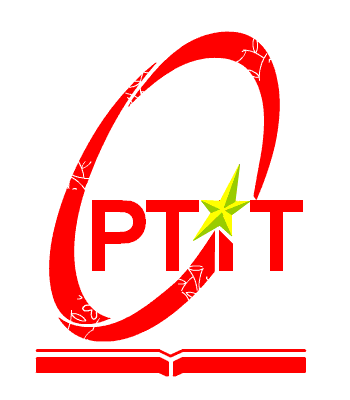


**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**----------**



**ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC**

**ĐỀ TÀI: Xây dựng hệ khuyến nghị sách trong bán hàng trực tuyến**

**sử dụng kỹ thuật lọc dựa trên nội dung**

**và kỹ thuật lọc cộng tác**

**Giảng viên hướng dẫn: Nguyễn Hoàng Anh**

**Sinh viên: Nguyễn Thị Lan Anh**

**Mã sinh viên: B16DCCN010**

**Lớp: D16CNPM1**

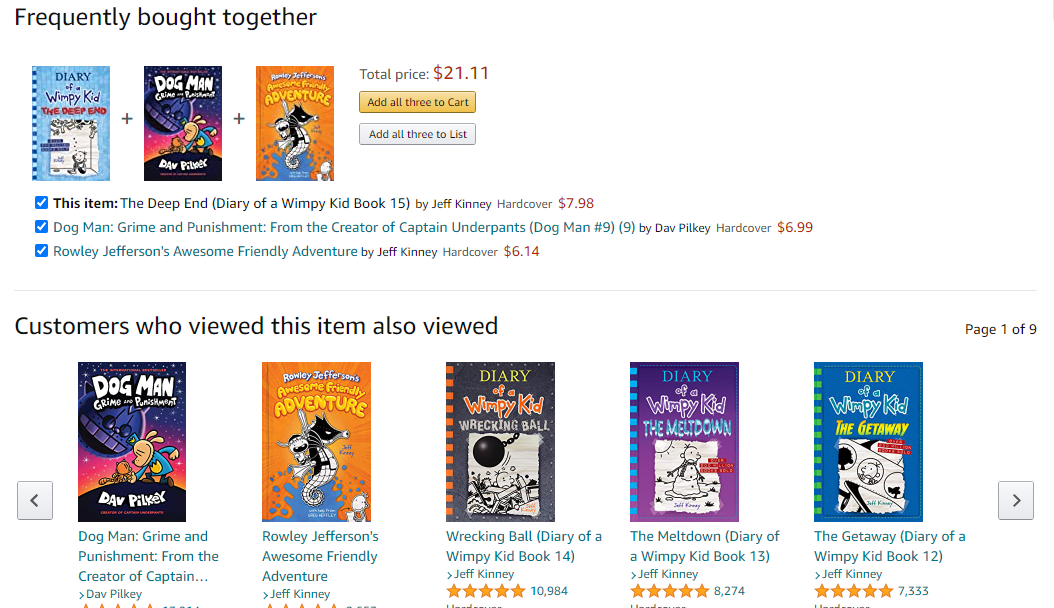
**Khóa: 2016-2021**

**Hệ đại học: Đại học chính quy**

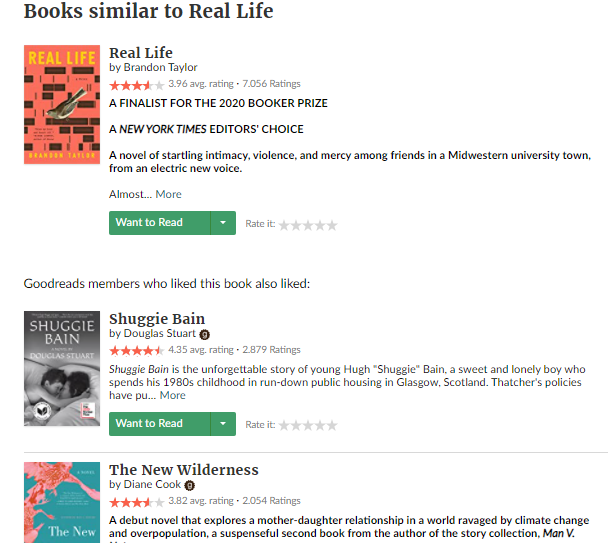
**Hà Nội, Tháng 11/2020**

CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ HỆ KHUYẾN NGHỊ VÀ HỆ KHUYẾN NGHỊ SÁCH TRONG BÁN HÀNG TRỰC TUYẾN

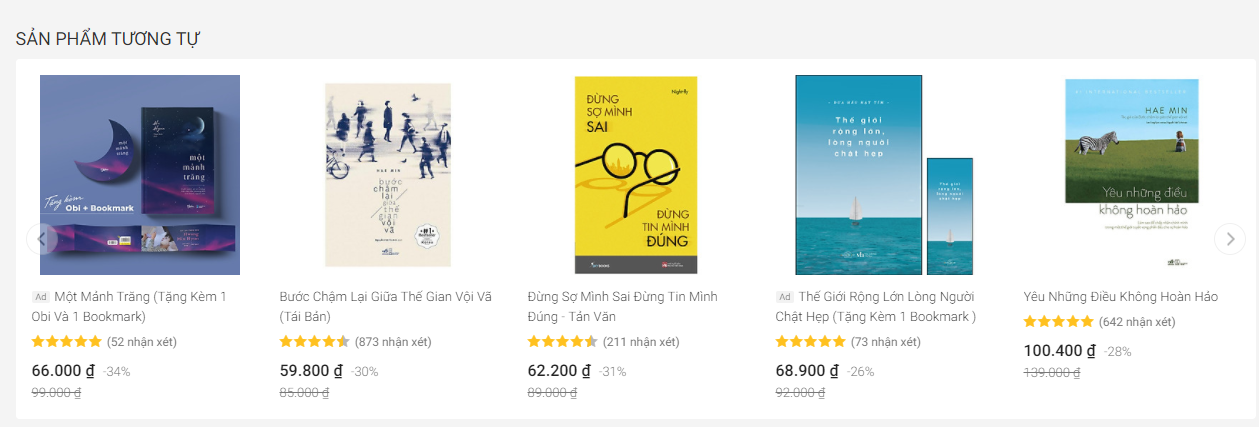
* 1. Giới thiệu hệ khuyến nghị – Recommender System (nêu được lý do chọn đề tài)
* Wikipedia định nghĩa, hệ thống khuyến nghị (Recommender System), hay còn gọi là hệ thống tư vấn là một hệ thống lọc thông tin nhằm dự đoán, đánh giá sở thích, mối quan tâm, nhu cầu của người dùng để đưa ra một hoặc nhiều mục, sản phẩm, dịch vụ mà người dùng có thể sẽ quan tâm với xác suất lớn nhất. Những gợi ý được cung cấp nhằm hỗ trợ cho người sử dụng đưa ra quyết định lựa chọn những sản phẩm, dịch vụ phù hợp với nhu cầu và thị hiếu của mình, chẳng hạn như: mua sản phẩm nào, nghe thể loại nhạc nào, hay tin tức nào nên đọc…
* Chức năng của hệ khuyến nghị:
* Đối với nhà cung cấp:
* Tăng số lượng mặt hàng bán ra cho các hệ thống thương mại điện tử: thay vì người dùng chỉ mua 1 sản phẩm mà họ cần, họ được gợi ý mua những sản phẩm ‘có thể họ cũng quan tâm’ mà bản thân họ không nhận ra. Bằng cách tìm ra những mối quan tâm ẩn của người dùng, hệ thống gợi ý làm gia tăng số lượng mặt hàng bán ra. Tương tự đối với các hệ thống phi thương mại, hệ thống gợi ý sẽ giúp người dùng tiếp cận với nhiều đối tượng thông tin hơn.
* Tăng sự hài lòng của người dùng: vai trò chủ đạo của hệ khuyến nghị là hiểu nhu cầu của người dùng, gợi ý cho họ những thứ họ cần. Họ sẽ tìm thấy các gợi ý thú vị, có hiệu quả, chính xác, gợi ý kịp thời và một giao diện đẹp có thể tối ưu việc sử dụng và tăng sự hài lòng của người dùng trong hệ thống.
* Tăng độ tin cậy, độ trung thực của người dùng: khi người dùng càng tin cậy vào hệ thống, đưa ra những đánh giá trung thực cho các sản phẩm, hệ thống càng mang lại cho người dùng những gợi ý chính xác hơn
* Đối với người sử dụng:
* Tìm ra một số sản phẩm tốt nhất: Hệ thống gợi ý tới người dùng một số sản phẩm được xếp hạng, từ đó người dùng có thể tìm được sản phẩm tốt nhất cho bản thân.
* Gợi ý liên tục: Thay vì tập trung vào gợi ý đơn, các hệ thống khuyến nghị tạo các gợi ý liên tục tới người dùng đến khi họ tìm được sản phẩm ưng ý
* Hệ thống khuyến nghị sẽ đưa ra các gợi ý dựa trên quá trình thu thập, xử lý và phân tích dữ liệu của người dùng. Dữ liệu đó được chia làm hai loại là phản hồi tường minh (explicit feedback) bằng cách yêu cầu người dùng phản hồi trực tiếp (đánh giá yêu thích, xếp hạng số sao, hành động thêm vào giỏ hàng,…) và phản hồi tiềm ẩn (implicit feedback) bằng cách tự động suy luận dựa trên những tương tác của người dùng với hệ thống (số lần nhấp chuột vào sản phẩm, thời gian quan sát…).
  1. Giới thiệu hệ khuyến nghị sách (các use case thực tế)
* Các trang web thương mại nói chung và website bán sách nói riêng hiện nay đều đã tích hợp hệ gợi ý.
* Quốc tế:
* Amazon.com



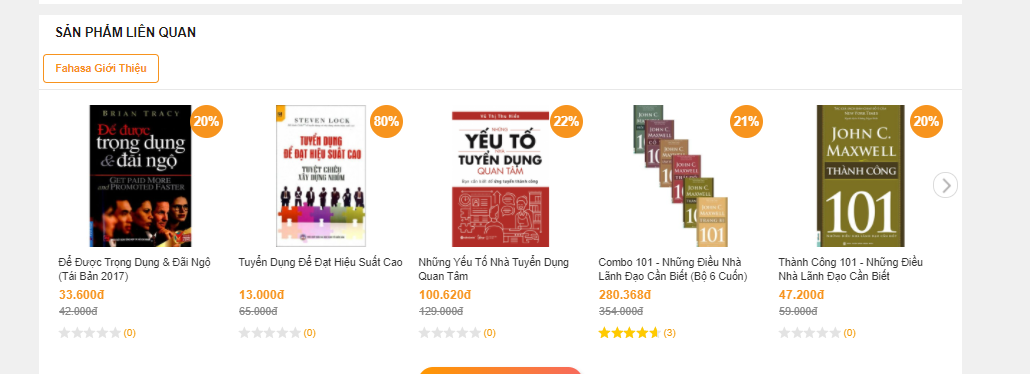
* Goodreads.com



* Tại Việt Nam:
* Tiki.vn



* Fahasa.com



* Nhamam.com.vn



* Lí do chọn đề tài:
  1. Phát biểu bài toán khuyến nghị sách (phân tích bài toán)
* Một cách hình thức, bài toán khuyến nghị hay bài toán tư vấn được phát biểu như sau:
* Gọi U là tập tất cả người dùng, I là tập tất cả các sản phẩm có thể tư vấn. Tập I có thể rất lớn, từ hàng trăm ngàn đến hàng triệu bản ghi.
* Hàm r(u, i) đo độ phù hợp hay hạng của sản phẩm i với người dùng u. Với mỗi người dùng u thuộc U, cần tìm sản phẩm i’ sao cho hàm r(u, i’) đạt giá trị lớn nhất:
* Trong hệ thống gợi ý, độ phù hợp của một sản phẩm thường được đo bằng điểm, ví dụ người dùng A đánh giá cuốn sách B được 4/5 sao. Tuy nhiên, độ phù hợp có thể là một hàm bất kỳ phụ thuộc vào ứng dụng cụ thể. Giá trị của hàm r có thể được xác định bởi người dùng hoặc tính toán bằng một công thức nào đó.
* Mỗi người dùng trong không gian U được xác định bởi 1 hồ sơ (profile). Hồ sơ này bao gồm thông tin người dùng (id, tên). Tương tự mỗi sản phẩm s trong không gian I cũng được xác định bởi 1 tập đặc trưng. Trong hệ thống gợi ý sách, đặc trưng của tập I có thể là: id, tên sách, mô tả nội dung, giá bán,….
* Quá trình xây dựng hệ gợi ý bao gồm các bước sau:
* Tập hợp thông tin (Information Recollection)
* Lựa chọn, chuyển đổi, cấu trúc hóa thông tin
* Biểu diễn thông tin
  1. Khó khăn và thách thức
  2. Các đóng góp của đồ án
  3. Kết luận chương

**CHƯƠNG 2: CÁC KỸ THUẬT KHUYẾN NGHỊ SÁCH (cơ sở lý thuyết)**

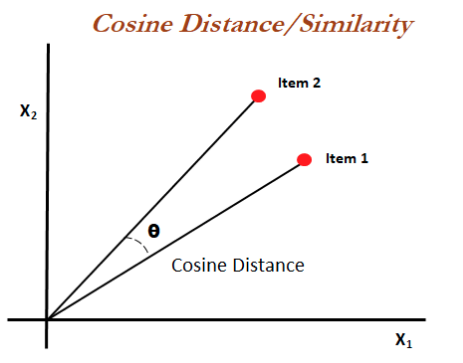
2.1. Kỹ thuật khuyến nghị sách lọc dựa trên nội dung

2.1.1. Khái niệm

* Lọc dựa trên nội dung (Content-based filtering) là phương pháp dựa trên những đặc điểm chính của item. Nói cách khác, thuật toán cho phương pháp này cố gắng gợi ý những item có những điểm giống với những item mà người dùng đã từng quan tâm trong quá khứ. Gợi ý dựa trên nội dung khai thác những khía cạnh có liên quan đến nội dung thông tin sản phẩm người dùng đã từng sử dụng hay truy cập trong quá khứ để tạo nên gợi ý hay nói cách khác gợi ý trên nội dung được dựa trên mô tả có sẵn về đối tượng. Các phương pháp tiếp cận cho lọc nội dung có nguồn gốc từ lĩnh vực truy vấn thông tin, trong đó mỗi sản phẩm được biểu diễn bằng một hồ sơ người dùng.
* Phương pháp gợi ý dựa trên nội dung dựa trên độ phù hợp r(u,v) của sản phẩm v với người dùng u được đánh giá dựa trên độ phù hợp r(u, ). Ví dụ, để gợi ý một cuốn sách cho người dùng u, hệ thống tư vấn sẽ tìm các đặc điểm của những cuốn sách được người dùng u đã xem.
* Hướng tiếp cận dựa trên nội dung bắt nguồn từ những nghiên cứu về thu thập thông tin và lọc thông tin. Những tiến bộ so với hướng tiếp cận cũ của thu thập thông tin là do việc sử dụng hồ sơ về người dùng. Hồ sơ này được xây dựng dựa trên những thông tin được người dùng cung cấp trực tiếp, hoặc gián tiếp
* Phần mô tả bên trên hiện tại em chưa phát triển, vì em cứ nghĩ hướng mình làm chuẩn là content based rồi, mà thực ra mới chỉ dựa và đối sánh mẫu.

2.1.2. Gợi ý dựa trên đối sánh mẫu

- Ý tưởng của kỹ thuật này là việc biến 2 văn bản cần so sánh thành 2 vector rồi từ đó tính toán khoảng cách giữa 2 vector đó.



- Công thức để tính khoảng cách của 2 vector trong không gian như sau:



- Đưa văn bản thành vector model để tính toán sử dụng kỹ thuật cơ bản tf-idf để đưa về dạng

- Đối sánh mẫu dựa trên tên sách và phần mô tả sách.

2.2. Kỹ thuật khuyến nghị sách dựa trên lọc cộng tác

2.2.1. Đưa ra các phương pháp con của lọc cộng tác sử dụng

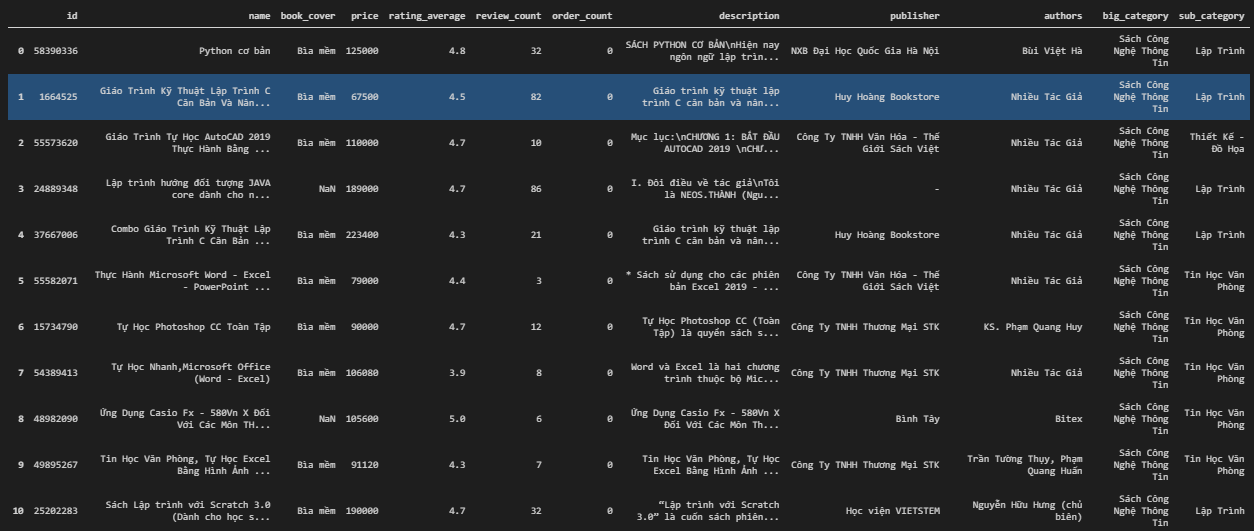
Item-item sử dụng thuật toán K-nearest neighbor, sử dụng các hàm đánh giá cosine, Euclid, manhatan để gợi ý

2.2.x. Phương pháp đánh giá

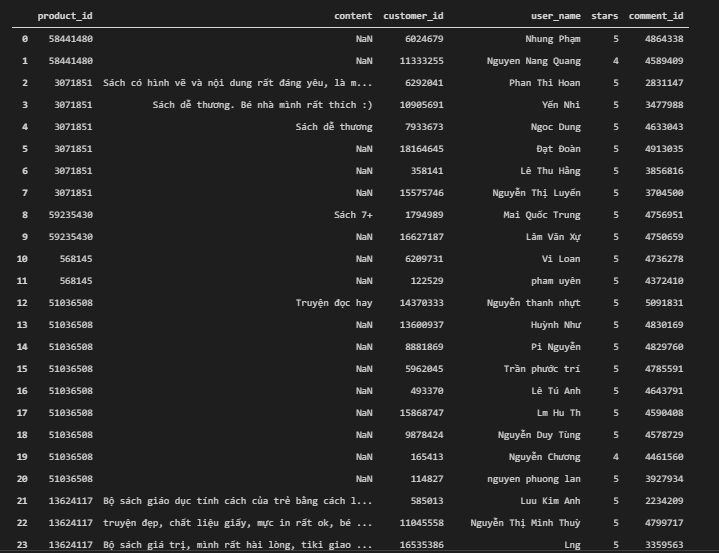
CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ VỚI DỮ LIỆU SÁCH

3.1. Dữ liệu thực nghiệm

* Dữ liệu thực nghiệm bao gồm 2 tập dữ liệu: 1 bộ dữ liệu thông tin sách và 1 bộ dữ liệu lưu trữ đánh giá bình luận của người dùng cho sách
* Dữ liệu sách gồm 10330 bản ghi sách, gồm các đặc trưng: id, tên, loại bìa, mô tả, giá bán, đánh giá trung bình, thể loại, thể loại con, tác giả, nhà xuất bản, số lượt bán



* Dữ liệu tương tác giữa người dùng và sách bao gồm 20849 bản ghi, tối thiểu số lượt đánh giá của 1 người là 10 để kết quả gợi ý chính xác hơn.



3.2. Xử lý dữ liệu

3.2.1. Tiền xử lý dữ liệu

* Đối với dữ liệu dạng text tiến hành xử lý, ví dụ một số trường trong dữ liệu sách như: phần mô tả sách, còn tồn tại các ký tự đặc biệt, thẻ html, từ dùng
* Làm sạch dữ liệu bằng cách loại bỏ một số dấu câu, kí tự không cần thiết (trừ dấu hỏi chấm), loại bỏ thẻ html
* Chuẩn hoá từ:Chuẩn hoá dạng ký tự viết hoa về ký tự viết thường.
* Loại bỏ từ dừng

3.2.2. Chia dữ liệu thực nghiệm

* Đối với bộ dữ liệu rating, chia dữ liệu thành 2 phần: 1 phần để huấn luyện mô hình (80%) và 1 phần để đánh giá (20%)

3.3. Thực nghiệm xây dựng mô hình tư vấn

3.3.1. Mô hình tư vấn sách theo kỹ thuật lọc dựa trên nội dung (thực hiện, kết quả, đánh giá)

3.3.1.1. Thực hiện tư vấn dựa trên đối sánh mẫu

3.3.2. Mô hình tư vấn sách theo kỹ thuật lọc cộng tác (thực hiện, kết quả, đánh giá)

3.4. So sánh và đánh giá các kết quả các phương pháp gợi ý

3.5. Kết luận chương

CHƯƠNG 4: ỨNG DỤNG MÔ HÌNH KHUYẾN NGHỊ SÁCH VÀO HỆ THỐNG BÁN HÀNG TRỰC TUYẾN

4.1. Mô tả hệ thống

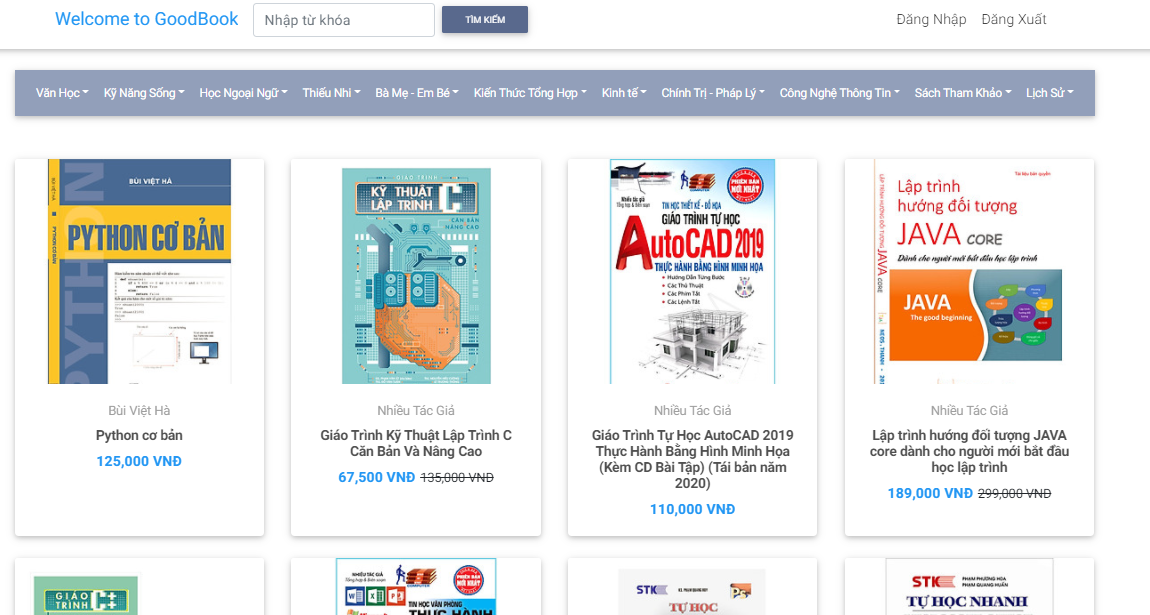
* Trang web được xây dựng với mục đích tìm hiểu, nghiên cứu hoạt động của hệ khuyến nghị
* Hệ thống giới thiệu và gợi ý sách cho phép khách hàng bất kỳ có thể tìm kiếm hoặc xem các sản phẩm theo thể loại, danh sách sách được ưa thích nhất …
* Hệ thống cho phép người dùng đăng nhập qua Facebook
* Trong quá trình xem sản phẩm, hệ thống sẽ gợi ý cho người dùng các loại sách trong quá trình chọn sản phẩm sử dụng kỹ thuật lọc theo nội dung và kỹ thuật lọc cộng tác, hiển thị các cuốn sách tương tự với cuốn sách mà người dùng đang xem sử dụng các thuộc tính tương tự về thể loại sách của sách đó.

4.2. Phân tích thiết kế hệ thống

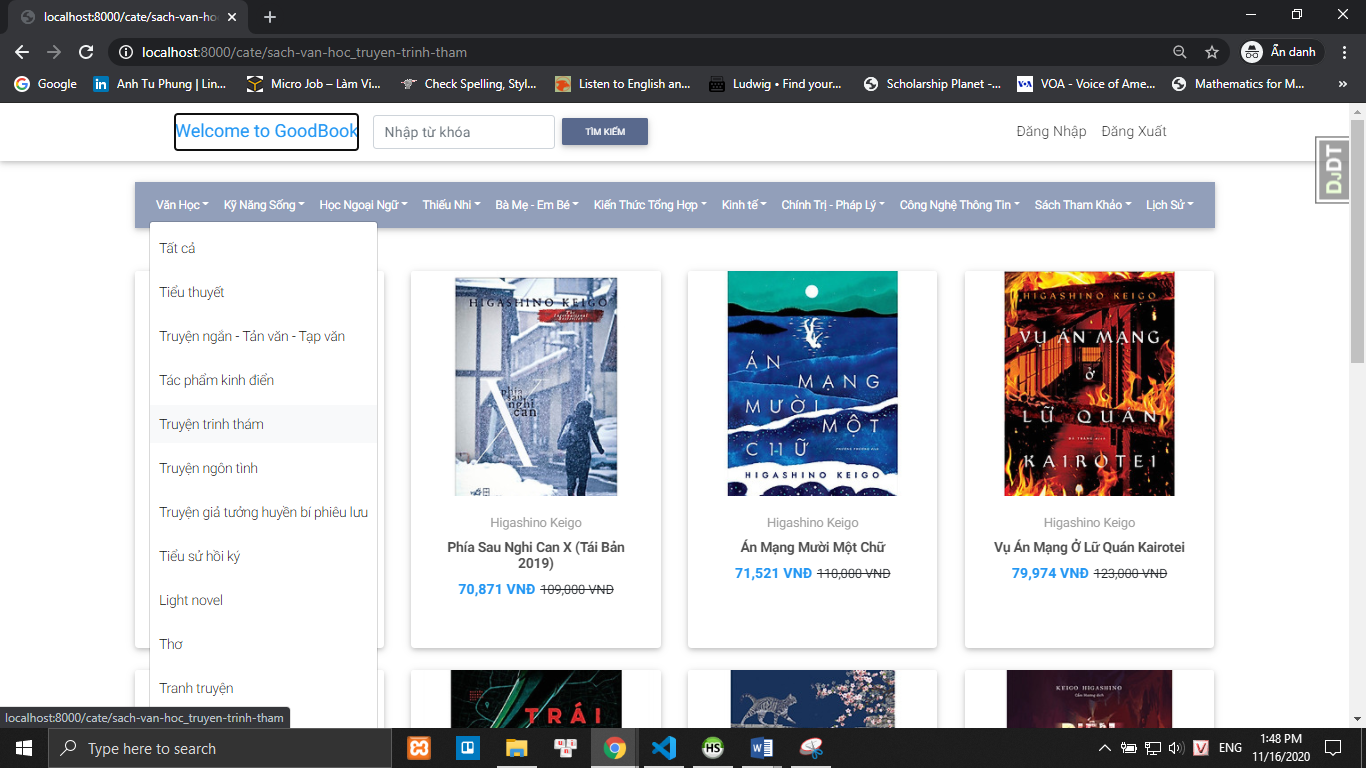
4.3. Thiết kế hệ thống (công nghệ, API, giao diện…)

Một số giao diện minh họa của hệ thống

* Trang chủ (mục đích hiển thị top bán chạy và gợi ý theo lọc cộng tác – tuy nhiên phần gợi ý lọc cộng tác vẫn còn chưa hoàn thiện)



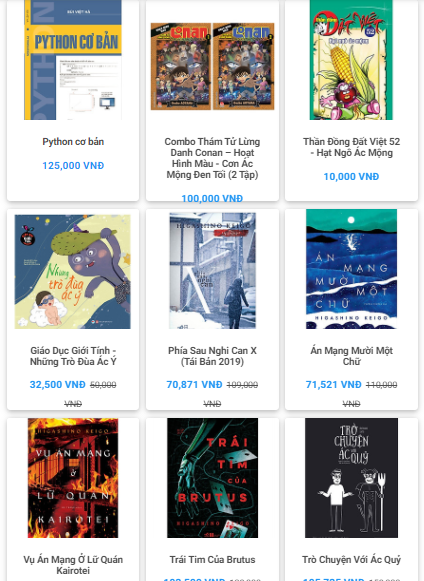
* Trang hiển thị sách theo phân loại.

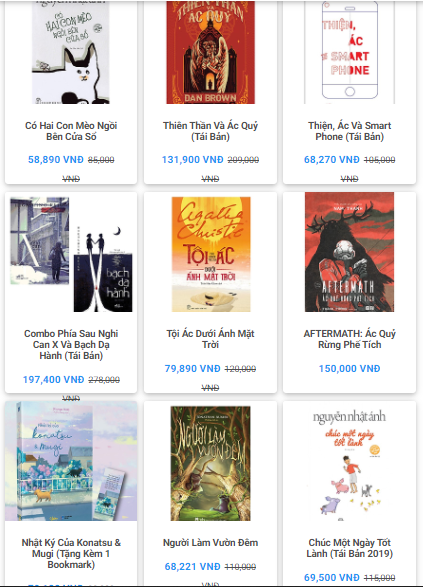


* Trang chi tiết sản phẩm: bao gồm thông tin chi tiết về sản phẩm và KẾT HỢP GỢI Ý THEO NỘI DUNG (gợi ý các sản phẩm có chỉ số cosine similarity theo 2 tiêu chí: Tiêu đề và Tóm tắt truyện lớn nhất)

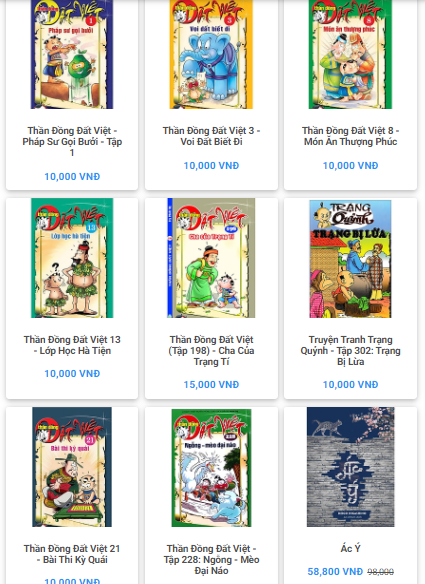


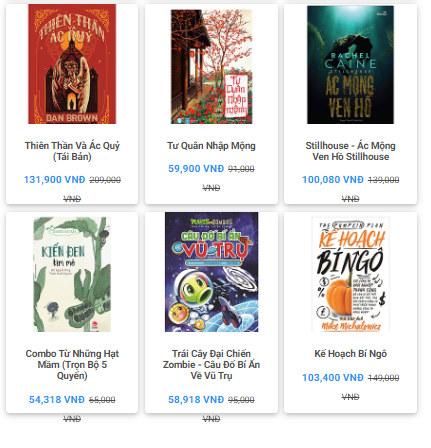
* Một số ví dụ gợi ý theo nội dung
* Quyển Ác Ý bên trên có nội dung gợi ý như sau:





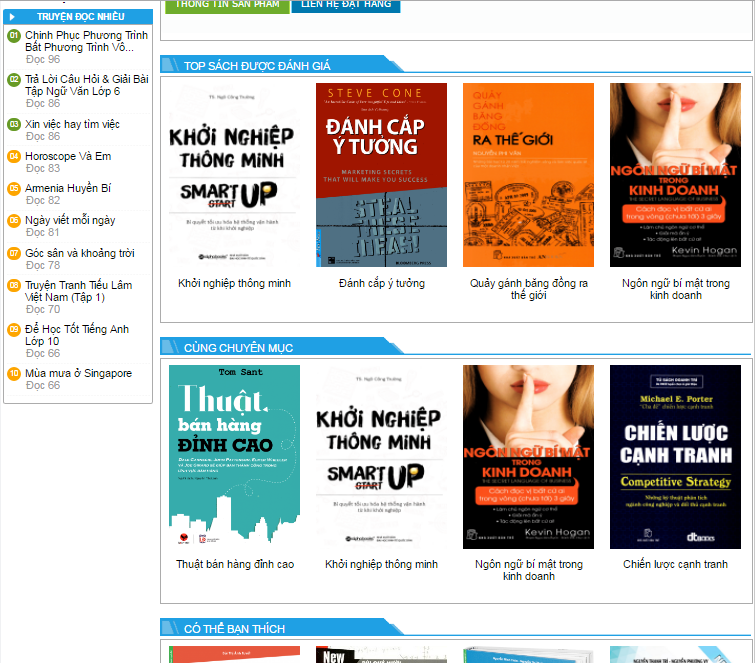
* Quyển Thần đồng Đất Việt 52 – Hạt Ngô Ác Mộng:





* Đánh giá: Vẫn còn mang tính chất ngẫu nhiên
* Giải pháp dự tính khắc phục: Loại bỏ hoàn toàn dấu Tiếng Việt rồi mới bắt đầu tính toán chỉ số cosine similarity.

Về giao diện, em mong muốn thiết kế được như hình dưới đây. Chia ra các mục

****