

BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ NÔNG NGHIỆP VÀ PTNT
TRƯỜNG ĐẠI HỌC THỦY LỢI



NGUYỄN TRUNG TUYẾN
BÙI TRUNG QUỐC

XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM

ĐỒ ÁN TRÍ TUỆ NHÂN TẠO

HÀ NỘI, NĂM 2024

BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ NÔNG NGHIỆP VÀ PTNT
TRƯỜNG ĐẠI HỌC THỦY LỢI

NGUYỄN TRUNG TUYẾN

BÙI TRUNG QUỐC

XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM

Ngành: Trí tuệ nhân tạo và khoa học dữ liệu

Mã số: -----

NGƯỜI HƯỚNG DẪN TS. Tạ Quang Chiểu

HÀ NỘI, NĂM 2024



TRƯỜNG ĐẠI HỌC THỦY LỢI
KHOA CÔNG NGHỆ THÔNG TIN

BẢN TÓM TẮT ĐỀ CƯƠNG ĐỒ ÁN TTNT

TÊN ĐỀ TÀI: XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM

Sinh viên thực hiện: Nguyễn Trung Tuyền

Bùi Trung Quốc

Lớp: 63TTNT

Giáo viên hướng dẫn: TS. Tạ Quang Chiêu

TÓM TẮT ĐỀ TÀI

Hệ thống đề xuất phim đóng vai trò quan trọng trong việc cung cấp trải nghiệm tốt hơn cho người dùng bằng cách gợi ý những bộ phim phù hợp với sở thích cá nhân. Với sự gia tăng nhanh chóng của các bộ phim và nội dung số, người dùng thường gặp khó khăn trong việc chọn lựa phim để xem. Các phương pháp truyền thống như dựa trên đánh giá của người dùng hoặc đề xuất ngẫu nhiên không đáp ứng được nhu cầu cá nhân hóa cao của người dùng.

Ngày nay, với sự phát triển của công nghệ thông tin, học máy và xử lý ngôn ngữ tự nhiên, nhiều phương pháp mới đã được áp dụng để giải quyết vấn đề này. Các phương pháp như Collaborative Filtering (lọc cộng tác) và Content-Based Filtering (lọc dựa trên nội dung) đã chứng minh hiệu quả trong việc cá nhân hóa các đề xuất dựa trên dữ liệu người dùng và nội dung phim. Nghiên cứu và tìm hiểu các phương pháp này không chỉ giúp hiểu rõ hơn về cách chúng hoạt động mà còn mở ra cơ hội ứng dụng chúng vào hệ thống gợi ý phim của riêng mình.

Trong đề tài môn học này, chúng em sẽ nghiên cứu và triển khai hệ thống đề xuất phim dựa trên phương pháp lọc dựa trên nội dung và lọc cộng tác.

CÁC MỤC TIÊU CHÍNH

Mục tiêu 1: Thu thập được các dữ liệu về phim cần sử dụng cho đề tài.

Mục tiêu 2: Xây dựng được hệ thống đề xuất phim sử dụng phương pháp lọc cộng tác và phương pháp lọc dựa trên nội dung.

Mục tiêu 3: Hoàn thành báo cáo tổng kết của đề tài.

KẾT QUẢ DỰ KIẾN

Xây dựng được hệ thống đề xuất phim hoàn chỉnh và chạy được trên giao diện website

MỤC LỤC

MỞ ĐẦU	1
CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN	3
1.1. Lý do chọn đề tài	3
1.2. Mục tiêu đề tài	3
1.2.1. Mục tiêu tổng quát	3
1.2.2. Mục tiêu cụ thể	3
1.3. Nhiệm vụ nghiên cứu	4
1.4. Đối tượng và phạm vi nghiên cứu	4
1.5. Phương pháp nghiên cứu	4
CHƯƠNG 2. TÌM HIỂU LÝ THUYẾT CÁC PHƯƠNG PHÁP	5
2.1. Phương pháp lọc dựa trên nội dung	5
2.2. Phương pháp lọc cộng tác	7
CHƯƠNG 3. XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM	12
3.1. Thu thập dữ liệu	12
3.2. Tiền xử lý dữ liệu	12
3.3. Trực quan hóa dữ liệu	13
3.4. Xây dựng hệ thống gợi ý phim	14
CHƯƠNG 4. KẾT LUẬN	21
4.1. Kết quả đạt được	21
4.2. Hướng phát triển	21
Tài liệu tham khảo	22

MỞ ĐẦU

Trí tuệ nhân tạo (Artificial Intelligence - AI) ngày càng được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau của đời sống. AI có thể được áp dụng trong ngành giải trí để tạo ra các hệ thống đề xuất phim dựa trên nội dung, giúp người dùng dễ dàng tìm thấy những bộ phim phù hợp với sở thích cá nhân. Việc xây dựng hệ thống đề xuất phim giúp tối ưu hóa trải nghiệm người dùng, giảm thời gian tìm kiếm và tăng cường sự hài lòng của khách hàng. Đặc biệt, trong bối cảnh thị trường giải trí ngày càng phát triển và phong phú, việc có một hệ thống đề xuất hiệu quả trở nên vô cùng quan trọng.

Các hệ thống đề xuất phim hiện đại thường sử dụng các kỹ thuật học máy và xử lý ngôn ngữ tự nhiên để phân tích và đánh giá nội dung phim. Các thuật toán này có khả năng học từ dữ liệu đầu vào như mô tả phim, thể loại, đạo diễn, diễn viên và phản hồi của người xem để đưa ra những đề xuất chính xác. Chẳng hạn, một bộ phim có thể được gợi ý dựa trên sự tương đồng về nội dung với những bộ phim mà người dùng đã xem và đánh giá cao.

Trên thế giới cũng đã có một số nghiên cứu trong việc sử dụng học máy, học sâu cũng như một số phương pháp trong xử lý ngôn ngữ tự nhiên để xây dựng hệ thống đề xuất sản phẩm, phim, sách,... Năm 2010, tác giả Paolo Cremonesi, Paolo Frasca, Francesco Tagliati đã xây dựng hệ thống đề xuất phim dựa trên phương pháp lọc cộng tác¹ và cũng đã đạt được kết quả khả quan. Đến năm 2018, hệ thống đề xuất phim dựa trên học máy học sâu được nhóm tác giả He Wang, Xiang Zhao, Wei-Wei Tu xây dựng có tên là Neural Collaborative Filtering². Các hệ thống đề xuất phim sử dụng phương pháp lọc cộng tác và lọc nội dung đã được triển khai rộng rãi trên nhiều nền tảng xem phim trực tuyến như Netflix, Amazon Prime, Disney+... Những hệ thống này không chỉ giúp người dùng tìm thấy những bộ phim yêu thích mà còn giúp các nền tảng giải trí tối ưu hóa việc phân phối nội dung và thu hút thêm người dùng mới.

Bố cục của đề tài được trình bày trong 04 chương như sau:

Chương 1: Giới thiệu tổng quan

Chương 2: Tìm hiểu lý thuyết các phương pháp

Chương 3: Xây dựng hệ thống đề xuất phim

Chương 4: Kết luận

CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN

1.1. Lý do chọn đề tài

Hệ thống đề xuất phim đóng vai trò quan trọng trong việc cung cấp trải nghiệm tốt hơn cho người dùng bằng cách gợi ý những bộ phim phù hợp với sở thích cá nhân. Với sự gia tăng nhanh chóng của các bộ phim và nội dung số, người dùng thường gặp khó khăn trong việc chọn lựa phim để xem. Các phương pháp truyền thống như dựa trên đánh giá của người dùng hoặc đề xuất ngẫu nhiên không đáp ứng được nhu cầu cá nhân hóa cao của người dùng.

Ngày nay, với sự phát triển của công nghệ thông tin, học máy và xử lý ngôn ngữ tự nhiên, nhiều phương pháp mới đã được áp dụng để giải quyết vấn đề này. Các phương pháp như Collaborative Filtering (lọc cộng tác) và Content-Based Filtering (lọc dựa trên nội dung) đã chứng minh hiệu quả trong việc cá nhân hóa các đề xuất dựa trên dữ liệu người dùng và nội dung phim. Nghiên cứu và tìm hiểu các phương pháp này không chỉ giúp hiểu rõ hơn về cách chúng hoạt động mà còn mở ra cơ hội ứng dụng chúng vào hệ thống gợi ý phim của riêng mình.

Trong đề tài môn học này, chúng em sẽ nghiên cứu và triển khai hệ thống đề xuất phim dựa trên phương pháp lọc dựa trên nội dung và lọc cộng tác.

1.2. Mục tiêu đề tài

1.2.1. Mục tiêu tổng quát

Mục tiêu tổng quát của đề tài này là nghiên cứu, phát triển và ứng dụng các phương pháp lọc dựa trên nội dung và lọc cộng tác trong hệ thống đề xuất phim. Nhằm nâng cao chất lượng và độ chính xác của các đề xuất phim từ đó mang lại trải nghiệm tốt hơn cho người dùng.

1.2.2. Mục tiêu cụ thể

Mục tiêu 1: Thu thập được các dữ liệu về phim cần sử dụng cho đề tài.

Mục tiêu 2: Xây dựng được hệ thống đề xuất phim sử dụng phương pháp lọc cộng tác và phương pháp lọc dựa trên nội dung.

Mục tiêu 3: Hoàn thành báo cáo tổng kết của đề tài.

1.3. Nhiệm vụ nghiên cứu

- Nghiên cứu lý thuyết phương pháp lọc cộng tác và phương pháp lọc dựa trên nội dung.
- Thu thập và tiền xử lý dữ liệu để có thể sử dụng cho hệ thống đề xuất phim.
- Phát triển hệ thống đề xuất phim sử dụng phương pháp lọc cộng tác và phương pháp lọc dựa trên nội dung.

1.4. Đối tượng và phạm vi nghiên cứu

Nghiên cứu tập trung vào các bộ phim và dữ liệu mô tả nội dung của phim bao gồm các thông tin như mô tả cốt truyện, thể loại, tác giả,... Bộ dữ liệu được thu thập từ trang web đánh giá phim nổi tiếng là IMDB.

1.5. Phương pháp nghiên cứu

Phương pháp lý thuyết: Nghiên cứu các thuật toán của phương pháp lọc cộng tác và phương pháp lọc dựa trên nội dung.

Phương pháp thực nghiệm: Phát triển và triển khai hệ thống sử dụng phương pháp lọc cộng tác và phương pháp lọc dựa trên nội dung.

CHƯƠNG 2. TÌM HIỂU LÝ THUYẾT CÁC PHƯƠNG PHÁP

Hệ thống đề xuất được thiết kế để đưa ra các đề xuất sản phẩm cho người dùng dựa trên sở thích, hành vi và dữ liệu tương tác trước đó của họ. Mục tiêu chính của các hệ thống này nhằm cải thiện trải nghiệm của người dùng bằng cách giúp họ tìm thấy những sản phẩm họ có thể thích mà không cần phải tìm kiếm quá nhiều.

Hầu hết các ông lớn trong lĩnh vực công nghệ đều áp dụng các hệ thống đề xuất này dưới một số hình thức khác nhau. Ví dụ như: Amazon đã sử dụng hệ thống đề xuất sản phẩm cho khách hàng; Facebook cũng đã ứng dụng hệ thống đề xuất vào chức năng gợi ý kết bạn trên nền tảng của họ; Netflix đã sử dụng hệ thống đề xuất này để đề xuất các video và quyết định sẽ phát video nào tiếp theo khi mà người dùng sử dụng chế độ tự động phát.

Hệ thống đề xuất được chia làm hai loại chính: Hệ thống đề xuất dựa vào phương pháp lọc cộng tác và hệ thống đề xuất phim dựa trên nội dung.

2.1. Phương pháp lọc dựa trên nội dung

Phương pháp lọc dựa trên nội dung (Content-Based Filtering)³ là một trong những kỹ thuật phổ biến trong hệ thống đề xuất. Kỹ thuật này dựa vào các đặc trưng và thông tin của các mục được đề xuất, chẳng hạn như mô tả, thể loại, diễn viên cho phim hoặc từ khóa, nội dung cho bài viết.

2.1.1. Phương pháp TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) là một kỹ thuật phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên và khai thác dữ liệu văn bản. Nó giúp chuyển đổi văn bản thành các vector số để dễ dàng sử dụng trong các mô hình học máy. TF-IDF kết hợp hai chỉ số: tần suất xuất hiện của từ trong một tài liệu (TF) và tần suất nghịch đảo của tài liệu chứa từ đó (IDF).

2.1.1.1. Term Frequency (TF)

TF đo lường tần suất xuất hiện của một từ trong một tài liệu cụ thể. Nó được tính bằng cách lấy số lần xuất hiện của từ trong tài liệu chia cho tổng số từ trong tài liệu đó. Công thức tính TF như sau:

$$TF(t, d) = \frac{\text{số lần từ } t \text{ xuất hiện trong tài liệu } d}{\text{tổng số từ trong tài liệu } d}$$

2.1.1.2. Inverse Document Frequency (IDF)

IDF đo lường tầm quan trọng của một từ trong toàn bộ tập hợp tài liệu. Nó được tính bằng logarithm của tổng số tài liệu chia cho số tài liệu chứa từ đó. Công thức tính IDF như sau:

$$IDF(t) = \log\left(\frac{\text{tổng số tài liệu}}{\text{số tài liệu chứa từ } t}\right)$$

2.1.1.3. TF-IDF

TF-IDF là tích của TF và IDF, giúp xác định tầm quan trọng của một từ trong một tài liệu cụ thể so với toàn bộ tập hợp tài liệu. Công thức TF-IDF như sau:

$$TFIDF = TF(t, d) * IDF(t)$$

Những từ có TF-IDF cao là những từ xuất hiện nhiều trong một tài liệu này và xuất hiện ít ở các tài liệu khác. Điều này ngụ ý rằng những từ này đặc biệt quan trọng với nội dung của một tài liệu nào đó vì nó thường không xuất hiện ở các tài liệu khác. Những từ này có thể là một từ khóa quan trọng hoặc chủ đề đặc biệt của tài liệu.

2.1.2. Độ đo Cosine

Cosine Similarity đo lường góc giữa hai vector trong không gian đa chiều. Giá trị của nó dao động từ -1 đến 1:

- Nếu là 1 nghĩa là hai vector hoàn toàn tương tự nhau.
- Nếu là 0 nghĩa là hai vector không tương quan.
- Nếu là -1 nghĩa là hai vector hoàn toàn đối lập nhau.

Để sử dụng được phương pháp lọc dựa trên nội dung cần phải tính giá trị TF-IDF để chuyển các đặc trưng từ ngôn ngữ tự nhiên về dạng vector để máy có thể hiểu được. Sau khi đã thu được ma trận nhờ TF-IDF, chúng ta sẽ tính độ đo Cosine để đánh giá mức tương quan của item người dùng chọn với các item có trong data.

Độ đo cosine là một độ đo để đánh giá mức độ tương đồng của hai vector trong không gian vector. Độ đo này được tính theo công thức dưới đây:

$$\text{Cosine Similarity} = \frac{A.B}{||A||.||B||}$$

Trong đó:

$A.B$ là tích vô hướng của hai vector A và B

$||A||$ và $||B||$ là độ dài của vector A và B

2.1.3. Ưu và nhược điểm của phương pháp lọc dựa trên nội dung

2.1.3.1. Ưu điểm

Cá nhân hóa cao: Hệ thống lọc dựa trên nội dung phân tích và hiểu sở thích của từng người dùng một cách chi tiết, từ đó đưa ra các đề xuất phim phù hợp nhất với từng cá nhân.

Không phụ thuộc vào dữ liệu từ người khác: Vì chỉ dựa trên dữ liệu và lịch sử của người dùng cụ thể, hệ thống này không cần thông tin từ người dùng khác.

Giải quyết vấn đề dữ liệu khởi tạo: Đối với người dùng mới, hệ thống có thể nhanh chóng học từ các phim mà họ đã xem để đưa ra đề xuất thay vì phải đợi dữ liệu từ nhiều người dùng.

2.1.3.1. Nhược điểm

Đề xuất thiếu đa dạng: Hệ thống chỉ đề xuất các phim có tính chất tương tự nhau dẫn đến sự đơn điệu và thiếu phong phú.

Yêu cầu thông tin chi tiết về phim: Hệ thống cần có thông tin chi tiết đầy đủ về các phim để so sánh và đưa ra đề xuất.

2.2. Phương pháp lọc cộng tác

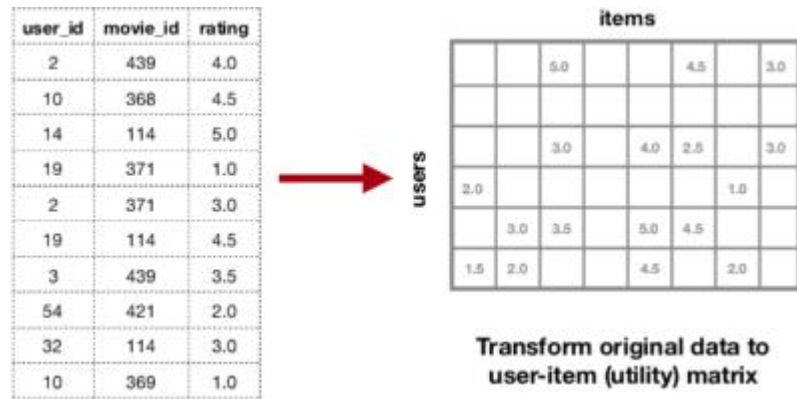
Phương pháp lọc cộng tác⁴ là một trong những kỹ thuật chính được sử dụng trong các hệ thống đề xuất dựa trên hành vi người dùng. Ý tưởng cơ bản của phương pháp này là dự đoán mức độ yêu thích của một user đối với một item dựa trên các user khác “gần giống” với user đang xét. Việc xác định độ “giống nhau” giữa các user có thể dựa vào mức độ quan tâm của các user này với các item khác mà hệ thống đã biết trong quá khứ.

Phương pháp này được chia làm ba phương pháp con khác: Lọc cộng tác dựa trên người dùng, lọc cộng tác dựa trên sản phẩm và Singular Value Decomposition(SVD).

Các phương pháp trên đều sử dụng Utility Matrix, được tạo thành từ dữ liệu. Utility Matrix là một ma trận biểu diễn các đánh giá của người dùng đối với các mục item. Đây là nền tảng của các phương pháp lọc cộng tác trong hệ thống đề xuất, nó giúp hệ thống đề xuất dự đoán rating của người dùng dựa trên dữ liệu hiện có.

Cấu trúc của Utility Matrix là một ma trận hai chiều với:

- Hàng (rows) và cột (columns): Tương ứng với người dùng và các phim.
- Giá trị: Tương ứng với đánh giá mà người dùng đã đưa ra cho một phim cụ thể. Nếu người dùng chưa đánh giá phim nào đó thì giá trị tương ứng của nó là 0.



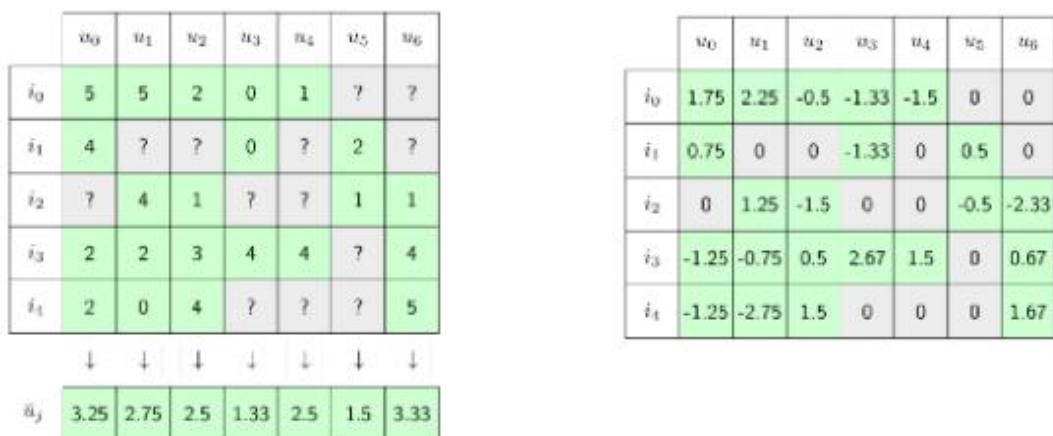
Hình 2.1. Minh họa ma trận Utility

2.2.1. Phương pháp lọc cộng tác dựa trên user (User-user Collaborative Filtering)

User - User Collaborative Filtering (CF) là một kỹ thuật đề xuất dựa trên sự tương đồng giữa các người dùng. Trong phương pháp này, chúng ta dự đoán đánh giá của người dùng dựa trên đánh giá của những người dùng tương tự. Đây là các bước cơ bản để thực hiện User-User CF:

Bước 1: Chuẩn hóa ma trận Utility

Tính trung bình đánh giá của mỗi người dùng, sau đó thực hiện chuẩn hóa đánh giá bằng cách trừ trung bình đánh giá của mỗi người dùng từ các đánh giá tương ứng. Điều này giúp loại bỏ sự thiên vị của mỗi người.



Hình 2.2. Hình ảnh ma trận Utility trước và sau khi chuẩn hóa

Bước 2: Tính độ tương đồng giữa các người dùng

Độ tương đồng giữa hai người dùng được tính bằng cách sử dụng độ đo Cosine.

Chi tiết về độ đo Cosine được nhắc đến trong đề tài này tại *phần 2.1.2*.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
u_0	1	0.83	-0.58	-0.79	-0.82	0.2	-0.38
u_1	0.83	1	-0.87	-0.40	-0.55	-0.23	-0.71
u_2	-0.58	-0.87	1	0.27	0.32	0.47	0.96
u_3	-0.79	-0.40	0.27	1	0.87	-0.29	0.18
u_4	-0.82	-0.55	0.32	0.87	1	0	0.16
u_5	0.2	-0.23	0.47	-0.29	0	1	0.56
u_6	-0.38	-0.71	0.96	0.18	0.16	0.56	1

Hình 2.3. Hình ảnh ma trận Utility sau khi đã tính độ tương đồng

Bước 3: Dự đoán giá trị còn thiếu của ma trận

Sau khi đã tính được độ tương đồng giữa các người dùng, công việc tiếp theo là dự đoán giá trị các đánh giá còn thiếu của ma trận bằng các bước sau đây:

- Tìm những người dùng đã đánh giá các mục cần dự đoán: Lấy tất cả người dùng đã đánh giá mục đó.
- Tính độ tương đồng giữa người dùng hiện tại và các người dùng đã đánh giá mục đó bằng cách sử dụng độ đo Cosine đã nêu ở trên.
- Chọn k người dùng có độ tương đồng cao nhất.
- Dự đoán giá trị còn thiếu bằng cách lấy trung bình có trọng số của các đánh giá từ những người dùng tương tự đã chọn ở trên theo công thức:

$$\hat{y}_{i,u} = \frac{\sum_{u_j \in \mathcal{N}(u,i)} \bar{y}_{i,u_j} \text{sim}(u, u_j)}{\sum_{u_j \in \mathcal{N}(u,i)} |\text{sim}(u, u_j)|}$$

Với: $\mathcal{N}(u,i)$: tập hợp những người dùng u_j đã đánh giá mục i và có độ tương đồng với người dùng u .

\bar{y}_{i,u_i} là đánh giá của người dùng u_j cho mục i .

$\text{Sim}(u,u_i)$ là độ tương đồng giữa người dùng u và u_i .

2.2.2. Phương pháp lọc cộng tác dựa trên item (Item - Item Collaborative Filtering)

Item-Item Collaborative Filtering là một kỹ thuật đề xuất dựa trên sự tương đồng giữa các phim. Phương pháp này thường được sử dụng nhiều hơn trong thực tế do một số ưu điểm so với User-User CF.

Ưu điểm của IF so với CF:

1. Số lượng phim ít hơn số lượng người dùng: Điều này làm cho ma trận nhỏ hơn, dễ lưu trữ và tính toán hơn.
2. Ma trận Utility thường ít sparse hơn theo hàng: Mỗi phim được nhiều người dùng đánh giá, do đó giá trị trung bình của mỗi hàng ít thay đổi hơn khi có thêm đánh giá mới.
3. Dễ cập nhật hơn: Vì số lượng phim ít hơn nên việc ma trận bị thay đổi thì cũng sẽ dễ cập nhật hơn.

Về cách hoạt động hay tính toán của phương pháp IF này cũng bao gồm các bước tương tự như phương pháp CF được nêu ở trên. Nhưng khác một điểm đó là thay vì thực hiện tính trung bình các đánh giá của mỗi người dùng thì IF sẽ tính trung bình các đánh giá của mỗi phim. Các bước sau đó làm tương tự CF.

2.2.3. Singular Value Decomposition(SVD)

SVD còn được gọi là một kỹ thuật trong hệ thống đề xuất. Được dùng để phân ra ma trận có thể áp dụng trên ma trận R để biểu diễn nó dưới dạng tích của ba ma trận khác. Cụ thể:

$$R = UEV^t$$

Trong đó:

R: Ma trận người dùng - phim, chứa các đánh giá của người dùng cho từng phim. Kích thước của ma trận này là $m \times n$, với m là số người dùng và n là số phim.

U: Ma trận $m \times k$, với k là số yếu tố tiềm ẩn đại diện cho người dùng

E: Ma trận đường chéo $k \times k$ chứa các giá trị kỳ dị

V^t: Ma trận $k \times n$, với k là số yếu tố tiềm ẩn đại diện cho các phim.

Quá trình phân rã ma trận hoạt động theo các bước sau:

- Xây dựng ma trận đánh giá R: Giả sử ở đây có ma trận R biểu diễn đánh giá của người dùng cho các phim. Nếu người dùng chưa đánh giá một phim nào đó thì giá trị tương ứng sẽ là 0 hoặc NaN.

- Áp dụng SVD lên ma trận R: SVD sẽ phân ra R thành 3 ma trận con như ở trên. Từ đó giúp xác định các yếu tố tiềm ẩn ảnh hưởng đến sự tương tác của người dùng với các phim liên quan.

- Tái tạo lại ma trận R: Bằng cách nhân 3 ma trận con lại để khôi phục lại ma trận R ban đầu nhưng đã được bổ sung các giá trị đánh giá bị thiếu.

Sau khi mà đã tính toán được ma trận R đầy đủ các giá trị, thì sẽ lấy các đánh giá lớn nhất của các phim vừa dự đoán ra. Từ đó sẽ đề xuất phim cho người dùng với nội dung phù hợp nhất.

CHƯƠNG 3. XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM

3.1. Thu thập dữ liệu

IMDb (Internet Movie Database)⁵ được thành lập năm 1990 bởi Col Needham là một trong những trang web hàng đầu và phổ biến nhất thế giới về cơ sở dữ liệu phim. IMDb. Tại trang web này cung cấp thông tin chi tiết về các bộ phim như là diễn viên, đạo diễn, nhà sản xuất, mô tả phim,...

Trong đề tài này, nhóm thực hiện thu thập dữ liệu về phim tại website IMDb: <https://www.imdb.com> bằng thư viện BeautifulSoup.

Các bước thu thập dữ liệu:

Bước 1: Cài đặt thư viện BeautifulSoup

Bước 2: Xem mã nguồn HTML của trang web IMDb để có thể biết được các thông tin cần lấy nằm trong thẻ HTML nào.

Bước 3: Viết code để tự động thu thập và lưu dữ liệu thô thu được vào file excel.

3.2. Tiền xử lý dữ liệu

Sau khi hoàn thành bước thu thập dữ liệu trên trang web IMDb, dữ liệu thô nhóm thu được gồm thông tin của hơn 5200 bộ phim, mỗi bộ phim chứa thông tin về: tên phim (Title), năm phát hành (Year), thời lượng phim (Runtime), thể loại (Genre), điểm đánh giá (Rating), mô tả phim (Description), đạo diễn (Director), diễn viên chính (Stars), số lượt votes (Votes), liên kết poster của phim (Img_link).

Nhưng từ dữ liệu đó chưa thể đem vào để xây dựng hệ thống đề xuất phim được do dữ liệu chưa được đồng nhất. Ví dụ như cột Year, có những phim năm phát hành là “1990”, một số phim lại là “1990 TV Movie”. Chính vì vậy mà cần phải tiền xử lý dữ liệu để loại bỏ các kí tự dư thừa, như ví dụ trên sẽ loại bỏ cụm từ “TV Movie” mà chỉ lấy “1990”.

	Title	Year	Runtime	Genre	Rating	Description	Director	Stars	Votes	Img_link
0	Sign of the Lion	1962	103 min	Drama	7.2	A French-American in Paris lives by sponging o...	Éric Rohmer	Éric Rohmer, Jess Hahn, Michèle Girardon, Van ...	2352.0	amazon.com/images/M/MV58OGZkMj...
1	Bogowie	2014	120 min	Biography, Drama	7.6	The early career of cardiac surgeon Zbigniew R...	Lukasz Palkowski	Lukasz Palkowski, Tomasz Kot, Piotr Glowacki, ...	7704.0	amazon.com/images/M/MV58MTg2YW...
2	WR: Mysteries of the Organism	1971	84 min	Comedy, Drama, Fantasy	6.7	An homage to the work of psychologist Wilhelm ...	Dusan Makavejev	Dusan Makavejev, Milena Dravic, Ivica Vidovic...	5423.0	amazon.com/images/M/MV58OTgyYW...
3	Macross Plus Movie Edition	1995	115 min	Animation, Action, Drama	7.5	Two rival test pilots strive to be the best th...	Shōji Kawamori	Shōji Kawamori, Yōji Moriyama, Shin'ichirō Wat...	689.0	amazon.com/images/M/MV58ODE5MD...
4	Visions of Light	1992	92 min	Documentary, History	7.7	Cameramen and women discuss the craft and art ...	Arnold Glassman	Arnold Glassman, Todd McCarthy, Stuart Samuels...	3336.0	amazon.com/images/M/MV58MTIwNj...
...
5229	Arthur the King	2024	107 min	Adventure, Drama	7.0	An adventure racer adopts a stray dog named Ar...	Simon Cellan Jones	Simon Cellan Jones, Mark Wahlberg, Simu Liu, J...	9165.0	amazon.com/images/M/MV58NzZzM2MD...
5230	Boy Kills World	2023	111 min	Action, Crime, Thriller	6.8	A fever dream action film that follows Boy, a ...	Moritz Mohr	Moritz Mohr, Bill Skarsgård, Jessica Rothe, M...	3014.0	amazon.com/images/M/MV58NmYyM2...
5231	Tuyen Dep Trai	2025	119 min	Drama	10.0	A story about a duck named duck	Tuyen	NaN	NaN	https://scontent.fhan18-1.fna.fbcdn.net/v/t13...
5232	Quoc Beautiful	2025	99	Action	8.0	Duck Goose Duck	Quoc	NaN	NaN	https://scontent.fhan18-1.fna.fbcdn.net/v/t1.6...
5233	New Man Haha	2026	100 min	Drama	7.0	dog, duck	NaN	NaN	NaN	https://scontent.fhan18-1.fna.fbcdn.net/v/t1.6...

5234 rows × 10 columns

Hình 3.1. Thông tin của dữ liệu thô

Tương tự như vậy, chi tiết tiền xử lý dữ liệu của mỗi cột như sau:

- Cột “Title”: loại bỏ các phim trùng lặp.
- Cột “Year”: loại bỏ các kí tự dư thừa, chỉ giữ lại các kí tự có dạng là năm.
- Cột “Runtime”: loại bỏ cụm từ “ min”.
- Cột “Genre”: do mỗi phim có số lượng thể loại khác nhau, nên xử lý dữ liệu chỉ giữ lại 3 thể loại đầu tiên nếu phim nào có nhiều hơn 3 thể loại.
- Cột “Stars”: cũng tương tự như cột Genre, nên cũng chỉ giữ lại 3 diễn viên chính đầu tiên của mỗi phim.

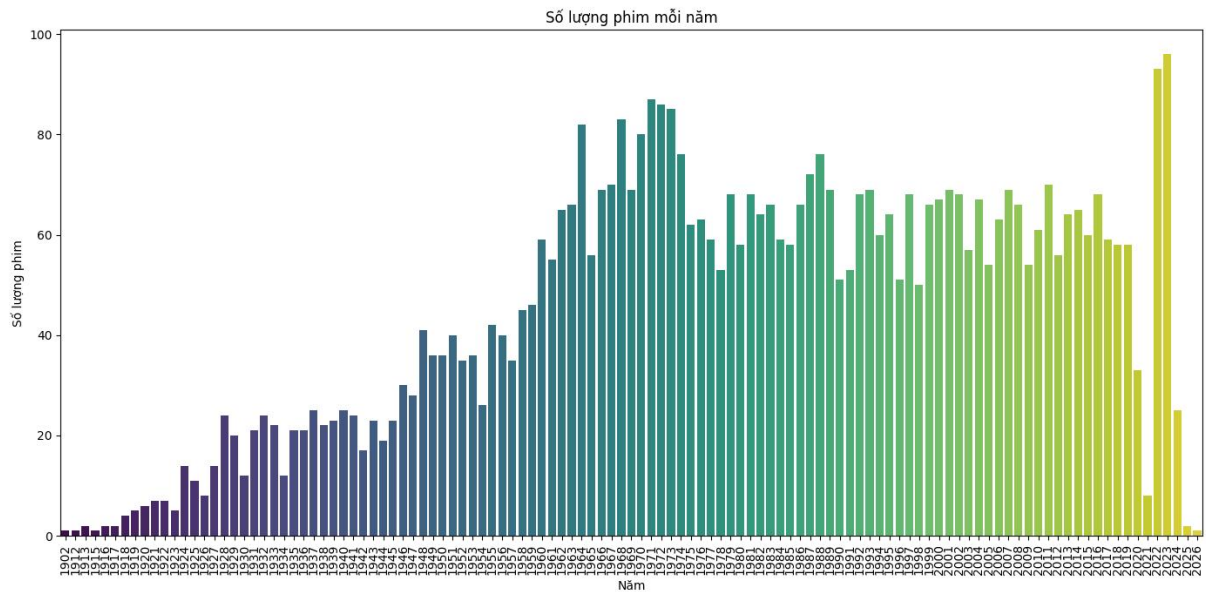
Sau khi đã hoàn thành các bước tiền xử lý dữ liệu trên, dữ liệu hoàn chỉnh thu được như hình dưới đây:

	Title	Year	Runtime	Genre	Rating	Description	Director	Stars	Votes	Img_link	Genre1	Genre2	Genre3	Stars1	Stars2	Stars3
0	Sign of the Lion	1962	103	Drama	7.2	A French-American in Paris lives by sponging o...	Éric Rohmer	Éric Rohmer, Jess Hahn, Michèle Girardon, Van ...	2352.0	amazon.com/images/M/MV58OGZkMj...	Drama	None	None	Éric Rohmer	Jess Hahn	Michèle Girardon
1	Bogowie	2014	120	Biography, Drama	7.6	The early career of cardiac surgeon Zbigniew R...	Lukasz Palkowski	Lukasz Palkowski, Tomasz Kot, Piotr Glowacki, ...	7704.0	amazon.com/images/M/MV58MTg2YW...	Biography	Drama	None	Lukasz Palkowski	Tomasz Kot	Piotr Glowacki
2	WR: Mysteries of the Organism	1971	84	Comedy, Drama, Fantasy	6.7	An homage to the work of psychologist Wilhelm ...	Dusan Makavejev	Dusan Makavejev, Milena Dravic, Ivica Vidovic...	5423.0	amazon.com/images/M/MV58OTgyYW...	Comedy	Drama	Fantasy	Dusan Makavejev	Milena Dravic	Ivica Vidovic
3	Macross Plus Movie Edition	1995	115	Animation, Action, Drama	7.5	Two rival test pilots strive to be the best th...	Shōji Kawamori	Shōji Kawamori, Yōji Moriyama, Shin'ichirō Wat...	689.0	amazon.com/images/M/MV58ODE5MD...	Animation	Action	Drama	Shōji Kawamori	Yōji Moriyama	Shin'ichirō Watanabe
4	Visions of Light	1992	92	Documentary, History	7.7	Cameramen and women discuss the craft and art ...	Arnold Glassman	Arnold Glassman, Todd McCarthy, Stuart Samuels...	3336.0	amazon.com/images/M/MV58MTIwNj...	Documentary	History	None	Arnold Glassman	Todd McCarthy	Stuart Samuels

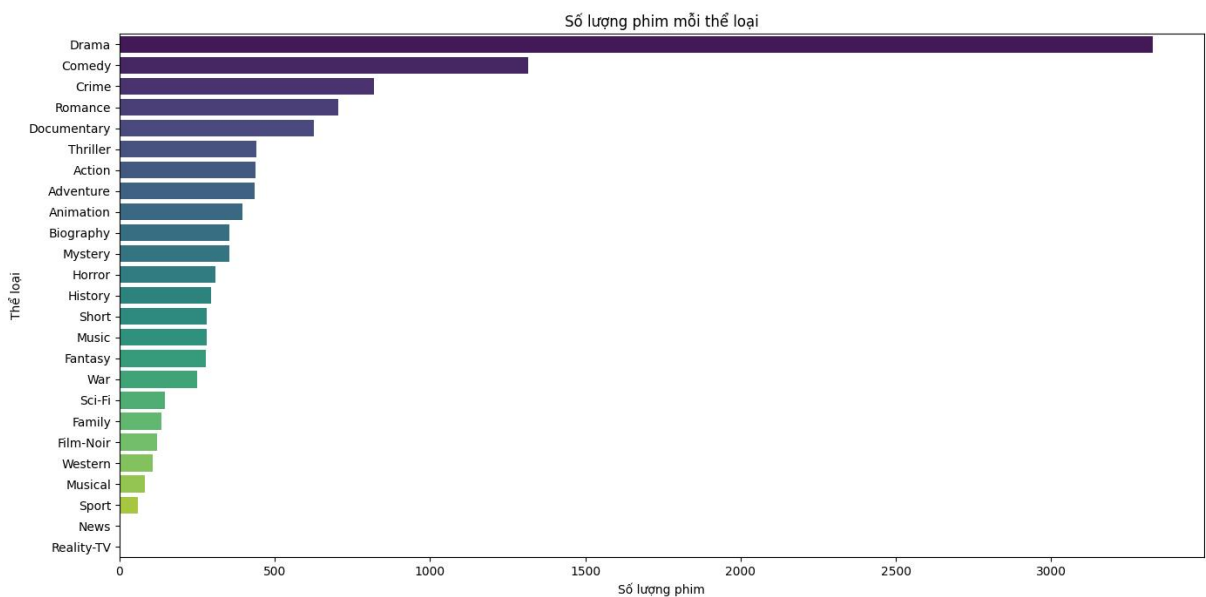
Hình 3.2. Thông tin của dữ liệu hoàn chỉnh

3.3. Trực quan hóa dữ liệu

Khi đã có dữ liệu hoàn chỉnh, nhóm thực hiện trực quan số lượng phim theo mỗi năm, theo thể loại, các hình dưới đây thể hiện điều đó:



Hình 3.3. Biểu đồ cột thể hiện số lượng phim theo năm



Hình 3.4. Biểu đồ cột thể hiện số lượng phim theo thể loại

Quan sát 2 hình 3.3 và hình 3.4, những năm 2022 và 2023 là có nhiều phim mới được phát hành nhất. Thể loại có nhiều phim nhất đó chính là Drama và ít nhất là thể loại Reality-TV.

3.4. Xây dựng hệ thống gợi ý phim

3.4.1. Xác định cảm xúc của mô tả phim

Để có thể cung cấp cho người dùng một đề xuất hợp lý nhất, nhóm thực hiện việc xác định cảm xúc của mô tả của phim. Trong đề tài này, nhóm sử dụng mô-đun

SentimentIntensityAnalyzer trong thư viện nltk để thực hiện xác định xem mạch cảm xúc của phim là tích cực, trung bình hay là tiêu cực khi xác định được cảm xúc của mô tả phim. Sau khi có được cảm xúc của mỗi phim, nhóm kết hợp so sánh với điểm đánh giá của phim để đề xuất với người dùng là có nên xem phim đó hay không. Ví dụ với một bộ phim có cảm xúc là tiêu cực nhưng điểm của nó lại cao chứng tỏ đây là phim thuộc thể loại kinh dị nên khi đề xuất sẽ được khuyến nghị thêm là “phim hay nên xem”.



```
nlTK.download('vader_lexicon')
sid = SentimentIntensityAnalyzer()

def label_sentiment(description):
    scores = sid.polarity_scores(description)
    compound_score = scores['compound']
    if compound_score >= 0.05:
        sentiment = "Positive"
    elif compound_score <= -0.05:
        sentiment = "Negative"
    else:
        sentiment = "Neutral"
    return sentiment
```

Hình 3.5. Hàm xác định cảm xúc của mô tả phim

3.4.2. Đề xuất phim

Để có thể đưa ra các phim có sự tương đồng cao và phù hợp với người dùng nhất thì nhóm chỉ sử dụng các thông tin: Title, Genre1, Genre2, Genre3, Stars1, Stars2, Stars3, Description. Từ những thuộc cần thiết đã chọn ở trên, nhóm sử dụng kỹ thuật TF-IDF để chuyển các đặc trưng đang có kiểu dữ liệu là chuỗi về thành vector để máy có thể hiểu được. Do trong mô tả của phim chứa các từ dừng hay là các từ vô nghĩa trong câu, nên cần sử dụng thêm bộ từ điển stop-word để loại bỏ chúng trong quá trình chuyển đổi sang vector.

Trong quá trình là đề tài, nhóm cũng đã thực hiện xây dựng một số phương pháp đề xuất như User - User, Item - Item, SVD nhưng sau quá trình chạy thực nghiệm cũng như tìm hiểu về lý thuyết nên nhóm quyết định xây dựng hệ thống với ba chức năng chính dưới đây:

1. Đề xuất phim dựa vào phim người dùng chọn có trong bộ data.

2. Đề xuất phim dựa vào nội dung hoặc một đoạn văn bản người dùng nhập vào.
3. Đề xuất phim cho user mới bằng phương pháp SVD

Với chức năng thứ nhất, hệ thống sẽ cho người dùng chọn tên một bộ phim có trong data. Từ thông tin về Title, Genre1, Genre2, Genre3, Stars1, Stars2, Stars3, Description sẽ tính độ đo cosine để so sánh độ tương quan giữa phim đầu vào và với các phim trong data. Dưới đây là hàm để thực hiện chức năng thứ nhất:

```
def recommend_movies(movie_title, top_n=5):
    idx = movies_use.index[movies_use['Title'] ==
movie_title].tolist()[0]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1],
reverse=True)
    sim_scores = sim_scores[1:top_n + 1]
    similar_movies_indices = [i[0] for i in sim_scores]
    similar_movies = movies_use.iloc[similar_movies_indices]
    return similar_movies
```

Hình 3.6. Hàm đề xuất phim của chức năng thứ nhất

Với chức năng thứ hai, người dùng sẽ nhập vào một đoạn văn có thể là tên phim hoặc là một đoạn nội dung nhỏ của phim. Thay vì như ở chức năng thứ nhất sẽ tính độ tương đồng của các thông tin của phim thì ở chức năng này chỉ tính độ tương đồng của nội dung người dùng nhập so sánh với các thông tin của phim trong data. Dưới đây là hàm thực hiện chức năng thứ hai:

```
def recommend_movies_based_on_text(input_text, top_n=5):
    input_tfidf = tfidf_vectorizer.transform([input_text])
    cosine_scores = linear_kernel(input_tfidf,
tfidf_matrix).flatten()
    top_indices = cosine_scores.argsort()[::-1][-top_n:]
    recommended_movies = movies_use.iloc[top_indices]
    return recommended_movies
```

Hình 3.7. Hàm đề xuất phim của chức năng thứ hai

Với chức năng thứ ba, do nhóm chỉ thu thập được dữ liệu về các bộ phim. Nên không thể xây dựng được hệ thống đề xuất phim bằng các phương pháp lọc cộng tác. Từ đó

nhóm quyết định thực hiện tự sinh data để có thể tiến hành xây dựng chức năng thứ ba này. Data tự sinh bao gồm 2000 user, mỗi user đánh giá ngẫu nhiên 70% số lượng phim có trong data thu thập được tại website IMDb. Điểm đánh giá cũng được đánh ngẫu nhiên từ 1-5.

```
new_ratings = session.get('new_ratings', [])

new_user_id = int(ratings['UserID'].max()) + 1
new_user = {'UserID': new_user_id, 'Gender': 'M', 'Age': 25,
'Occupation': 1, 'Zip-code': '00000'}
users.loc[len(users)] = new_user

for user_id, movie_id, rating in new_ratings:
    ratings.loc[len(ratings)] = [user_id, movie_id, rating]

ratings_merged = ratings[['UserID', 'MovieID', 'Rating']]

if 'Rating' not in ratings_merged.columns:
    raise ValueError("Rating column not found in merged
DataFrame.")
reader = Reader(rating_scale=(1, 5))
data = Dataset.load_from_df(ratings_merged, reader)
trainset = data.build_full_trainset()
model = SVD()
model.fit(trainset)
recommendations = recommend_movies(new_user_id, model,
movies)

return render_template('recommendations.html',
user_id=new_user_id, recommendations=recommendations)
```

Hình 3.8. Đoạn code thực hiện thêm user mới và đề xuất cho user mới

Bằng việc sử dụng dữ liệu tự sinh này và bộ dữ liệu về phim đã thu thập được, nhóm đã xây dựng được hệ thống đề xuất phim gồm 3 chức năng con bao gồm cả 2 phương pháp đó là đề xuất dựa trên nội dung và đề xuất bằng phương pháp lọc cộng tác.

3.4.2. Xây dựng giao diện web

Nhằm đem lại cho người dùng những trải nghiệm tốt nhất có thể, nhóm đã sử dụng thư viện Flask để xây dựng giao diện web. Dưới đây là giao diện người dùng:

The screenshot shows the initial state of the web application. At the top, the title "MOVIE RECOMMENDATION" is centered. Below it, there are two input fields: "Enter a movie name:" and "Enter a sentence:". The "Enter a movie name:" field contains the text "The Whip and the Body". To the right of the "Enter a movie name:" field is a "New user" button. Below each input field is a green "Recommend" button.

Hình 3.9. Hình ảnh giao diện chính của hệ thống






The screenshot shows the web application after a recommendation. The title "MOVIE RECOMMENDATION" is centered. Below it, there are two input fields: "Enter a movie name:" and "Enter a sentence:". The "Enter a movie name:" field contains the text "The Whip and the Body". To the right of the "Enter a movie name:" field is a "New user" button. Below each input field is a green "Recommend" button. Below the "Recommend" buttons, the text "Recommended Movies: The Whip and the Body" is centered. Below this text, there are five movie posters arranged in a row. Each poster has a title, year, runtime, and rating. Below each poster is a short description and a recommendation status.

Movie Title	Year	Runtime	Rate	Recommendation Status
Kill, Baby... Kill!	1966	85 min	6.9	Không được hay cho
Il demonio	1963	98 min	7.2	Không được hay cho
Black Sunday	1960	87 min	7.1	Không được hay cho
The Evil Eye	1963	86 min	7.0	Không được hay cho
Hellraiser	1987	94 min	6.9	Không được hay cho

Hình 3.10. Hình ảnh hệ thống sau khi người dùng chọn đề xuất phim bằng chức năng thứ nhất

MOVIE RECOMMENDATION

Recommended Movies:

				
Sleeping Man - 1996	Open Hearts - 2002	Dwelling in the Fuchun Mountains - 2019	Old Joy - 2006	Feelings of Mountains and Waters - 1988
103 min - rate: 7.1	113 min - rate: 7.5	150 min - rate: 7.2	76 min - rate: 6.8	19 min - rate: 7.5
Ever since an accident in the mountains outside town, Takui's slept in a coma, his neighbors care for him as new events occur every day.	A Dogme film about an engaged couple that is torn apart after the man is paralyzed in an accident, and the woman falls in love with the husband of the woman who caused the accident.	A feature film shot over the course of two years intended to capture the changing of the seasons along the river in a town in the Fuyang district of Hangzhou city.	Two old pals reunite for a camping trip in Oregon's Cascade Mountains.	A short film about a young girl who cares for a wise old scholar in exchange for guqin lessons.
Không được hay cho	Phim hay nên xem	Không được hay cho	Không được hay cho	Phim hay nên xem

Hình 3.11. Hình ảnh hệ thống sau khi người dùng chọn đề xuất phim bằng chức năng thứ hai

Create New User and Rate Movies

Select a movie:

Sign of the Lion
▼

Rating:

1

▼

Add Rating



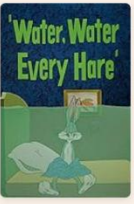

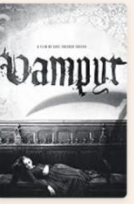
Current Ratings

Submit Ratings

Hình 3.12. Hình ảnh hệ thống cho phép người dùng nhập lịch sử đánh giá phim

Trong hình 3.12, chúng ta giả sử một user đã từng xem một loạt các phim và đã đánh giá các phim đó. Nên sẽ cho user nhập lại các đánh giá, từ các đánh giá mà người dùng nhập vào mà đưa qua kỹ thuật SVD để tính toán và thực hiện đề xuất phim.

Recommendations for New User

				
Hitch-Hike - 1977	The Lacemaker - 1977	Water, Water Every Hare - 1952	The Bridge - 1959	Vampyr - 1932
104 min - rate: 6.8	107 min - rate: 7.5	7 min - rate: 7.9	103 min - rate: 8.0	75 min - rate: 7.4
A bickering couple driving cross-country pick up a murderous hitchhiker who threatens to kill them unless they take him to a sanctuary. In return he agrees to split some bank loot he has on him.	A reserved young woman moves into an apartment with a young student she met while on vacation.	Bugs Bunny's rabbit hole floods, causing him to float to the laboratory of an evil scientist who wants to use his brain for a robot.	In 1945, Germany is being overrun, and nobody is left to fight but teenagers.	A drifter obsessed with the supernatural stumbles upon an inn where a severely ill adolescent girl is slowly becoming a vampire.
Không được hay cho lắm	Phim hay	Phim hay nên xem	Phim hay nên xem	Không được hay cho lắm

[Back](#)

Hình 3.13. Hình ảnh hệ thống sau khi người dùng chọn đề xuất phim bằng chức năng thứ ba

CHƯƠNG 4. KẾT LUẬN

4.1. Kết quả đạt được

Trong đề tài này, nhóm đã xây dựng được hệ thống đề xuất phim dựa trên nội dung và phương pháp lọc cộng tác với việc sử dụng bộ dữ liệu nhóm đã thu thập. Khi người dùng chỉ cần chọn tên một phim hoặc nhập vào nội dung sẽ đề xuất ra các bộ phim có tính tương đồng cao nhất. Kết hợp thêm việc tự sinh dữ liệu về user, nhóm cũng đã xây dựng được hệ thống đề xuất với phương pháp lọc cộng tác. Từ đó nhóm đã thực hiện tạo một giao diện web, giúp cho trải nghiệm của người dùng được tốt nhất có thể.

4.2. Hướng phát triển

Một hướng phát triển khá hay trong thời gian tới của nhóm đó là xây dựng một website xem phim hoàn chỉnh. Khi đó sẽ có dữ liệu thật về user để nhóm có thể phát triển hoàn chỉnh hệ thống đề xuất phim dựa vào phương pháp lọc cộng tác.

Vì thời gian có hạn và trình độ hiểu biết của nhóm còn hạn chế nên đề tài này không thể tránh khỏi những thiếu sót. Nhóm rất mong nhận được sự góp ý của thầy để đề tài của nhóm được hoàn thiện nhất có thể.

Chúng em xin chân thành cảm ơn thầy!

Tài liệu tham khảo

- ¹ Paolo Cremonesi, Paolo Frasca, Francesco Tagliati, "*Collaborative Filtering for Recommending Movies*", 2010.
- ² He Wang, Xiang Zhao, Wei-Wei Tu, "*Neural Collaborative Filtering for Movie Recommendations*", 2018.
- ³ MachineLearningcoban, *Bài 23: Content-based Recommendation Systems*, <https://machinelearningcoban.com/2017/05/17/contentbasedrecommendersys/> (Truy cập ngày 15/05/2024).
- ⁴ MachineLearningcoban, *Bài 24: Neighborhood-Based Collaborative Filtering*, <https://machinelearningcoban.com/2017/05/24/collaborativefiltering/> (Truy cập ngày 16/05/2024).
- ⁵ Wikipedia, *Internet Movie Database*, https://vi.wikipedia.org/wiki/Internet_Movie_Database. (Truy cập ngày 16/05/2024)