Hệ Thống Gợi Ý Phim Dựa Trên Phương Pháp Lọc Nội Dung

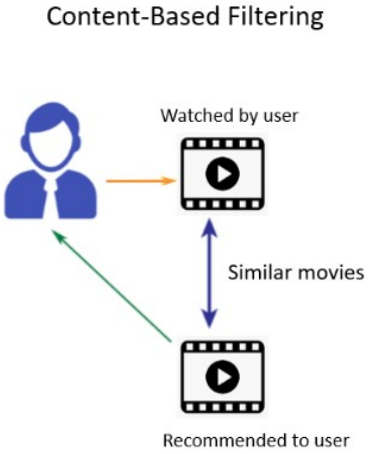
Trương Nhật Quang, Đỗ Tuấn Trực
Trương Lưu Song Tâm, Nguyễn Đức Tài
Trường Đại Học Công Nghệ Thông Tin, Thành Phố Hồ Chí Minh, Việt Nam
Đại học quốc gia, Thành Phố Hồ Chí Minh, Việt Nam
{22521207, 22521548, 22521294, 22521277}@gm.uit.edu.vn
GVHD: TS.Nguyễn Gia Tuấn Anh, CN.Trần Quốc Khánh

***Tóm tắt*—Phim ảnh là một phần không thể thiếu trong cuộc sống hàng ngày, bởi vì nó không chỉ mang lại niềm vui mà còn tạo ra những giây phút thư giãn, giải trí cho mọi người. Chính vì thế trong bài báo này, chúng tôi sẽ giới thiệu đến một hệ thống đề xuất phim tiên tiến, nhằm giúp người dùng dễ dàng tìm kiếm và lựa chọn bộ phim phù hợp với sở thích của mình. Đó là hệ thống khuyến nghị dựa trên nội dung (Content-Based Filtering) là một trong những hệ thống phổ biến trong việc gợi ý phim có thể phục vụ tốt cho cả người dùng mới và người dùng hiện tại. Được thiết kế để mang đến sự mới mẻ, đồng thời tạo ra các gợi ý phim phù hợp với sở thích riêng biệt của từng người dùng. Bằng cách khai thác thông tin từ cơ sở dữ liệu phim, hệ thống của chúng tôi tạo ra những gợi ý phim đa dạng và hấp dẫn, giúp người dùng khám phá ra những bộ phim mới mẻ mà họ chưa từng biết đến. Chúng tôi tin rằng hệ thống này sẽ mang đến cho người dùng một trải nghiệm xem phim độc đáo và thú vị.**

***Từ khóa*: Hệ thống gợi ý, Phim, lọc theo nội dung, Máy học.**

1 Giới Thiệu

Kể từ khi được phát minh, Internet đã phát triển nhanh chóng và tiếp tục phát triển mỗi ngày. Sự phong phú của thông tin có sẵn trực tuyến đã khiến việc truy cập thông tin phù hợp theo ý muốn của người dùng trở nên khó khăn hơn bao giờ hết.[**1**] Với sự bùng nổ của ngành công nghiệp điện ảnh và sự đa dạng ngày càng tăng của các thể loại phim, việc tìm kiếm những bộ phim phù hợp có thể trở thành một nhiệm vụ mất thời gian và phức tạp.
Điều này làm cho hệ thống gợi ý phim trở thành một công cụ quan trọng để giúp người dùng dễ dàng tìm kiếm và khám phá các bộ phim được chúng tôi đề cập trong bài nghiên cứu này bởi vì nó có thể giúp người dùng dễ dàng tìm kiếm những bộ phim một cách thuận tiện và hiệu quả.
Hệ thống gợi ý phim của chúng tôi sử dụng phương pháp lọc nội dung (Content-based Filtering) dựa trên các yếu tố như tên phim, thể loại, nội dung và các từ khóa liên quan trong trường hợp người dùng muốn khám phá về một chủ đề cụ thể như "đua xe", "robot". Dễ dàng và tiện lợi, người dùng chỉ cần cung cấp cho chúng tôi một số thông tin về những bộ phim mà bạn thích và hệ thống sẽ tự động tạo ra danh sách gợi ý phim dựa trên sở thích đó. Từ những bộ phim hành động gay cấn đến những bộ phim lãng mạn ngọt ngào, chúng tôi sẽ giúp bạn khám phá những tác phẩm điện ảnh tuyệt vời mà người dùng có thể chưa biết đến.
Phương pháp lọc nội dung không chỉ giúp bạn tìm thấy những bộ phim mới dựa trên sở thích cá nhân mà còn giúp người dùng khám phá những tác phẩm có cùng phong cách hoặc chủ đề mà người dùng yêu thích. Nhờ vào khả năng phân tích chi tiết và chính xác, hệ thống đảm bảo mang đến cho người dùng những đề xuất phim phù hợp và hấp dẫn nhất.



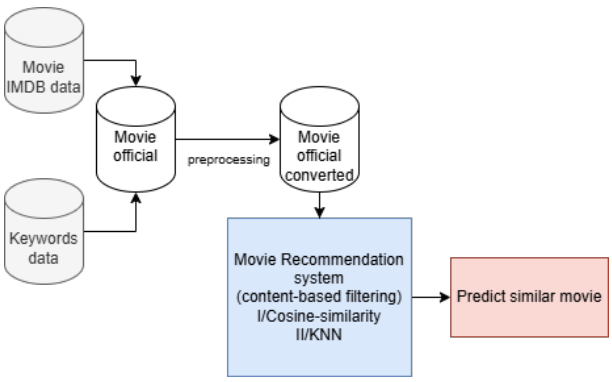
Hình 1: Mô tả quy trình Content-based Filtering recommendation movie system của nhóm

Động lực đằng sau việc áp dụng phương pháp lọc cộng tác là các đề xuất được thực hiện dựa trên sự tương tác của người dùng tương tự. Nó giải quyết

vấn đề quá tải nội dung bằng cách lọc qua thư viện phim khổng lồ để gợi ý cho người dùng những lựa chọn có khả năng thu hút sự quan tâm của họ, từ đó cải thiện trải nghiệm người dùng tổng thể.
Một cách tổng quát, nghiên cứu này có các đóng góp:

- **Nâng cao chất lượng dịch vụ:** Hệ thống gợi ý phim này đóng góp vào việc nâng cao chất lượng dịch vụ hơn nữa có thể giải quyết tình trạng mà người dùng không thể quyết định chọn xem gì do có quá nhiều lựa chọn.
- **Hỗ trợ nhà sản xuất và phân phối:** Ngoài ra, hệ thống cũng giúp các nhà sản xuất và phân phối phim hiểu rõ hơn về sở thích của khán giả, từ đó có chiến lược sản xuất và marketing hiệu quả hơn.
- **Mở rộng ứng dụng:** Hơn thế nữa, không chỉ phát triển ở trên website phim mà có thể mở rộng ra những ý tưởng góp phần tối ưu hơn các trang web cũng sử dụng hệ thống gợi ý như Youtube, Tiktok, Facebook,...thay vì đề xuất những bộ phim phù hợp, thì có thể gợi ý cho người dùng những video, bài viết phù hợp cho người dùng. Thậm chí có thể loại bỏ những nội dung nhạy cảm lách luật gây hại cho người dùng trong tương lai.

Trong phần còn lại bài báo này, chúng tôi sẽ trình bày tiếp đến **Phần 2 (Các Nghiên cứu liên quan)** và các cơ sở lý thuyết được áp dụng, sau đó là **Phần 3 (Bộ dữ liệu)** sử dụng trong nghiên cứu này, **Phần 4 (Cấu hình cho nghiên cứu)** gồm các ý tưởng được để thực hiện thiết kế hệ thống dựa trên content-based filtering, **Phần 5 (Các thực nghiệm nghiên cứu và kết quả đánh giá)**. Cuối cùng **Phần 6 (Kết luận và định hướng trong tương lai)**.



Hình 2: Mô tả quy trình Content-based Filtering recommendation movie system của nhóm

2 Các Nghiên cứu liên quan

A. Tham khảo các bài nghiên cứu

Hệ thống đề xuất phim đã thu hút sự quan tâm của nhiều nhà nghiên cứu với mục đích khám phá, phát triển và hoàn thiện các phương pháp gợi ý phim phù hợp với sở thích của người dùng. Một trong những phương pháp phổ biến được sử dụng là lọc dựa trên nội dung (Content-based filtering). Nghiên cứu của Reddy và các cộng sự đã đề xuất một hệ thống gợi ý phim dựa trên thể loại mà người dùng có thể quan tâm, bài nghiên cứu này đã phân tích rõ ràng cả ưu điểm và nhược điểm của phương pháp này[**2**].
Bên cạnh đó, các nền tảng phim nổi tiếng như Netflix cũng đã áp dụng các thuật toán tiên tiến để đưa ra các gợi ý phim phù hợp với sở thích của người dùng. Nghiên cứu "Movie Recommendation: Netflix uses algorithm for recommending movies according to their interest" cũng như các nền tảng khác như Hotstar, SonyLIV, Voot, ALTBalaji, v.v. cũng được tham khảo để cung cấp cái nhìn tổng quan về cách các hệ thống này hoạt động[**3**].
Các nghiên cứu trên đã khám phá và áp dụng các thuật toán học máy tiên tiến. Mặc dù nhóm không đạt đến trình độ phân tích như các bài báo trên nhưng cũng dựa trên các nghiên cứu liên quan để hoàn thiện bài nghiên cứu hơn.
Trong quá trình phát triển, nhóm nghiên cứu đã tham khảo nhiều nguồn khác nhau, bao gồm cả trang Kaggle, với mục tiêu tối ưu hóa phần code[**4**][**5**][**6**][**7**]. Một cải tiến cụ thể mà nhóm đã thực hiện là tối ưu hóa việc input bằng cách cung cấp một thanh tìm kiếm cho phép người dùng nhập bất kỳ từ nào liên quan đến chủ đề phim ảnh, giúp hệ thống trở nên linh hoạt và tiện dụng hơn.
Mặc dù các phương pháp hiện có đều có ưu và nhược điểm riêng, nhưng chúng vẫn chưa thể giải quyết triệt để tất cả các vấn đề về bảo mật, năng lượng và trải nghiệm người dùng. Nhóm nghiên cứu hy vọng rằng với những cải tiến và tối ưu hóa liên tục, hệ thống gợi ý phim sẽ ngày càng hoàn thiện và đáp ứng tốt hơn nhu cầu của người dùng.

B. Cơ sở lý thuyết

1/TF-IDF (Tần số nghịch đảo tần số thuật ngữ) là một kỹ thuật sử dụng trong phân tích dùng để đánh giá tầm quan trong của một từ trong bộ dữ liệu. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào tầng suất xuất hiện hiện trong dữ liệu.
TF (Term Frequency): Tần suất xuất hiện của từ

trong văn bản. Được tính bằng cách lấy số lần từ xuất hiện trong văn bản chia cho tổng số từ trong văn bản
IDF (Inverse Document Frequency): Nghịch đảo tần suất của văn bản, giúp đánh giá tầm quan trọng của một từ. Khi tính toán TF, tất cả các từ được coi như có độ quan trọng bằng nhau. Nhưng một số từ như "is", "of" và "that" thường xuất hiện rất nhiều lần nhưng độ quan trọng là không cao. Như thế chúng ta cần giảm độ quan trọng của những từ này xuống
Công thức tính TF-IDF của một từ trong một văn bản là tích của TF và IDF của từ đó.

$$TF - IDF(t, d, D) = tf(t, d) \times idf(t, D) \quad (1)$$

trong đó $tf(t, d)$ là tần suất thuật ngữ của thuật ngữ t trong tài liệu d và $idf(t, D)$ là tần suất tài liệu nghịch đảo của thuật ngữ t trong bộ sưu tập tài liệu D.

$$TF(t) = \frac{frequencyoccurrenceoftermtindocument}{totalnumberoftermsindocument}$$

$$IDF(t) = \log_{10}\left(\frac{totalnumberofdocument}{numberofdocumentscontainingtermt}\right)$$

Sự tương đồng giữa các vector có thể được tính bằng hai phương pháp:

1. Độ tương đồng Cosine **(Cosine similarity)**

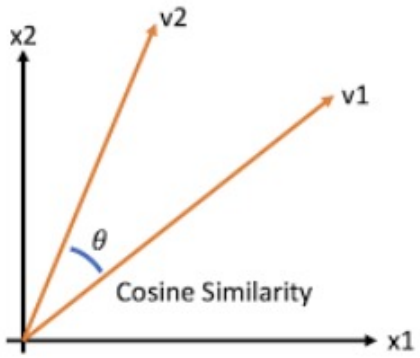
2. Khoảng cách Euclide **(Euclidian distance)**

2/ Cosine Similarity Độ tương đồng cosin giữa hai đối tượng đo lường góc cosin giữa hai vector (đối tượng). Nó so sánh hai tài liệu trên một thang đo chuẩn hóa.. Điều này có thể được thực hiện bằng cách tìm tích vô hướng giữa hai vector.

Công thức tính tương đồng cosin giữa hai vector $\mathbf{a}$ và $\mathbf{b}$ được cho bởi:

$$Cosinesimilarity(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \quad (2)$$

trong đó $\mathbf{a} \cdot \mathbf{b}$ là tích vô hướng của hai vector $\mathbf{a}$ và $\mathbf{b}$, và $\|\mathbf{a}\|$ và $\|\mathbf{b}\|$ là các norm Euclid của vector $\mathbf{a}$ và $\mathbf{b}$, tương ứng.



Hình 3: Mô tả cosine similarity

Cosine similarity có giá trị nằm trong khoảng từ -1 đến 1, trong đó 1 chỉ ra rằng hai văn bản giống nhau hoàn toàn, 0 có nghĩa là không có sự tương tự, và -1 cho thấy hai văn bản hoàn toàn khác nhau. Hình 1 trên cho thấy, góc giữa v1 và v2 là. Góc giữa hai vectơ càng nhỏ thì độ tương đồng càng lớn. Điều đó có nghĩa là nếu góc giữa hai vectơ nhỏ thì chúng gần như giống nhau và nếu góc giữa hai vectơ lớn thì các vectơ rất khác nhau.
3/ K-Nearest Neighbors (KNN) K-Nearest Neighbors (KNN) là một thuật toán học máy có giám sát đơn giản nhưng mạnh mẽ. KNN hoạt động bằng cách tìm kiếm (K) điểm dữ liệu gần nhất trong tập huấn luyện để dự đoán đầu ra cho một điểm dữ liệu mới. Điểm mạnh của KNN là nó không cần giả định về phân phối của dữ liệu, điều này làm cho thuật toán trở nên linh hoạt và dễ sử dụng. Tuy nhiên, KNN cũng có nhược điểm là nó nhạy cảm với nhiễu do chỉ dựa vào thông tin của các điểm dữ liệu gần nhất. Trong nghiên cứu này, chúng ta sẽ tập trung vào việc sử dụng khoảng cách Euclidean trong thuật toán KNN. Khoảng cách Euclidean là một phép đo khoảng cách giữa hai điểm trong không gian Euclidean, và thường được sử dụng trong thuật toán KNN để tìm kiếm các điểm dữ liệu gần nhất. **Công thức Khoảng cách Euclidean:**
Khoảng cách Euclidean giữa hai điểm dữ liệu $x = (x_1, x_2, \ldots, x_n)$ và $y = (y_1, y_2, \ldots, y_n)$ được tính theo công thức:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

3 Bộ Dữ Liệu

A. Thu thập dữ liệu

1/Nguồn thu thập Dữ liệu cho bài nghiên cứu này được lấy dữ liệu từ website: https://www.imdb.com/list/ls098063263/ (tính cập nhật tới 2023) bằng công cụ Python-beautifulsoup để tạo ra bộ dữ liệu imdb.csv và keywordsmovie.csv là bộ dữ liệu được lấy từ Kaggle.com có tính cập nhật 2017. Dữ liệu imdb chứa thông tin về các bộ phim và được lưu trữ trong tập tin imdb.csv(có kích thước là 1000 dòng) tính cập nhật tới 2023. Bộ dữ liệu này gồm các cột chính:

- **Name of movie**: Là tiêu đề của bộ phim. (object)
- **Year of release**: Là năm phát hành của bộ phim. (int64)
- **Watchtime**: Là thời lượng của bộ phim tính bằng phút. (int64)
- **Genre**: Là mô tả thể loại của bộ phim. (object)
- **Classification**: Là phân loại của bộ phim cho lứa tuổi. (ví dụ: PG, R). (object)
- **Movie Rating**: Là điểm đánh giá của bộ phim của người dùng. (float64)
- **Metascore**: Là điểm metascore của bộ phim. (int64)
- **Votes**: Là số lượng phiếu bầu cho bộ phim. (int64)
- **Gross collection**: Là tổng doanh thu của bộ phim.(int64)
- **Des**: Là mô tả của bộ phim. (object)

Dữ liệu imdb chứa thông tin về keywords của các bộ phim được lưu trữ trong tập tin keywordsmovie.csv(có kích thước là 45465 dòng) tính cập nhật tới 2017. Với bộ dữ liệu do chỉ lấy keywords nên sẽ quan tâm tới hai cột là tittle(tên phim) và keywords:

- **tittle**: Là tiêu đề của bộ phim và giống với Name of movie (object)
- **Keywords**: Là mô tả các từ khóa của từng bộ phim. (object)

Việc có chung cột title(Name of movie) trong cả hai bộ dữ liệu là rất quan trọng vì nó đóng vai trò như một khóa ngoại (foreign key) liên kết hai bảng này với nhau. Giúp tối ưu hóa hiệu suất truy vấn và cải thiện tốc độ xử lý dữ liệu. Điều này đặc biệt quan trọng khi làm việc với các bộ dữ liệu lớn và phức tạp trong các ứng dụng như hệ thống gợi ý phim.
Sau đó nhóm đã gộp hai bộ dữ liệu gồm: imdb.csv và keywords.csv để tạo thành một bộ dữ liệu duy nhất có tên movieoffical.csv gồm 1000 bộ phim chứa những cột thiết yếu: "Name of movie", "Year of release", "Watchtime", "Genre", "Classification", "Movie Rating", "Metascore", "Votes", "Gross collection", "Des", "keywords". Với title(Name of movie) là khóa ngoại (foreign key).

B. Tiền xử lý

Qua quá trình nghiên cứu, chúng tôi nhận thấy rằng dữ liệu từ khóa của các bộ phim sau năm 2017 thường không có sẵn. Để khắc phục vấn đề này, chúng tôi đã sử dụng mô hình TF-IDF kết hợp với cosine similarity để dự đoán từ khóa cho những bộ phim này. Quá trình thực hiện được chia thành các bước sau:
Chúng tôi đã tạm thời chia dữ liệu thành hai tập: withkeyword.csv và withoutkeyword.csv Sau đó, chúng tôi sử dụng TfidfVectorizer để chuyển đổi mô tả và tên của các bộ phim trong cả hai tập dữ liệu thành các vector TF-IDF. Vector TF-IDF biểu diễn mỗi bộ phim dưới dạng một vector trong không gian nhiều chiều, với mỗi chiều tương ứng với một từ trong từ điển của tất cả các từ xuất hiện trong tất cả các bộ phim.
Tiếp theo, chúng tôi tính toán độ tương tự cosine giữa mô tả của các bộ phim trong withoutkeyword.csv và withkeyword.csv Độ tương tự cosine là một phép đo sự tương tự giữa hai vectơ, dựa trên góc giữa chúng. Trong trường hợp này, nó cho chúng tôi biết mức độ tương tự giữa mô tả của mỗi bộ phim chưa có từ khóa và mô tả của tất cả các bộ phim đã có từ khóa.
Cuối cùng, chúng tôi dự đoán từ khóa cho mỗi bộ phim trong withoutkeyword.csv bằng cách lấy từ khóa của bộ phim từ withkeyword.csv mà mô tả của nó tương tự nhất với so với của bộ phim không có từ khóa. Chúng tôi thực hiện việc này bằng cách tìm bộ phim có từ khóa mà độ tương tự cosine của nó với bộ phim không có từ khóa là lớn nhất.
Kết quả cuối cùng là một DataFrame withoutkeyword đã được cập nhật, trong đó mỗi bộ phim giờ

đây đều có từ khóa dự đoán dựa trên mô tả và tên của nó. Đây là một cách hiệu quả để tự động tạo từ khóa cho các bộ phim dựa trên thông tin đã có từ các bộ phim khác.

Bảng 1: **Các bộ phim có sẵn keyword**

Tên phim	Keywords
The Avengers (2012)	new york, shield, marvel comic, superhero, based on comic, alien invasion, superhero team, aftercreditsstinger, duringcreditsstinger, marvel cinematic universe
Avengers: Age of Ultron (2015)	marvel comic, sequel, superhero, based on comic, vision, superhero team, duringcreditsstinger, marvel cinematic universe, 3d

Bảng 2: **Các bộ phim chưa có keywords đã được phân tích và gán keywords**

Tên phim	Keywords
Avengers: Infinity War (2018)	mutant, supernatural powers, marvel comic, superhero, based on comic, superhuman, apocalypse, superhero team, world domination, aftercreditsstinger, 1980s
Avengers: Endgame (2019)	civil war, war, marvel comic, sequel, superhero, based on comic, imax, aftercreditsstinger, duringcreditsstinger, marvel cinematic universe, 3d

4 Cấu hình cho nghiên cứu

Phần cứng:

- **CPU**: Bộ vi xử lý trung tâm Intel Core i7 .
- **RAM**: 16GB DDR4.
- **Storage**: 512GB SSD.
- **GPU**:NVIDIA GeForce GTX 3050

Phần mềm:

- **Hệ điều hành** : Hệ điều hành Windows 11.
- **Môi trường phát triển**: Jupyter Notebook và Visual Studio Code.
- **Ngôn ngữ lập trình**: Python.
- **Thư viện**: Pandas, Scikit-learn, nltk, fuzzywuzzy, matplotlib, seaborn.

5 Mô hình và kết quả

A. Cài đặt mô hình

Kết hợp các thuộc tính: Mã sau đó kết hợp các thuộc tính"Name of movie", "Genre", "keywords"và ‘des’ thành một chuỗi duy nhất. Chuỗi này sau đó được lưu trữ trong cột ‘combined’.
Vector hóa: Mã sau đó sử dụng TfidfVectorizer để chuyển đổi các chuỗi trong cột ‘combined’ thành vector TF-IDF. TfidfVectorizer là một phương pháp biểu diễn văn bản dưới dạng vector số, trong đó mỗi số đại diện cho trọng số của một từ trong văn bản. Trọng số này được tính toán dựa trên tần suất xuất hiện của từ đó trong văn bản cũng như trong toàn bộ tập dữ liệu.
Tham số stopwords yêu cầu công cụ bỏ qua các từ có trong danh sách từ dừng tiếng Anh, là những từ không mang nhiều ý nghĩa, như "the", "is", "and", ... Bằng cách này, công cụ có thể tập trung vào những từ có liên quan và mang lại nhiều thông tin cho việc phân tích thông tin bộ phim.
Input: Dữ liệu được nhập vào có thể là các thuộc tính "Name of movie", "Genre", "keywords".
1. Xây dựng mô hình tính toán sự tương đồng Cosine Similarity : Sử dụng cosine_similarity để tính toán ma trận tương đồng giữa các vectơ TF-IDF.. Sự tương đồng cosine là một phép đo tương đồng giữa hai vectơ không gian nhiều chiều. Nó được tính bằng cách lấy tích vô hướng của hai vectơ và chia cho tích của độ dài của hai vector.
2. Xây dựng mô hình K-Nearest Neighbors (KNN): Sử dụng K-NearestNeighbors từ thư viện sklearn.neighbors. Mô hình này được cấu hình để sử dụng khoảng cách Euclidean(đo khoảng cách giữa hai điểm trong không gian nhiều chiều, và thường được sử dụng trong các thuật toán học máy như KNN) và thuật toán ‘brute’(sẽ duyệt qua tất cả các điểm dữ liệu, tính toán khoảng cách từ mỗi điểm dữ liệu đến điểm dữ liệu cần tìm, và chọn ra K điểm có khoảng cách gần nhất). Mô hình này sau đó được huấn luyện với ma trận TF-IDF, tức là nó học cách tìm các hàng (tức là các bộ phim) gần

nhất với một hàng nhất định dựa trên khoảng cách Euclidean giữa các vectơ TF-IDF của chúng.
Output: Cuối cùng, mã sẽ hiển thị kết quả đề xuất dưới dạng danh sách đề suất những bộ phim có sự liên quan tới thuộc tính đã nhập

B. Kết quả

1. Mô hình Cosine Similarity

	Name of movie	Year of release	Genre	Similarity Score
782	The Fast and the Furious	2001	Action, Crime, Thriller	0.392096
722	2 Fast 2 Furious	2003	Action, Crime, Thriller	0.380914
408	Fast & Furious 6	2013	Action, Adventure, Crime	0.346331
3	Fast X	2023	Action, Adventure, Crime	0.288712
318	Furious 7	2015	Action, Crime, Thriller	0.261671
550	Fast & Furious	2009	Action, Crime, Thriller	0.261211
480	Fast Five	2011	Action, Crime, Thriller	0.247087
537	The Final Destination	2009	Horror, Thriller	0.205030
797	Gone in 60 Seconds	2000	Action, Crime, Thriller	0.202658
95	Ford v Ferrari	2019	Action, Biography, Drama	0.198483

Hình 5: Input "car racing"

Mô tả và từ khóa: Từ khóa "car racing"liên quan trực tiếp đến các phim trong series Fast & Furious hoặc có nhiều phim có từ khóa "car", "racing"trong mô tả hoặc từ khóa của chúng. Do nó không trùng khớp với tên phim nào nên sẽ không có bộ phim nào làm chuẩn cho các gợi ý phim.
'The Fast and the Furious' (0.392096), '2 Fast 2 Furious' (0.380914), 'Fast & Furious 6' (0.346331), 'Fast X' (0.288712), 'Furious 7' (0.261671), 'Fast & Furious' (0.261211), 'Fast Five' (0.247087), và các bộ phim ngoài series 'The Final Destination' (0.205030), 'Gone in 60 Seconds' (0.202658), and 'Ford v Ferrari' (0.198483).
Kết quả dựa trên tần suất các từ khóa "car", "racing"trong bộ dữ liệu phim nên 'similarity score' càng cao thì tần suất xuất hiện xuất hiện trong mô tả hoặc từ khóa của từng bộ phim càng nhiều.

Name of movie	Year of release	Genre	Similarity Score
Fast & Furious	2009	Action, Crime, Thriller	0.975572
Fast & Furious 6	2013	Action, Adventure, Crime	0.418437
Fast & Furious Presents: Hobbs & Shaw	2019	Action, Adventure, Thriller	0.413925
The Fast and the Furious	2001	Action, Crime, Thriller	0.345145
2 Fast 2 Furious	2003	Action, Crime, Thriller	0.300724
Fast Five	2011	Action, Crime, Thriller	0.282506
Furious 7	2015	Action, Crime, Thriller	0.280640
Fast X	2023	Action, Adventure, Crime	0.203205
Gone in 60 Seconds	2000	Action, Crime, Thriller	0.186833
The Fate of the Furious	2017	Action, Crime, Thriller	0.186599

Hình 4: Input "fast & furious"

Khớp chính xác: 'Fast & Furious' (2009) có điểm tương đồng cao nhất (0.975572) vì truy vấn "fast&furious"khớp chính xác với tiêu đề phim và trở thành phim làm chuẩn.
Các phim liên quan: Các phim khác trong series Fast & Furious cũng có điểm tương đồng cao do mô tả khá giống nhau. Ví dụ: 'Fast & Furious 6' (0.418437), 'Fast & Furious Presents: Hobbs & Shaw' (0.413925), 'The Fast and the Furious' (0.345145), '2 Fast 2 Furious' (0.300724), 'Fast Five' (0.282506), 'Furious 7' (0.280640), 'Fast X' (0.203205), 'The Fate of the Furious' (0.186599). Do không chứa trực tiếp cụm từ 'Fast & Furious', mà chỉ liên quan về mặt nội dung và là một phần của series nên có 'similarity score' thấp hơn so với 'Fast & Furious' (2009)
Mặc dù phim này là một phần của series, nhưng tiêu đề và từ khóa của nó không trực tiếp chứa cụm từ 'Fast & Furious', do đó vector TF-IDF của nó có thể ít tương đồng hơn với truy vấn so với các phim có tiêu đề chứa 'Fast & Furious'. Đối với 'Gone in 60 Seconds'(0.186833) là vì nó có liên quan đến nội dung car racing nên được đề xuất.

Name of movie	Year of release	Genre	Similarity Score
Fast & Furious	2009	Action, Crime, Thriller	0.487248
2 Fast 2 Furious	2003	Action, Crime, Thriller	0.446577
The Fast and the Furious	2001	Action, Crime, Thriller	0.427835
Fast & Furious 6	2013	Action, Adventure, Crime	0.406033
Furious 7	2015	Action, Crime, Thriller	0.297529
Fast & Furious Presents: Hobbs & Shaw	2019	Action, Adventure, Thriller	0.289654
Fast X	2023	Action, Adventure, Crime	0.269917
Fast Five	2011	Action, Crime, Thriller	0.231002
The Fate of the Furious	2017	Action, Crime, Thriller	0.194213
The Final Destination	2009	Horror, Thriller	0.134205

Hình 6: Input "car racing fast & furious"

Name of movie	Year of release	Genre	Similarity Score
Fast & Furious	2009	Action, Crime, Thriller	0.488757
2 Fast 2 Furious	2003	Action, Crime, Thriller	0.449952
The Fast and the Furious	2001	Action, Crime, Thriller	0.430217
Fast & Furious 6	2013	Action, Adventure, Crime	0.409101
Furious 7	2015	Action, Crime, Thriller	0.304113
Fast & Furious Presents: Hobbs & Shaw	2019	Action, Adventure, Thriller	0.293546
Fast X	2023	Action, Adventure, Crime	0.275236
Fast Five	2011	Action, Crime, Thriller	0.235554
The Fate of the Furious	2017	Action, Crime, Thriller	0.200873
Gone in 60 Seconds	2000	Action, Crime, Thriller	0.139081

Hình 7: Input "car racing fast & furious action"

Ngoài ra còn thể kết hợp cả từ khóa, tên phim và thể loại trong truy vấn tìm kiếm, tức là đang cung cấp thêm thông tin cho hệ thống để tìm kiếm các

bộ phim tương tự. Tuy nhiên, điều này cũng có thể làm giảm số lượng kết quả tìm kiếm, vì nó giới hạn kết quả chỉ cho các bộ phim phù hợp với tất cả các tiêu chí đã nhập.
Ở hình 5 và hình 6 khi input mô tả rõ ràng hơn "car racing fast & furious action"và "car racing fast & furious"gần như kết quả tương đồng nhau từ xếp hạng đến "similarity score"
'The Fast and the Furious' (khoảng 0.48): Phim này có điểm tương đồng cao nhất trong đoạn văn so với các phim khác trong series, điều này cho thấy mô tả và từ khóa của phim rất sát với từ khóa "car racing", "Fast & Furious"và "action".
Xếp sau đó là '2 Fast 2 Furious' (khoảng 0.44), 'The Fate and the Furious' (0.427-0.43),'Fast & Furious 6' (0.409-0.406), 'Fast X' (0.275-0.269), 'Furious 7' (0.304-0.297), 'Fast & Furious' (khoảng 0.48), 'Fast Five' (xấp xỉ 0.23), 'The Fate of the Furious' (0.200-0.194). Các phim ngoài series 'Fast & Furious': 'Gone in 60 Seconds' (0.139081), 'The Final Destination' (0.134205): Điểm tương đồng rất thấp, cho thấy phim này hầu như không liên quan đến từ khóa "car racing"và "Fast & Furious".

2. Mô hình K-Nearest Neighbors (KNN)

Name of movie	Year of release	Genre	Distance
Robots	2005	Animation, Adventure, Comedy	0.947110
I, Robot	2004	Action, Mystery, Sci-Fi	0.960347
Real Steel	2011	Action, Drama, Sci-Fi	1.051595
Transformers: Age of Extinction	2014	Action, Adventure, Sci-Fi	1.158090
Big Hero 6	2014	Animation, Action, Adventure	1.236624
The Day the Earth Stood Still	2008	Adventure, Drama, Sci-Fi	1.253967
WALL·E	2008	Animation, Adventure, Family	1.270625
Transformers	2007	Action, Adventure, Sci-Fi	1.296347
Transformers: Rise of the Beasts	2023	Action, Adventure, Sci-Fi	1.306453
Sing 2	2021	Animation, Adventure, Comedy	1.308129

Hình 8: Input "robot"

Phim về robot: 'Robots' (2005) và 'I, Robot' (2004) có khoảng cách nhỏ nhất (0.947110 và 0.960347) vì chúng có từ khóa "robot"trong tiêu đề và nội dung, đồng thời tập trung vào chủ đề robot.
Phim có yếu tố robot:'Real Steel'(1.051595), 'Transformers: Age of Extinction' (1.158090), 'Big Hero 6' (1.236624), 'The Day the Earth Stood Still' (1.253967), 'WALL·E'(1.270625), 'Transformers'(1.296347), 'Transformers: Rise of the Beasts' (1.306453). Khi bạn nhập từ khóa, hệ thống sẽ tìm kiếm các bộ phim có từ khóa trong mô tả, thể loại, từ khóa hoặc tên. Tuy nhiên, kết quả tìm kiếm có thể rộng hơn và khoảng cách Euclidean có thể cao hơn, vì từ khóa có thể xuất hiện ở nhiều nơi khác nhau và không nhất thiết phải liên quan mật thiết đến nội dung của bộ phim.'

Name of movie	Year of release	Genre	Distance
Transformers	2007	Action, Adventure, Sci-Fi	0.201345
Transformers: The Last Knight	2017	Action, Adventure, Sci-Fi	1.198529
Transformers: Dark of the Moon	2011	Action, Adventure, Sci-Fi	1.222117
Transformers: Age of Extinction	2014	Action, Adventure, Sci-Fi	1.222804
Transformers: Rise of the Beasts	2023	Action, Adventure, Sci-Fi	1.223172
Transformers: Revenge of the Fallen	2009	Action, Adventure, Sci-Fi	1.228981
Toy Story	1995	Animation, Adventure, Comedy	1.292620
I, Robot	2004	Action, Mystery, Sci-Fi	1.308799
The Day the Earth Stood Still	2008	Adventure, Drama, Sci-Fi	1.309299
Toy Story 4	2019	Animation, Adventure, Comedy	1.311730

Hình 9: Input "Transformers"

Khớp chính xác: 'Transformers' (2007) có khoảng cách nhỏ nhất (0.201345) vì truy vấn "Transformers"khớp chính xác với tiêu đề phim và trở thành phim làm chuẩn.
Các phim liên quan: Các phim khác trong series Transformers như 'Transformers: The Last Knight' (1.198529), 'Transformers: Dark of the Moon' (1.222117), 'Transformers: Age of Extinction' (1.222804), 'Transformers: Rise of the Beasts' (1.223172), 'Transformers: Revenge of the Fallen' (1.228981). Các phim được sắp xếp dựa trên độ tương đồng với từ khóa tìm kiếm. Khoảng cách tương đồng (KNN Distance) được tính dựa trên mô tả của phim và từ khóa bạn nhập. Các phim được xếp hạng dựa trên khoảng cách tương đồng, với giá trị càng thấp thể hiện sự tương đồng càng cao.
Những phim như 'Toy Story' (1995), 'I, Robot' (2004), 'The Day the Earth Stood Still' (1.309299), 'Toy Story 4' (1.311730) có khoảng cách lớn hơn các bộ phim liên quan, cho thấy chúng ít liên quan hơn so với các phim trong series Transformers. Tuy nhiên, chúng vẫn có yếu tố liên quan đến truy vấn như một trong ba thể loại của phim làm chuẩn 'Transformers' (2007) thể loại Action, Adventure, Sci-Fi.

Name of movie	Year of release	Genre	Distance
I, Robot	2004	Action, Mystery, Sci-Fi	0.947110
Real Steel	2011	Action, Drama, Sci-Fi	0.960347
Transformers: Age of Extinction	2014	Action, Adventure, Sci-Fi	1.051595
Robots	2005	Animation, Adventure, Comedy	1.158090
Big Hero 6	2014	Animation, Action, Adventure	1.236624
The Day the Earth Stood Still	2008	Adventure, Drama, Sci-Fi	1.253967
WALL·E	2008	Animation, Adventure, Family	1.270625
Transformers	2007	Action, Adventure, Sci-Fi	1.296347
Transformers: Rise of the Beasts	2023	Action, Adventure, Sci-Fi	1.306453
Sing 2	2021	Animation, Adventure, Comedy	1.308129

Hình 10: Input "robot Transformers"

Mô tả và từ khóa: Từ khóa "robot transformers"và "robot transformers sci-fi"liên quan trực tiếp đến

	Name of movie	Year of release	Genre	Distance
0	Transformers: Age of Extinction	2014	Action, Adventure, Sci-Fi	0.927604
1	Transformers: Rise of the Beasts	2023	Action, Adventure, Sci-Fi	0.966285
2	Transformers: Revenge of the Fallen	2009	Action, Adventure, Sci-Fi	1.024485
3	Transformers	2007	Action, Adventure, Sci-Fi	1.058810
4	I, Robot	2004	Action, Mystery, Sci-Fi	1.134221
5	Transformers: The Last Knight	2017	Action, Adventure, Sci-Fi	1.137672
6	Real Steel	2011	Action, Drama, Sci-Fi	1.141398
7	Transformers: Dark of the Moon	2011	Action, Adventure, Sci-Fi	1.142060
8	Moon Man	2022	Comedy, Sci-Fi	1.211183
9	Sing 2	2021	Animation, Adventure, Comedy	1.260679

Hình 11: Input "robot Transformers sci-fi"

các phim trong series Transformers hoặc các phim có từ khóa "robot"hoặc "transformers"trong mô tả hoặc từ khóa của chúng. Dưới đây là phân tích chi tiết

'Transformers: Age of Extinction' (0.919 - 0.927) phim này có điểm tương đồng cao nhất trong cả hai bảng, cho thấy mô tả và từ khóa của phim rất sát với từ khóa "robot transformers"và "robot transformers sci-fi". Tiếp theo là các bộ phim liên quan tới series phim "Transformers"'Transformers: Rise of the Beasts'(0.963 - 0.966), 'Transformers: Revenge of the Fallen'(1.021828 - 1.024485), 'Transformers'(1.074961 - 1.058810), 'Transformers: Dark of the Moon'(1.153941 - 1.142060), 'Transformers: The Last Knight'(1.163391 - 1.137672), Bên cạnh đó cũng có các phim ngoài series Transformers: 'I, Robot'(1.150107 - 1.134221), 'Real Steel'(1.156827 - 1.141398), 'Sing 2'(1.236795 - 1.260679), 'Moon Man'(1.238776 - 1.211183).

Trong thuật toán tính khoảng cách (distance) và độ tương đồng (similarity), các yếu tố như từ khóa, mô tả phim và thể loại phim được cân nhắc. Phim "I, Robot"và "Real Steel"có các yếu tố về robot rất mạnh trong mô tả và từ khóa, khiến chúng có khoảng cách gần với từ khóa "robot transformers", mặc dù chúng không thuộc series Transformers. Điều này giải thích tại sao phim như 'I, Robot' và 'Real Steel' có thể xếp trên một số phim trong series Transformers dựa trên khoảng cách tính toán.

C. Thảo luận

Hệ thống gợi ý dựa trên nội dung (Content-Based Recommendation System) là hệ thống mạnh mẽ để cá nhân hóa đề xuất cho người dùng. dựa trên dữ liệu các bộ phim sẵn có đã đưa ra những đề xuất phim các nhân hóa hiệu quả. Các dòng phim phổ biến như Marvel, DC, Transformers, Batman và Fast & Furious thường xuất hiện trong kết quả gợi ý. Điều này mang lại sự đa dạng cho người dùng, vì họ có nhiều lựa chọn từ các dòng phim này.

Qua hai thí nghiệm ở trên áp dụng cho lần lượt mô hình Cosine Similarity, KNN,.. mặc dù cho hai tên phim, "keywords"khác nhau và nó sử dụng hai cách đo khác nhau (similarity score và Distance) nhưng chung quy lại thì nó đều in ra danh sách những bộ phim có mối quan hệ mật thiết với input. Quan sát khá thú vị cho thấy do dòng phim Fast & Furious khá phổ biến nên khi input "car racing"thì sẽ đưa ra top gợi đa số là của Fast & Furious. Ta thấy được sự đa dạng rõ ràng khi input "robot"thì trong tập dữ liệu còn có rất nhiều bộ phim có liên quan đến robot khác nên Transformers cũng chưa chắc là sự ưu tiên trong top danh sách được đề xuất.

Hạn Chế: Hiệu suất của hệ thống có thể thay đổi khi áp dụng với các bộ dữ liệu khác hoặc trong các tình huống thực tế, chẳng hạn như: vì nó chủ yếu dựa vào dữ liệu về các bộ phim nên khó khăn trong việc đề xuất những bộ phim với nội dung rất khác so với những gì người dùng đã xem trước đó, dẫn đến việc gợi ý có thể trở nên nhàm chán nếu không có sự đa dạng hóa.

Về vấn đề input nó bắt buộc yêu cầu chính tả, nếu như từ nhập input bị sai chính tả thì hệ thống sẽ đưa ra gợi những bộ phim sai lệch và thậm chí là không có danh sách bộ phim nào. Từ đó làm mất đi tính chính xác của hệ thống gợi ý các bộ phim. Nếu bạn nhập "Tranformer," hệ thống gợi ý của bạn sẽ không tìm ra "Transformers" vì chúng là hai từ khác nhau. "Tranformer" không phải là một từ tiếng Anh thông dụng hoặc đúng chính tả, trong khi "Transformers" là tên của một loạt phim robot thương hiệu Transformers.

	Name of movie	Year of release	Genre	Similarity Score
1000	Snow White and the Seven Dwarfs	1937	Animation, Adventure, Family	0.0
328	22 Jump Street	2014	Action, Comedy, Crime	0.0
341	Dracula Untold	2014	Action, Drama, Fantasy	0.0
340	Annabelle	2014	Horror, Mystery, Thriller	0.0
339	Stand by Me Doraemon	2014	Animation, Comedy, Drama	0.0
338	Into the Woods	2014	Adventure, Comedy, Drama	0.0
337	Ode to My Father	2014	Drama, War	0.0
336	Fury	2014	Action, Drama, War	0.0
335	The Fault in Our Stars	2014	Drama, Romance	0.0
334	The Equalizer	2014	Action, Crime, Thriller	0.0

Hình 12: Lỗi chính tả

Bộ keywords của những bộ phim sau 2017 được dự đoán từ mô hình TF-IDF kết hợp với cosine similarity tuy vẫn đảm bảo được chính xác nếu input những keywords phổ biến, nhưng nếu xét kỹ hơn ở ví dụ đã cho Avengers: Endgame(2019) vẫn có những keywords không có tính chính xác cao như civival war, hay ở Avengers: Infinity War (2018) là

1980s. keywords Transformers: Rise of the Beasts là egypt, sun, chaos, symbol, artifact, transformers, tank, robot, imax, duringcreditsstinger trong số đó có egypt, sun, chao có thể chưa chuẩn xác...

6 Kết luận và định hướng tương lai

A. Kết luận

Hệ thống gợi ý phim dựa trên nội dung (Content-Based Recommendation System) đã chứng minh được tính hiệu quả và tiềm năng trong việc cung cấp các đề xuất phim phù hợp với sở thích cá nhân của người dùng. Với việc tận dụng các đặc điểm nội dung của phim như thể loại, tên phim và từ khóa mô tả, hệ thống có khả năng phân tích và hiểu rõ hơn về sở thích của từng người dùng.

Giúp mở rộng tầm nhìn của họ tới những tác phẩm mới mà họ có thể chưa từng biết đến nhưng có nội dung tương đồng với những bộ phim họ đã xem và yêu thích. Một lợi thế quan trọng của hệ thống dựa trên nội dung là khả năng hoạt động hiệu quả ngay cả khi có ít thông tin về người dùng, vì nó chủ yếu dựa vào dữ liệu về các bộ phim. Tuy nhiên, hệ thống Content-Based Recommendation System cũng có một số hạn chế. Một trong số đó là khả năng hạn chế của việc phân tích nội dung. Mặc dù hệ thống có thể tạo ra các gợi ý dựa trên yếu tố nội dung, nhưng nó có thể bỏ qua những yếu tố khác như cảm xúc và sự phát triển của nhân vật. Điều này có thể dẫn đến việc hệ thống không thực sự hiểu sâu thị hiếu của người dùng và đề xuất những bộ phim không hoàn toàn phù hợp.

Để tối ưu hóa hệ thống này, cần liên tục cập nhật và mở rộng cơ sở dữ liệu nội dung phim, cũng như kết hợp với các phương pháp gợi ý khác như Collaborative Filtering để đạt được độ chính xác và tính đa dạng cao hơn trong các đề xuất. Bằng việc liên tục cải tiến và kết hợp các phương pháp khác nhau, hệ thống gợi ý phim có thể mang lại trải nghiệm giải trí tốt nhất cho người dùng, từ đó tăng cường sự hài lòng và gắn kết của họ với dịch vụ.

B. Định hướng tương lai

Mặc dù hệ thống gợi ý phim dựa trên nội dung (Content-Based Recommendation System) đã chứng minh được hiệu quả trong việc cung cấp các đề xuất phim phù hợp với sở thích cá nhân của người dùng, vẫn còn nhiều lĩnh vực cần được cải thiện và khám phá để nâng cao hiệu suất và độ chính xác của hệ thống. Dưới đây là một số hướng nghiên cứu và phát triển tiềm năng

Tự động sửa lỗi chính tả trong hệ thống gợi ý có thể được thực hiện thông qua việc sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP).

Sử dụng thư viện xử lý ngôn ngữ tự nhiên: Các thư viện như NLTK (Natural Language Toolkit) hoặc SpaCy có thể được sử dụng để phân tích và xử lý văn bản. Các công cụ này có thể giúp tách từ, kiểm tra chính tả và đề xuất các từ đúng chính tả.

Kiểm tra từ gần giống nhất: Khi người dùng nhập một từ sai chính tả, hệ thống có thể kiểm tra từ gần giống nhất trong từ điển và đề xuất từ đúng. Ví dụ: Nếu người dùng nhập "Tranformer,"hệ thống có thể đề xuất "Transformers."

Xây dựng mô hình học máy để nhận diện và sửa lỗi chính tả như Naive Bayes, SVM, hoặc mạng neural để nhận diện lỗi chính tả. Mạng neural, đặc biệt là mạng LSTM, thường hoạt động tốt trong việc này.

Tối ưu hóa các thuật toán và phương pháp để tăng độ chính xác và hiệu quả của việc dự đoán các từ khóa.

Sử dụng mô hình học máy: Xây dựng một mô hình học máy để nhận diện và sửa lỗi chính tả. Các mô hình như Random Forest hoặc Gradient Boosting có thể được huấn luyện để dự đoán mức độ quan trọng của keywords dựa trên các đặc trưng của keywords và ngữ cảnh.

Kết hợp các phương pháp: có thể kết hợp nhiều phương pháp trên để đánh giá mức độ quan trọng của từ khóa một cách tổng quát hơn. Ví dụ, bạn có thể sử dụng TF-IDF để xác định các từ khóa ban đầu, sau đó sử dụng attention scores từ BERT hoặc TextRank để tinh chỉnh mức độ quan trọng của các keywords này.

Phát triển hệ thống gợi ý lai (Hybrid Recommendation System): Hệ thống gợi ý lai có thể kết hợp nhiều phương pháp khác nhau, bao gồm Content-Based, Collaborative Filtering, và Knowledge-Based Systems. Có thể giúp tận dụng ưu điểm của các phương pháp. Điều này sẽ giúp nâng cao độ chính xác và tính đa dạng của các gợi ý, đồng thời khắc phục các hạn chế của từng phương pháp riêng lẻ. Việc kết hợp này sẽ giúp hệ thống đề xuất các phim không chỉ dựa trên nội dung mà còn dựa trên hành vi và sở thích của những người dùng có thị hiếu tương tự.

Tuy nhiên, hệ thống không phải lúc nào cũng hoàn hảo. Có thể xảy ra trường hợp hệ thống đưa ra đề xuất không chính xác hoặc không phù hợp. Do đó, trong nghiên cứu này hệ thống chỉ đóng vai trò hỗ

trợ cho cái quyết định của con người.

Acknowledgment

Cảm ơn Thầy TS.Nguyễn Gia Tuấn Anh và Thầy CN.Trần Quốc Khánh đã giúp nhóm hoàn thiện bài nghiên cứu.

References

[1] M. Fu, H. Qu, Z. Yi, L. Lu, and Y. Liu, "A novel deep multi-criteria collaborative filtering model for recommendation systems," Knowledge-Based Systems, vol. 187, Article 104811, 2020.
[2] S. R. S. Reddy, N. Sravani, K. Subramanyam, S. Ashok and B. Venkatesh, "Content-based movie recommendation system using genre correlation," in Smart Intelligent Computing and Applications, Smart Innovation, Systems and Technologies, In: S. C. Satapathy (Ed.), vol. 105, Singapore: Springer, pp. 391–397, 2019.
[3] Movie Recommendation: Netflix uses algorithm for recommending movies according to their interest. Other such platforms that provide recommendations include hotstar, sonyLIV, voot, ALTBalaji etc
[4] R. Banik, "Movie Recommender Systems," https://www.kaggle.com/code/rounakbanik/movie-recommender-systems/notebook.
[5] BK Excel, "How to Build a Movie Recommendation System," https://medium.com/@bkexcel2014/eff2d4e60d24.
[6] K. Cross, "Netflix's Movie Recommendation System," https://www.kaggle.com/code/krishcross/netflix-s-movie-recommendation-system/notebook#Content-Based-Filtering.
[7] A. Kamr, "Recommendation System from Zero to Hero," https://www.kaggle.com/code/abdokamr/recommendation-system-form-zero-to-hero/notebook#Recommendation-system-Collaborative-Filtering.