

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN



BÁO CÁO

ĐỒ ÁN MÔN HỌC KHAI THÁC DỮ LIỆU TRUYỀN THÔNG XÃ HỘI

**Đề tài: KHUYẾN NGHỊ THỰC PHẨM
CHO NGƯỜI DÙNG TRÊN SHOPEE**

Lớp: IE403.M21

Giảng viên hướng dẫn: ThS. Nguyễn Văn Kiệt

Sinh viên thực hiện: Hồ Đình Long – 18521022

Đỗ Hùng Dũng – 185200629

Nguyễn Thanh Tường Vi – 18521636

TP. Hồ Chí Minh, 06/2022

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

....., ngày.....tháng.....năm 2021

Người nhận xét

(Ký tên và ghi rõ họ tên)

LỜI CẢM ƠN.

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và quý Thầy Cô khoa Khoa học và Kỹ thuật thông tin đã đã tạo điều kiện, môi trường để chúng em có cơ hội nghiên cứu và thực hiện đồ án môn học Khai thác dữ liệu truyền thông xã hội. Đặc biệt, chúng em xin gửi lời cảm ơn đến Thầy Nguyễn Văn Kiệt (Giảng viên hướng dẫn môn học) đã tận tình giúp đỡ, chỉ bảo để nhóm chúng em có thể hoàn thành đề tài này.

Xuất phát từ mục đích học tập, cũng như tìm hiểu thêm về các phương pháp khuyến nghị, chúng em đã thực hiện đồ án với đề tài: **“Khuyến nghị thực phẩm cho người dùng trên shopee”** cho môn học Khai thác dữ liệu truyền thông xã hội.

Theo đó, phạm vi kiến thức là vô cùng rộng lớn. Vậy nên, với những giới hạn về kiến thức và thời gian, trong quá trình tìm hiểu và hoàn thành đề tài sẽ không tránh khỏi thiếu sót, chúng em rất mong nhận được nhận xét từ Thầy và lời góp ý từ Thầy chính là động lực để nhóm em có thể hoàn thiện hơn nữa những kiến thức của mình.

Chúng em xin chân thành cảm ơn Thầy!

MỤC LỤC

MỤC LỤC.....	4
DANH MỤC CÁC HÌNH ẢNH	5
Chương 1. TỔNG QUAN	6
1.1 Giới thiệu	6
1.2 Mô tả bài toán	6
1.3 Mô tả bộ dữ liệu.....	8
1.3.1 Giới thiệu về nguồn dữ liệu	8
1.3.2 Mô tả các thuộc tính trong từng file của bộ dữ liệu.....	9
Chương 2. XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU	11
2.1 Xử lý dữ liệu	11
2.1.1 Phương pháp lọc dựa trên nội dung.....	11
2.1.2 Phương pháp lọc cộng tác.....	11
2.2 Phân tích dữ liệu	12
Chương 3. CÀI ĐẶT THỰC NGHIỆM	14
3.1 Phương pháp lọc cộng tác.....	14
3.2 Phương pháp lọc dựa trên nội dung.....	14
Chương 4. ĐÁNH GIÁ KẾT QUẢ	15
4.1 Độ đo đánh giá.....	15
4.1.1 MSE (Mean-square error).....	15
4.1.2 RMSE (Root-mean-square error)	15
4.1.3 MAE (Mean absolute error).....	15
4.1.4 MEDAE (Median absolute error)	16
4.2 Chia dữ liệu train/test.....	16
4.3 Kết quả.....	16
4.3.1 Đối với phương pháp lọc cộng tác.....	16
4.3.2 Đối với phương pháp lọc dựa trên nội dung.....	17
Chương 5. KẾT LUẬN	18
TÀI LIỆU THAM KHẢO	19

DANH MỤC CÁC HÌNH ẢNH

Hình 1.2.1 Minh họa bài toán khuyến nghị thực phẩm cho người dùng shopee sử dụng phương pháp lọc cộng tác.....	7
Hình 1.2.2 Minh họa bài toán khuyến nghị thực phẩm cho người dùng shopee sử dụng phương pháp lọc dựa trên nội dung.....	8
Hình 2.1.2 Hình minh họa dữ liệu xếp hạng của người dùng với 88528 đánh giá.	12
Hình 2.2 Thống kê tỉ lệ đánh giá theo thang điểm 1 - 5.....	13
Hình 3.1 Quy trình thực nghiệm cho phương pháp lọc cộng tác	14

Chương 1. TỔNG QUAN

1.1 Giới thiệu

Hệ khuyến nghị là một trong những ứng dụng phổ biến trong khoa học dữ liệu. Nhiều công việc, đề tài đã được thực hiện về chủ đề này. Mỗi quan tâm và nhu cầu ngày càng cao vì tốc độ phát triển nhanh chóng của internet và vấn đề quá tải thông tin. Giúp người dùng đối phó với tình trạng này và cung cấp các đề xuất, nội dung và các dịch vụ cá nhân trở nên cần thiết đối với các doanh nghiệp và các nền tảng trực tuyến.

Xu hướng mua hàng thực phẩm online đã gia tăng mạnh trong thời gian gần đây. Nhất là khi chúng ta phải giãn cách xã hội do đại dịch Covid-19 thì số lượng người mua thực phẩm qua các trang thương mại điện tử tăng lên mạnh mẽ. Nhu cầu thực phẩm tăng cao, cùng với đó là lượng sản phẩm càng ngày càng đa dạng để đáp ứng nhu cầu khách hàng. Nhưng không phải mặt hàng nào người dùng cũng biết, cũng có thể tiếp cận được.

Hiểu được vấn đề này, nhóm chúng tôi đã quyết định tiến hành xây dựng mô hình khuyến nghị sử dụng phương pháp lọc cộng tác và lọc dựa trên nội dung để nhận biết thực phẩm từ lịch sử, sở thích của người dùng.

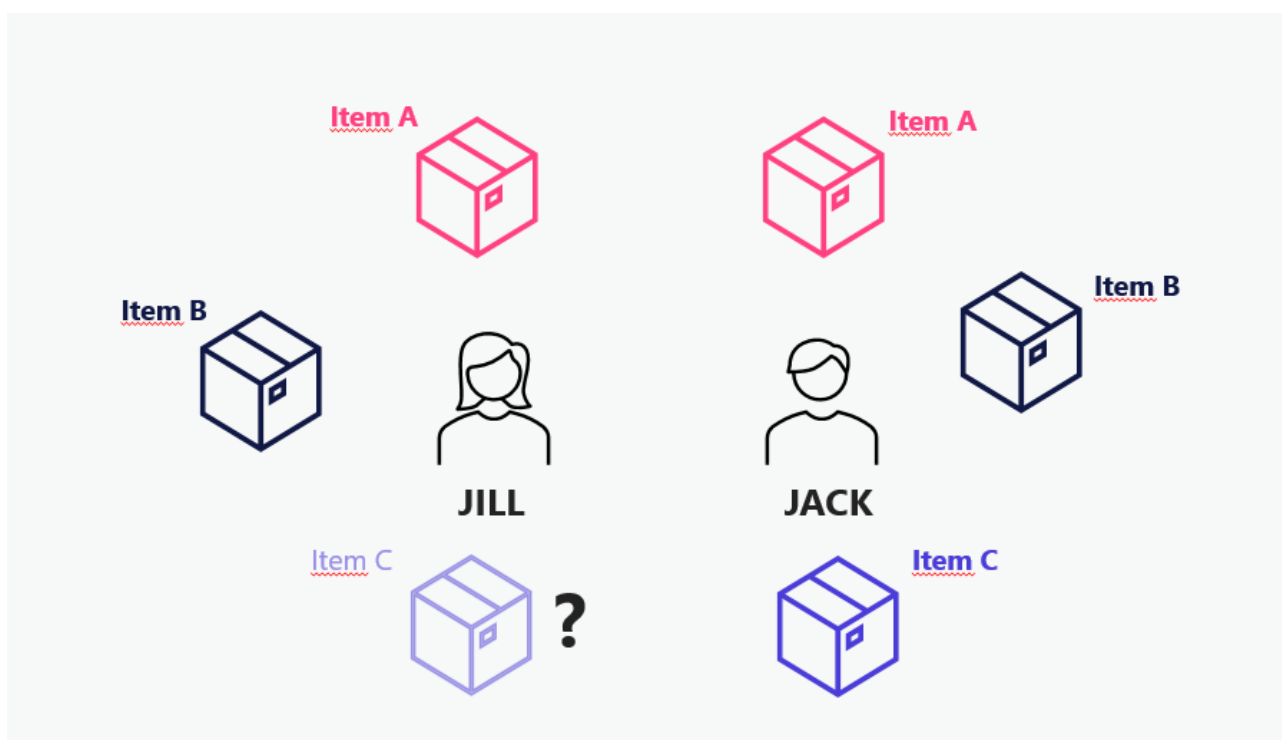
Lọc cộng tác là thuật toán thành công nhất trong lĩnh vực hệ thống đề xuất. Hệ thống khuyến nghị là một hệ thống thông minh có thể giúp người dùng tìm thấy các mặt hàng mà họ quan tâm. Nó sử dụng các kỹ thuật khai phá dữ liệu và lọc thông tin. Lọc cộng tác tạo ra các đề xuất cho người dùng dựa trên sở thích của những người láng giềng của họ. Nhưng nó có độ chính xác và khả năng mở rộng kém.

Lọc dựa trên nội dung là phương pháp đơn giản nhất trong các hệ thống khuyến nghị. Đặc điểm của phương pháp này là việc xây dựng mô hình cho mỗi người dùng không phụ thuộc vào các người dùng khác và có thể được coi như bài toán hồi quy hay phân lớp với dữ liệu huấn luyện là cặp dữ liệu (item profile, rating) mà người dùng đó đã đánh giá. item profile không phụ thuộc vào người dùng, nó thường phụ thuộc vào các đặc điểm mô tả của mặt hàng hoặc cũng có thể được xác định bằng cách yêu cầu người dùng gắn tag.

1.2 Mô tả bài toán

- Đầu vào: Dữ liệu người dùng đã mua sản phẩm và đánh giá của họ.
- Đầu ra: Khuyến nghị những sản phẩm phù hợp với sở thích của người dùng.

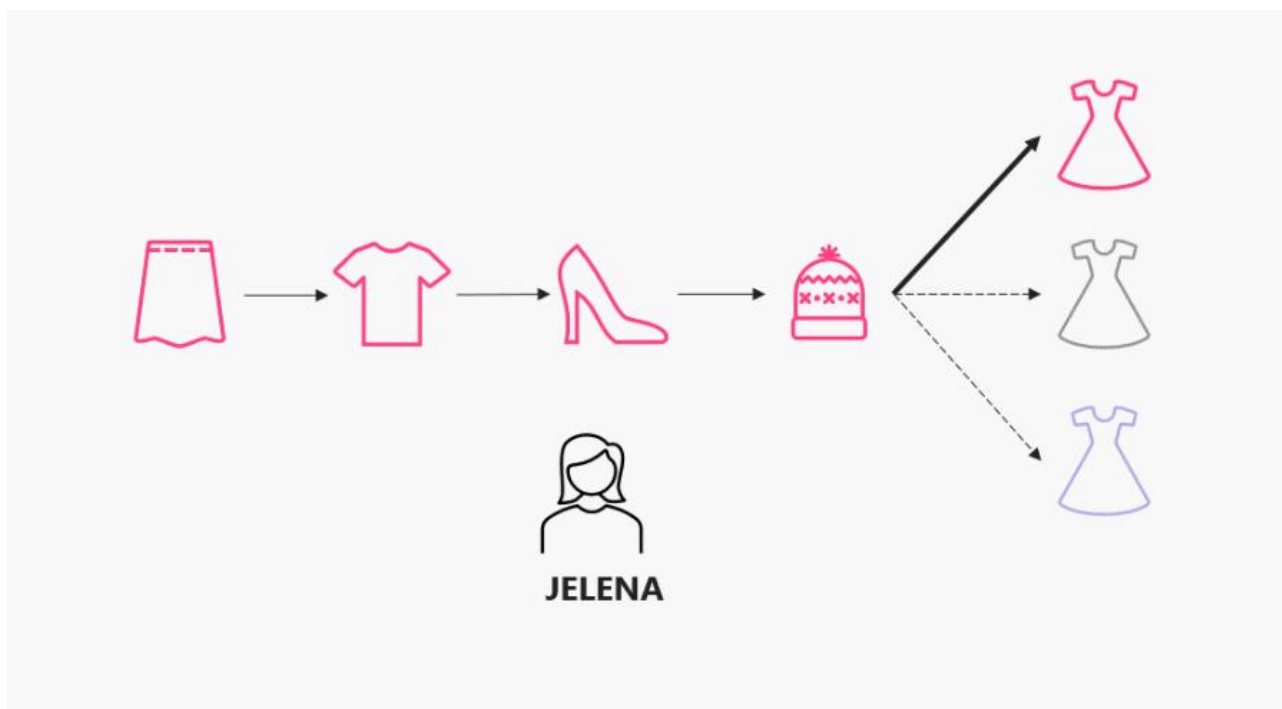
Cho ví dụ của 2 khách hàng - Jack và Jill (Hình 1.2.1). Nếu Jill mua các mặt hàng A và B, Jack đã mua các mặt hàng A, B và C, khi đó, Jack và Jill đã cùng mua và đánh giá 2 mặt hàng, có khả năng cao là Jill cũng thích mặt hàng C. Vì vậy, theo phương pháp lọc cộng tác, hệ thống sẽ khuyến nghị mặt hàng C cho Jill. Mặt khác, lọc cộng tác có một số vấn đề phổ biến, đó là khi một mặt hàng mới xuất hiện, nó không có tương tác. Điều này có nghĩa là nó sẽ không bao giờ xuất hiện trong danh sách khuyến nghị cho người dùng.



Hình 1.0.1 Minh họa bài toán khuyến nghị thực phẩm cho người dùng shopee sử dụng phương pháp lọc cộng tác.

Phương pháp lọc dựa trên nội dung dựa vào giả định rằng những gì khách hàng thích/mua trong quá khứ, có thể sẽ được thích/mua trong tương lai. Nó sử dụng thông tin meta của các thuộc tính và hồ sơ của các lựa chọn ưa thích của người dùng. Xét ví dụ của Jelena, người thường mua quần áo của cô ấy trực tuyến (Hình 1.2.2). Trong vài tháng qua, Jelena đã mua một số mặt hàng trực tuyến. Đầu tiên, cô mua cho mình một chiếc váy màu hồng, sau đó một vài ngày, cô mua một chiếc áo phông màu hồng, sau đó là giày cao gót màu hồng, rồi một chiếc mũ màu hồng. Rõ ràng là Jelena thích quần áo màu hồng, một đặc trưng phổ biến mà tất cả các mặt hàng cùng có. Rất có khả năng Jelena sẽ thích một chiếc đầm màu hồng hơn màu đen hay màu xanh lam. Vì vậy, theo cách tiếp cận dựa trên nội dung, hệ thống sẽ khuyến nghị một chiếc đầm màu hồng cho Jelen. Mặt khác, mô hình dựa trên nội dung cũng

gặp vấn đề tương tự lọc cộng tác. Khi người dùng mới xuất hiện, họ không có sản phẩm mua trước đó.



Hình 1.0.2 Minh họa bài toán khuyến nghị thực phẩm cho người dùng shopee sử dụng phương pháp lọc dựa trên nội dung.

1.3 Mô tả bộ dữ liệu

1.3.1 Giới thiệu về nguồn dữ liệu

Bộ dữ liệu chúng tôi sử dụng được thu thập từ trang thương mại điện tử Shopee, dùng công cụ Selenium. Dữ liệu là thông tin và đánh giá của người dùng đối với những sản phẩm mà họ đã mua. Với chủ đề là thực phẩm, chúng tôi thu thập dữ liệu dựa trên 28 từ khóa bao gồm thực phẩm chức năng, thực phẩm bảo vệ sức khỏe, màu thực phẩm, đồ ăn vặt, khô, bánh tráng, sữa, trứng, khô cá, khô bò, ruốc, xúc xích, cơm cháy, khô mực, sấy, rong biển, snack, chà bông, hạt dẻ, hạt điều, thịt heo, khô rim, khô gà, đậu, thịt xông khói, tai heo, khô cá, thực phẩm chay. Đây là những từ khóa phổ biến trên trang thương mại điện tử.

Bộ dữ liệu chứa 93037 dòng dữ liệu và 11 thuộc tính, bao gồm thông tin của 746 sản phẩm và sở thích của 111566 người dùng khác nhau được thu thập tại trang thương mại điện tử shopee.

Bộ dữ liệu chứa các file:

IE403.M21 – Khai thác dữ liệu truyền thông xã hội

- full.csv chứa danh sách tất cả các sản phẩm được người dùng mua với mã shop, tên shop, mã sản phẩm, thông tin sản phẩm giá sản phẩm, tổng số bình luận, rating trung bình của sản phẩm, rating của người dùng, mã người dùng và bình luận của người dùng tương ứng.
- ratings.csv chứa danh sách từng người dùng đã đánh giá cho từng sản phẩm và tỉ lệ đánh giá của họ.

1.3.2 Mô tả các thuộc tính trong từng file của bộ dữ liệu.

- File full.csv.

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
1	Shop_id	Số nguyên	Thể hiện mã shop
2	Shop_name	Chuỗi ký tự	Thể hiện tên shop
3	Prod_id	Số nguyên	Thể hiện mã sản phẩm
4	Prod_name	Chuỗi ký tự	Thể hiện tên sản phẩm
5	Description_Prod	Chuỗi ký tự	Thể hiện thông tin sản phẩm
6	Price	Số thực	Thể hiện giá sản phẩm
7	Total_coments	Số nguyên	Thể hiện tổng số bình luận
8	Avg_rating	Số thực	Thể hiện rating trung bình của sản phẩm
9	User_rating	Số thực	Thể hiện rating của người dùng
10	User_id	Số thực	Thể hiện mã của người dùng
11	User_review	Chuỗi ký tự	Thể hiện bình luận của người dùng

- File ratings.csv.

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
1	User_id	Số thực	Thể hiện mã của người dùng
2	Prod_id	Số nguyên	Thể hiện mã sản phẩm
3	User_rating	Số thực	Thể hiện xếp hạng của người dùng

Chương 2. XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU

2.1 Xử lý dữ liệu

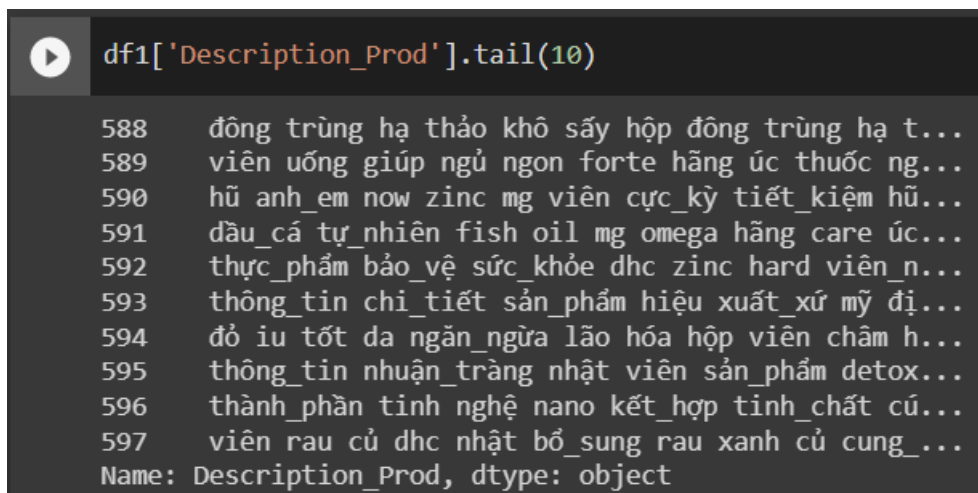
2.1.1 Phương pháp lọc dựa trên nội dung

Đối với phương pháp lọc dựa trên nội dung, nhóm sử dụng file full.csv và thực hiện với 2 cột là Description_Prod và Prod_Name trong file này.

Vì dữ liệu văn bản có nhiều vấn đề cần được xử lý, nên nhóm đã tiến hành tiền xử lý dữ liệu trước khi đưa vào mô hình khuyến nghị.

Những thao tác tiền xử lý bao gồm:

- Xóa hashtags, emojis, HTML, URL, stopwords, number.
- Biến đổi câu viết tắt thành câu hoàn chỉnh, loại bỏ spam.
- Chuẩn hóa dấu câu, tách từ, tạo negation, intensification.



```
df1['Description_Prod'].tail(10)
```

```
588  đông trùng hạ thảo khô sấy hộp đông trùng hạ t...
589  viên uống giúp ngủ ngon forte hãng úc thuốc ng...
590  hũ anh_em now zinc mg viên cực kỳ tiết_kiệm hũ...
591  đầu cá tự_nhiên fish oil mg omega hãng care úc...
592  thực_phẩm bảo_vệ sức_khỏe dhc zinc hard viên_n...
593  thông_tin chi_tiết sản_phẩm hiệu xuất_xứ mỹ đị...
594  đỏ iu tốt da ngăn_ngừa lão hóa hộp viên châm h...
595  thông_tin nhuận_tràng nhật viên sản_phẩm detox...
596  thành_phần tinh nghệ nano kết_hợp tinh_chất củ...
597  viên rau củ dhc nhật bổ_sung rau xanh củ cung...
Name: Description_Prod, dtype: object
```

Hình 2.1.1: Minh họa dữ liệu mô tả sản phẩm sau khi được tiền xử lý

2.1.2 Phương pháp lọc cộng tác

Đối với phương pháp lọc cộng tác, nhóm sử dụng file ratings.csv và thực hiện xử lý dữ liệu trên file này.

Vì số lượng đánh giá của người dùng chênh lệch nhau nhiều, việc sử dụng toàn bộ lượt đánh giá không đem lại kết quả chính xác cao hơn và còn gây lãng phí tài nguyên hệ thống,. Do đó, nhóm chỉ lấy những đánh giá của người dùng có từ 10 xếp hạng trở lên để phục vụ cho đồ án môn học này. Kết quả thu được thể hiện trong hình 2.1.2 bên dưới:

	User_id	Prod_id	User_Rating
0	9463240.0	15940877040	5.0
1	320629185.0	15940877040	5.0
2	774700227.0	15940877040	5.0
3	112133605.0	15940877040	5.0
4	91525994.0	15940877040	5.0
...
93026	576874288.0	13825733816	5.0
93030	229090458.0	6333550166	5.0
93031	121312494.0	6333550166	5.0
93032	98702403.0	6333550166	5.0
93033	98702403.0	6333550166	5.0

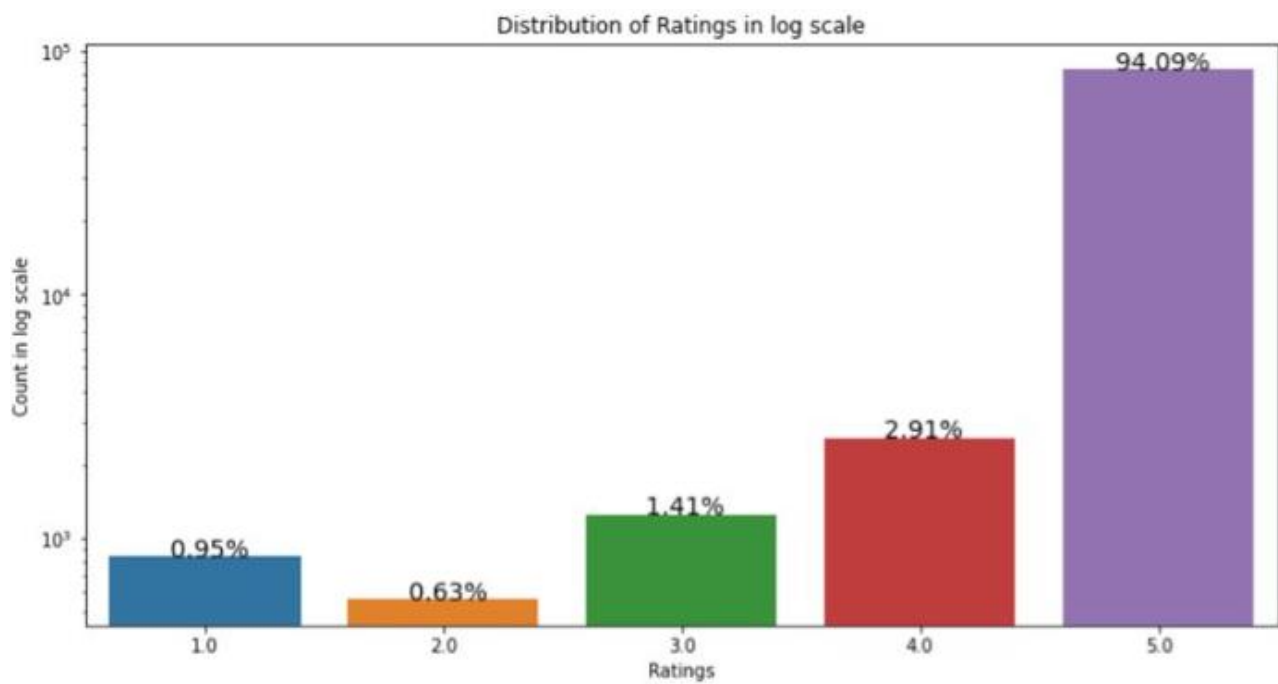
88528 rows × 3 columns

Hình 0.1.2 Hình minh họa dữ liệu xếp hạng của người dùng với 88528 đánh giá.

Dữ liệu của 2730 người dùng đánh giá cho 746 sản phẩm được dùng cho phương pháp lọc cộng tác.

2.2 Phân tích dữ liệu

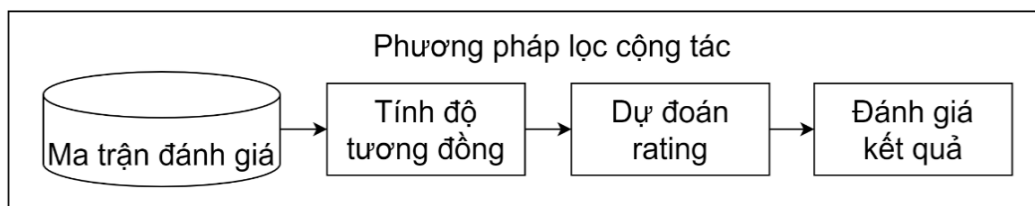
Đối với việc phân tích dữ liệu từ các lượt đánh giá, nhóm sử dụng thư viện seaborn và phương thức countplot() để vẽ biểu đồ thống kê tỉ lệ đánh giá từ người dùng. Sau khi thực hiện thống kê, kết quả cho thấy thang điểm 5 nhận được nhiều đánh giá nhất từ phía người dùng với 94.09%. Ngược lại, thang điểm 1 và 2 được rất ít người dùng sử dụng để đánh giá cho một sản phẩm với tỉ lệ thấp hơn 1%. Kết quả cho thấy, phần lớn người dùng đều ưa thích sản phẩm mà họ đã mua.



Hình 0.2 Thống kê tỉ lệ đánh giá theo thang điểm 1 - 5

Chương 3. CÀI ĐẶT THỰC NGHIỆM

3.1 Phương pháp lọc cộng tác



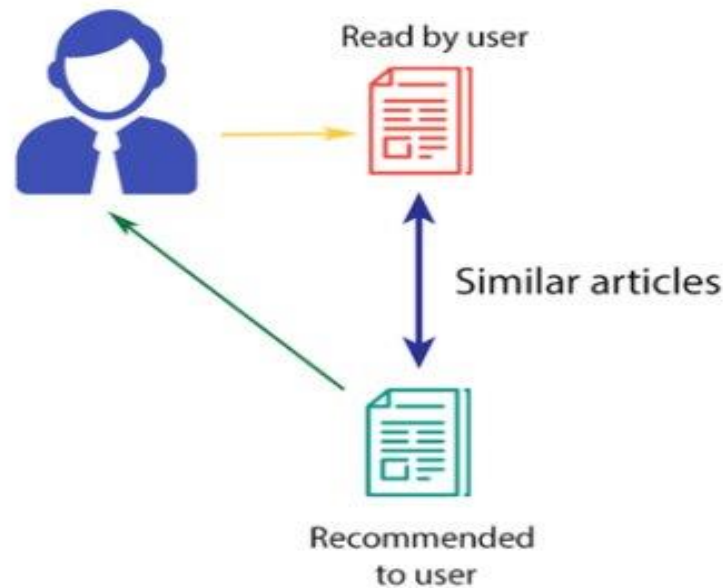
Hình 3.1 Quy trình thực nghiệm cho phương pháp lọc cộng tác

Cosine Similarity Độ tương tự cosin là một phép tính toán học ta biết sự giống nhau giữa hai vectơ A và B. Thực tế, ta đang tính cosin của góc theta giữa hai vectơ này. Hàm trả về giá trị giữa -1 (cho biết các vectơ hoàn toàn đối lập) đến 1 (cho biết cùng một vectơ). 0 cho thấy sự thiếu tương quan giữa các vectơ và các giá trị trung gian cho biết mức độ tương đồng trung gian. Hàm tương tự cosine tăng độ phức tạp tuyến tính khi tăng kích thước của A và B (lưu ý rằng A và B có cùng kích thước, n). Tích số chấm của A và B sẽ yêu cầu thêm $n + t$ phép tính nếu thêm t thêm giá trị vào A và B, và độ lớn của mỗi giá trị này cũng sẽ tăng tuyến tính. Cho đến nay, không có rắc rối trong tính toán phức tạp. Tuy nhiên, thuật toán thực hiện tính toán độ tương đồng cosine giữa mỗi cặp phim có thể có. Nếu có k sản phẩm, thì ta cần thực hiện các phép tính k^2 . Cuối cùng, nhóm xác định hệ khuyến nghị là một chức năng lấy thông tin đầu vào của người dùng, thực hiện tính điểm similarity cosine, sau đó đề xuất 10 sản phẩm hàng đầu tương tự nhất theo sở thích của người dùng.

3.2 Phương pháp lọc dựa trên nội dung

Không giống như lọc cộng tác chỉ dựa vào tương tác giữa người dùng với sản phẩm, lọc dựa trên nội dung sử dụng hồ sơ của người dùng và sản phẩm hoặc dựa vào nội dung, thuộc tính của những item tương tự như item mà người dùng đã chọn trong quá khứ. Ý tưởng của phương pháp dựa trên nội dung là cố gắng xây dựng một mô hình, dựa trên các thuộc tính giải thích các tương tác giữa người dùng và sản phẩm được quan sát.

Ví dụ: nếu người dùng xem một bộ phim hài có sự tham gia của diễn viên A, hệ thống sẽ giới thiệu cho họ những bộ phim cùng thể loại hoặc có sự tham gia của cùng một diễn viên hoặc cả hai. Đầu vào để xây dựng hệ thống đề xuất dựa trên nội dung là các thuộc tính phim.



Hình 3.2: Mô tả phương pháp lọc dựa trên nội dung

Chương 4. ĐÁNH GIÁ KẾT QUẢ

4.1 Độ đo đánh giá

Để đánh giá mức độ hiệu quả của các thuật toán, nhóm sử dụng 4 độ đo:

4.1.1 MSE (Mean-square error)

Sai số bình phương trung bình (MSE) là chênh lệch bình phương trung bình giữa các giá trị dự đoán và giá trị quan sát.

$$MSE = \frac{\sum (f_i - y_i)^2}{N}$$

4.1.2 RMSE (Root-mean-square error)

RMSE được tính bằng cách lấy căn bậc hai của sai số bình phương trung bình (MSE).

$$RMSE = \sqrt{\frac{\sum (f_i - y_i)^2}{N}}$$

4.1.3 MAE (Mean absolute error)

Trung bình của sai biệt tuyệt đối (MAE) dùng để đo độ lớn trung bình của các lỗi trong một tập hợp các dự đoán mà không cần xem xét hướng của chúng.

$$MAE = \frac{\sum abs(f_i - y_i)}{N}$$

4.1.4 MEDAE (Median absolute error)

Trung vị của sai biệt tuyệt đối.

$$MEDAE = Median(abs(f_i - y_i))$$

Hai độ đo MAE và MEDAE được dùng để đo lường sai biệt giữa giá trị thực tế (y_i) và tiên lượng của mô hình (f_i), giá trị tuyệt đối được dùng để tránh sai lầm trong trường hợp mô hình đồng thời có nguy cơ đánh giá quá cao và quá thấp, dẫn đến việc sai số > 0 và < 0 triệt tiêu lẫn nhau.

4.2 Chia dữ liệu train/test

Đối với phân chia tập dữ liệu train/test, nhóm chia được train/test cho phương pháp lọc cộng tác. Với tập 2730 người dùng, nhóm sử dụng pivot-table để tạo ma trận đánh giá bao gồm 1 cột User_id (hiển thị id của người dùng) và 1 hàng Prod_id (hiển thị id của sản phẩm tương ứng mà người dùng đánh giá). Thang điểm đánh giá của người dùng cho sản phẩm là xếp hạng từ 1 – 5.

Đối với mỗi người dùng nhóm tiến hành xóa 4 rating mà người dùng đã xếp hạng và thêm 4 rating đó vào tập test.

4.3 Kết quả

4.3.1 Đối với phương pháp lọc cộng tác

Kết quả thu được thông qua 4 độ đo được thể hiện dưới bảng sau:

Phương pháp	RMSE	MSE	MAE	MEDAE
User-based CF	4.9120	24.1278	4.8846	5.0
Top-k User-based CF	4.9119	24.1275	4.8845	5.0
Item-based CF	4.9111	24.1197	4.8836	5.0
Top-k Item-based CF	4.9111	24.1196	4.8836	5.0

Nhóm sử dụng 2 phương pháp user-based và item-based, ngoài ra thực hiện tính top-k user-based và top-k item-based (k=40) Kết quả đạt được tốt nhất khi sử dụng top-k item-based với độ lỗi thấp nhất.

4.3.2 Đối với phương pháp lọc dựa trên nội dung

Đối với lọc dựa trên nội dung, nhóm thực sử dụng trên 2 phương pháp là sử dụng 2 loại ma trận vector là TFIDF và Word2vec.

Ở cả 2 phương pháp, kết quả đề xuất đều tương đối tốt:

```
Product name users need to recommend to them:  
Hủ 200gr ruốc xấy ngon số 1 TÂY NINH dùng chung bánh  
18 500g Ruốc Sấy Tây Ninh Giò  
13 {Ruốc gà thịt tươi 100g}Hạnh Gà/ Ruốc Gà - chà bông...  
16 Đồ Ăn Nhanh Ruốc Cá Chép,Chà Bông Cá  
19 100gr CHÀ BÔNG TÔM/ RUỐC TÔM ĐÀ NẴNG - siêu  
11 Hủ 300G Ruốc Sấy Hành Phi Trụng
```

Hình 4.3.2.1: Kết quả dựa trên ma trận TFIDF

```
Product name users need to recommend to them:  
Hủ 200gr ruốc xấy ngon số 1 TÂY NINH dùng chung bánh  
13 500g Ruốc Sấy Tây Ninh Giò  
16 Quả dâu tây sấy dẻo  
8 100GR Bò khô xé sợi thơm ngon hàng 1  
18 100gr CHÀ BÔNG TÔM/ RUỐC TÔM ĐÀ NẴNG - siêu  
11 [HÀ NỘI] Xúc Xích Sụn Gà Cay Ăn Liền 1 gói 4 v...
```

Hình 4.3.2.2: Kết quả dựa trên ma trận Word2vec

Chương 5. KẾT LUẬN

Trong bài báo này, chúng tôi đã xây dựng được mô hình khuyến nghị cho người dùng dựa trên 2 phương pháp phổ biến là lọc cộng tác và lọc dựa trên nội dung. Trong đó mô hình mang lại kết quả tốt nhất với RMSE đạt 0.78.

Ưu điểm khi làm đề tài này là có hệ thống crawl dữ liệu dễ dàng trên trang thương mại điện tử, dữ liệu đa dạng, có đầy đủ các thuộc tính để sử dụng cho phương pháp lọc cộng tác và lọc dựa trên nội dung.

Về khó khăn, dữ liệu thu thập còn phức tạp và khó xử lý. Các đặc trưng của sản phẩm cần thực hiện nhiều bước tiền xử lý và chuẩn hóa. Shopee thường xuyên chặn khi crawl dữ liệu. Ngoài ra, nhóm gặp khó khăn khi sự trùng lặp người dùng đánh giá cho các sản phẩm khác nhau có khối lượng dữ liệu rất thấp dẫn đến việc hiệu suất mô hình chưa cao. Nhóm vẫn chưa tính được độ đo cho phương pháp lọc cộng tác.

Trong tương lai, chúng tôi muốn phát triển và cải thiện hệ thống khuyến nghị hơn nữa bằng cách sử dụng các phương pháp học sâu để huấn luyện trên bộ dữ liệu lớn hơn. Tiếp theo, chúng tôi muốn xây dựng API có sẵn để mọi người có thể dễ dàng sử dụng.

TÀI LIỆU THAM KHẢO

- [1] ("Recommendation systems in ecommerce: How it works? | BE-terna", 2022)
- [2] Pathak, A. (2019, March 9). Recommendation systems: User-based collaborative filtering using n nearest neighbors. Medium. Retrieved December 29, 2021, from <https://medium.com/sfu-big-data/recommendation-systems-user-based-collaborative-filtering-using-n-nearest-neighbors-bf7361dc24e0>
- [3] Rakesh4real. (2019, July 30). Evaluating recommendation systems -part 2. Medium. Retrieved December 29, 2021, from <https://medium.com/fnplus/evaluating-recommender-systems-with-python-code-ae0c370c90be>