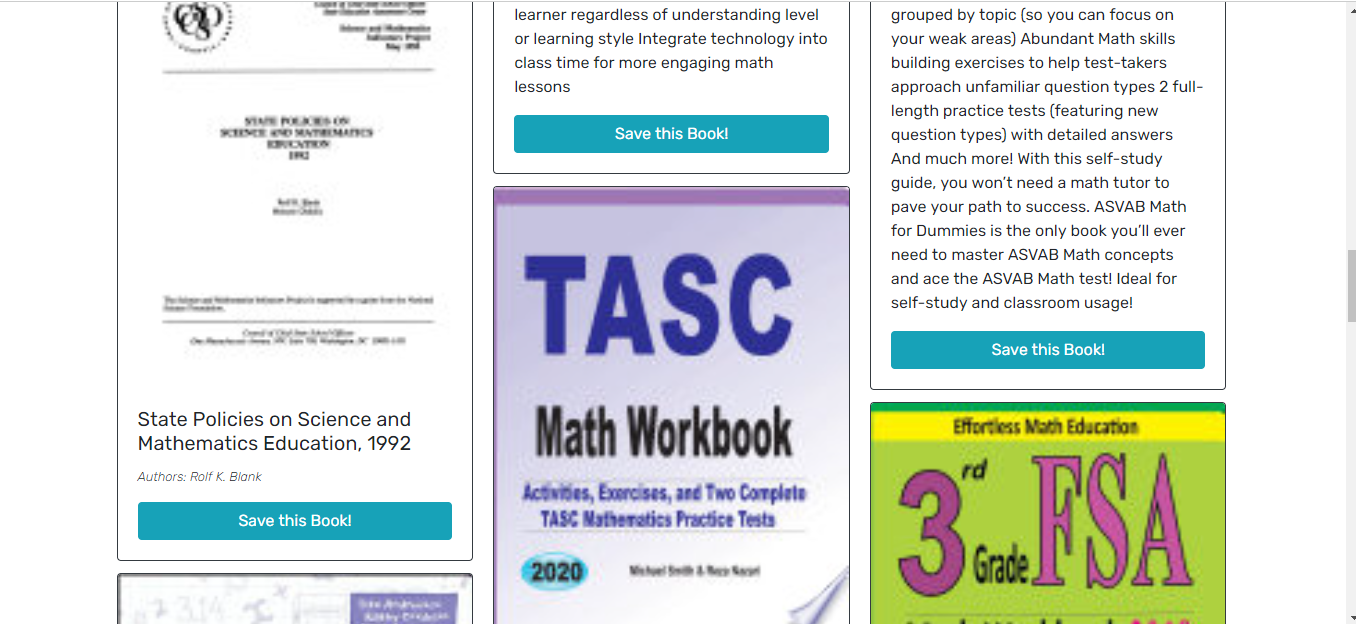
Để có trải nghiệm một cách đầy đủ các chức năng của trang web người dùng có thể đăng ký nếu chưa có tài khoản hoặc đăng nhập nếu đã có tài khoản. Để thực hiện tìm kiếm người dùng nhập vào nội dung tìm kiếm ở thanh tìm kiếm “Search for a book” sau đó Submit Search. Hệ thống sẽ trả về 10 kết quả sát với nội dung tìm kiếm của người dùng nhất và sắp xếp theo thứ tự từ cao đến thấp. Để thuận tiện cho việc tìm kiếm lại cuốn sách đã được tìm kiếm trước đó người dùng có thể click vào “Save this Book!” dưới phần mô tả để lưu vào thư viện sách mà bạn yêu thích. Hoặc cũng có thể xóa khỏi thư viện sách của bạn.

Một số hình ảnh minh họa:

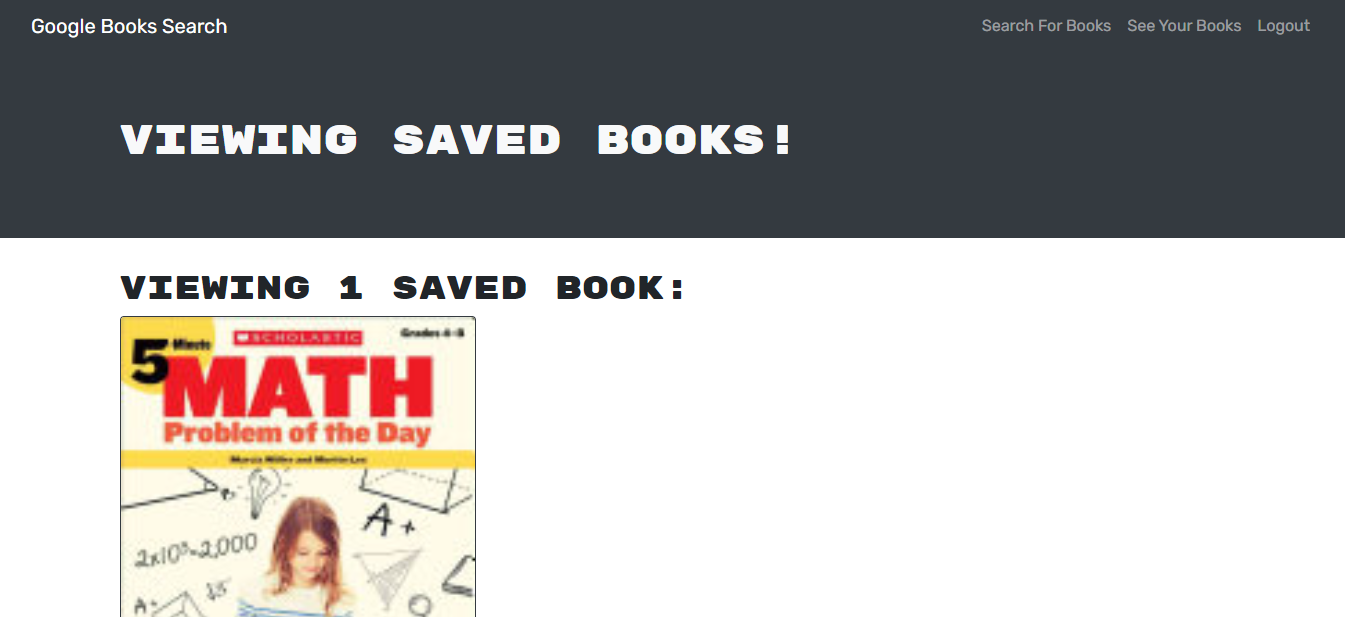


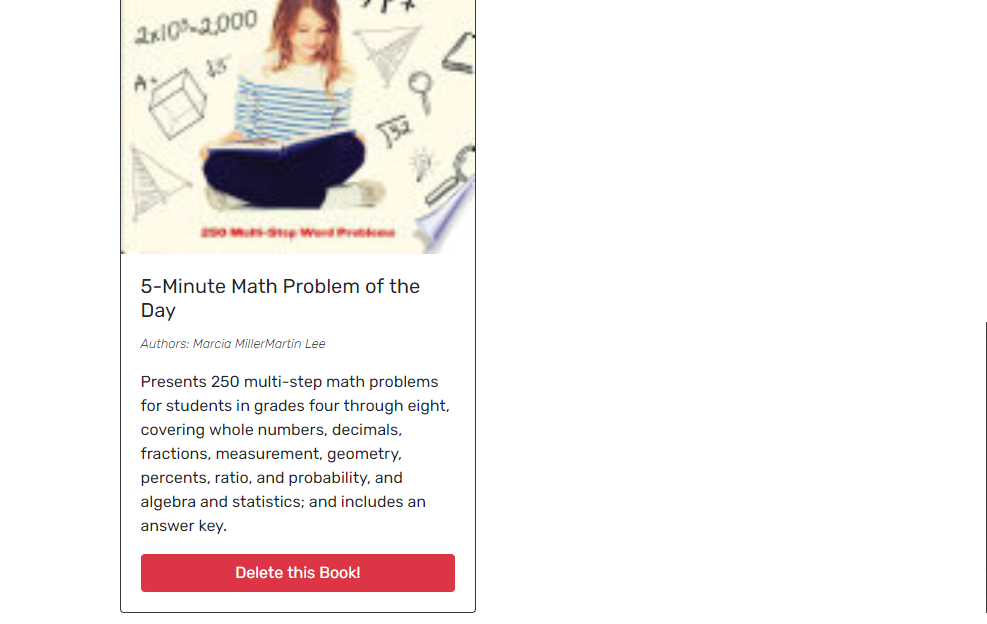
**Hình 2.** Giao diện đăng nhập





**Hình 3.** Kết quả tìm kiếm





**Hình 4**. Giao diện nơi lưu trữ sách yêu thích

## **Bài toán**

* Đầu vào:
* Tập dữ liệu d = {d1, d2, …, d100} chứa 100 bài viết liên quan đến các chủ đề khác nhau.
* Truy vấn q là từ khóa tìm kiếm của người dùng.
* Đầu ra:
* Tập kết quả d’ = {d’1, d’2, …, d’10} chứa 10 bài viết phù hợp với từ khóa (chủ đề) nhập vào.

1. **Phương pháp**

Có rất nhiều kỹ thuật được áp dụng để xây dựng một máy tìm kiếm thông tin. Mỗi kỹ thuật đều có một vai trò riêng không thể thiếu và nó quyết định chức năng và mục đích sinh ra của máy tìm kiếm. Tuy nhiên, có ba kỹ thuật cơ bản nhất cần phải có khi xây dựng một máy tìm kiếm đó là: crawler, indexing (kỹ thuật đánh chỉ mục) và kỹ thuật xử lý truy vấn.

1. **Crawler**

* Hệ thống crawler và lưu trữ dữ liệu trực tiếp trên google api dưới dạng các chuỗi JSON.



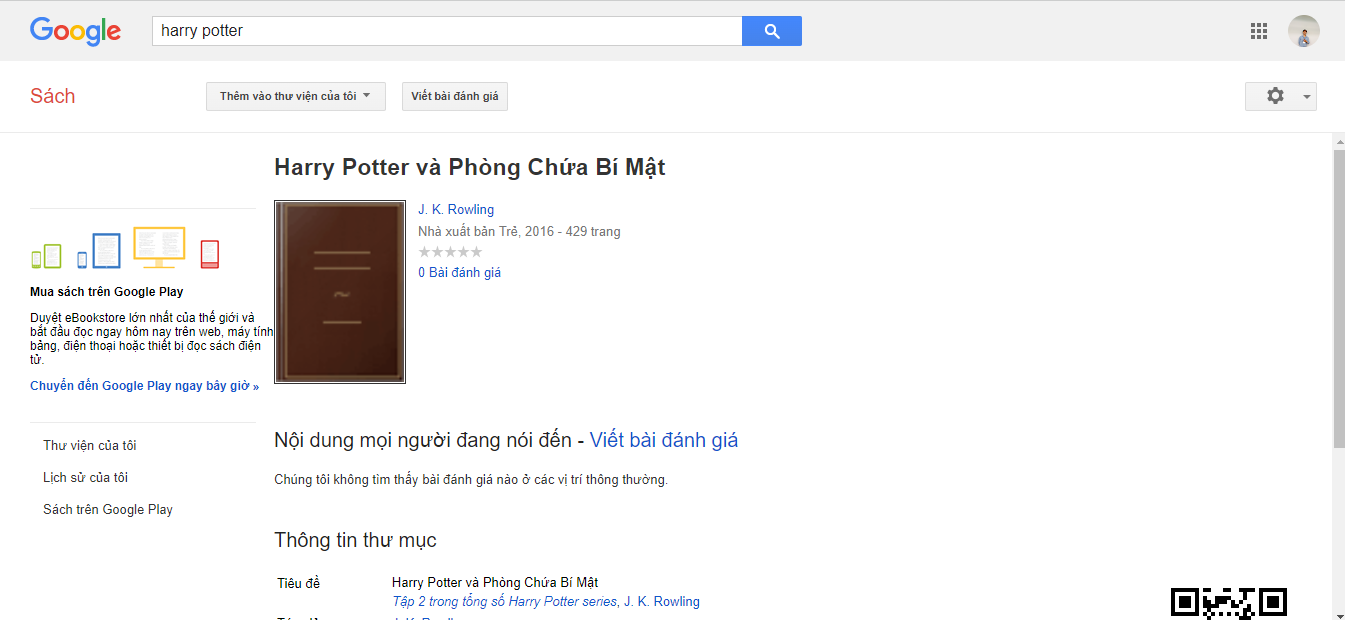
**Hình 5.** Dữ liệu được lưu trữ dưới dạng các chuỗi JSON

* Googlebot sẽ lần lượt đi từ liên kết này đến các liên kết khác và thu thập tất cả các dữ liệu về trang web đầu tiên cùng tất cả các trang có liên quan đến trang đó.
* Ví dụ Googlebot bắt đầu thu thập dữ liệu từ liên kết ở seftlink:

<https://www.googleapis.com/books/v1/volumes?q=harry+potter>

* Việc của nó sẽ tiếp tục truy cập vào các liên kết đã được lưu trữ chẳng hạn:

"previewLink": "http://books.google.com.vn/books?id=3KgoyQEACAAJ&dq=harry+potter&hl=&cd=1&source=gbs\_api"



**Hình 6.** Kết quả truy cập vào liên kết trong previewLink

* Ở file SearchBook.js sẽ gọi phương thức searchGoogleBooks để xử lý tìm kiếm

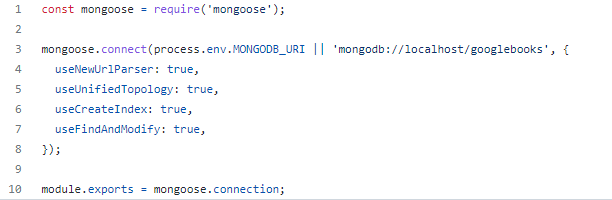


1. **Indexing**

Cơ sở dữ liệu được lập chỉ mục , thường được gọi là **IndexedDB**, có thể tạo một loại **cơ sở dữ liệu** từ JavaScript và lưu nó trong trình duyệt. So với localStorage, nó mạnh hơn ở chỗ nó có thể lưu trữ các đối tượng cũng như các chuỗi ký tự và có chức năng tìm kiếm.

Trước đây, đặc tả của cơ sở dữ liệu phía máy khách (WebSQL) sử dụng SQL đã được phát triển, nhưng sau đó nó đã bị loại bỏ và IndexedDB được chuẩn bị như một API cơ sở dữ liệu mới theo hướng không sử dụng SQL.

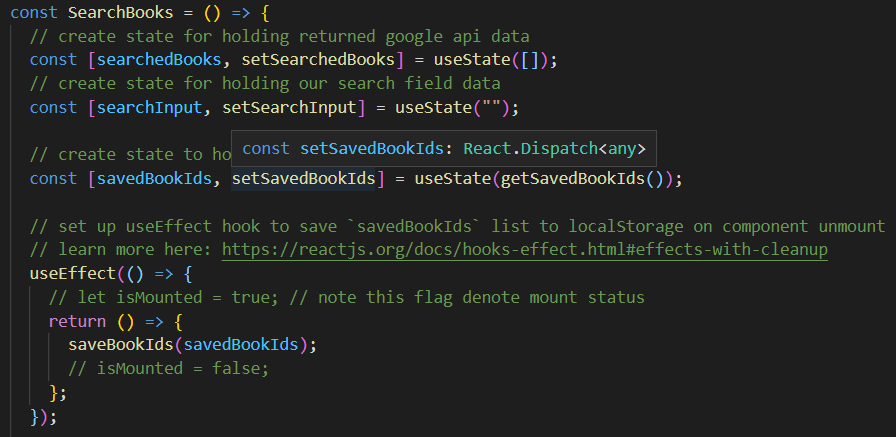
Sử dụng thư viện Mongoose của JavaScript, gọi phương thức useCreateIndex: true, dùng để đánh chỉ mục.



**Hình 7**. Chỉ mục trong [book-search-engine](https://github.com/yulduzetta/book-search-engine)/[server](https://github.com/yulduzetta/book-search-engine/tree/main/server)/[config](https://github.com/yulduzetta/book-search-engine/tree/main/server/config)/**connection.js**

1. **Ranking**

* Khi tìm kiếm thì hệ thống sẽ đặt trạng thái của dữ liệu ở file SearchBooks.js

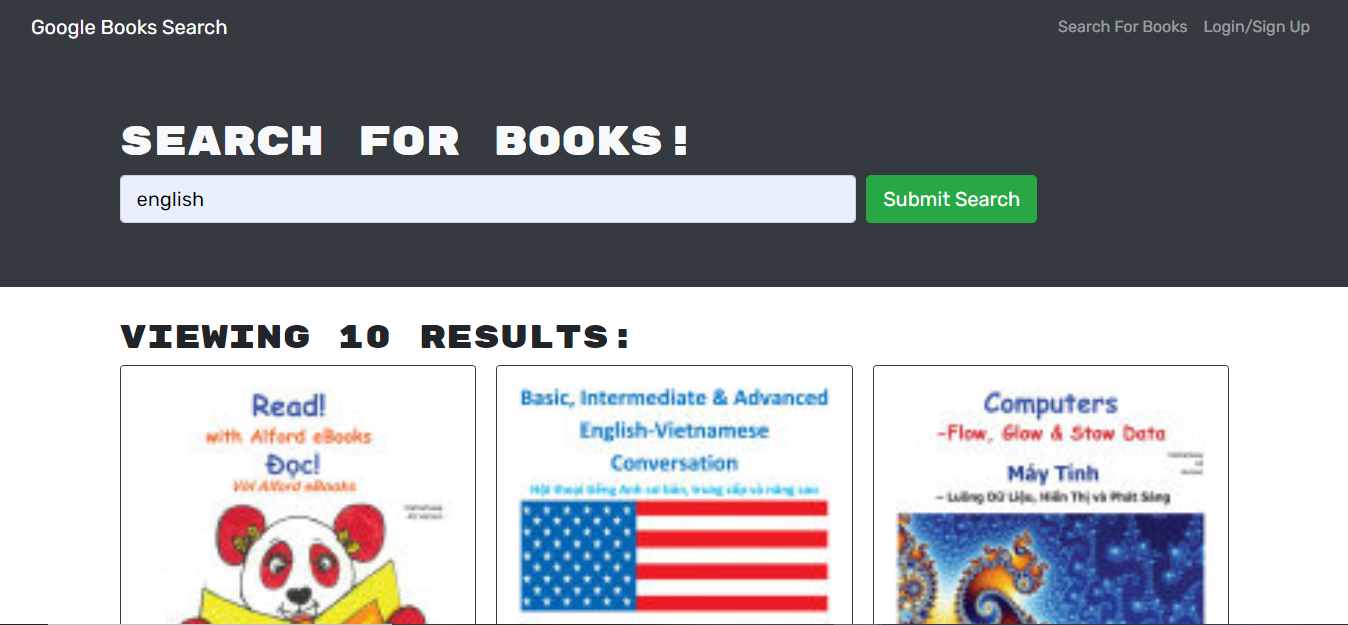


* Sau đó sẽ map với dữ liệu được lưu để trả về danh sách các kết quả được ranking dựa vào id

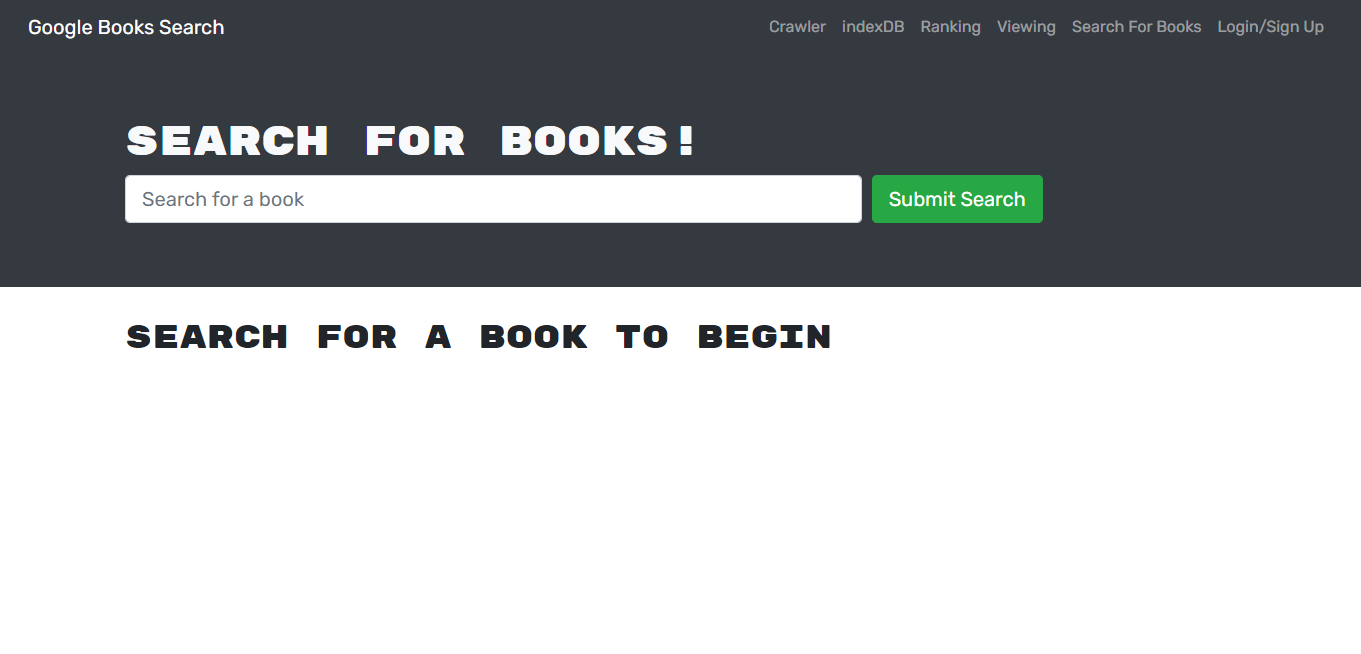


1. **Viewing**

* Hiển thị kết quả tìm kiếm



**Hình 8.** Hiển thị 10 kết quả đầu tiên theo xếp hạng Ranking



**Hình 9.** Màn hình menu cho 3 chức năng Crawler, Ranking, Viewing

1. **Một số module khác**
2. Check login/logout

([book-search-engine](https://github.com/yulduzetta/book-search-engine)/[client](https://github.com/yulduzetta/book-search-engine/tree/main/client)/[src](https://github.com/yulduzetta/book-search-engine/tree/main/client/src)/[utils](https://github.com/yulduzetta/book-search-engine/tree/main/client/src/utils)/**auth.js)**

1. Lưu trữ dữ liệu cho nút “Save Your Book”

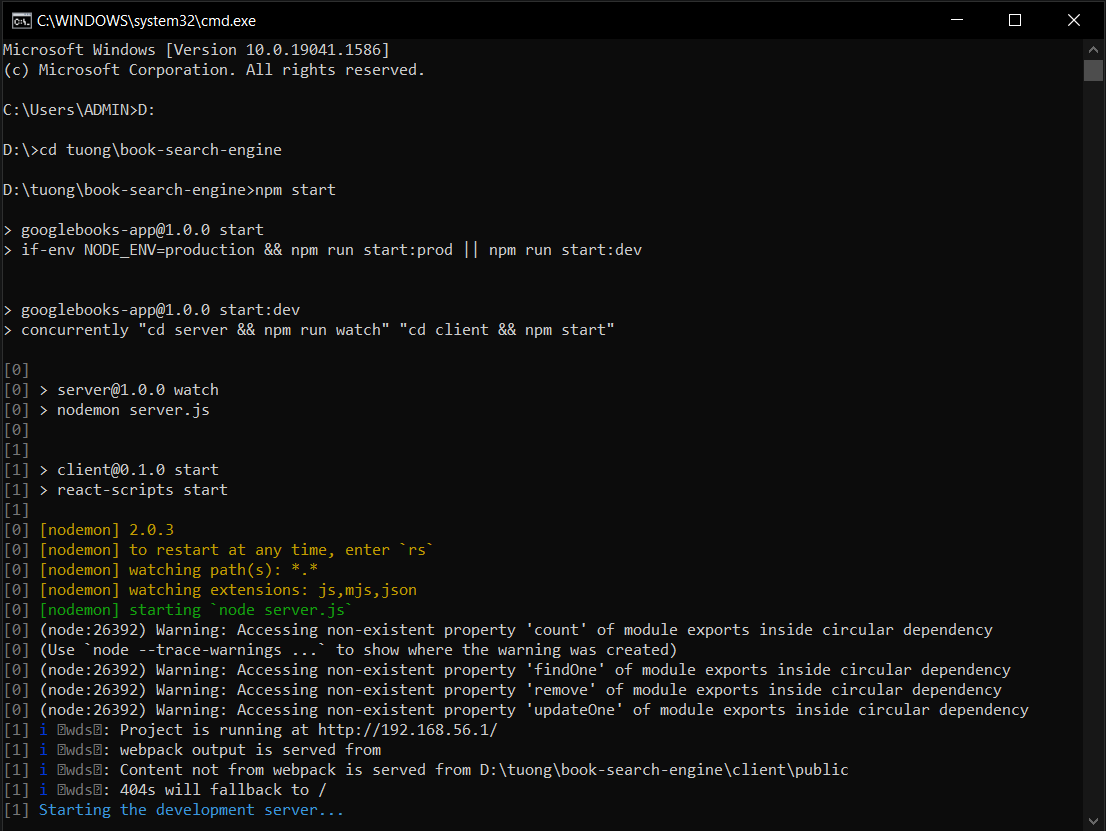
([book-search-engine](https://github.com/yulduzetta/book-search-engine)/[client](https://github.com/yulduzetta/book-search-engine/tree/main/client)/[src](https://github.com/yulduzetta/book-search-engine/tree/main/client/src)/[utils](https://github.com/yulduzetta/book-search-engine/tree/main/client/src/utils)/**localStorage.js)**

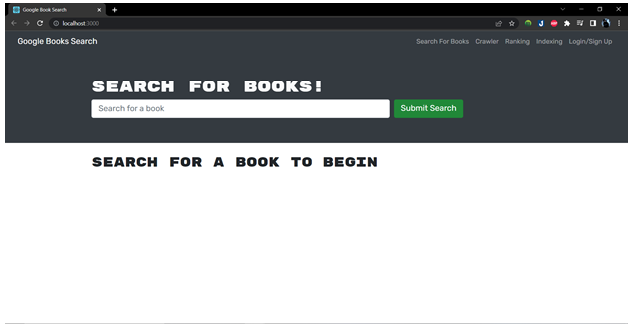
1. Truy vấn dữ liệu: ([book-search-engine](https://github.com/yulduzetta/book-search-engine)/[client](https://github.com/yulduzetta/book-search-engine/tree/main/client)/[src](https://github.com/yulduzetta/book-search-engine/tree/main/client/src)/[utils](https://github.com/yulduzetta/book-search-engine/tree/main/client/src/utils)/queries.js)
2. **Các bước cài đặt**
3. **Công cụ**

* Visual Studio Code (viết, chỉnh sửa code)
* Cài đặt môi trường Nodejs: Chọn phiên bản LTS 
* Tải và cài đặt Git Bash (dùng để install code từ github)

1. **Các bước cài đặt**

* Tạo 1 Folder để lưu trữ project
* Thực hiện theo các bước sau:
  + Mở cmd và truy cập đến thư mục đã tạo
  + Nhập lệnh: git init => git clone <https://github.com/yulduzetta/book-search-engine> => cd book-search-engine => npm install => npm start
* Kết quả:





**Hình 10.** Giao diện ứng dụng sau khi được cài đặt

1. **Kết luận**
2. **Kết quả đạt được**

* Tìm hiểu và nắm rõ được các nội dung cơ bản của máy tìm kiếm.
* Chạy demo được một mô hình đơn giản của máy tìm kiếm : Google Book Search.
* Nắm được cơ bản các thành phần, cách cài đặt, kết quả của máy tìm kiếm.

1. **Hạn chế**

* Chưa nắm rõ được các module ranking viewing của mô hình Google Book Search.
* Chưa triển khai được các module riêng biệt của máy tìm kiếm.

1. **Hướng phát triển**

* Tiếp tục tìm kiếm, làm rõ các module của máy tìm kiếm.
* Tiếp tục phát triển mô hình máy tìm kiếm.
* Mở rộng nội dung của mô hình.
* Có thể thiết kế lại giao diện thành một trang web hoàn chỉnh.

1. **Tài liệu tham khảo**