**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO MÔN HỌC**

**CHUYÊN ĐỀ CÔNG NGHỆ PHẦN MỀM**

**ĐỀ TÀI: NGHIÊN CỨU VỀ RECOMMENDATION SYSTEM**

**Nhóm sinh viên thực hiện**:

1. Võ Minh Quân - 5951071083

2. Trần Thế Vĩnh - 5751071050

3. Nguyễn Hoàng Hiếu - 5751071013

4. Nguyễn Văn Long - 5951071051

**Lớp** : CÔNG NGHỆ THÔNG TIN

**Khoá** :K59

Tp. Hồ Chí Minh, năm 2021

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO MÔN HỌC**

**CHUYÊN ĐỀ CÔNG NGHỆ PHẦN MỀM**

**ĐỀ TÀI: NGHIÊN CỨU VỀ RECOMMENDATION SYSTEM**

**Nhóm sinh viên thực hiện**:

1. Võ Minh Quân - 5951071038

2. Trần Thế Vĩnh - 5751071050

3. Nguyễn Hoàng Hiếu - 5751071013

4. Nguyễn Văn Long - 5951071051

**Lớp** : CÔNG NGHỆ THÔNG TIN

**Khoá** :K59

Tp. Hồ Chí Minh năm 2021

**MỤC LỤC**

[CHƯƠNG 1. TỔNG QUAN 6](#_Toc84302249)

[1.1 Giới thiệu 6](#_Toc84302250)

[1.2 Hệ thống Recommendation System 7](#_Toc84302251)

[1.2.1 Các khái niệm chính 7](#_Toc84302252)

[1.2.2 Thông tin phản hồi và hai dạng bài toán chính 8](#_Toc84302253)

[1.3 Các kỹ thuật chính 9](#_Toc84302254)

[1.3.1 Lọc cộng tác 9](#_Toc84302255)

[1.3.2 Lọc dựa trên nội dung 11](#_Toc84302256)

[1.3.3 Hệ thống gợi ý lai (Hybrid recommender systems) 13](#_Toc84302257)

[1.3.4 Các kỹ thuật không cá nhân hóa 16](#_Toc84302258)

[1.4 Hệ thống gợi ý tin tức 16](#_Toc84302259)

[CHƯƠNG 2. ÁP DỤNG THUẬT TOÁN GỢI Ý VỚI BỘ DỮ LIỆU THỰC TẾ 17](#_Toc84302260)

[2.1 Giới thiệu về Dataset 17](#_Toc84302261)

[2.2 Simple Recommender 18](#_Toc84302262)

[2.2.1 Đọc tập dữ liệu 18](#_Toc84302263)

[2.2.2 Quyết định điểm số để xếp hạng phim 19](#_Toc84302264)

[2.2.3 Tính điểm cho từng bộ phim 21](#_Toc84302265)

[2.2.4 Xếp hạng cho từng bộ phim dựa vào điểm số. 22](#_Toc84302266)

[2.3 Đề xuất dựa trên nội dung (Content – Baesd Recommender) 22](#_Toc84302267)

[2.3.1 Xây dựng ma trận 23](#_Toc84302268)

[2.3.2 Tính toán độ tương đồng giữa hai bộ phim 24](#_Toc84302269)

[2.3.3 Xây dựng chức năng gợi ý 25](#_Toc84302270)

[2.3.4 Đề xuất dựa trên nội dung, thể loại, diễn viên, từ khóa, đoàn phim. 27](#_Toc84302271)

[CHƯƠNG 3. KẾT LUẬN 34](#_Toc84302272)

[3.1 Kết quả đạt được 34](#_Toc84302273)

[3.3 Phương hướng phát triển 34](#_Toc84302274)

[TÀI LIỆU THAM KHẢO 35](#_Toc84302275)

**DANH MỤC HÌNH ẢNH**

[Hình 1. 1 Sơ đồ minh họa 6](#_Toc84303410)

[Hình 1. 2 Hệ thống gợi ý sản phẩm của Amazon 7](#_Toc84303411)

[Hình 1. 3 Ma trận biểu diễn dữ liệu 8](file:///D:\download\Document\word\TRƯỜNG%20ĐẠI%20HỌC%20GIAO%20THÔNG%20VẬN%20TẢI.docx#_Toc84303412)

[Hình 2. 1 Tải tệp dữ liệu 20](#_Toc84302214)

[Hình 2. 2 Dữ liệu trong movies\_metadata.csv 20](#_Toc84302215)

[Hình 2. 3 Dữ liệu mới 21](#_Toc84302216)

[Hình 2. 4 Xếp hạng trung bình 21](#_Toc84302217)

[Hình 2. 5 Số phếu bầu hơn 90% 22](#_Toc84302218)

[Hình 2. 6 Lọc các bộ phim trên 160 phiếu 22](#_Toc84302219)

[Hình 2. 7 Hàm tính điểm 23](#_Toc84302220)

[Hình 2. 8 Tính điểm cho mỗi bộ phim 23](#_Toc84302221)

[Hình 2. 9 Xếp hạng các bộ phim 24](#_Toc84302222)

[Hình 2. 10 Các bước xây dựng ma trận TF-IDF 25](#_Toc84302223)

[Hình 2. 11 Một số từ trong tfidf 25](#_Toc84302224)

[Hình 2. 12 Biểu diễn toán học của cosin similarty 26](#_Toc84302225)

[Hình 2. 13 Điểm tương đồng giữa các bộ phim 26](#_Toc84302226)

[Hình 2. 14 Hàm gợi ý 27](#_Toc84302227)

[Hình 2. 15 Kết quả gợi ý 28](#_Toc84302228)

[Hình 2. 16 Kết quả gợi ý 28](#_Toc84302229)

[Hình 2. 17 Hợp nhất dữ liệu 29](#_Toc84302230)

[Hình 2. 18 Kết quả hợp nhất 29](#_Toc84302231)

[Hình 2. 19 Lấy tên đạo diễn trong chuỗi json 30](#_Toc84302232)

[Hình 2. 20 Lấy tên thể loại, từ khóa, diễn viên 30](#_Toc84302233)

[Hình 2. 21 Kết quả sau khi lọc 30](#_Toc84302234)

[Hình 2. 22 Loại bỏ khoảng cách 31](#_Toc84302235)

[Hình 2. 23 Gom dữ liệu vào feature mới 31](#_Toc84302236)

[Hình 2. 24 Kết quả gom 32](#_Toc84302237)

[Hình 2. 25 Tạo ma trận và tính sự tương đồng 32](#_Toc84302238)

[Hình 2. 26 Hàm gợi ý 33](#_Toc84302239)

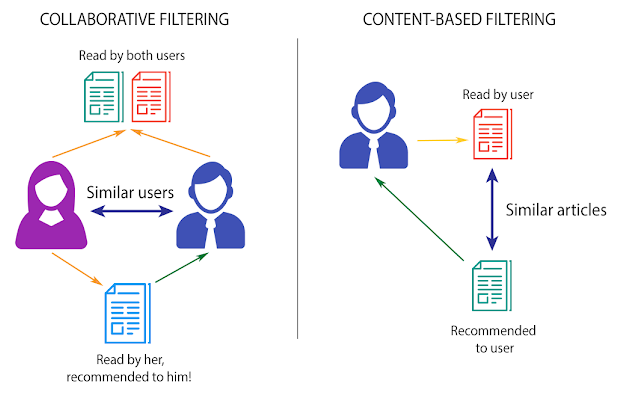
[Hình 2. 27 Kết quả cuối cùng 33](#_Toc84302240)

[Hình 2. 28 Kết quả cuối cùng 34](#_Toc84302241)

# CHƯƠNG 1. TỔNG QUAN

## 1.1 Giới thiệu

Hệ thống gợi ý (Recommender Systems / Recommendation Systems) là một dạng của hệ thống lọc thông tin (information filtering), nó được sử dụng để dự đoán sở thích (preferences) hay xếp hạng (rating) mà người dùng có thể dành cho một mục thông tin (item) nào đó mà họ chưa xem xét tới trong quá khứ (item có thể là bài hát, bộ phim, đoạn video clip, sách, bài báo,..).



Hình 1. 1 Sơ đồ minh họa

Ví dụ, trong hệ thống bán hàng trực tuyến (chẳng hạn như Amazon), nhằm tối ưu hóa khả năng mua sắm của khách hàng (user), người ta quan tâm đến việc những khách hàng nào đã ‘yêu thích’ những sản phẩm (item) nào bằng cách dựa vào dữ liệu quá khứ của họ (dữ liệu này có thể là xếp hạng mà người dùng đã bình chọn trên sản phẩm, thời gian duyệt (browse) trên sản phẩm, số lần click chuột trên sản phẩm,..) từ đó hệ thống sẽ dự đoán được người dùng có thể thích sản phẩm nào và đưa ra những gợi ý phù hợp cho họ. Hình 1 là một ví dụ minh họa cho hệ thống gợi ý bán hàng của Amazon.

Ngoài lĩnh vực thương mại điện tử như đã thấy ở ví dụ trên, hiện tại RS cũng được ứng dụng khá thành công trong nhiều lĩnh vực khác như trong giải trí: gợi ý bài hát cho người nghe (ví dụ, hệ thống của LastFM - www.last.fm), gợi ý phim ảnh (ví dụ, hệ thống của Netflix - www.netflix.com), gợi ý các video clip (ví dụ, hệ thống của YouTube - www.youtube.com); trong giáo dục và đào tạo (gợi ý nguồn tài nguyên học tập như sách, bài báo, địa chỉ web,… cho người học).

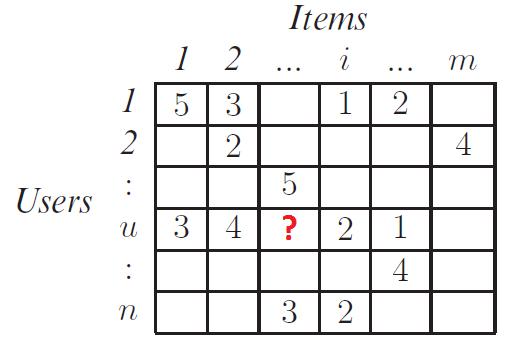


Hình 1. 2 Hệ thống gợi ý sản phẩm của Amazon

Hệ thống gợi ý không chỉ đơn thuần là một dạng Hệ thống thông tin mà nó còn là cả một lĩnh vực nghiên cứu hiện đang rất được các nhà khoa học quan tâm. Kể từ năm 2007 đến nay, hàng năm đều có hội thảo chuyên về hệ thống gợi ý của ACM (ACM RecSys) cũng như các tiểu bang dành riêng cho RS trong các hội nghị lớn khác như ACM KDD, ACM CIKM,.

## 1.2 Hệ thống Recommendation System

### 1.2.1 Các khái niệm chính

Trong Recommendation System, thông thường người ta quan tâm đến ba thông tin chính là người dùng (user), mục tin (item, item có thể là sản phẩm, bộ phim, bài hát, bài báo,.. tùy hệ thống), và phản hồi (feedback) của người dùng trên mục tin đó (thư ờng là các xếp hạng/đánh giá – rating biểu diễn mứcđộthích/quan tâm của họ). Các thông tin nàyđược biểu diễn thông qua một ma trận như trong Hình 2. Ở đó, mỗi dòng là một user, mỗi cột là một item, và mỗi ô là một giá trị phản hồi (ví dụ, xếp hạng) biểu diễn “mức độ thích” của user trên item tương ứng. Các ô có giá trị là những item mà các user đã xếp hạng trong quá khứ. Những ô trống là những item chưa được xếp hạng (điều đáng lưu ý là mỗi user chỉ xếp hạng cho một vài item trong quá khứ, do vậy có rất nhiều ô trống trong ma trận này – còn gọi là ma trận thưa – sparse matrix).

Hình 1. 3 Ma trận biểu diễn dữ liệu

Nhiệm vụ chính của RS là dựa vào các ô đã có giá trị trong ma trận trên (dữ liệu thu được từ quá khứ), thông qua mô hình đã đư ợc xây dựng, RS dự đoán các ô còn trống (của user hiện hành), sau đó sắp xếp kết quả dự đoán (ví dụ, từ cao xuống thấp) và chọn ra Top-N items theo thứ tự, từ đó gợi ý chúng cho người dùng.

### 1.2.2 Thông tin phản hồi và hai dạng bài toán chính

Trong Recommendation System, giá trị phản hồi (feedback) *rui* của mỗi người dùng trên mục tin sẽ được ghi nhận lại để làm cơ sở cho việc dự đoán các giá trị kế tiếp. Tùy theo hệ thống mà giá trị này sẽ có ý nghĩa khác nhau, ví dụ nó có thể dùng để đo độ “phù hợp” hay “mức độ thích” (thường là các đánh giá trên các sản phẩm) trong các hệ thống thương mại điện tử hay “năng lực/kết quả thực hiện” của người dùng trong các hệ thống e-learning.

Giá trị *rui* có thể được xác định một cách tường minh (explicit feedbacks) như thông qua việc đánh giá/xếp hạng (ví dụ, rating từ  đến ; hay like (1) và dislike (0),…) mà người dùng *u* đã bình chọn cho item *i*; hoặc *rui* có thể được xác định một cách không tường minh (implicit feedbacks) thông qua số lần click chuột, thời gian mà *u* đã duyệt/xem *i*,…

Có 2 dạng bài toán chính trong RS là *dự đoán xếp hạng (rating prediction)* của các hệ thống có phản hồi tường minh như đã trình bày ở trên và *dự đoán mục thông tin* (item prediction/recommendation) là việc xác định xác suất mà người dùng thích mục tin tương ứng.

## 1.3 Các kỹ thuật chính

### 1.3.1 Lọc cộng tác

Một cách tiếp cận để thiết kế các hệ thống recommender được sử dụng rộng rãi là lọc cộng tác. Các phương pháp lọc cộng tác dựa trên việc thu thập và phân tích một lượng lớn thông tin về hành vi, hoạt động hoặc sở thích của người dùng và dự đoán những gì người dùng sẽ thích dựa trên sự tương đồng của họ với người dùng khác. Một lợi thế quan trọng của phương pháp lọc cộng tác là nó không dựa vào nội dung phân tích máy và do đó nó có khả năng đề xuất chính xác các mục phức tạp như phim mà không yêu cầu “hiểu biết” về mục đó. Nhiều thuật toán đã được sử dụng để đo lường sự giống nhau của người dùng hoặc sự tương đồng về mặt hàng trong các hệ thống giới thiệu.

Ví dụ, cách tiếp cận hàng xóm gần nhất (k-nearest neighbor) và Pearson Correlation được Allen triển khai lần đầu tiên.

Lọc cộng tác dựa trên giả định rằng những người đã đồng ý trong quá khứ sẽ đồng ý trong tương lai và rằng họ sẽ thích các loại mặt hàng tương tự như họ thích trong quá khứ.

Khi xây dựng mô hình từ hành vi của người dùng, sự phân biệt thường được thực hiện giữa các hình thức thu thập dữ liệu rõ ràng và tiềm ẩn.

Ví dụ về thu thập dữ liệu rõ ràng bao gồm:

* Yêu cầu người dùng xếp hạng một mục trên thang trượt.
* Yêu cầu người dùng tìm kiếm.
* Yêu cầu người dùng xếp hạng một bộ sưu tập các mục từ yêu thích đến ít yêu thích nhất.
* Trình bày hai mục cho một người dùng và yêu cầu anh ta / cô ấy chọn một trong số chúng tốt hơn.
* Yêu cầu người dùng tạo danh sách các mục mà anh / cô ấy thích.

Ví dụ về thu thập dữ liệu ngầm bao gồm:

* Quan sát các mục mà người dùng xem trong cửa hàng trực tuyến.
* Phân tích thời gian xem mục / người dùng.
* Lưu giữ một bản ghi các mục mà người dùng mua trực tuyến.
* Lấy danh sách các mục mà người dùng đã nghe hoặc xem trên máy tính của họ.
* Phân tích mạng xã hội của người dùng và khám phá những lượt thích và không thích tương tự.

Hệ thống recommender so sánh dữ liệu đã thu thập với dữ liệu tương tự và khác nhau được thu thập từ những người khác và tính toán danh sách các mục được đề xuất cho người dùng. Một số ví dụ thương mại và phi thương mại được liệt kê trong bài viết về các hệ thống lọc cộng tác .

Một trong những ví dụ nổi tiếng nhất về lọc cộng tác là lọc cộng tác theo từng mục (những người mua x cũng mua y), một thuật toán được phổ biến rộng rãi bởi hệ thống gợi ý của Amazon.com. Các ví dụ khác bao gồm:

* Như đã đề cập chi tiết ở trên, Last.fm đề xuất âm nhạc dựa trên so sánh thói quen nghe của những người dùng tương tự, trong khi Readgeek so sánh xếp hạng sách cho các đề xuất.
* Facebook , MySpace , LinkedIn và các mạng xã hội khác sử dụng tính năng lọc cộng tác để giới thiệu bạn bè, nhóm và các kết nối xã hội khác (bằng cách kiểm tra mạng kết nối giữa người dùng và bạn bè của họ). Twitter sử dụng nhiều tín hiệu và tính toán trong bộ nhớ để giới thiệu cho người dùng của họ rằng họ nên “theo dõi”.

Các phương pháp lọc cộng tác thường gặp phải ba vấn đề: Cold Start, khả năng mở rộng và sự thưa thớt (sparsity).

* Cold Start: Các hệ thống này thường yêu cầu một lượng lớn dữ liệu hiện có của người dùng để đưa ra các đề xuất chính xác.
* Khả năng mở rộng: Trong nhiều môi trường mà các hệ thống này đưa ra các khuyến nghị, có hàng triệu người dùng và sản phẩm. Do đó, một lượng lớn công suất tính toán thường là cần thiết để tính toán các gợi ý.
* Sparsity: Số lượng các mặt hàng được bán trên các trang web thương mại điện tử lớn là cực kỳ lớn. Những người dùng tích cực nhất sẽ chỉ đánh giá một tập con nhỏ của cơ sở dữ liệu tổng thể. Do đó, ngay cả những mặt hàng phổ biến nhất cũng có rất ít xếp hạng.

Một loại thuật toán lọc cộng tác cụ thể sử dụng hệ số ma trận hóa (matrix factorization), kỹ thuật xấp xỉ ma trận cấp thấp (low-rank matrix approximation).

Các phương pháp lọc cộng tác được phân loại là bộ lọc cộng tác dựa trên bộ nhớ và dựa trên mô hình. Một ví dụ nổi tiếng về các phương pháp dựa trên bộ nhớ là thuật toán dựa trên người dùng và các phương pháp dựa trên mô hình là Kernel-Mapping Recommender.

### 1.3.2 Lọc dựa trên nội dung

Một cách tiếp cận phổ biến khác khi thiết kế hệ thống recommender là lọc nội dung . Phương pháp lọc dựa trên nội dung dựa trên mô tả về mặt hàng và hồ sơ về các tùy chọn của người dùng.

Trong hệ thống gợi ý dựa trên nội dung, từ khóa được sử dụng để mô tả các mục và hồ sơ người dùng được xây dựng để chỉ ra loại mục mà người dùng này thích. Nói cách khác, các thuật toán này cố gắng đề xuất các mục tương tự với các mục mà người dùng đã thích trong quá khứ (hoặc đang kiểm tra trong hiện tại). Cụ thể, các mục đề cử khác nhau được so sánh với các mục được đánh giá trước đây bởi người dùng và các mục phù hợp nhất được đề xuất. Cách tiếp cận này có nguồn gốc từ việc thu thập thông tin và nghiên cứu lọc thông tin .

Để tóm tắt các tính năng của các mục trong hệ thống, một thuật toán trình bày mục được áp dụng. Một thuật toán được sử dụng rộng rãi là **biểu diễn tf – idf**(còn được gọi là biểu diễn không gian vectơ).

Để tạo hồ sơ người dùng, hệ thống chủ yếu tập trung vào hai loại thông tin:

1. Một mô hình ưu tiên của người dùng.

2. Lịch sử tương tác của người dùng với hệ thống gợi ý.

Về cơ bản, các phương thức này sử dụng một hồ sơ mặt hàng (ví dụ, một tập hợp các thuộc tính và tính năng rời rạc) mô tả mục trong hệ thống. Hệ thống tạo hồ sơ dựa trên nội dung của người dùng dựa trên vectơ trọng số của các đối tượng địa lý. Trọng số biểu thị tầm quan trọng của từng tính năng đối với người dùng và có thể được tính từ các vectơ nội dung được xếp hạng riêng lẻ bằng nhiều kỹ thuật. Các phương pháp đơn giản sử dụng các giá trị trung bình của vector hạng mục trong khi các phương pháp phức tạp khác sử dụng các kỹ thuật máy học như Bayesian Classifiers , phân tích cụm , cây quyết định và mạng thần kinh nhân tạo (artificial neural networks) để ước tính xác suất người dùng sẽ thích mục đó.

Phản hồi trực tiếp từ người dùng, thường dưới dạng nút thích hoặc không thích , có thể được sử dụng để gán trọng số cao hơn hoặc thấp hơn về tầm quan trọng của các thuộc tính nhất định (sử dụng phân loại Rocchio hoặc các kỹ thuật tương tự khác).

Một vấn đề quan trọng với lọc dựa trên nội dung là liệu hệ thống có thể tìm hiểu các tùy chọn của người dùng từ hành động của người dùng liên quan đến một nguồn nội dung hay không và sử dụng chúng trên các loại nội dung khác. Khi hệ thống bị hạn chế đề xuất nội dung cùng loại với người dùng đang sử dụng, giá trị từ hệ thống đề xuất thấp hơn đáng kể so với các loại nội dung khác từ các dịch vụ khác có thể được đề xuất. Ví dụ: giới thiệu các bài viết tin tức dựa trên việc duyệt tin tức hữu ích nhưng sẽ hữu ích hơn nhiều khi bạn có thể đề xuất âm nhạc, video, sản phẩm, cuộc thảo luận, v.v. từ các dịch vụ khác nhau dựa trên duyệt tin tức.

Pandora Radio là một ví dụ về hệ thống giới thiệu dựa trên nội dung phát nhạc có các đặc điểm tương tự như một bài hát do người dùng cung cấp làm hạt giống ban đầu. Ngoài ra còn có một số lượng lớn các hệ thống gợi ý dựa trên nội dung nhằm cung cấp các đề xuất phim, một vài ví dụ như Rotten Tomatoes , Internet Movie Database , Jinni , Rovi Corporation và Jaman . Các hệ thống gợi ý giới thiệu tài liệu liên quan nhằm mục đích cung cấp các đề xuất tài liệu cho các nhà nghiên cứu. Các chuyên gia y tế công cộng đã nghiên cứu các hệ thống gợi ý để cá nhân hóa giáo dục sức khỏe và các chiến lược phòng ngừa.

### 1.3.3 Hệ thống gợi ý lai (Hybrid recommender systems)

Nghiên cứu gần đây đã chứng minh rằng một phương pháp lai, kết hợp lọc cộng tác và lọc dựa trên nội dung có thể hiệu quả hơn trong một số trường hợp. Các phương pháp lai có thể được thực hiện theo nhiều cách:

* Bằng cách đưa ra các dự đoán dựa trên nội dung và dựa trên lọc cộng tác riêng biệt và sau đó kết hợp chúng.
* Bằng cách thêm các khả năng dựa trên nội dung vào phương pháp cộng tác (và ngược lại).
* Bằng cách thống nhất các phương pháp tiếp cận thành một mô hình.

Một số nghiên cứu thực nghiệm so sánh hiệu suất của phương pháp lai với các phương pháp cộng tác thuần túy và chứng minh rằng các phương pháp lai có thể cung cấp các khuyến nghị chính xác hơn các phương pháp thuần túy. Những phương pháp này cũng có thể được sử dụng để khắc phục một số vấn đề thường gặp trong hệ thống gợi ý như Cold Start và vấn đề thưa thớt.

Netflix là một ví dụ tốt về việc sử dụng các hệ thống hybrid recommender. Trang web đưa ra các đề xuất bằng cách so sánh thói quen xem và tìm kiếm của những người dùng tương tự (ví dụ: lọc cộng tác) cũng như bằng cách cung cấp những bộ phim có chung đặc điểm với những bộ phim mà người dùng đánh giá cao (lọc dựa trên nội dung).

Một loạt các kỹ thuật đã được đề xuất làm cơ sở cho các hệ thống gợi ý: các kỹ thuật hợp tác (collaborative), dựa trên nội dung (content-based), dựa trên kiến ​​thức (knowledge-based) và nhân khẩu học (demographic techniques). Mỗi kỹ thuật này đều có những thiếu sót, như vấn đề Cold Start cho các hệ thống cộng tác và dựa trên nội dung (phải làm gì với người dùng mới với ít xếp hạng) và tắc nghẽn kỹ thuật tri thức (knowledge engineering bottleneck) trong các phương pháp dựa trên tri thức . Một hệ thống gợi ý lai là một hệ thống trong đó kết hợp nhiều kỹ thuật với nhau để đạt được một số sức mạnh tổng hợp giữa chúng.

* Cộng tác – Collaborative: Hệ thống tạo đề xuất chỉ sử dụng thông tin về hồ sơ xếp hạng cho những người dùng hoặc mục khác nhau. Các hệ thống cộng tác định vị “người dùng/mục” ngang hàng với lịch sử xếp hạng tương tự như người dùng hoặc mục hiện tại và tạo đề xuất sử dụng vùng lân cận này. Các thuật toán dựa trên người dùng và dựa trên hàng gần nhất có thể được kết hợp để giải quyết vấn đề Cold Start và cải thiện kết quả đề xuất.
* Dựa trên nội dung – Content-based: Hệ thống tạo đề xuất từ ​​hai nguồn: các tính năng liên quan đến sản phẩm và xếp hạng mà người dùng đã cung cấp cho họ. Đề xuất dựa trên nội dung coi đề xuất là sự cố phân loại người dùng cụ thể và tìm hiểu trình phân loại cho lượt thích và không thích của người dùng dựa trên các tính năng của sản phẩm.
* Nhân khẩu học – demographic techniques: Trình giới thiệu nhân khẩu học cung cấp các đề xuất dựa trên hồ sơ nhân khẩu học của người dùng. Sản phẩm được đề xuất có thể được sản xuất cho các mục nhân khẩu học khác nhau, bằng cách kết hợp xếp hạng của người dùng trong các mục đó.
* Dựa trên tri thức – knowledge-based: Trình giới thiệu dựa trên kiến ​​thức gợi ý các sản phẩm dựa trên các suy luận về nhu cầu và sở thích của người dùng. Kiến thức này đôi khi sẽ chứa kiến ​​thức chức năng rõ ràng về cách các tính năng sản phẩm nhất định đáp ứng nhu cầu của người dùng.

Thuật ngữ Hybrid recommender systems được sử dụng ở đây để mô tả bất kỳ hệ thống recommender nào kết hợp nhiều kỹ thuật đề xuất với nhau để tạo dữ liệu đầu ra của nó.

**Có bảy kỹ thuật lai cơ bản (hybridization techniques):**

* Có trọng số (Weighted): Điểm số của các thành phần đề xuất khác nhau được kết hợp theo số lượng.
* Chuyển đổi (Switching): Hệ thống chọn giữa các thành phần đề xuất và áp dụng hệ thống được chọn.
* Hỗn hợp (Mixed): Các khuyến nghị từ những người giới thiệu khác nhau được trình bày cùng nhau để đưa ra đề xuất.
* Kết hợp tính năng (Feature Combination): Các tính năng được lấy từ các nguồn tri thức khác nhau được kết hợp với nhau và được đưa ra cho một thuật toán gợi ý duy nhất.
* Tính năng tăng cường (Feature Augmentation): Một kỹ thuật gợi ý được sử dụng để tính toán một tính năng hoặc tập hợp các tính năng, sau đó là một phần của đầu vào cho kỹ thuật tiếp theo.
* Cascade: Các khuyến nghị được ưu tiên nghiêm ngặt, với những ưu tiên thấp hơn phá vỡ các mối quan hệ trong việc tính điểm của những người cao hơn.
* Cấp độ meta (Meta-level): Một kỹ thuật đề xuất được áp dụng và tạo ra một số loại mô hình, sau đó là đầu vào được sử dụng bởi kỹ thuật tiếp theo.

### 1.3.4 Các kỹ thuật không cá nhân hóa

Trong nhóm kỹ thuật này, do chúng khá đơn giản, dễ cài đặt nên nên thường được các website/hệ thống tích hợp vào, gồm cả các website thương mại, website tin tức, hay giải trí. Chẳng hạn như trong các hệ thống bán hàng trực tuyến, người ta thường gợi ý các sản phẩm được xem/mua/bình luận/.. nhiều nhất; gợi ý các sản phẩm mới nhất; gợi ý các sản phẩm cùng loại/ cùng nhà sản xuất/..; gợi ý các sản phẩm được mua/chọn cùng nhau. Một ví dụ khá điển hình là thông qua luật kết hợp (như Apriori), Amazon đã áp dụng khá thành công để tìm ra các sản phẩm hay được mua cùng nhau.

Tuy vậy, bất lợi của các phương pháp này là không cá nhân hóa cho từng người dùng, nghĩa là tất cả các user đều được gợi ý giống nhau khi chọn cùng sản phẩm.

## 1.4 Hệ thống gợi ý tin tức

Các cổng tin tức phổ biến, như Google News, Yahoo! News, The New York Times, Washington Post, cùng với nhiều cổng thông tin khác đã thu hút được sự chú ý ngày càng tăng từ một lượng lớn độc giả trên internet. Các hệ thống khuyến nghị tin tức trực tuyến đã được các nhà nghiên cứu đề cập đến trong những năm qua, bằng cách sử dụng nhiều phương pháp khác nhau: lọc dựa trên nội dung, lọc cộng tác và phương pháp lai kết hợp.

Một số thách thức đối với hệ khuyến nghị tin tức có thể kể đến:

* Hồ sơ người dùng thưa thớt - phần lớn độc giả là ẩn danh và họ thực sự chỉ đọc một vài câu chuyện từ toàn bộ kho lưu trữ. Điều này dẫn đến mức độ thưa thớt cực cao trong ma trận bài viết - người dùng, vì người dùng thường theo dõi rất ít thông tin về hành vi trong quá khứ của họ, nếu có.
* Số lượng bài viết tăng nhanh - hàng trăm bài viết mới được thêm vào hàng ngày trong các cổng tin tức (ví dụ: hơn 300 bài trên trang The New York Times). Điều này làm nghiêm trọng vấn đề cold-start, vì đối với các bài viết mới, ta không có nhiều tương tác trong quá khứ để có thể dựa vào đó và đề xuất chúng. Đối với các công cụ tổng hợp tin tức, các vấn đề về khả năng mở rộng có thể phát sinh, vì một khối lượng lớn các bài báo sẽ làm quá tải web trong khoảng thời gian giới hạn.
* Thời gian sống của bài viết - giá trị thông tin phân rã theo thời gian. Điều này đặc biệt đúng trong lĩnh vực tin tức, vì hầu hết người dùng quan tâm đến thông tin mới. Vì vậy, mỗi bài viết ​​sẽ có thời hạn sử dụng ngắn. Thị hiếu của người dùng liên tục thay đổi - chủ đề tin tức được quan tâm không ổn định như trong lĩnh vực giải trí. Một số sở thích của người dùng thay đổi theo thời gian, trong khi một số chủ đề khác vẫn ổn định. Mối quan tâm hiện tại của người dùng trong một phiên truy cập có thể bị ảnh hưởng bởi bối cảnh của phiên đó (ví dụ: địa điểm, thời gian truy cập) hoặc bởi các bối cảnh chung (ví dụ: tin nóng hoặc các sự kiện quan trọng).

# CHƯƠNG 2. ÁP DỤNG THUẬT TOÁN GỢI Ý VỚI BỘ DỮ LIỆU THỰC TẾ

## 2.1 Giới thiệu về Dataset

Các tệp tập dữ liệu chứa siêu dữ liệu cho tất cả 45.000 phim được liệt kê trong Full MovieLens Dataset. Tập dữ liệu bao gồm các phim được phát hành vào hoặc trước tháng 7 năm 2017. Tập dữ liệu này ghi lại các điểm đặc trưng như dàn diễn viên, đoàn làm phim, từ khóa cốt truyện, ngân sách, doanh thu, áp phích, ngày phát hành, ngôn ngữ, công ty sản xuất, quốc gia, số phiếu bầu của TMDB và số phiếu trung bình.

Tập dữ liệu này bao gồm các tệp sau:

* movies\_metadata.csv : Tệp này chứa thông tin về ~ 45.000 phim có trong tập dữ liệu Full MovieLens. Các tính năng bao gồm áp phích, phông nền, ngân sách, thể loại, doanh thu, ngày phát hành, ngôn ngữ, quốc gia sản xuất và công ty.
* keywords.csv : Chứa các từ khóa cốt truyện phim cho các phim MovieLens. Có sẵn ở dạng đối tượng JSON được xâu chuỗi.
* credits.csv : Bao gồm thông tin về diễn viên và đoàn làm phim cho tất cả các bộ phim. Có sẵn ở dạng đối tượng JSON được xâu chuỗi.
* links.csv : Tệp này chứa ID TMDB và IMDB của tất cả các phim có trong tập dữ liệu Full MovieLens.
* links\_small.csv : Chứa ID TMDB và IMDB của một tập con nhỏ gồm 9.000 phim của Tập dữ liệu đầy đủ.
* rating\_small.csv : Tập hợp con gồm 100.000 xếp hạng từ 700 người dùng trên 9.000 phim.

## 2.2 Simple Recommender

Simple Recommender là hệ thống cơ bản đề xuất các mục hàng đầu dựa trên một số liệu hoặc điểm số nhất định.

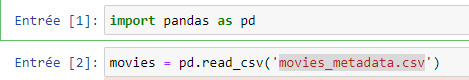
Sau đây là các bước liên quan:

* Quyết định chỉ số hoặc điểm số để xếp hạng phim.
* Tính điểm cho mỗi bộ phim.
* Sắp xếp phim dựa trên điểm số và đưa ra kết quả hàng đầu.

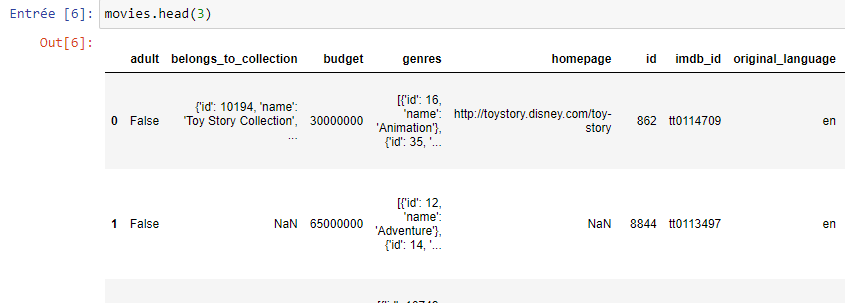
Trong phần này chúng ta sử dụng bộ dữ liệu movies\_metadata.csv trong bộ dữ liệu Full MovieLens Dataset.

### 2.2.1 Đọc tập dữ liệu

Để đọc tập dữ liệu trước hết cần phải thêm thư viện pandas. Thư viện pandas chủ yếu được sử dụng để thao tác và phân tích dữ liệu. Nó đại diện cho dữ liệu của bạn ở định dạng hàng-cột.



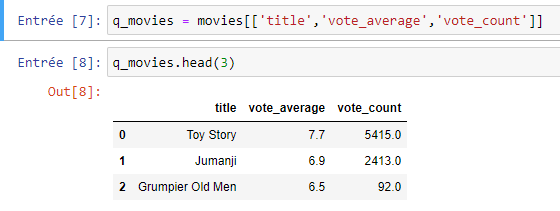
Hình 2. 1 Tải tệp dữ liệu



Hình 2. 2 Dữ liệu trong movies\_metadata.csv

### 2.2.2 Quyết định điểm số để xếp hạng phim

Tạo dữ liệu q\_movies mới độc lập với dữ liệu ban đầu.



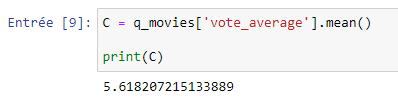
Hình 2. 3 Dữ liệu mới

Công thức xếp hạng phim được biểu diễn như sau:

Trong phương trình trên,

* v là số phiếu bầu cho phim;
* m là số phiếu tối thiểu cần thiết để được liệt kê trong biểu đồ;
* R là điểm đánh giá trung bình của bộ phim;
* C là bình chọn trung bình trên toàn bộ báo cáo.

Bước đầu tiên, hãy tính giá trị của C, xếp hạng trung bình trên tất cả các phim bằng cách sử dụng hàm mean().



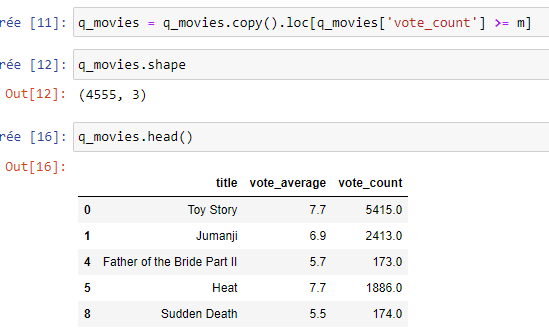
Hình 2. 4 Xếp hạng trung bình

Số phiếu bình chọn nhiều hơn ít nhất 90% các bộ phim trong danh sách.



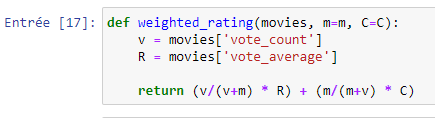
Hình 2. 5 Số phếu bầu hơn 90%

Lọc các bọ phim có phiếu bầu lớn hơn hoặc bằng 160.

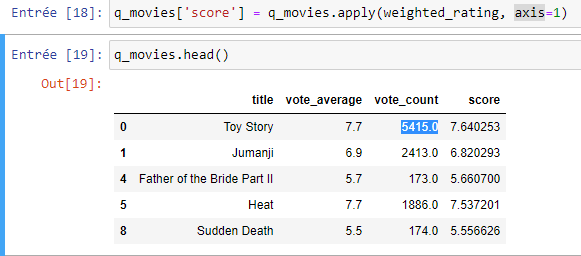


Hình 2. 6 Lọc các bộ phim trên 160 phiếu

### 2.2.3 Tính điểm cho từng bộ phim

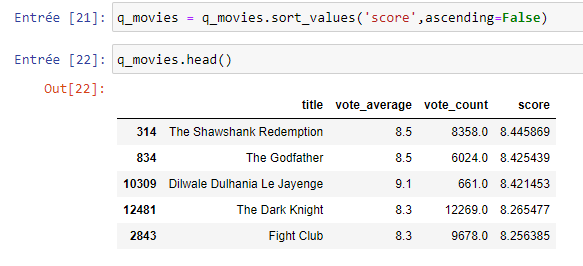


Hình 2. 7 Hàm tính điểm



Hình 2. 8 Tính điểm cho mỗi bộ phim

### 2.2.4 Xếp hạng cho từng bộ phim dựa vào điểm số.



Hình 2. 9 Xếp hạng các bộ phim

## 2.3 Đề xuất dựa trên nội dung (Content – Baesd Recommender)

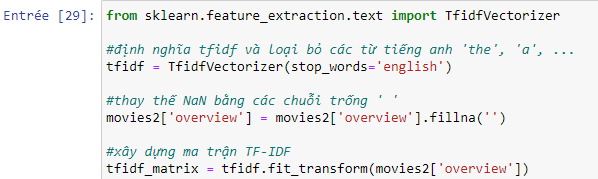
Đề xuất dựa trên nội dung là hệ thống đề xuất các bộ phim có nội dung tương tự như một bộ phim cụ thể.

### 2.3.1 Xây dựng ma trận

Tính toán Term Frequency-Inverse Document Frequency vectơ (TF-IDF) cho mỗi tài liệu. Điều này sẽ cung cấp một ma trận trong đó mỗi cột đại diện cho một từ trong từ vựng tổng quan (tất cả các từ xuất hiện trong ít nhất một tài liệu) và mỗi cột đại diện cho một bộ phim.

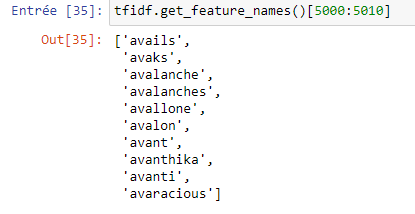
Scikit-learning cung cấp một lớp TfIdfVectorizer tích hợp tạo ra ma trận TF-IDF.

* Nhập mô-đun Tfidf bằng scikit-learning;
* Loại bỏ các từ dừng như 'the', 'an', v.v. vì chúng không cung cấp bất kỳ thông tin hữu ích nào về chủ đề;
* Thay thế các giá trị not-a-number bằng một chuỗi trống;
* Cuối cùng, xây dựng ma trận TF-IDF trên dữ liệu.



Hình 2. Các bước xây dựng ma trận TF-IDF

Một số từ vựng trong tập dữ liệu

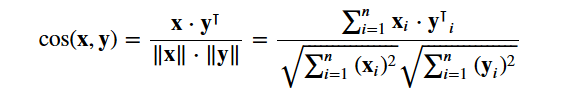


Hình 2. 11 Một số từ trong tfidf

### 2.3.2 Tính toán độ tương đồng giữa hai bộ phim

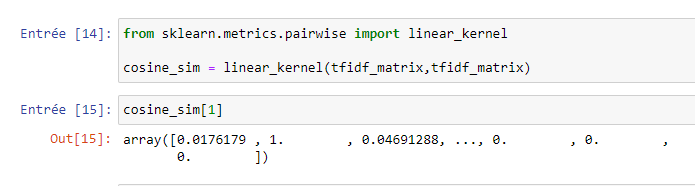
Sử dụng cosine similarity để tính toán một đại lượng số biểu thị sự giống nhau giữa hai bộ phim.

Về mặt toán học, nó được định nghĩa như sau:



Hình 2. Biểu diễn toán học của cosin similarty

Vì đã sử dụng vectơ TF-IDF, việc tính tích số chấm giữa mỗi vectơ sẽ trực tiếp cung cấp điểm tương tự cosin. Do đó, sử dụng sklearn's linear\_kernel() thay cosine\_similarities() vì nó nhanh hơn.

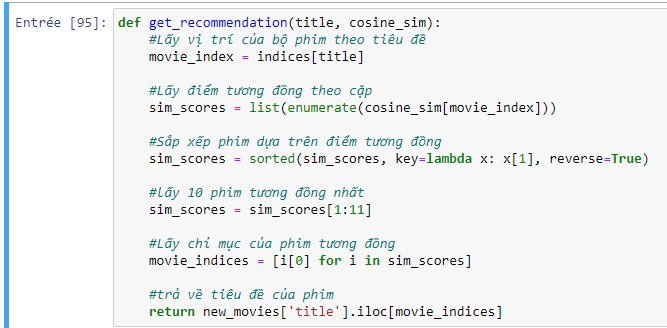


Hình 2. Điểm tương đồng giữa các bộ phim

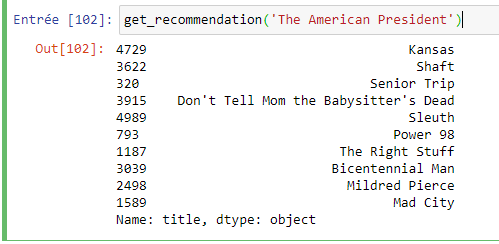
### 2.3.3 Xây dựng chức năng gợi ý

Các bước xây dựng:

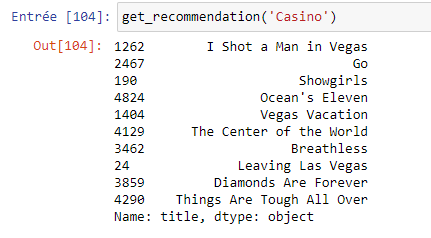
* Nhận chỉ số của bộ phim với tiêu đề của nó.
* Nhận danh sách điểm tương đồng cosine của bộ phim cụ thể đó với tất cả các bộ phim. Chuyển nó thành một danh sách các bộ giá trị trong đó phần tử đầu tiên là vị trí của nó và phần tử thứ hai là điểm tương tự.
* Sắp xếp danh sách các bộ giá trị nói trên dựa trên điểm số tương tự
* Nhận 10 yếu tố hàng đầu của danh sách này. Bỏ qua yếu tố đầu tiên vì nó đề cập đến bản thân (bộ phim giống nhất với một bộ phim cụ thể là chính bộ phim đó).
* Trả lại các tiêu đề tương ứng với chỉ số của các phần tử trên cùng.



Hình 2. Hàm gợi ý



Hình 2. 15 Kết quả gợi ý



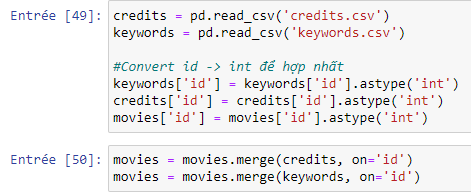
Hình 2. 16 Kết quả gợi ý

### 2.3.4 Đề xuất dựa trên thể loại, diễn viên, từ khóa, đoàn phim.

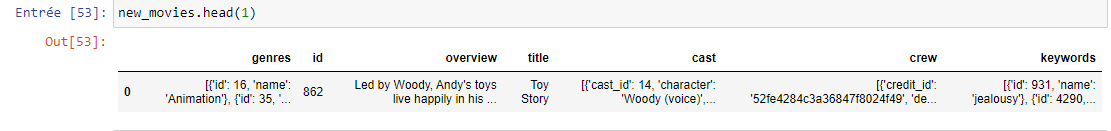
Trong khi hệ thống đã thực hiện tốt công việc tìm kiếm các bộ phim có mô tả cốt truyện tương tự, chất lượng của các đề xuất không phải là tuyệt vời. "The Dark Knight Rises" trả lại tất cả các phim Batman trong khi nhiều khả năng những người đã thích phim đó có xu hướng thưởng thức các phim khác của Christopher Nolan. Đây là thứ mà hệ thống hiện tại không thể nắm bắt được.

Vì vậy, xây dựng hệ thống giới thiệu dựa trên siêu dữ liệu sau: 3 diễn viên hàng đầu, đạo diễn, thể loại liên quan và từ khóa cốt truyện phim.

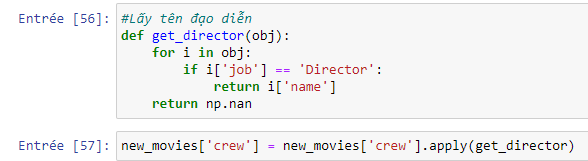
Dữ liệu từ khóa, diễn viên và đoàn phim không có sẵn trong tập dữ liệu hiện tại, vì vậy bước đầu tiên sẽ là tải và hợp nhất chúng vào DataFrame.



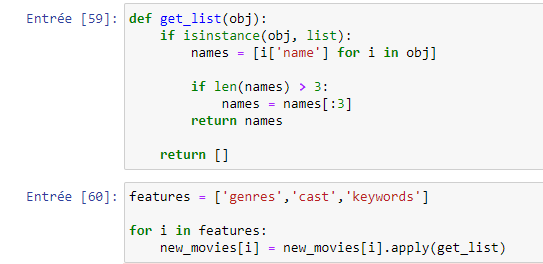
Hình 2. Hợp nhất dữ liệu



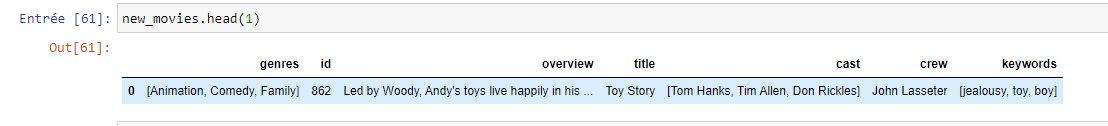
Hình 2. Kết quả hợp nhất



Hình 2. Lấy tên đạo diễn trong chuỗi json

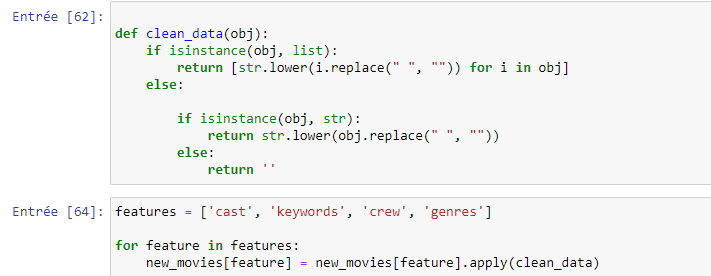


Hình 2. Lấy tên thể loại, từ khóa, diễn viên

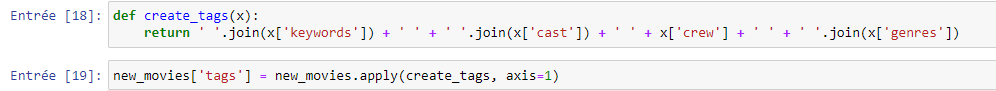


Hình 2. Kết quả sau khi lọc

Chuyển đổi tên và các trường hợp từ khóa thành chữ thường và loại bỏ tất cả các khoảng cách giữa chúng.



Hình 2. Loại bỏ khoảng cách

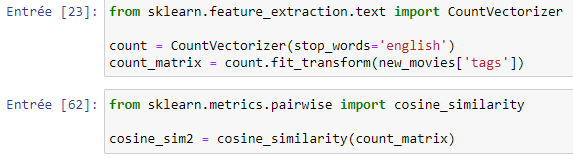


Hình 2. Gom dữ liệu vào feature mới

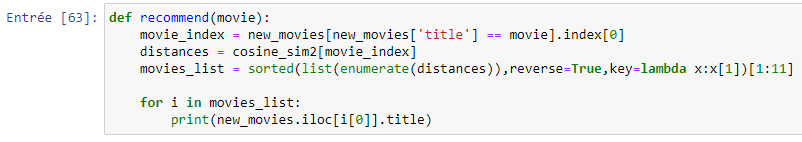


Hình 2. Kết quả gom

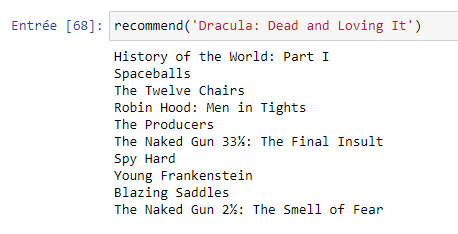
Sử dụng CountVectorizer() thay vì TF-IDF. Điều này là do không muốn giảm bớt sự hiện diện của diễn viên / đạo diễn nếu anh ấy hoặc cô ấy đã tham gia diễn xuất hoặc đạo diễn trong các bộ phim tương đối nhiều hơn.



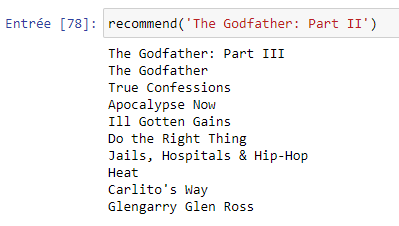
Hình 2. 25 Tạo ma trận và tính sự tương đồng



Hình 2. Hàm gợi ý



Hình 2. Kết quả cuối cùng



Hình 2. Kết quả cuối cùng

# CHƯƠNG 3. KẾT LUẬN

## 3.1 Kết quả đạt được

Tìm hiểu về hệ thống đề xuất, phương thức hoạt động và các cách đề xuất của hệ thống.

Thành công học cách xây dựng mô hình cơ bản của hệ thống khuyến nghị đơn giản và dựa trên nội dung.

## Phương hướng phát triển

Áp dụng những kỹ thuật đề xuất khác nhau như:

* Hybrid Recommender
* Neighborhood-Based Collaborative Filtering
* Matrix Factorization Collaborative Filtering

# TÀI LIỆU THAM KHẢO

[1] <https://github.com/gabrielspmoreira/chameleon_recsys>

[2] <https://machinelearningcoban.com/>

[3] <https://vi.wikipedia.org/>

[4] <https://viblo.asia/p/lam-the-nao-de-xay-dung-mot-recommender-system-rs-phan-1-aWj53V2Gl6m>