|  |  |
| --- | --- |
| **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

**NGUYỄN TRUNG TUYẾN - BÙI TRUNG QUỐC 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

|  |  |
| --- | --- |
| logo | TRƯỜNG ĐẠI HỌC THUỶ 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 Tuyến  Bùi Trung Quốc |
| *Lớp:* | 63TTNT |
| *Giáo viên hướng dẫn:* | TS. Tạ Quang Chiểu |

**TÓM TẮT ĐỀ TÀI**

Hệ thống đề xuất phim dựa trên nội dung 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. Một trong những phương pháp hiệu quả là sử dụng TF-IDF (Term Frequency-Inverse Document Frequency) để xây dựng bộ hồ sơ cho từng phim, giúp phân tích và gợi ý phim dựa trên nội dung.

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

**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 dựa trên nội dung sử dụng TF-IDF.

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](#_Toc7616)

[CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN 2](#_Toc7442)

[1.1. Lý do chọn đề tài 2](#_Toc2012)

[1.2. Mục tiêu đề tài 2](#_Toc27498)

[1.2.1. Mục tiêu tổng quát 2](#_Toc31343)

[1.2.2. Mục tiêu cụ thể 2](#_Toc28672)

[1.3. Nhiệm vụ nghiên cứu 2](#_Toc30814)

[1.4. Đối tượng và phạm vi nghiên cứu 3](#_Toc28991)

[1.5. Phương pháp nghiên cứu 3](#_Toc20670)

[CHƯƠNG 2. TÌM HIỂU LÝ THUYẾT CÁC PHƯƠNG PHÁP 4](#_Toc2081)

[2.1. Kỹ thuật TF-IDF 4](#_Toc17218)

[2.2. Phương pháp lọc dựa trên nội dung 4](#_Toc26997)

[CHƯƠNG 3. XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM 6](#_Toc15815)

[3.1. Thu thập dữ liệu 6](#_Toc17158)

[3.2. Tiền xử lý dữ liệu 6](#_Toc26586)

[3.3. Trực quan hóa dữ liệu 7](#_Toc27162)

[3.4. Xây dựng hệ thống gợi ý phim 8](#_Toc9681)

[CHƯƠNG 4. KẾT LUẬN 12](#_Toc11574)

[1.1. Kết quả đạt được 12](#_Toc8604)

[1.2. Hướng phát triển 12](#_Toc13848)

[Tài liệu tham khảo 13](#_Toc12103)

**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, các hệ thống đề xuất phim dựa trên nội dung đã được triển khai rộng rãi trên nhiều nền tảng xem phim trực tuyến như Netflix, Amazon Prime, và 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

## Lý do chọn đề tài

Hệ thống đề xuất phim dựa trên nội dung 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. Một trong những phương pháp hiệu quả là sử dụng TF-IDF (Term Frequency-Inverse Document Frequency) để xây dựng bộ hồ sơ cho từng phim, giúp phân tích và gợi ý phim dựa trên nội dung.

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

## Mục tiêu đề tài

## Mục tiêu tổng quát

Nghiên cứu và ứng dụng các kỹ thuật học máy trong việc xây dựng hệ thống đề xuất phim dựa trên nội dung nhằm cung cấp các gợi ý phim phù hợp với sở thích người dùng.

### 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 dựa trên nội dung sử dụng TF-IDF.

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

### Nhiệm vụ nghiên cứu

- Nghiên cứu các khái niệm và phương pháp liên quan đến TF-IDF và hệ thống đề xuất 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 dựa trên nội dung sử dụng TF-IDF.

## 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.

## Phương pháp nghiên cứu

Phương pháp lý thuyết: Tìm hiểu và nghiên cứu các mô hình TF-IDF và các thuật toán học máy liên quan đến hệ thống đề xuất dựa trên nội dung.

Phương pháp thực nghiệm: Xây dựng và thử nghiệm hệ thống đề xuất phim dựa trên nội dung sử dụng TF-IDF.

**CHƯƠNG 2. TÌM HIỂU LÝ THUYẾT CÁC PHƯƠNG PHÁP**

**2.1. Kỹ thuật 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. 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:

**2.1.2. 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:

**2.1.3. 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:

**2.2. Phương pháp lọc dựa trên nội dung**

Phương pháp lọc dựa trên nội dung (Content-Based Filtering) 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.

Để sử dụng được phương pháp này 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 mực 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:

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*

**CHƯƠNG 3. XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM**

**3.1. Thu thập dữ liệu**

IMDb(Internet Movie Database) 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”.



*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.

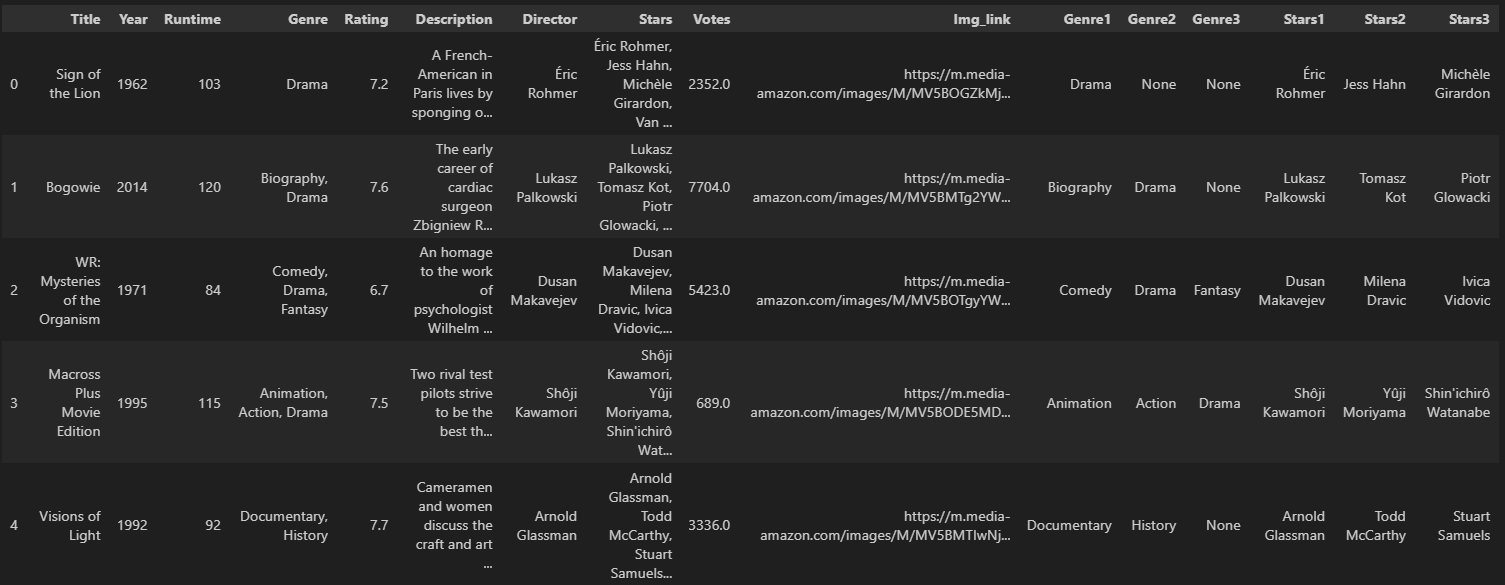
- Cột “Year”: loại bỏ các kí tự dư thừa, chỉ giữ lại các kí tự có dạng là năm.

- 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.

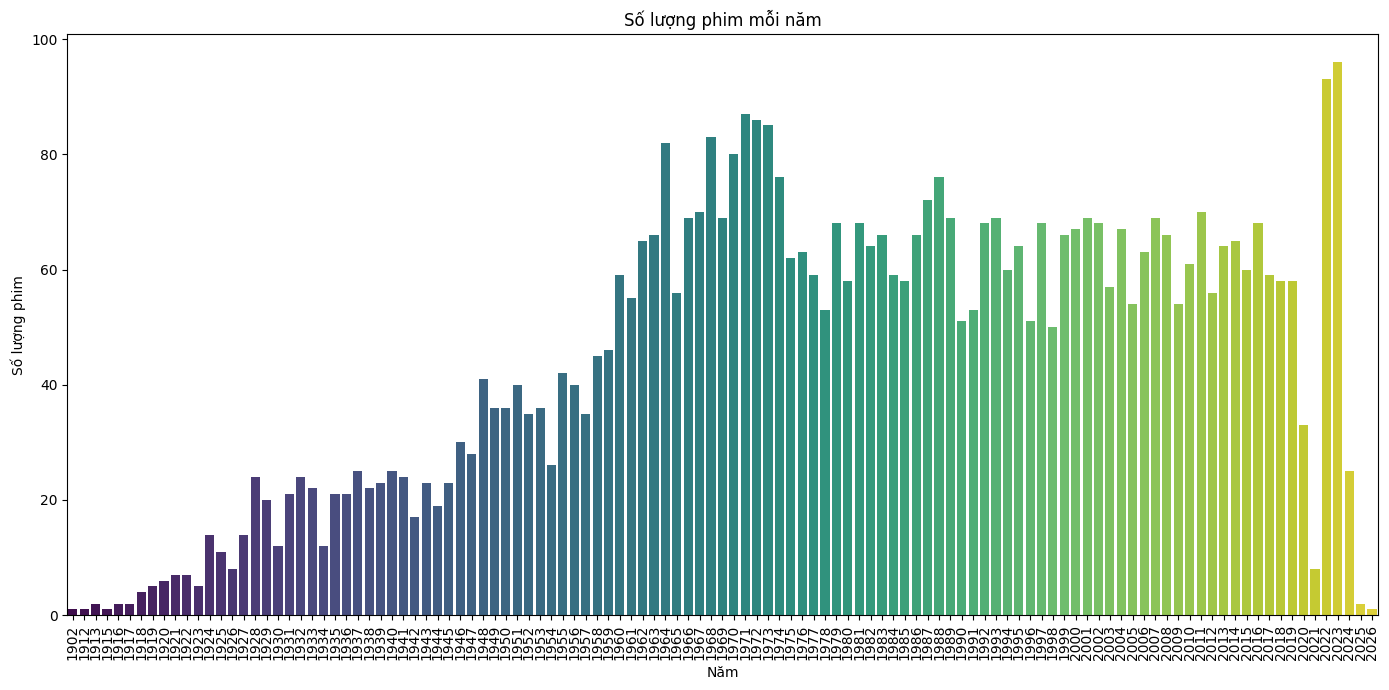
Sau khi đã hoàn thành các bước tiền xử lý dữ liệu trên, dữ liệu hoàn chỉnh thu được như hình dưới đây:



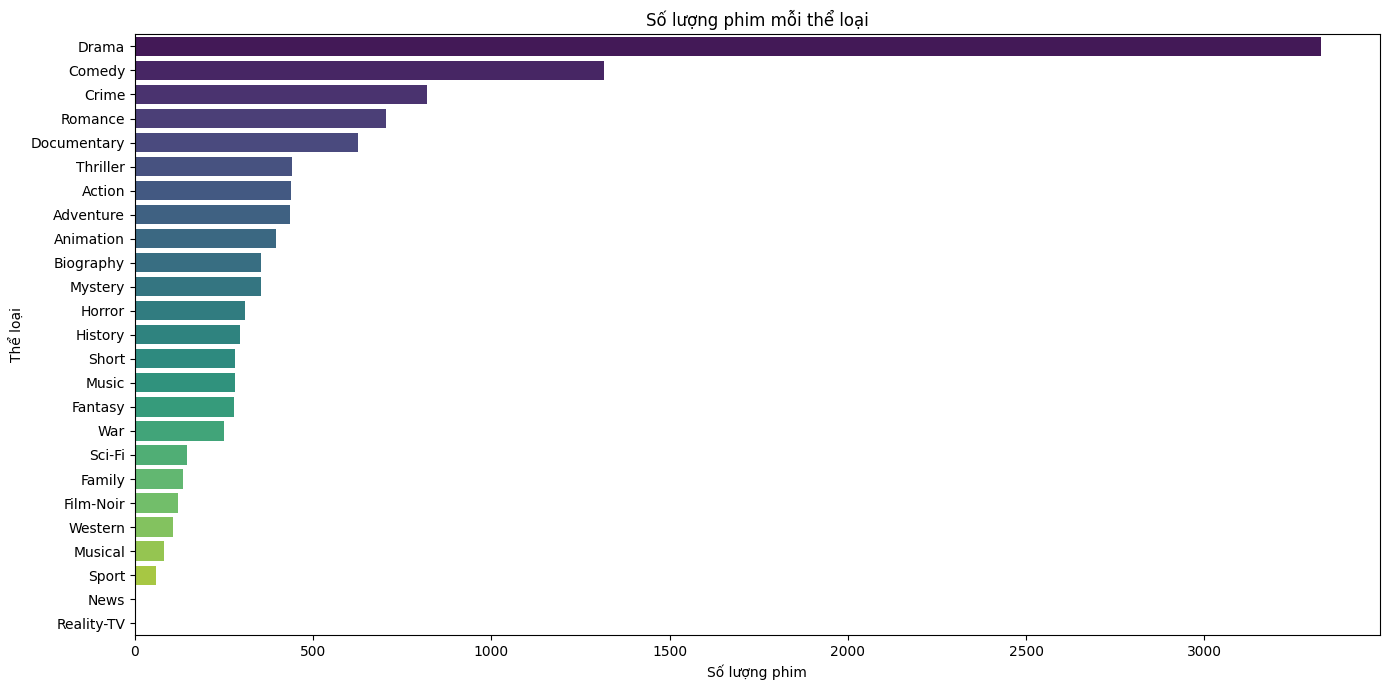
*Hình 3.2. Thông tin của dữ liệu hoàn chỉnh*

**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”.



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

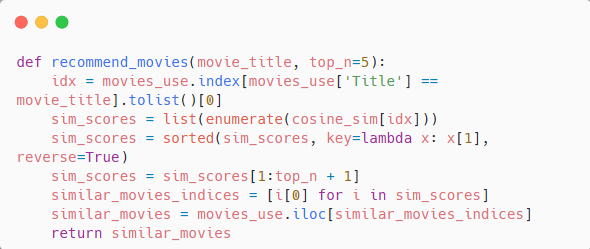
**3.4.2. Đề xuất phim**

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.

Ở hệ thống này, nhóm xây dựng hai chức năng chính:

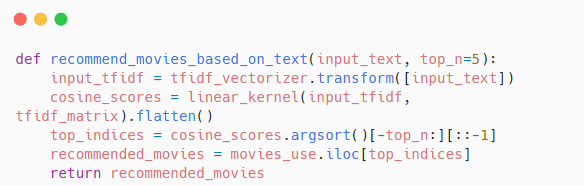
1. *Đề xuất phim dựa vào phim người dùng chọn có trong bộ data.*
2. *Đề xuất phim dựa vào nội dung hoặc một đoạn văn bản người dùng nhập vào.*

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:



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

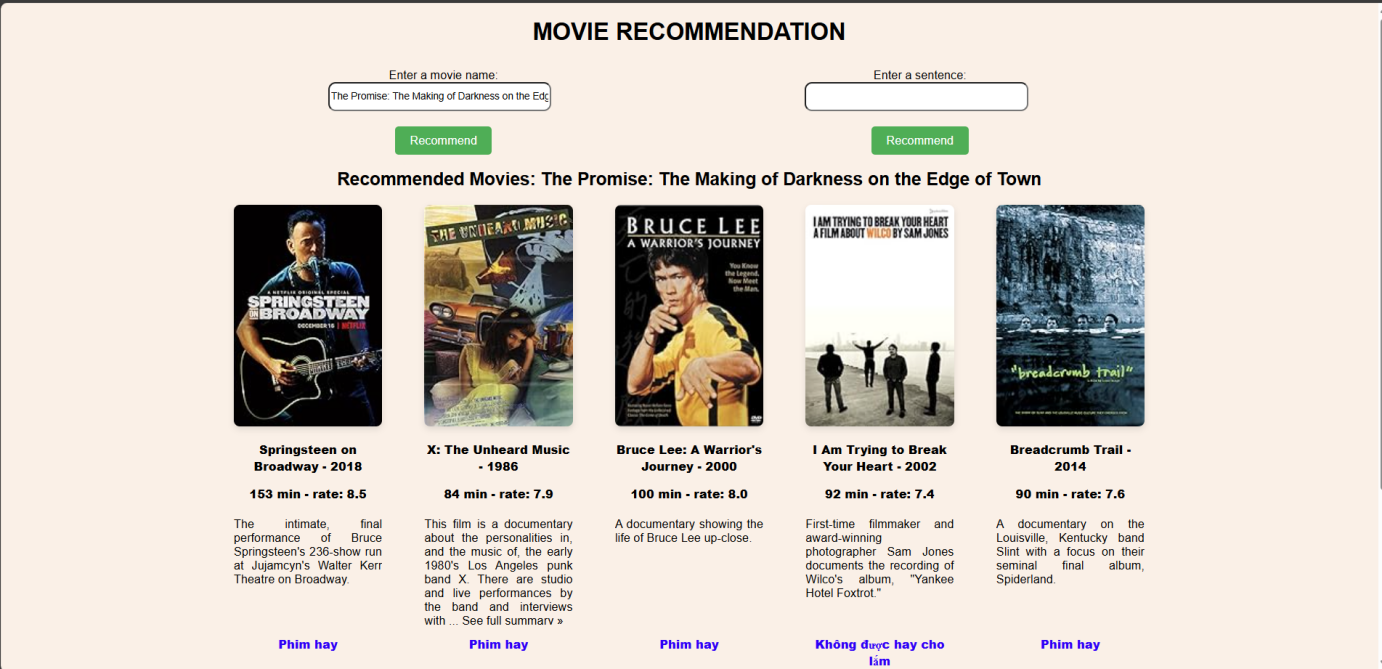
Với chức năng thứ hai, người dùng sẽ nhập vào một đoạn văn có thể là tên phim hoặc là một đoạn nội dung nhỏ của phim. Thay vì như ở chức năng thứ nhất sẽ tính độ tương đồng của các thông tin của phim thì ở chức năng này chỉ tính độ tương đồng của nội dung người dùng nhập so sánh với các thông tin của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



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

**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:



**CHƯƠNG 4. KẾT LUẬN**

## Kết quả đạt được

Trong đề tài này, nhóm đã xây dựng được hệ thống đề xuất phim dựa trên nội dung với việc sử dụng bộ dữ liệu nhóm thu thập. Khi người dùng chỉ cần chọn tên một phim hoặc nhập vào nội dung sẽ trả ra các bộ phim có tính tương đồng cao nhất. Cũng như việc xây dựng được giao diện web, giúp cho trải nghiệm cuả người dùng được tốt nhất có thể.

## Hướng phát triển

Do bộ dữ liệu nhóm thu thập được không có thông tin về user, nên chỉ có thể sử dụng phương pháp đề xuất phim dựa trên nội dung. 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 thêm dữ liệu về user để nhóm có thể bổ sung thêm đề xuất phim dựa vào phương pháp lọc cộng tác.

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

Thay mặt nhóm, em xin chân thành cảm ơn!

# 

# Tài liệu tham khảo