HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Khoa Công Nghệ Thông tin



**Phát triển các hệ thống dựa trên tri thức**

**Đề tài: Hệ thống gợi ý phim**

**Giảng viên**

**Trần Tiến Công**

**Sinh viên thực hiện**

**Nhóm 11**

**Lê Khắc Nam B18DCCN424**

**Phạm Anh Khiêm B18DCCN314**

**Nguyễn Khả Thiết B18DCCN639**

**Hà Nội, 2022**

**MỞ ĐẦU**

Kể từ khi xuất hiện đến nay, phim ảnh vẫn luôn đóng một vai trò quan trọng, có thể khẳng định là không thể thay thế trong đời sống con người. Tuy là một ngành giải trí nhưng giá trị của phim ảnh truyền tải đến con người luôn mang những giá trị nhân văn to lớn. Cũng vì thế mà trong nhiều thập kỷ gần đây, các nhà sản xuất đã cho ra đời vô số những tác phẩm phim ảnh với mọi nội dung, chủ đề,… mà đáp ứng nhu cầu người xem. Nhưng sự phát triển mạnh mẽ về số lượng đó khiến những khán giá yêu phim lạc vào trong mê cung phim, không biết chọn lựa những bộ phim nào để xem phù hợp với yêu cầu, đề tài mà họ muốn.

Để đáp ứng nhu cầu này, một số forum, hội nhóm và những người đánh giá phim xuất hiện nhằm giải đáp những thắc mắc của người xem, đồng thời giới thiệu, gợi ý các bộ phim. Tuy nhiên những giải pháp này vẫn có những mặt hạn chế.

Tốc độ phản hồi chậm: khi một người cần tìm bộ phim về chủ đề mà mình muốn, họ sẽ đặt câu hỏi và chờ cộng đồng hoặc reviewer nào đó trả lời qua văn bản hoặc video. Hay khi bất kì người nào đó tìm kiếm về những bộ phim, họ sẽ lang thang trên internet. Việc chờ đợi, tìm kiếm này tốn không ít thời gian.

Lượng thông tin mang lại ít: Vì đây là cách làm thủ công, con người không ai có thể xem hết những bộ phim, và đưa ra gợi ý cho người khác khi được hỏi được. Vì vậy có thể bỏ sót những bộ phim hay cho người khác.

Ngoài ra trong những năm gần đây xảy ra đại dịch lớn covid trên toàn thế giới cũng kiến nhu cầu giải trí của con người dần chuyển dịch hướng online internet thay vì ra rạp xem phim thì mọi người lựa chọn xem phim ở nhà thông qua các ứng dụng, website và đây cũng là 1 cơ hội lớn để các app và các website giải trí phát triển đồng thời cũng là 1 sự cạnh tranh giữa các ứng dụng đòi hỏi phải đưa ra được trải nghiệm tốt nhất cho người dùng, với sự phát triển của công nghệ, lượng dữ liệu về phim được thu thập trong nhiều năm, vì vậy nghiên cứu tập trung vào việc xây dựng các hệ thống gợi ý phim là 1 vấn đề tiên quyết để tăng tối ưu trải nghiệm và thu hút người dùng.

Việc xây dựng hệ thống gợi ý phim giúp cải thiện thời gian tìm kiếm phim của người dùng, tăng trải nghiệm sử dụng dịch vụ.

Đồng thời đây cũng là yếu tố giúp các công ty, tổ chức cung cấp dịch vụ phim ảnh có thể dựa vào hệ thống có thể giữ chân, thu hút khách hàng mới và cạnh tranh với các công ty, tổ chức trong cùng lĩnh vực.

Trong đề tài này chúng tôi xin đề xuất phương pháp nghiên cứu phát triển hệ thống gợi ý phim trên nền tảng website dựa trên các thuật toán lọc để đề xuất các bộ phim phù hợp hơn đối với người dùng, góp phần tăng trải nghiệm của người dùng trên website.

Nội dung nghiên cứu của đề tài bao gồm những nội dung chính sau đây:

**Nội dung 1**: Khảo sát nghiên cứu tổng quan các hệ thống gợi ý phim cùng với các mô hình kỹ thuật đang được sử dụng ở Việt Nam và trên thế giới

**Nội dung 2**: Tìm kiếm, đánh giá và phân tích bộ dữ liệu để phục vụ cho hệ thống gợi ý phim.

* Tìm kiếm bộ dữ liệu phù hợp cho hệ thống
* Đánh giá, phân tích bộ dữ liệu để lựa chọn các đặc trưng phù hợp
* Craw thêm thông tin cần thiết cho hệ thống

**Nội dung 3**: Nghiên cứu xây dựng, thiết kế mô hình gợi ý phim dựa trên bộ dữ liệu ở trên.

* Nghiên cứu khảo sát các mô hình gợi ý hiện có
* Lựa chọn các đặc trưng thích hợp của các bộ dữ liệu để đưa vào mô hình triển khai
* Xây dựng thuật toán, mô hình gợi ý cho bài toán gợi ý phim

**Nội dung 4**: Nghiên cứu xây dựng hệ thống để lưu trữ dữ liệu người dùng, dữ liệu hệ thống và tích hợp mô hình đã xây dựng ở trên (phần Back-end).

* Nghiên cứu các framework , ngôn ngữ lập trình, thư viện giúp lưu trữ dữ liệu có thể tích hợp mô hình, thuật toán và cách tích hợp
* Xây dựng ứng dụng lưu trữ dữ liệu và tích hợp mô hình, thuật toán gợi ý vào hệ thống

**Nội dung 5**: Nghiên cứu xây dựng giao diện để người dùng tương tác với hệ thống và hiển thị kết quả cho người dùng (phần Front-end) và kết nối với phần xử lý dữ liệu (Back-end).

* Nghiên cứu các ngôn ngữ lập trình, framework, thư viện giúp tạo giao diện hiển thị cho người dùng
* Nghiên cứu các hệ thống, các cách kết nối phần giao diện người dùng với phần lưu trữ dữ liệu hệ thống
* Xây dựng hệ thống giao diện người dùng
* Xây dựng kết nối giữa hệ thống giao diện người dùng và hệ thống lưu trữ dữ liệu

**Nội dung 6**: Thử nghiệm đánh giá chất lượng hệ thống.

**Nội dung 7:** Viết báo cáo

Mục lục

[**Chương 1** 6](#_Toc118959201)

[**Khảo sát nghiên cứu tổng quan các hệ thống gợi ý phim cùng với các mô hình kỹ thuật đang được sử dụng ở Việt Nam và trên thế giới** 6](#_Toc118959202)

[***1.1.*** ***Nghiên cứu về sự phát triển của điện ảnh và xu hướng giải trí qua phim ảnh hiện nay của Việt Nam và thế giới*** 6](#_Toc118959203)

[***1.2.*** ***Tổng quan về các hệ thống gợi ý phim cùng với các mô hình kỹ thuật được đang được áp dụng tại Việt Nam và trên thế giới*** 9](#_Toc118959204)

[***1.2.1.*** ***Tổng quan về các hệ thống gợi ý phim cùng với các mô hình kỹ thuật đang được áp dụng trên thế giới.*** 9](#_Toc118959205)

[***1.2.2.*** ***Tổng quan về các hệ thống gợi ý phim cùng với các mô hình kỹ thuật đang được áp dụng ở Việt Nam*** 11](#_Toc118959206)

[**Chương 2** 13](#_Toc118959207)

[**Tìm kiếm, đánh giá và phân tích bộ dữ liệu để phục vụ cho hệ thống gợi ý phim** 13](#_Toc118959208)

[***2.1.*** ***Tìm kiếm bộ dữ liệu phù hợp cho hệ thống*** 13](#_Toc118959209)

[***2.2 Đánh giá phân tích bộ dữ liệu để để lựa chọn các đặc trưng phù hợp*** 15](#_Toc118959210)

[***2.3: Thu thập thêm thông tin cần thiết cho hệ thống*** 16](#_Toc118959211)

[**Chương 3** 20](#_Toc118959212)

[**Nghiên cứu xây dựng, thiết kế mô hình gợi ý phim dựa trên bộ dữ liệu ở trên.** 20](#_Toc118959213)

[*3.1 Khảo sát nghiên cứu tổng quan các mô hình gợi ý hiện có* 20](#_Toc118959214)

[***3.1.1 Content-based Recommendation – Hệ thống gợi ý dựa trên nội dung*** 21](#_Toc118959215)

[***3.1.2 Collaborative-filtering Recommendation – Hệ thống gợi ý dựa trên lọc cộng tác.*** 24](#_Toc118959216)

[***3.1.3 Hybrid Recommender Systems – Hệ thống gợi ý kết hợp*** 28](#_Toc118959217)

[***3.2*** ***Lựa chọn các đặc trưng thích hợp của các bộ dữ liệu để đưa vào mô hình triển khai*** 30](#_Toc118959218)

[***3.3*** ***Xây dựng mô hình thuật toán cho bài toán gợi ý phim*** 31](#_Toc118959219)

[***3.3.1 Xây dựng thuật toán cho bài toán trên phương pháp Content – Based Filtering ( gợi ý phim dựa trên việc lọc nội dung)*** 31](#_Toc118959220)

[***3.3.2 Xây dựng thuật toán cho bài toán trên phương pháp Collaborative – Filtering (gợi ý phim dựa trên việc lọc cộng tác)*** 37](#_Toc118959221)

[**Chương 4** 42](#_Toc118959222)

[**Nghiên cứu xây dựng hệ thống để lưu trữ dữ liệu người dùng, dữ liệu hệ thống và tích hợp mô hình đã xây dựng ở trên (phần Back-end).** 42](#_Toc118959223)

[*4.1 Nghiên cứu các framework , ngôn ngữ lập trình, thư viện giúp lưu trữ dữ liệu có thể tích hợp mô hình, thuật toán và cách tích hợp.* 42](#_Toc118959224)

[***4.1.1 Framework, ngôn ngữ lập trình, thư viện để cài đặt hệ thống***. 42](#_Toc118959225)

[***4.1.2 Nghiên cứu cách tích hợp mô hình thuật toán vào hệ thống.*** 45](#_Toc118959226)

[*4.2 Xây dựng mô hình lưu trữ dữ liệu và tích hợp mô hình thuật toán gơi ý phim (backend)* 45](#_Toc118959227)

[***4.2.1 Thiết kế mô hình lưu trữ dữ liệu cho hệ thống*** 45](#_Toc118959228)

[***4.2.2 Cấu hình router cho các đường dẫn của trang web*** 47](#_Toc118959229)

[***4.2.3 Các hàm chức năng trong hệ thống*** 48](#_Toc118959230)

[**Chương 5:** 52](#_Toc118959231)

[**Nghiên cứu xây dựng giao diện để người dùng tương tác với hệ thống và hiển thị kết quả cho người dùng (phần Front-end) và kết nối với phần xử lý dữ liệu (Back-end).** 52](#_Toc118959232)

[*5.1 Nghiên cứu các ngôn ngữ lập trình, framework, thư viện giúp tạo giao diện hiển thị cho người dùng* 52](#_Toc118959233)

[*5.2 Nghiên cứu các hệ thống, các cách kết nối phần giao diện người dùng với phần lưu trữ dữ liệu hệ thống* 54](#_Toc118959234)

[*5.3: Xây dựng hệ thống giao diện của hệ thống và kết hợp kết nối hệ thống giao diện với hệ thống xử lý dữ liệu* 55](#_Toc118959235)

[***Chương 6*** 62](#_Toc118959236)

[**Thử nghiệm đánh giá chất lượng hệ thống.** 62](#_Toc118959237)

[6.1 Đánh giá sự chính xác của hệ thống 62](#_Toc118959238)

[***6.1.1 Đánh giá độ mất mát của hệ thống*** 62](#_Toc118959239)

[**6.1.2 Đánh giá độ chính xác của hệ thống collaborative filtering** 63](#_Toc118959240)

[6.2 Đánh giá sự ổn định của hệ thống 63](#_Toc118959241)

[**Kết luận và Định hướng nghiên cứu trong tương lai** 65](#_Toc118959242)

# **Chương 1**

# **Khảo sát nghiên cứu tổng quan các hệ thống gợi ý phim cùng với các mô hình kỹ thuật đang được sử dụng ở Việt Nam và trên thế giới**

* 1. ***Nghiên cứu về sự phát triển của điện ảnh và xu hướng giải trí qua phim ảnh hiện nay của Việt Nam và thế giới***

Từ khi ra đời đến nay, dịch vụ giải trí luôn là một phần không thể thiếu  
trong cuộc sống, đóng vai trò đắc lực trong việc thỏa mãn các nhu cầu về tinh thần của con người. Khi đời sống ngày càng được nâng cao, nhu cầu giải trí cũng ngày một đa dạng phong phú, đòi hỏi các ngành dịch vụ giải trí phải phát triển nhanh chóng để kịp thời đáp ứng các nhu cầu đó. Tuy là một ngành giải trí nhưng giá trị của phim ảnh truyền tải đến con người luôn mang những giá trị nhân văn to lớn. Ở một số nước, dịch vụ giải trí đã phát triển ở mức độ cao, trở thành một trong những ngành trọng điểm của nền kinh tế, đóng góp một phần quan trọng trong tổng sản phẩm công nghiệp của toàn nền kinh tế.

Trong số các loại hình dịch vụ giải trí, không thể không kể đến ngành công  
nghiệp điện ảnh, ngành được coi là loại hình nghệ thuật ra đời muộn nhất  
nhưng đã có những bước phát triển nhanh chóng để trở thành một trong những ngành lớn mạnh và mang lại lợi nhuận lớn nhất hiện nay. Lịch sử điện ảnh tuy chưa được lâu xong đã chứng kiến sự ra đời của những tác phẩm kinh điển, có sức ảnh hướng lớn đến thị hiếu thưởng thức nghệt huật cũng như phương thức tiếp cận các dịch vụ giải trí cuat người tiêu dung. Các bộ phim bom tấn đem về nguồn lợi nhuận khổng lồ cũng như các kinh đô điện ảnh nổi tiếng trên thế giới giờ đây đã trở thành một phần tất yếu trong đời sống nghệ thuật. Sự phát triển của loại hình dịch vụ giải trí này, tuy nhiên lại không thống nhất và đồng đều ở mỗi nước khác nhau.

Trên thế giới khu vực phải nói đến đầu tiên khi nói về sự phát triển của nền giải trí điện ảnh là hollywood. **Hollywood** là một khu của thành phố [Los Angeles](https://vi.wikipedia.org/wiki/Los_Angeles), [California](https://vi.wikipedia.org/wiki/California), [Hoa Kỳ](https://vi.wikipedia.org/wiki/Hoa_K%E1%BB%B3), nằm về phía tây bắc của thành phố này. Được biết đến như một trung tâm lịch sử [điện ảnh](https://vi.wikipedia.org/wiki/%C4%90i%E1%BB%87n_%E1%BA%A3nh), Hollywood đại diện cho ngành giải trí và [điện ảnh của Hoa Kỳ](https://vi.wikipedia.org/wiki/%C4%90i%E1%BB%87n_%E1%BA%A3nh_Hoa_K%E1%BB%B3). Ngày nay, những ngành phục vụ cho [công nghiệp điện ảnh](https://vi.wikipedia.org/wiki/C%C3%B4ng_nghi%E1%BB%87p_%C4%91i%E1%BB%87n_%E1%BA%A3nh) của nước này cũng được mở rộng ra những vùng lân cận như [Burbank](https://vi.wikipedia.org/wiki/Burbank,_California) và [Westside](https://vi.wikipedia.org/w/index.php?title=T%C3%A2y_Los_Angeles&action=edit&redlink=1), nhưng những ngành quan trọng như [biên tập](https://vi.wikipedia.org/w/index.php?title=Bi%C3%AAn_t%E1%BA%ADp&action=edit&redlink=1), [kỹ xảo](https://vi.wikipedia.org/w/index.php?title=K%E1%BB%B9_x%E1%BA%A3o_h%C3%ACnh_%E1%BA%A3nh&action=edit&redlink=1), hậu sản xuất và [ánh sáng](https://vi.wikipedia.org/w/index.php?title=K%E1%BB%B9_thu%E1%BA%ADt_%C3%A1nh_s%C3%A1ng_%C4%91i%E1%BB%87n_%E1%BA%A3nh&action=edit&redlink=1) trong phim ảnh vẫn được duy trì tại Hollywood. Vào thời cao điểm giữa [thập niên 1940](https://vi.wikipedia.org/wiki/Th%E1%BA%ADp_ni%C3%AAn_1940), các xưởng phim ở Hollywood mỗi năm cho ra đời tới khoảng 400 bộ phim với lượng khán giả mỗi tuần khoảng 90 triệu người.

Ở khu vực Châu Á, ngành điện ảnh có phần tụt hậu so với các nước phương Tây, tuy nhiên gần đây, cũng đã có những bước tiến đáng  
kể, bắt kịp với nền điện ảnh thế giới về cả phương diện nghệ thuật cũng như doanh thu.

Tại Việt Nam, tuy vậy, ngành điện ảnh lại chưa phát triển xứng tầm với tiềm năng hiện có. Trong xu thế chung của thế giới ngày càng coi trọng phát triển các ngành dịch vụ, Việt nam tất yếu phải tìm cách thúc đẩy ngành dịch vụ nói chung và ngành dịch vụ điện ảnh nói riêng. Trong những điều kiện thuận lợi cũng như những hạn chế của Việt Nam, cần phát triển ngành điện ảnh theo hướng đi nào để tận dụng được tối đa các nguồn lực của đất nước cũng như nâng cao trình độ cung cấp ngành dịch vụ giải trí ở Việt Nam.

Sự ra đời của truyền hình tạo ra sự cạnh tranh lớn với Điện ảnh. Để tồn tại, các hãng phim đa dạng các thể loại phim. Những phim hoạt hình đi vào lịch sử như Sleeping Beauty (1959), 101 Dalmations (1961) được ra đời thời kỳ này. Và thể loại phim kinh dị cũng được phát triển trong thời gian này.  
 Từ thập niên 1980 trở đi, sự phát triển của công nghệ VCR đã tạo ra một loạt các phim ―bom tấn với nhiều kỹ xảo hấp dẫn công chúng xem phim. Những ―bon tấn đầu tiên phải kể đến là Jaws, hay Star Wars. Năm 2009 được coi là năm làm ăn phát đạt của ngành công nghiệp điện ảnh thế giới với hàng loạt những bộ phim bom tấn và mang lại doanh thu kỷ lục. Công nghệ làm phim 3D đã tạo một bước nhảy đột phá về loại hình giải trí này. Tuy nhiên giao đoạn này nền công nghiệp Điện ảnh gặp phải sự cạnh tranh lớn của một công nghệ mới, băng VCR.

Tuy nhiên với sự bùng phát của dịch bệnh covid trong những năm gần đây cùng với sự phát triển mạnh mẽ của internet toàn cầu đã và đang dịch chuyển xu hướng giải trí của người dùng dần sang hướng onine (from home) đây là 1 cơ hội lớn để các app và các website giải trí phát triển.

Theo Kantar Media Việt Nam, có tới 84% người Việt có độ tuổi từ 15 – 54 sử dụng mạng xã hội mỗi ngày, và trong khảo sát cùng lĩnh vực của Nielsen, có 90% người được hỏi trả lời khảo sát họ xem truyền hình trực tiếp hàng tuần. Bên cạnh đó, theo tổ chức nghiên cứu Muvi, doanh thu thị trường OTT\* Đông Nam Á từ năm sau có thể đạt đến mức 650 triệu đô mỗi năm. Đây là một cơ hội lớn cho các công ty ở Việt Nam và nước ngoài đầu tư và phát triển ở hiện tại và trong tương lai tại thị trường Việt Nam.

Những lợi thế của việc phát triển các website, ứng dụng xem phim trực tuyến so với việc xem phim tại rạp và các kênh truyền hình truyền thống có thể kể đến như là ít tốn kém hơn so với việc xem phim tại rạp, có thể xem bất cứ lúc nào và ở đâu chỉ cần có thiết bị kết nối mạng và số lượng những bộ phim cũng đa dạng lựa chọn hơn so với các kênh truyền hình truyền thống.

đồng thời cũng là 1 sự cạnh tranh giữa các ứng dụng đòi hỏi phải đưa ra được trải nghiệm tốt nhất cho người dùng.

Bên cạnh về nội dung phim cuốn hút, hấp dẫn thì với kho phim khổng lồ như hiện nay thì chúng ta cần phải có 1 hệ thống đưa ra được những bộ phim có thể phù hợp với người dùng thay vì đưa ra vô vàn bộ phim nhưng lại không đúng với sở thích của người dùng làm cho trải nghiệm của người dùng trở nên kém đi.

Khi chưa có các hệ thống ứng dụng xem phim thì khi người dùng muốn xem bộ phim nào mới thì người xem phim cần phải đi tìm kiếm thông tin về từng bộ phim thông qua các forum, hội nhóm và những người đánh giá phim và với cách này thì có rất nhiều vấn đề như là: Tốc độ phản hồi chậm: khi một người cần tìm bộ phim về chủ đề mà mình muốn, họ sẽ đặt câu hỏi và chờ cộng đồng hoặc reviewer nào đó trả lời qua văn bản hoặc video. Hay khi bất kì người nào đó tìm kiếm về những bộ phim, họ sẽ lang thang trên internet. Việc chờ đợi, tìm kiếm này tốn không ít thời gian. Lượng thông tin mang lại ít: Vì đây là cách làm thủ công, con người không ai có thể xem hết những bộ phim, và đưa ra gợi ý cho người khác khi được hỏi được. Vì vậy có thể bỏ sót những bộ phim hay cho người khác.

Ngày nay, với sự phát triển của công nghệ, lượng dữ liệu về phim được thu thập trong nhiều năm, dần dần có nhiều nghiên cứu tập trung vào việc xây dựng các hệ thống gợi ý phim để tăng tối ưu trải nghiệm và thu hút người dùng.

Không chỉ ở thế giới mà cả ở Việt Nam và tất cả các nước khác, hầu hết các ứng dụng website trực tuyến đều đang sử dụng các hệ thống gợi ý sản phẩm đến người dùng tăng tối đa trải nghiệm người dùng nhằm tằng doanh thu cho công ty đồng thời cạnh tranh với các công ty khác.

* 1. ***Tổng quan về các hệ thống gợi ý phim cùng với các mô hình kỹ thuật được đang được áp dụng tại Việt Nam và trên thế giới***

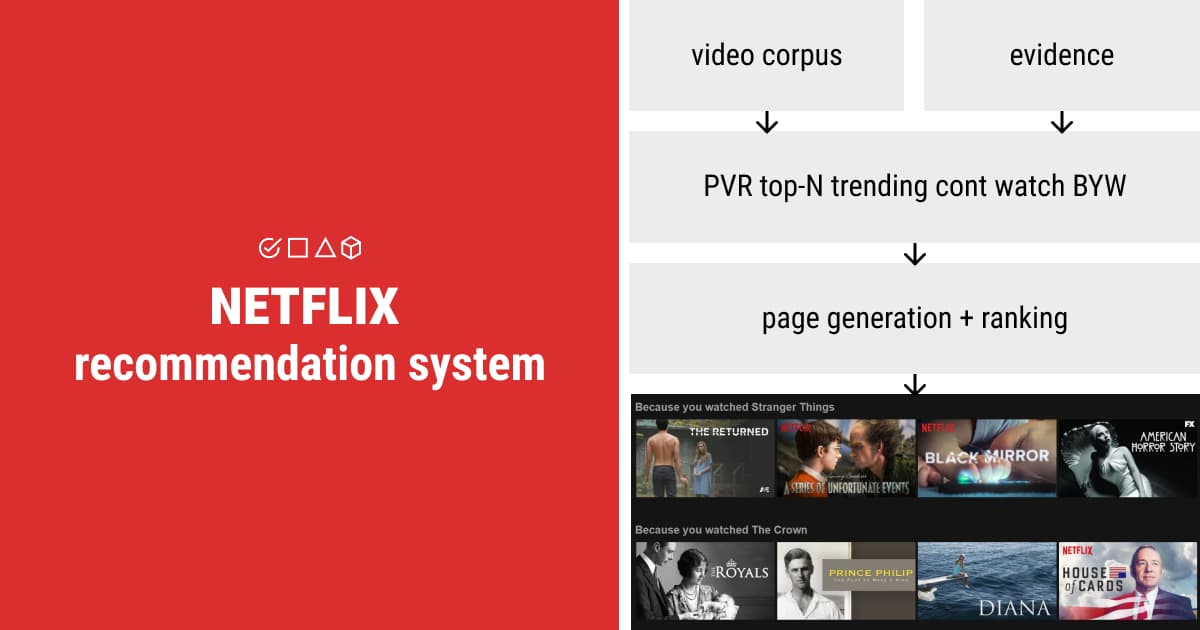
Hiện nay trên thế giới cũng như Việt Nam hầu như ứng dụng, hệ thống nào cũng đang sử dụng các hệ thống gợi ý cho người dùng không chỉ trong lĩnh vực phim ảnh mà còn rất nhiều lĩnh vực khác gần như là hầu hết các trang trực tuyến đều có.

* + 1. ***Tổng quan về các hệ thống gợi ý phim cùng với các mô hình kỹ thuật đang được áp dụng trên thế giới.***

Ngày nay tốc độ phát triển của mạng internet là cực kỳ mạnh mẽ cùng với nhu cầu giải trí của con người ngày càng cao nên các dịch vụ giải trí ngày càng được chú trọng và xem phim trực tuyến cũng là 1 lĩnh vực đang rất phát triển trên thế giới. Biết được xu hướng này các công ty đã và đang đầu tư rất mạnh mẽ vào lĩnh vực này.

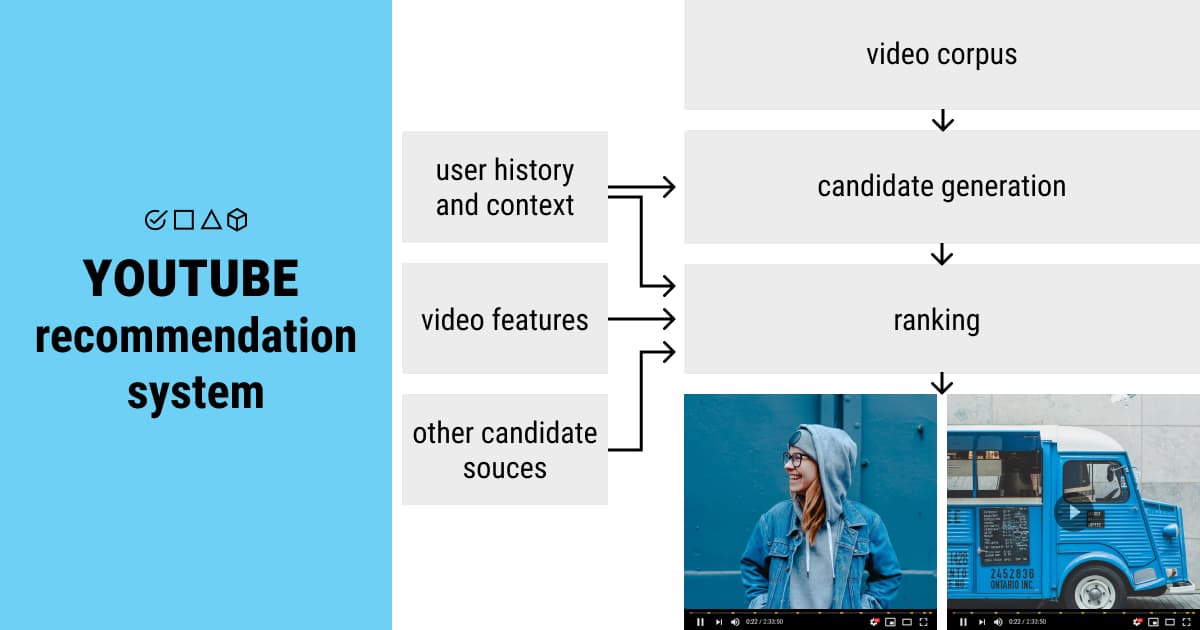
Hiện nay trên thế giới đang có rất nhiều ứng dụng website đang rất thành công trên lĩnh vực xem phim trực tuyến này có thể kể đến các ứng dụng lớn như là Netflix, Youtube,...

Netflix: Nền tảng xem film trực tuyến được thành lập ngày 29/08/1997, ở Scotts Valley, California, United States. Công ty bắt đầu dịch vụ đăng ký trả phí từ năm 1999. Tính đến năm 2009, Netflix sở hữu tập hợp 100.000 tựa đề DVD và hơn 10 triệu lượt người đăng ký. Vào ngày 25 tháng 2 năm 2007, Netflix phân phối chiếc DVD thứ 1 tỷ. Tính đến giữa tháng 4 năm 2021, Netflix báo cáo có hơn 208 triệu lượt người sử dụng trên toàn cầu, bao gồm 73 triệu người dùng tại Hoa Kỳ và [Canada](https://vi.wikipedia.org/wiki/Canada)  Công ty lần đầu đưa ra dịch vụ [truyền dữ liệu](https://vi.wikipedia.org/wiki/Truy%E1%BB%81n_d%E1%BB%AF_li%E1%BB%87u) tại thị trường quốc tế vào năm 2010 và tiếp tục mở rộng sang 130 lãnh thổ mới vào tháng 1 năm 2016, trong đó có [Việt Nam](https://vi.wikipedia.org/wiki/Vi%E1%BB%87t_Nam). Cho năm tài chính 2018, Netflix đã báo cáo lợi nhuận của Mỹ $ 1,21 tỷ, với doanh thu hàng năm của Mỹ $ 15,8 tỷ, tăng khoảng 116% so với chu kỳ tài chính trước. Cổ phiếu của Netflix giao dịch ở mức hơn 400 $ cho mỗi cổ phiếu tại mức giá cao nhất của nó vào năm 2018, và vốn hóa thị trường của nó đạt đến một giá trị trên Mỹ $ 180 tỷ trong tháng 6 năm 2018. Netflix xếp hạng 261 trên 2018 Fortune 500 danh sách các công ty lớn nhất Hoa Kỳ theo doanh thu.  Netflix được công bố là cổ phiếu tốt nhất số một trong những năm 2010, với tổng lợi nhuận là 3.693%. Bên cạnh các nội dung hấp dẫn mà NETFLIX sở hữu thì với số lượng phim khổng lồ (hàng ngàn bộ phim) thì NETFLIX cần có 1 hệ thống gợi ý các bộ phim phù hợp cho người dùng và phương pháp mà NETFLIX đang sử dụng đó là phương pháp Collaborative Filtering được ứng dụng trên nền thư viện Tensorflow.



***Hình 1. Workflow recommend system của NETFLIX***

YouTube là một [nền tảng chia sẻ video trực tuyến](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_v%E1%BB%A5_l%C6%B0u_tr%E1%BB%AF_video) của [Mỹ](https://vi.wikipedia.org/wiki/M%E1%BB%B9) có trụ sở chính tại [San Bruno, California](https://vi.wikipedia.org/wiki/San_Bruno,_California). Nền tảng này được tạo ra vào tháng 2 năm 2005 bởi ba nhân viên cũ của [PayPal](https://vi.wikipedia.org/wiki/PayPal) - [Chad Hurley](https://vi.wikipedia.org/wiki/Chad_Hurley), [Steve Chen](https://vi.wikipedia.org/wiki/Steve_Chen) và [Jawed Karim](https://vi.wikipedia.org/wiki/Jawed_Karim) - đã được [Google](https://vi.wikipedia.org/wiki/Google) mua lại vào tháng 11 năm 2006 với giá 1,65 tỷ [đô la](https://vi.wikipedia.org/wiki/%C4%90%C3%B4_la_M%E1%BB%B9) [Mỹ](https://vi.wikipedia.org/wiki/Hoa_K%E1%BB%B3) và hiện hoạt động như một trong những [công ty con](https://vi.wikipedia.org/wiki/C%C3%B4ng_ty_con) của Google. YouTube là trang web được truy cập nhiều thứ hai sau [Google Tìm kiếm](https://vi.wikipedia.org/wiki/Google_T%C3%ACm_ki%E1%BA%BFm), theo xếp hạng của [Alexa Internet](https://vi.wikipedia.org/wiki/Alexa_Internet). Trang web cho phép người dùng [tải lên](https://vi.wikipedia.org/wiki/T%E1%BA%A3i_l%C3%AAn), xem, chia sẻ, thêm vào danh sách phát, [báo cáo](https://vi.wikipedia.org/wiki/B%C3%A1o_c%C3%A1o) và nhận xét về [video](https://vi.wikipedia.org/wiki/Video), [đăng ký](https://vi.wikipedia.org/wiki/%C4%90%C4%83ng_k%C3%BD) người dùng khác và sử dụng công nghệ [WebM](https://vi.wikipedia.org/wiki/WebM), [H.264](https://vi.wikipedia.org/wiki/H.264)/[MPEG-4 AVC](https://vi.wikipedia.org/wiki/MPEG-4_AVC) và [Adobe Flash Player](https://vi.wikipedia.org/wiki/Adobe_Flash_Player) để hiển thị nhiều video [đa phương tiện](https://vi.wikipedia.org/wiki/%C4%90a_ph%C6%B0%C6%A1ng_ti%E1%BB%87n) do người dùng và [doanh nghiệp](https://vi.wikipedia.org/wiki/Doanh_nghi%E1%BB%87p) tạo ra. Nội dung có sẵn bao gồm video clip, đoạn chương trình truyền hình, [video âm nhạc](https://vi.wikipedia.org/wiki/Video_%C3%A2m_nh%E1%BA%A1c), [phim tài liệu](https://vi.wikipedia.org/wiki/Phim_t%C3%A0i_li%E1%BB%87u) ngắn và [tài liệu](https://vi.wikipedia.org/wiki/T%C3%A0i_li%E1%BB%87u), bản ghi âm, [đoạn giới thiệu phim](https://vi.wikipedia.org/wiki/Trailer) và các nội dung khác như viết [blog](https://vi.wikipedia.org/wiki/Blog) bằng video, video sáng tạo ngắn và video [giáo dục](https://vi.wikipedia.org/wiki/Gi%C3%A1o_d%E1%BB%A5c). Doanh thu tính đến năm 2020 là 19,8 tỷ đô la Mỹ (2020). Tuy đây không hẳn là một nền tảng xem phim trực tuyến mà nó còn là một nên tảng chia sẻ, đăng lại các video ngắn nhưng nó cũng như các nền tảng xem phim trực tuyến khác. Nó có một hệ thống gợi ý các bộ phim, video đến người dùng dùng dựa trên mạng học sâu DNN rất hiệu quả.



***Hình 2: Workflow recommend system của Youtube***

* + 1. ***Tổng quan về các hệ thống gợi ý phim cùng với các mô hình kỹ thuật đang được áp dụng ở Việt Nam***

Trong nước hiện có khá nhiều công ty. tổ chức thực hiện nghiên cứu và ứng dụng bài toán gợi ý phim dựa trên nội dung liên quan và đánh giá của người dùng trong việc phát triển lĩnh vực giải trí này do những năm gần đây nhu cầu giải trí của người dân Việt Nam cũng đang chuyển dịch rất nhiều theo hướng online và đây cũng là một hướng phát triển rất tiềm năng trong tương lai mà các công ty có thể hướng đến.

Các ứng dụng hệ thống có thể kể đến trong lĩnh vực này như là ứng dụng VIEON, FPT Play, ...

**VieON** là một [ứng dụng OTT](https://vi.wikipedia.org/wiki/%E1%BB%A8ng_d%E1%BB%A5ng_OTT) [video theo yêu cầu](https://vi.wikipedia.org/wiki/Video_on_demand) được cung cấp bởi Dzones Network trực thuộc công ty Đất Việt VAC tại Việt Nam. Ứng dụng được phát hành vào năm 2020, với hơn 140 kênh truyền hình trong nước và quốc tế, ngoài ra còn có một số bộ [phim truyền hình](https://vi.wikipedia.org/wiki/Phim_truy%E1%BB%81n_h%C3%ACnh), [chương trình truyền hình](https://vi.wikipedia.org/wiki/Ch%C6%B0%C6%A1ng_tr%C3%ACnh_truy%E1%BB%81n_h%C3%ACnh) trong nước và quốc tế như [Hàn Quốc](https://vi.wikipedia.org/wiki/Phim_truy%E1%BB%81n_h%C3%ACnh_H%C3%A0n_Qu%E1%BB%91c), [Trung Quốc](https://vi.wikipedia.org/wiki/Phim_truy%E1%BB%87n_truy%E1%BB%81n_h%C3%ACnh_Trung_Qu%E1%BB%91c) hay [điện ảnh Hollywood](https://vi.wikipedia.org/wiki/%C4%90i%E1%BB%87n_%E1%BA%A3nh_Hoa_K%E1%BB%B3). Ứng dụng được lên kế hoạch phát triển vào năm 2016, sau 4 năm nghiên cứu, phát triển và thử nghiệm cùng đối tác tư vấn chiến lược từ Mỹ là BCG Digital Ventures. VieON chính thức được tập đoàn Đất Việt VAC phát hành vào ngày 15 tháng 6 năm 2020. Đất Việt VAC cũng không giấu giếm tham vọng sẽ biến ứng dụng này trở thành kênh giải trí số 1 dành cho người Việt.

Luôn luôn đổi mới, sáng tạo và dẫn đầu xu hướng công nghệ truyền thông giải trí, VieON tiếp tục đầu tư phát triển Hệ thống gợi ý (recommendation system) với các công nghệ AI, Machine Learning tiên tiến nhất nhằm gợi ý các nội dung phù hợp với từng người xem. 70% lượt xem đến từ Hệ thống gợi ý chính là con số ấn tượng minh chứng cho những nỗ lực của đội ngũ VieON. Ứng dụng từ các mô hình máy học, VieON đã phát triển hệ thống AI Indexing có khả năng phân tích nội dung video sử dụng công nghệ trí thông minh nhân tạo (AI).

**FPT Play** thuộc quyền sở hữu của công ty FPT (Công ty Cổ phần Viễn thông FPT) có trụ sở chính thuộc Tòa nhà FPT Tower, số 10 Phạm Văn Bạch, [Cầu Giấy](https://vi.wikipedia.org/wiki/C%E1%BA%A7u_Gi%E1%BA%A5y), [Hà Nội](https://vi.wikipedia.org/wiki/H%C3%A0_N%E1%BB%99i) .Ngày 8 tháng 9 năm 2021, [FPT Play](https://fptso1vn.com/fpt-play/) và [Truyền hình FPT](https://fptso1vn.com/combo-internet-truyen-hinh/) công bố hợp nhất thương hiệu, trở thành Dịch vụ truyền hình FPT với tên thương hiệu là FPT Play. Ứng dụng hỗ trợ lên đến 150 kênh truyền hình khác nhau bao gồm các kênh trong nước như VTV3, VTV6, VTVCab, VTC, HTV,…và các kênh nước ngoài đặc sắc như Cartoon Network, CNN, WB, HBO…. Hơn 57.000 giờ phim bom tấn chiếu rạp & hơn 1.000 phim bộ đầy đủ các quốc gia: Việt Nam, Hàn Quốc, Hoa Kỳ, Trung Quốc… đáp ứng đầy đủ các yêu cầu. Kho phim được cập nhất liên tục hàng giờ. Cũng như các nền tảng xem phim trực tuyến khác FPT Play cũng có thuật toán gợi ý riêng của họ để tăng trải nghiệm của người dùng. Thuật toán này hiện chưa tìm hiểu được

**Chương 2**

**Tìm kiếm, đánh giá và phân tích bộ dữ liệu để phục vụ cho hệ thống gợi ý phim**

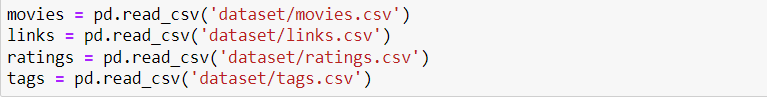
* 1. ***Tìm kiếm bộ dữ liệu phù hợp cho hệ thống***

Để có thể hoàn thiện được một mô hình, hệ thống gợi phim hiệu quả, phù hợp cho người dùng trước hết chúng ta cần phải có 1 bộ dữ liệu phù hợp cho hệ thống.

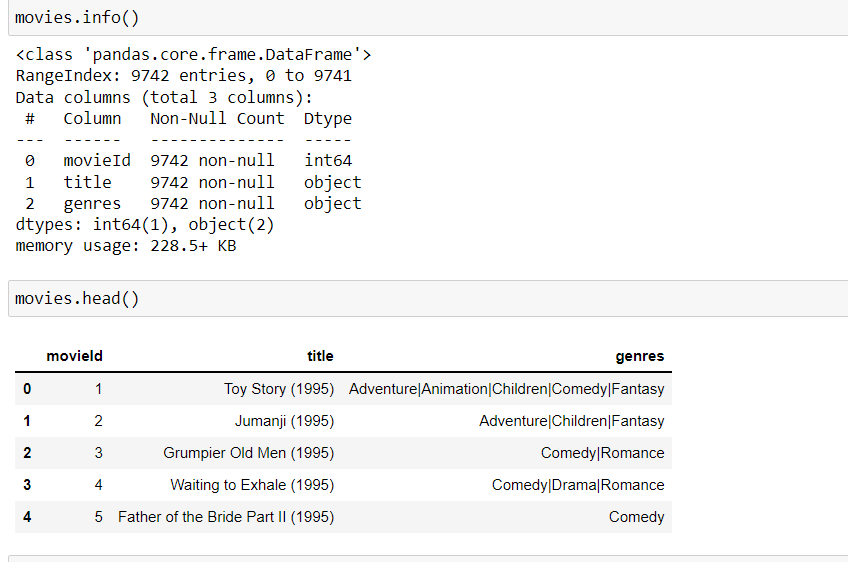
Ở đây chúng tôi đã tìm được một tập dataset là các file csv thô mô tả các dữ liệu về các bộ phim.

Đẻ phát triền hệ thống gợi ý phim này chùng ta lựa chọn sử dụng bộ dữ liệu về hơn 9000 bộ phim của movieLens cùng với lượng lớn đánh giá của các người dùng thử nghiệm khác mà họ đã thu thập.

Link dữ liệu drive: [https://drive.google.com/file/d/1WWQCl9w52M1sXNWd4JSKL7q-HHywk03p/view](https://drive.google.com/file/d/1WWQCl9w52M1sXNWd4JSKL7q-HHywk03p/view%20)



Bộ dữ liệu này bao gôm 4 bảng dữ liệu được lưu trữ trong 4 files .csv chứa thông tin về các bộ film cũng như là các lượt đánh giá của các người dùng khác.

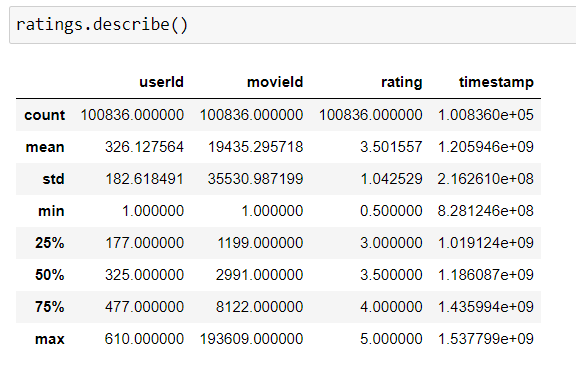


***Hình 3: File movies.csv chứa thông tin về các bộ phim***



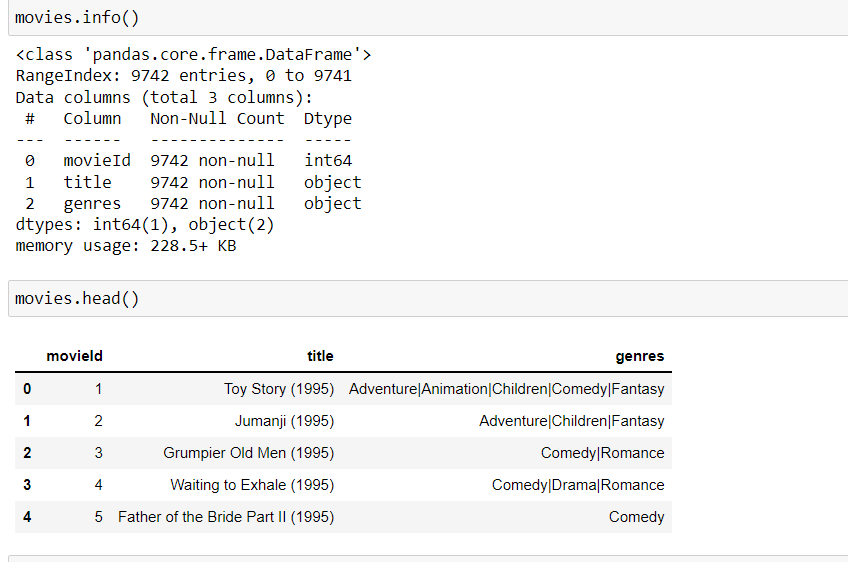
***Hình 4: File ratings.csv chứa thông tin về các lượt rating***

***2.2 Đánh giá phân tích bộ dữ liệu để để lựa chọn các đặc trưng phù hợp***



***Hình 5: Mô tả về bộ dữ liệu ratings***

Trong bộ dữ liệu này cho chúng ta thấy rằng có 610 user thử nghiệm và hơn 9000 bộ phim được đánh giá có hơn 100000 lượt đánh giá từ các người dùng đối với các bộ phim với đánh giá cao nhất là 5.0 và thấp nhất là 0.5 vì vậy có thể thấy rằng nếu bộ phim được đánh giá trên 2.5 thì có thể nói rằng người dùng có cảm tình với bộ phim này và ngược lại là không từ đó có thể đưa ra các gợi ý phù hợp với người dùng.



Đối với dữ liệu về các bộ phim thì chúng ta thấy có các trường như movieId, title, genres với 3 trường này thì chưa đủ để chúng ta có thế xây dựng được hệ thống một cách tốt nhất vì vậy chúng ta cần lấy thêm 1 số thông tin khác về các bộ phim như là overview của bộ phim.., ngoài ra chúng ta cần lấy thêm 1 poster của bộ phim để đưa ra được kết quả cùng với poster của bộ phim để người dùng có 1 cái nhìn tổng quan về bộ phim.

Hầu hết những bộ dữ liệu này đã được làm sạch khá kỹ nên chúng ta không cần làm sạch bộ dữ liệu này nữa.

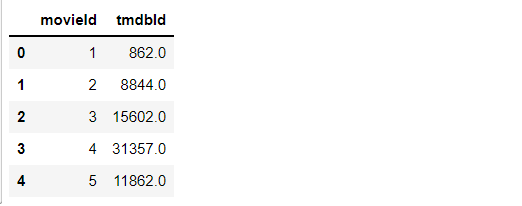
Các đặc trưng sẽ dùng khi triển khai mô hình sẽ là title, genres và overview của bảng movie và các đặc trưng userId, movieId và rating trong bảng rating.

Tóm lại ở bảng movies chúng ta cần thêm 1 trường dữ liệu overview và 1 trường poster của các bộ phim để trực quan hóa khi đưa ra kết quả.

***2.3: Thu thập thêm thông tin cần thiết cho hệ thống***

Như chúng ta đã phân tích ở trên chúng ta cần thêm 2 trường nữa cho bảng dữ liệu movie là overview để phân tích bộ phim cùng với trường poster để hiển thị bộ phim cho người dùng. Chúng ta sẽ tiến hành triển khai craw thêm thông tin cho các bộ phim.

Để tiến hành thu thập thêm thông tin của các bộ phim chúng ta cần sử dụng thêm dữ liệu của bảng links để lấy ID của các bộ phim trên trang the movie database



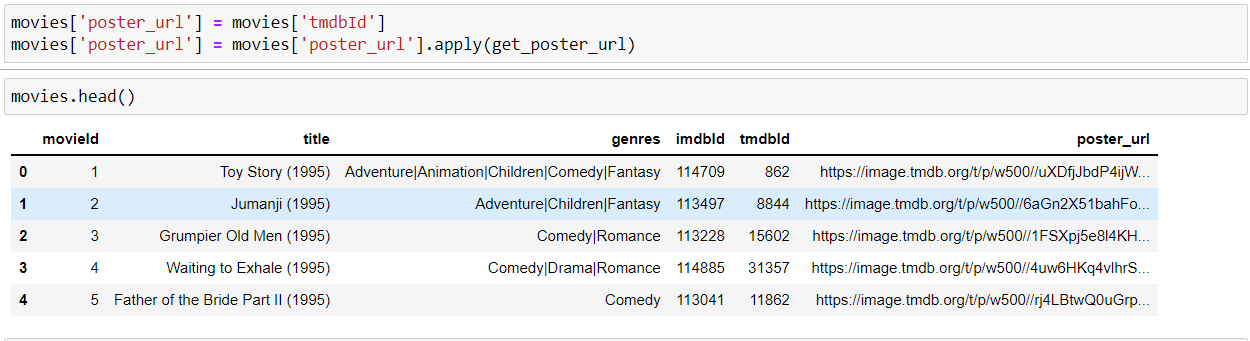
***Hình 6: dữ liệu bảng links***

Sử dụng bảng links lấy được tmdbId của từng bộ phim và sau đó viết hàm request đến trang the movie DB để lấy thêm thông tin cần thiết cho từng bộ phim.

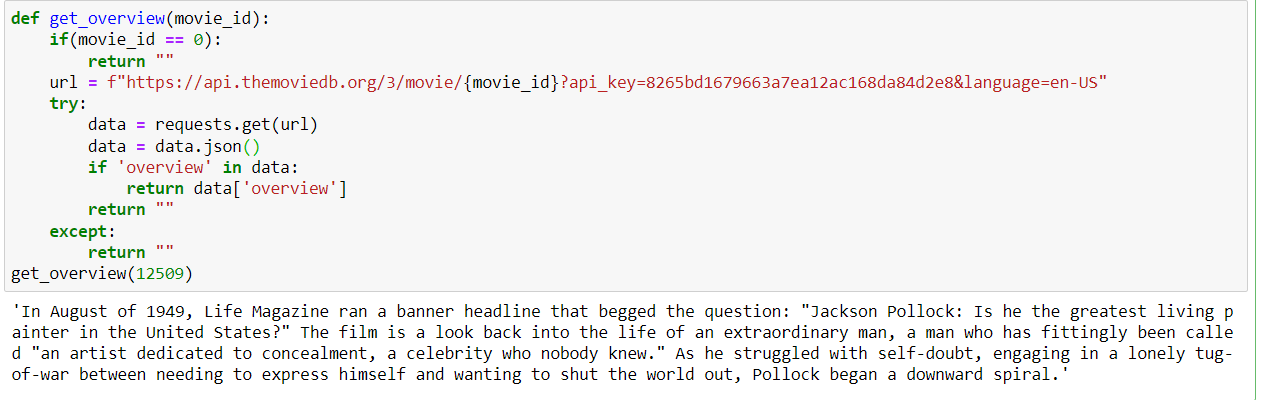
Dưới đây là 2 hàm mà chúng ta viết để lấy được overview và poster\_url của từng bộ phim.



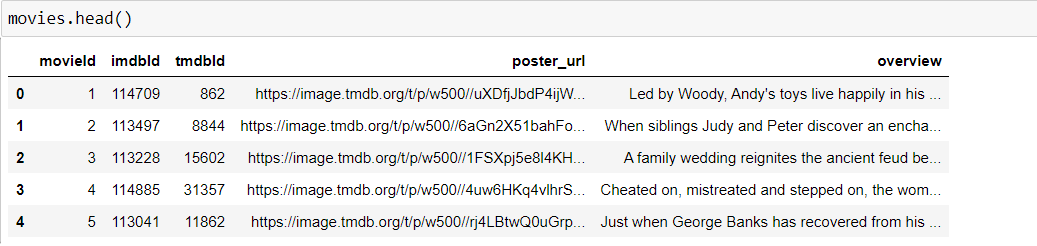
***Hình 7: Hàm lấy dữ liệu poster\_url của phim thông qua dữ liệu tmdbId của từng bộ phim***



Kết quả thu được sau khi gọi đến hàm get\_poster\_url(tmdbId): và gộp kết quả với bảng movies



***Hình 8: Hàm lấy overview của các bộ phim thông qua tmdbId cửa từng bộ phim***



Đây là kết quả cuối cùng sau khi thu thập thêm thông tin của các bộ phim và đây cũng sẽ là dữ liệu thô cuối cùng mà chúng ta sẽ sử dụng trong bài toán gợi ý phim mà chúng ta đang hướng tới.

**Chương 3**

**Nghiên cứu xây dựng, thiết kế mô hình gợi ý phim dựa trên bộ dữ liệu ở trên.**

## *3.1 Khảo sát nghiên cứu tổng quan các mô hình gợi ý hiện có*

Một **hệ thống gợi ý** (hay còn gọi là *hệ thống khuyên dùng*, [tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh): **recommendation system**) là một lớp con của [hệ thống lọc thông tin](https://vi.wikipedia.org/w/index.php?title=H%E1%BB%87_th%E1%BB%91ng_l%E1%BB%8Dc_th%C3%B4ng_tin&action=edit&redlink=1), tìm kiếm dự đoán "đánh giá" hoặc "ưa thích" của người dùng với một sản phẩm hoặc đối tượng nào đó. Hệ thống gợi ý chủ yếu dùng trong các ứng dụng thương mại.

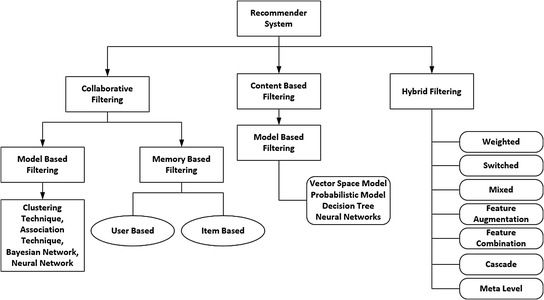
Các hệ thống gợi ý được tận dụng trong nhiều lĩnh vực, chẳng hạn như trình tạo danh sách phát cho video và âm nhạc của một số ứng dụng như Netflix, YouTube và Spotify. Sau khi người xem một vài video hoặc nghe một số bài hát, hệ thống gợi ý sẽ nhận diện sở thích người dùng và mức độ tương tác với video/bài hát để tạo ra một danh sách gợi ý theo chủ đề và mức độ liên quan mà người dùng có thể ưa thích. Hệ thống gợi ý còn áp dụng cho các dịch vụ gợi ý sản phẩm ở Amazon, các gợi ý nội dung cho các nền tảng xã hội như Facebook hay Twitter. Các hệ thống này có thế một phần giúp người dùng đưa ra quyết định chính xác, phù hợp nhất.

Hệ thống khuyến nghị là một chủ đề nghiên cứu rất nóng trong những năm gần đây. Nhiều nhà nghiên cứu đã nêu ra rất nhiều cách tiếp cận khuyến nghị khác nhau. Danh mục nổi tiếng nhất của những cách tiếp cận này là:

• Content-based Recommendation – Hệ thống gợi ý dựa trên nội dung.

• Collaborative-filtering Recommendation – Hệ thống gợi ý dựa trên lọc cộng tác.

• Hybrid Recommendation – Hệ thống gợi ý lai.



***Hình 8: Phân loại các loại hệ thống recommendation***

***3.1.1 Content-based Recommendation – Hệ thống gợi ý dựa trên nội dung***

Gợi ý dựa trên nội dung là một cách tiếp cận quan trọng trong hệ thống gợi ý. Ý tưởng cơ bản là giới thiệu các mặt hàng tương tự với những gì người dùng thích trước đây. Nhiệm vụ cốt lõi của hệ thống giới thiệu dựa trên nội dung là tính toán sự giống nhau giữa các mục. Có rất nhiều phương pháp để lập mô hình vật phẩm và

nổi tiếng nhất là Mô hình Không gian Vector . Mô hình trích xuất các từ khóa của mặt hàng và tính weight bằng TF-IDF. Ví dụ: đặt làm từ khóa thứ i của mục , là trọng số của cho , khi đó nội dung của , có thể được xác định là:



Như chúng ta đã nói trước đây, hệ thống giới thiệu dựa trên nội dung gợi ý các mặt hàng tương tự với những gì người dùng thích trước đây. Vì vậy, thị hiếu của người dùng có thể được mô phỏng theo lịch sử của những gì người dùng thích. Coi *ContentBaseProfile(u)* là vectơ tùy chọn của người dùng u, định nghĩa là:



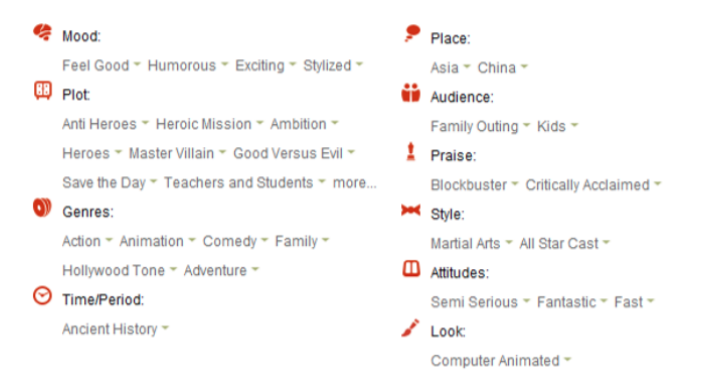
*N (u)* là những gì người dùng bạn thích trước đây. Sau khi tính toán vector *Content(.)* Và vectơ tùy chọn nội dung *ContentBaseProfile(.)*  Của tất cả người dùng, với bất kỳ người dùng nào là *u* và một mục *d*, cách người dùng thích mục đó được xác định là sự giống nhau giữa *ContentBaseProfile(u)* và *Content(d):*



Sử dụng từ khóa để lập mô hình mặt hàng là một bước quan trọng đối với nhiều hệ thống giới thiệu. Nhưng việc trích xuất từ khóa của một mặt hàng cũng là một vấn đề khó khăn, đặc biệt là trong lĩnh vực truyền thông, vì rất khó để trích xuất từ khóa văn bản từ video. Để giải quyết loại vấn đề này, có hai cách chính. Một là để các chuyên gia gắn thẻ các mặt hàng và một là cho phép người dùng gắn thẻ chúng.

Đại diện của các hệ thống được gắn thẻ tags chuyên gia là Pandora cho âm nhạc và Jinni cho phim. Hãy lấy Jinni làm ví dụ, các nhà nghiên cứu của Jinni đã xác định hơn 900 tags là thể loại phim và họ để các chuyên gia điện ảnh tạo tags cho họ. Các thẻ này thuộc các danh mục khác nhau, bao gồm thể loại phim, cốt truyện, thời gian, địa điểm và dàn diễn viên.

Hình 9 của Jinni, là các tags của bộ phim Kung Fu Panda. Như chúng ta có thể thấy từ hình, các thẻ của Kung Fu Panda được chia thành mười loại hoàn toàn, Mood, Plot, Genres, Time, Place, Audience, Praise, Style, Attitudes and Look. Các thẻ này chứa tất cả các khía cạnh của thông tin phim, có thể mô tả một bộ phim rất chính xác.

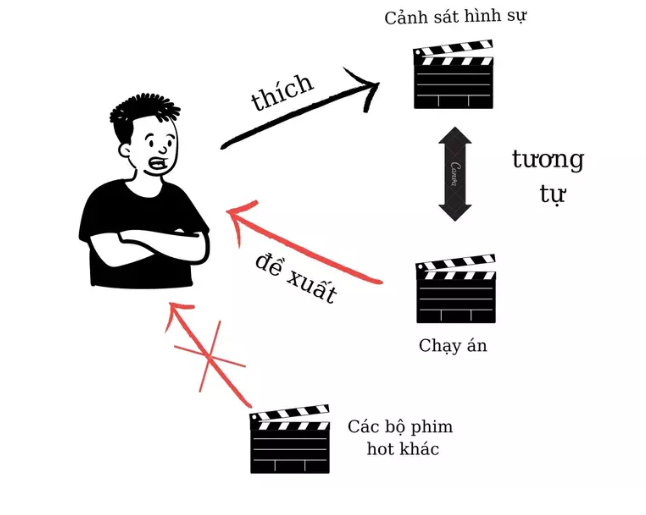


***Hình 9: Thẻ cho bộ phim kung fu panda***

So với hệ thống được gắn thẻ chuyên gia, hệ thống được gắn thẻ người dùng được áp dụng rộng rãi hơn.

Các trang web đại diện là Delicious và Flickr. Đặc điểm của hệ thống do người dùng gắn thẻ là các thẻ đa dạng hơn so với hệ thống được gắn thẻ chuyên gia. Nhưng điểm yếu là các thẻ có chất lượng kém hơn, thậm chí có rất nhiều thẻ sai. Vì vậy, trong hệ thống gắn thẻ người dùng, có hai vấn đề chính, một là gợi ý thẻ, nghĩa là khi người dùng gắn thẻ một mục, hệ thống có thể gợi ý một số thẻ tương đối để anh ta lựa chọn. Mục đích đầu tiên là tạo sự thuận tiện cho người dùng và thứ hai là nó có thể tăng chất lượng của thẻ. Một câu hỏi khác là làm thế nào để giới thiệu các mục dựa trên thẻ (khuyến nghị dựa trên thẻ. Sau khi các mặt hàng được gắn thẻ, cách tiếp cận gợi ý đơn giản nhất là sử dụng các thẻ làm từ khóa của mặt hàng và gợi ý bằng thuật toán dựa trên nội dung.

Sau khi tính toán được sự tương đồng của các sản phẩm mình đã thích trước đây với danh sách các sản phẩm khác có trong hệ thống chúng ta sẽ sắp xếp chúng theo danh sách tăng dần của sự tương đồng và lấy ra các sản phẩm có độ tương đồng lớn nhất để gợi ý cho người dùng.



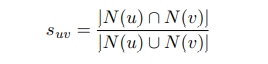
***Hình 10: Mô hình gợi ý dựa trên nội dung của các sản phẩm – Content base Recommendation***

***3.1.2 Collaborative-filtering Recommendation – Hệ thống gợi ý dựa trên lọc cộng tác.***

Collaborative-filtering Recommendation – Gợi ý dựa trên lọc cộng tác là thuật toán nổi tiếng nhất trong các hệ thống gợi ý. Thuật toán này mô hình hóa thị hiếu của người dùng theo lịch sử hành vi của người dùng. GroupLens đã xuất bản bài báo đầu tiên về tính năng lọc cộng tác và bài báo đã nêu ra tính năng lọc cộng tác dựa trên người dùng. Vào năm 2000, Amazon đã đưa ra phương pháp lọc cộng tác dựa trên mục trong bài báo của họ. Hai thuật toán này rất nổi tiếng trong các hệ thống giới thiệu doanh nghiệp.

***3.1.2.1 User-based collaborative-filtering – Hệ thống lọc cộng tác dựa trên người dùng.***

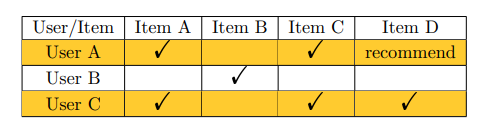
Trong lọc cộng tác dựa trên người dùng, người dùng sẽ thích các mặt hàng được người dùng có cùng sở thích với họ thích. Vì vậy, bước đầu tiên của lọc cộng tác dựa trên người dùng là tìm những người dùng có cùng sở thích. Trong lọc cộng tác, những người dùng được coi là giống nhau khi họ thích các mục tương tự. Nói một cách đơn giản, người dùng cho trước *u* và *v*, *N (u)* và *N (v)* là các mục được thiết lập bởi u và v tương ứng. Vì vậy, sự giống nhau của u và v có thể được định nghĩa đơn giản là:



***[Phương trình 2.4]***

Có rất nhiều thuật toán tương tự, Phương trình 2.4 là một trong số đó. Khả năng yêu thích của người dùng *u* đối với mặt hàng tôi có thể được tính bằng:





***Hình 11: Ví dụ User-Base Collaborative filtering***

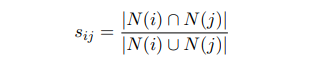
Hình 11 là một ví dụ về gợi ý dựa trên sự cộng tác người dùng User-Base Collaborative Filtering. Theo lịch sử quan tâm của Người dùng A, chỉ Người dùng C mới có thể là hàng xóm của anh ta, vì vậy Mục D sẽ được đề xuất cho Người dùng A.

***3.1.2.2 Item-based collaborative-filtering – Hệ thống lọc cộng tác dựa trên sản phẩm Items.***

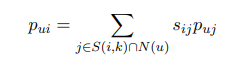
Tính năng lọc cộng tác dựa trên sản phẩm items thì khác, nó giả định rằng người dùng sẽ thích các mặt hàng tương tự với các mặt hàng mà người dùng đã thích trước đó.

Vì vậy, bước đầu tiên của lọc cộng tác dựa trên sản phẩm – Item-based collaborative filtering là tìm ra các mục tương tự với những gì người dùng đã thích trước đó. Điểm cốt lõi của lọc cộng tác dựa trên sản phẩm – Item-based collaborative filtering là tính toán mức độ giống nhau của hai mục. Item-based collaborative cho rằng các mặt hàng được nhiều người dùng giống nhau thích thì chúng càng giống nhau.

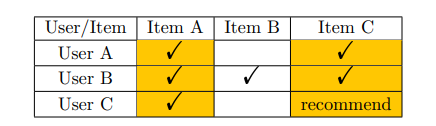
Giả sử N (i) và N (j) là tập người dùng thích i và j tương ứng. Vì vậy, sự giống nhau của i và j có thể được định nghĩa là:



Khả năng yêu thích của người dùng u của mặt hàng tôi có thể được tính bằng:



Bảng 2.2 là một ví dụ Item-Based collaborative filtering hệ thống gợi ý lọc cộng tác dựa trên sản phẩm (item). Theo lịch sử quan tâm của tất cả người dùng cho Mặt hàng A, những người thích Mặt hàng A cũng thích Mặt hàng C, vì vậy chúng tôi có thể kết luận rằng Mặt hàng A tương tự với Mặt hàng C. Trong khi Người dùng C thích Mặt hàng A, vì vậy chúng ta có thể suy ra rằng có lẽ Người dùng C cũng thích Mục C

******

***Hình 12: Ví dụ Items-Base Collaborative Filtering***

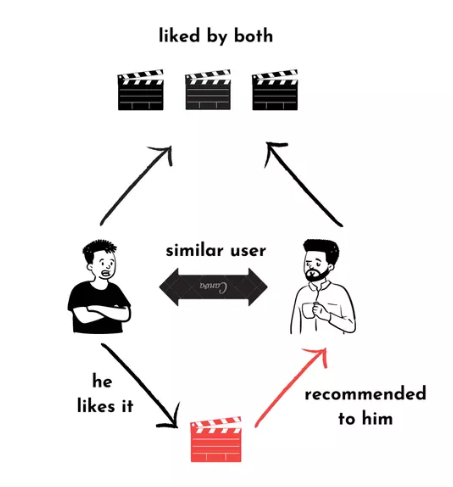
Thuật toán lọc cộng tác dựa trên người dùng và dựa trên sản phẩm (User-based and Item-based collaborative-filtering algorithms) đều là thuật toán dựa trên vùng lân cận, ngoài ra còn có rất nhiều thuật toán lọc cộng tác khác. Hoffman đã nêu ra Mô hình lớp tiềm ẩn (Latent Class Model) trong bài báo này, mô hình kết nối người dùng và mặt hàng theo lớp tiềm ẩn, cho rằng người dùng sẽ không quan tâm đến các mặt hàng một cách trực tiếp. Thay vào đó, người dùng quan tâm đến một số danh mục có chứa các sản phầm, do đó, mô hình sẽ học cách tạo các danh mục theo hành vi của người dùng.

Trên mô hình lớp tiềm ẩn (Latent Class Model) , các nhà nghiên cứu đã đưa ra Mô hình phân rã ma trận (Matrix Decomposition Model), được gọi là Mô hình yếu tố tiềm ẩn (Latent Factor Model). Có rất nhiều mô hình dựa trên sự phân rã ma trận và chúng chủ yếu đến từ Cuộc thi Giải thưởng Netflix, chẳng hạn như RSVD , SVD ++ , v.v...

Bên cạnh Mô hình phân rã ma trận (Matrix Decomposition Model), Mô hình đồ thị (Graph Model) được ứng dụng rộng rãi trong lọc cộng tác. Baluja đã giới thiệu mô hình biểu đồ đồng xem đằng sau thuật toán gợi ý của YouTube trong và cũng đưa ra một thuật toán quảng bá trên biểu đồ để đo lường mức độ thích một mặt hàng của người dùng.

Tài liệu này nghiên cứu cách tăng khả năng tìm kiếm của kết quả đề xuất bằng cách phân tích mối liên kết giữa các node trong biểu đồ. Mirza đã nghiên cứu một cách có hệ thống các vấn đề về đề xuất dựa trên mô hình biểu đồ và chỉ ra bản chất của đề xuất là kết nối người dùng và sản phẩm (items).

Biểu đồ là phương pháp tự nhiên cho điều đó. nghiên cứu các thuật toán tương tự giữa các nút của biểu đồ và so sánh độ chính xác khuyến nghị của các thuật toán khác nhau.



***Hình 13: Mô hình gợi ý lọc cộng tác - Collaborative-filtering Recommendation***

### ***3.1.3 Hybrid Recommender Systems – Hệ thống gợi ý kết hợp***

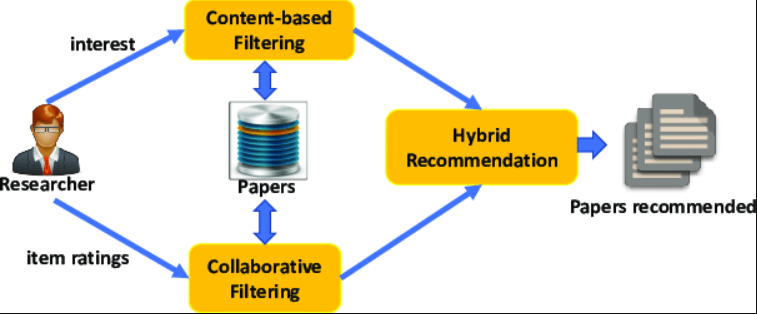
***Hybrid Recommender Systems –*** Hệ thống gợi ý kết hợp hiện đang ngày càng phổ biến hơn. Kết hợp lọc cộng tác (Collabrative Filtering) và lọc dựa trên nội dung (Content-Base Filtering) có thể hiệu quả hơn trong nghiên cứu gần đây.

Có nhiều cách để thực hiện các hệ thống ***Hybrid Recommender Systems – Hệ thống gợi ý kết hợp***: chỉ cần tổng hợp kết quả của các gợi ý Colarative filtering và Conten-Base Filtering, thêm khả năng Collabrative Filtering vào phương pháp Content-Base Filtering.

Có bảy phương pháp lai:

* Weighted: Thêm điểm số từ các thành phần gợi ý khác nhau (recommender components.)
* Switching: Chuyển đổi các thành phần gợi ý khác nhau
* Mixed: Hiển thị kết quả gợi ý các hệ thống khác nhau.
* Features Combination: Trích xuất các đặc trưng từ các nguồn khác nhau và kết hợp chúng như một đầu vào duy nhất.
* Feature Augmentation: Tính toán các đặc trưng theo một recommender và đưa kết quả sang bước tiếp theo.
* Cascade: Tạo kết quả thô bằng kỹ thuật gợi ý và gợi ý trên đầu kết quả trước đó.
* Meta-level: Sử dụng mô hình do một recommender tạo ra làm đầu vào của kỹ thuật gợi ý khác.

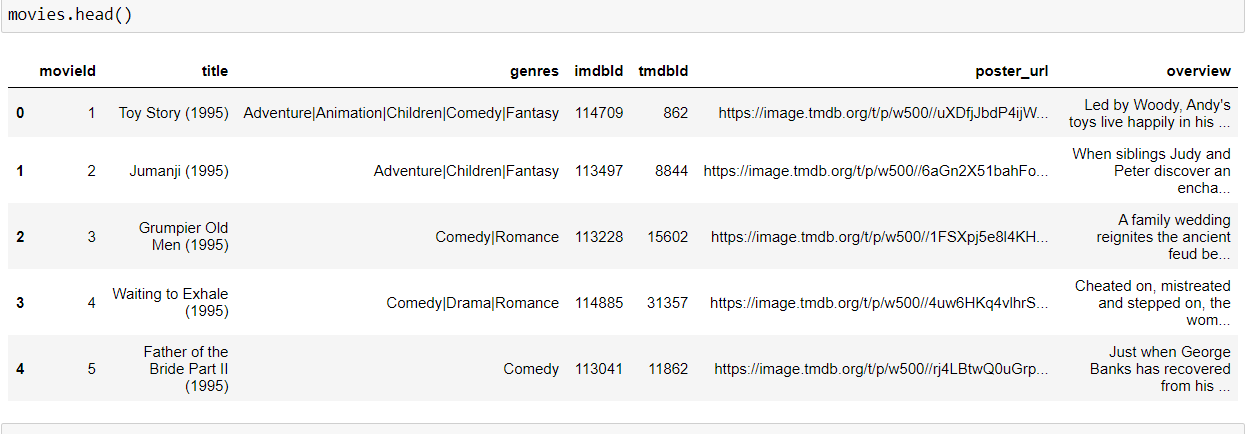
Mặc dù có nhiều cách kết hợp về mặt lý thuyết, nhưng nó không phải lúc nào cũng hiệu quả cho một vấn đề cụ thể. Nguyên tắc quan trọng nhất của khuyến cáo lai là tránh hoặc tạo ra điểm yếu của mọi kỹ thuật gợi ý đơn lẻ.



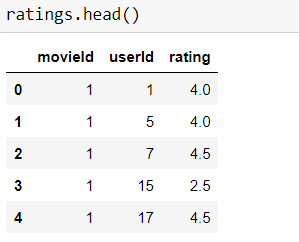
***Hình 14: Hình tổng quan hệ thống Hybrid Recommender Systems – Hệ thống gợi ý kết hợp***

* 1. ***Lựa chọn các đặc trưng thích hợp của các bộ dữ liệu để đưa vào mô hình triển khai***

Trước khi đi vào triển khai mô hình chúng ta sẽ nhìn lại bộ dữ liệu để lựa chọn ra những đặc trưng, thuộc tính phù hợp cho mô hình mà chúng ta sẽ triển khai.



Bảng movies chứa thông tin về các bộ phim mà chủng ta đã lấy thêm thông tin từ trước bao gồm các trường như là movieId, title, genres, imdbId, tmdbId, poster\_url, overview.



Bảng ratings chứa thông tin đánh giá phim của người dùng bao gồm các trường movieId, userId, rating.

Với những mô hình kỹ thuật đã được giới thiệu ở trên phần ***3.1***  chúng ta sẽ lựa chọn 2 phương pháp để xây dựng mô hình trong bài toán gợi ý phim này là 2 phương pháp Content – Base Filtering Recommend gợi ý dựa trên lọc nội dung của các bộ phim và phương pháp Collaborative Filtering recommend lọc cộng tác giữa các bộ phim và người dùng cùng với sự đánh giá của họ.

Tóm lại chúng ta cần các trường dữ liệu sau.

Đối với phương pháp Content-Based Filtering – Lọc dựa theo nội dung chúng ta sẽ lựa chọn các trường như là title, genres và overview của bảng movies. Còn đối với phương pháp Collaborative Filtering chúng ta sẽ sử dụng các trường movieId, userId, rating của bảng ratings

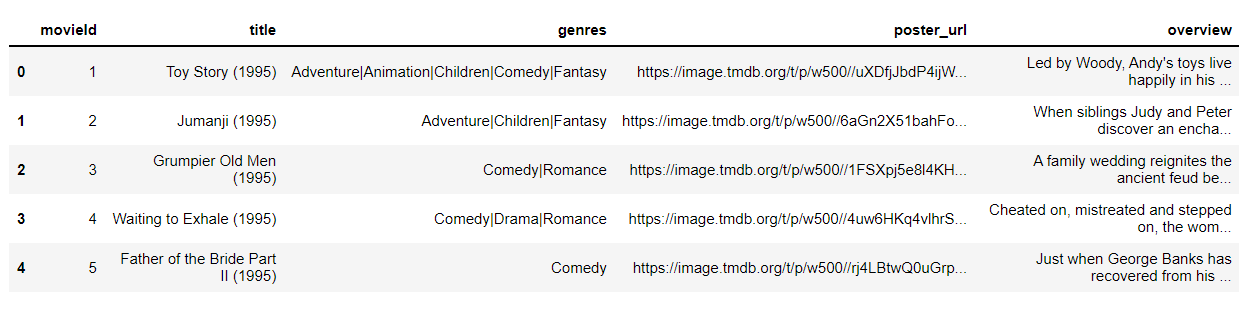
* 1. ***Xây dựng mô hình thuật toán cho bài toán gợi ý phim***

### ***3.3.1 Xây dựng thuật toán cho bài toán trên phương pháp Content – Based Filtering ( gợi ý phim dựa trên việc lọc nội dung)***

Đối với phương pháp này, theo chúng ta đã phân tích tại mục ***3.2*** chúng ta sẽ sử dụng các trường dữ liệu của bảng movies: titl, genres. overview.

Các bước tiến hành xây dựng thuật toán:

* Bước 1: Load dữ liệu và tiền xử lý dữ liệu.

Trước khi đi vào xây dựng thuật toán chúng ta cần phải tiền xử lý cho tập dữ liệu đầu và

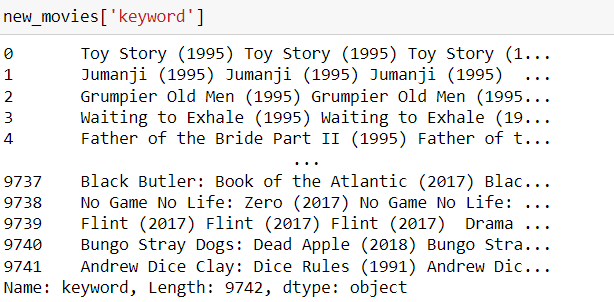
Ta thấy ở bộ dữ liệu trên trường genres và trường overview dữ liệu vẫn chưa được chuẩn hóa( đối với trường genres chúng ta cần tách các từ ra khỏi nhau đã được ngăn cách bằng dấu ‘|’ và cũng chưa được đồng bộ chữ hoa và chữ thường, còn trường overview vẫn chưa được đồng bộ chữ hoa chữ thường)

- Tiến hành tách các từ trong trường genres



- Tạo 1 trường mới ‘keyword’ để kết hợp dữ liệu của các bảng lại với nhau bằng cách gộp các dữ liệu của các trọng số của các trường khác nhau: 3 x title + 2 x genres + overview

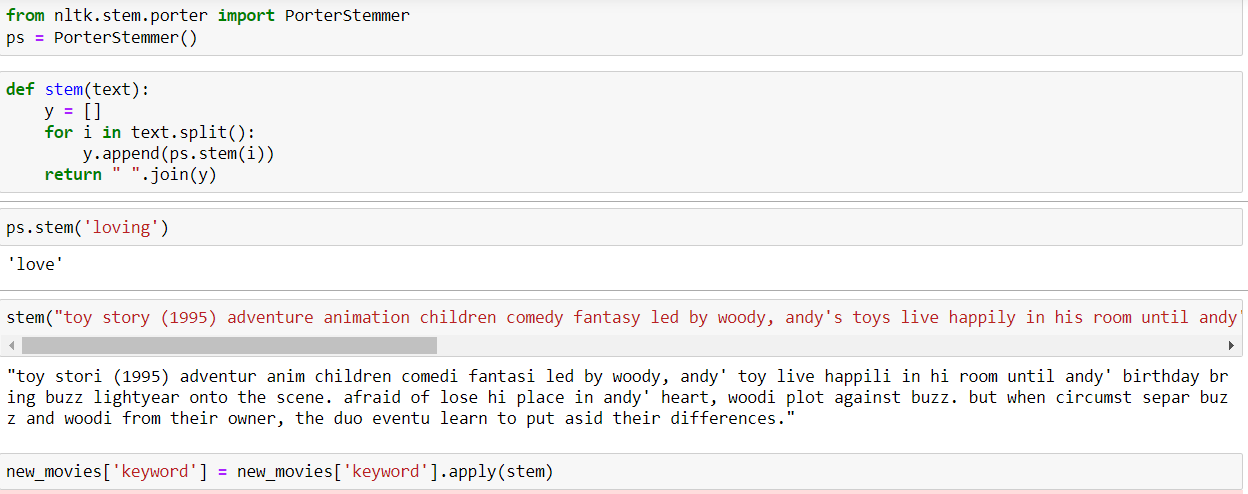




Sau khi kết hợp các trường lại với nhau ta được 1 trường mới là trường keyword chứa thông tin kết hợp và sau đó chúng ta chỉ cần làm việc trên trường này.

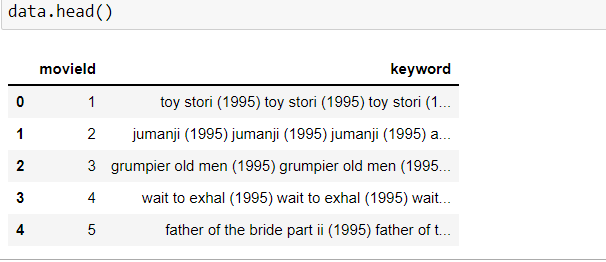
Sau khi kết hợp các trường lại với nha ta thấy 1 vấn đề đó là số từ sẽ rất lớn và sẽ có các từ dồng nghĩa với nhau sẽ khó tạo ra 1 dictionary chuẩn và làm cho quá trình lọc sẽ bị nhiễu.

Tạo hàm cắt các từ tiếng anh biến thể thành các từ nguyên mẫu đơn giản hơn ví dụ như loving, loved, love đểu chuyển thành love.



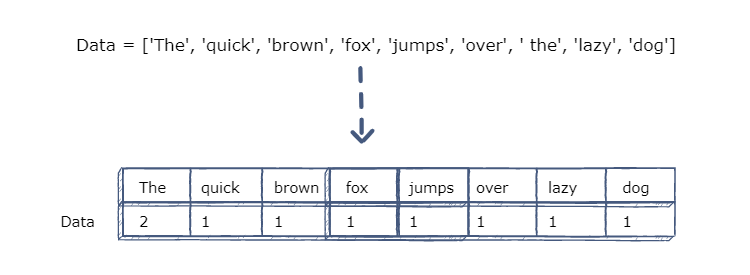
Như vậy chúng ta đã xong phần tiền xử lý cho phần dữ liệu cho thuật toán Content – Based Filtering (Lọc dựa theo nội dung phim, title, genres và overview)

* Bước 2: Tạo ma trận vector đặc trưng cho từng bộ phim dựa trên keyword mà chúng ta đã tạo ở bước tiền xử lý



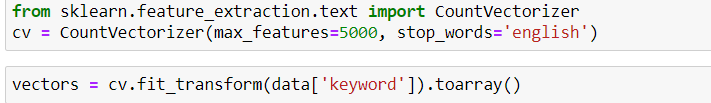
***Hình 15: Dữ liệu cho thuật toán Content – Based Filtering trong bài toán***

Tiếp theo chúng ta sẽ trích rút đặc trưng bằng cách thống kê các từ xuất hiện trong từng keyword của từng bộ phim.



***Hình 16. Tổng quan trích rút đặc trưng của các keyword***

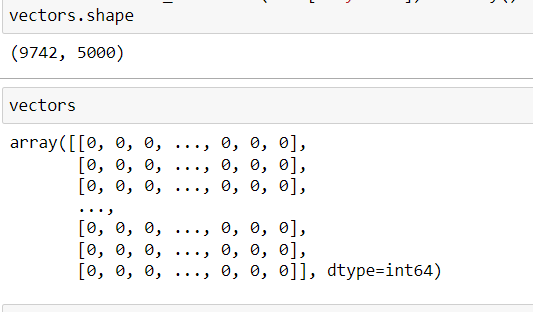
Sử dụng lớp CounterVectorizer của thư viện sklearn để trích rút đặc trưng của từng keyword của từng bộ phim



Chúng ta sẽ set max\_feature = 5000 cho lớp này để giới hạn số từng mà chúng ta mong muốn và ở đây chúng ta sẽ lấy ra 5000 xuất hiện nhiều nhất trong tất cả các keyword của các bộ phim.

Ngoài ra chúng ta cũng set stop\_word = ‘english’ để loại bỏ đi các từ gây nhiễu và không có giá trị trong tập dữ liệu như là a, the, an, is,...

Sau khi thực hiện fit tập dữ liệu cho lớp CounterVectorizer chúng ta sẽ nhận được 1 matrix 2 chiều mỗi hàng là 1 vector đặc trưng của từng bộ phim và chiều dài của mỗi vector đặc trưng là số kì tự được trích rút trong toàn bộ các keyword của các bộ phim.

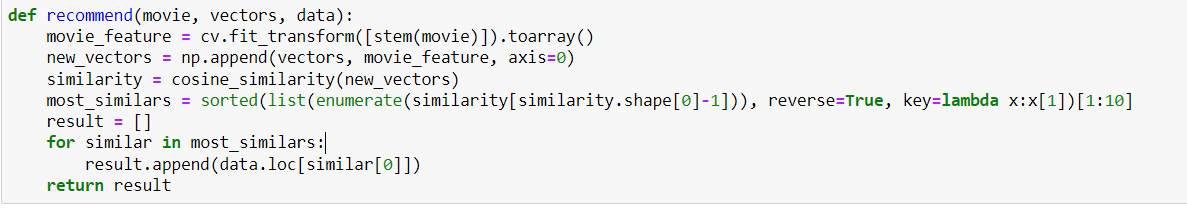


***Hình 17: matrix cho tất cả các vector đặc trưng của từng bộ phim***

* Bước 3: Đưa ra gợi ý cho người dùng

Sau khi chúng ta có matrix vector đặc trưng của các bộ phim chúng ta tiến hành đưa ra gợi ý cho người dùng mỗi khi người dùng có nhu cầu.

Xây dựng hàm đưa ra gợi ý cho người dùng.

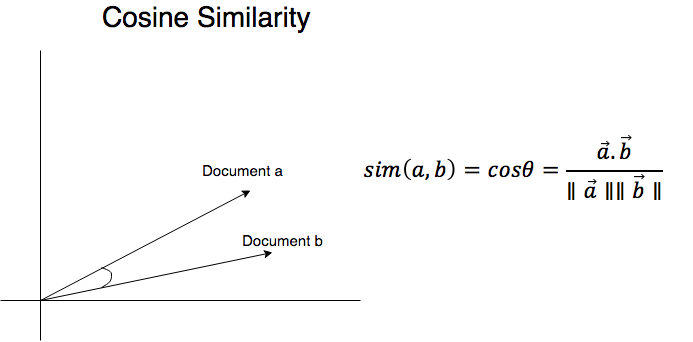


Hàm recommend yêu cầu 3 tham số đầu vào bao gồm movie: tên, mô tả hoặc bất cứ cái gì mà người dùng muốn tìm. Vectors: ma trận đặc trưng của các bộ phim đã trích rút được, data dữ liệu thực của các bộ phim.

Chúng ta sẽ tiến hành trích rút vector đặc trưng của yêu cầu của người dùng.

Gộp vector vừa trích rút được vào vectors đặc trưng của các bộ phim đã trích rút từ trước.

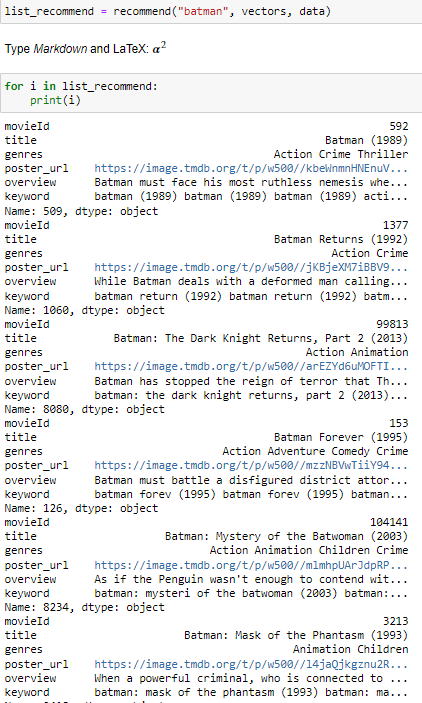
Sử dụng hàm cosine\_similarity để tính khoảnh cách giữa các vector đặc trưng giá trị này càng lớn thì các bộ phim sẽ càng có độ tương đồng cao giá trị này được nằm trong khoảng từ [-1, 1].



***Hình 18: công thức tính khoảng cách cosine\_similarity***

Sorted: sắp xếp giá trị độ tương đồng của các vector theo vector đặc trưng cuối cùng mà chúng ta vừa trích rút đây là vector đặc trưng của yêu cầu người dùng, sau khi sắp xếp chúng ta lấy ra được chỉ số index của các vector có giá trị cosine cao nhất đối với vector này.

Sử dụng bảng data để đưa ra được dữ liệu thực của các bộ phim thông qua các chỉ index của các vector lấy được ở trên và trả về kết quả cho người dùng.



***Hình 19: Kết quả của thử nghiệm phương pháp content-based filtering***

Vì phương pháp này chỉ dựa trên việc lọc nội dung và đưa ra các bộ phim có nội dung, keyword tương đồng nên chúng tôi chưa có sự đánh giá cho thuật toán này

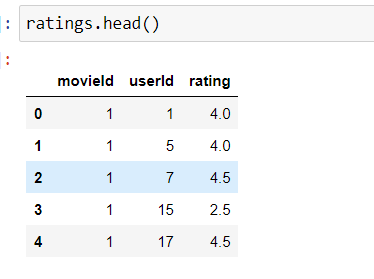
### ***3.3.2 Xây dựng thuật toán cho bài toán trên phương pháp Collaborative – Filtering (gợi ý phim dựa trên việc lọc cộng tác)***

Đối với phương pháp này chúng ta sẽ sử dụng các thuộc tính của bảng ratings cho dữ liệu của thuật toán bao gồm các trường như là movieId, userId, và ratings.

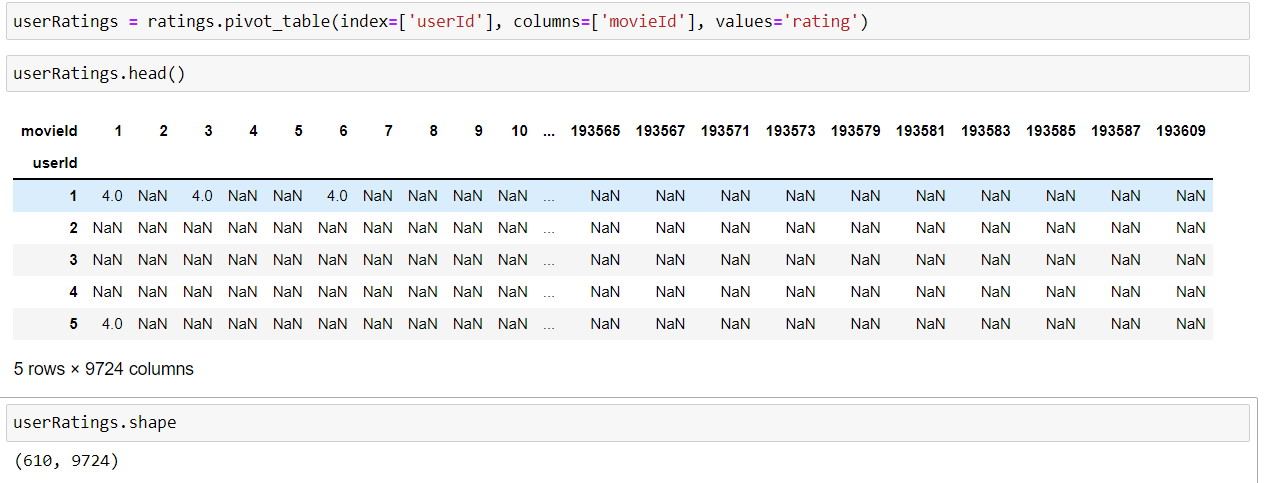
Và ngoài ra chúng ta sẽ sử dụng bảng movies để đưa ra được các dữ liệu về các bộ phim thực.

Các bước xử lý và xây dựng thuật toán:

* Bước 1: đánh giá tổng quan và tiền xử lý bộ dữ liệu

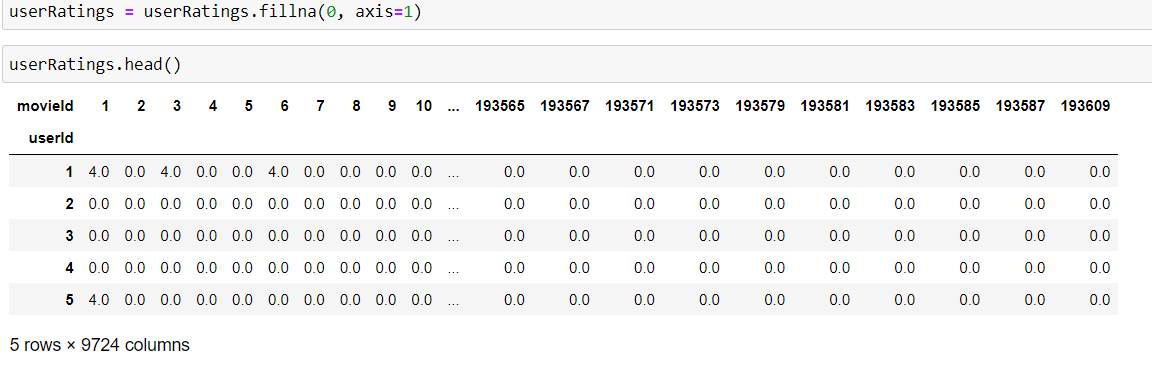
\

Dữ liệu bảng ratings phục vụ cho thuật toán mà chúng ta sẽ triển khai



Chuyển đổi dữ liệu sang dạng bảng với mối quan hệ giữa userId, movieId, và rating dể trực quan hóa dữ liệu

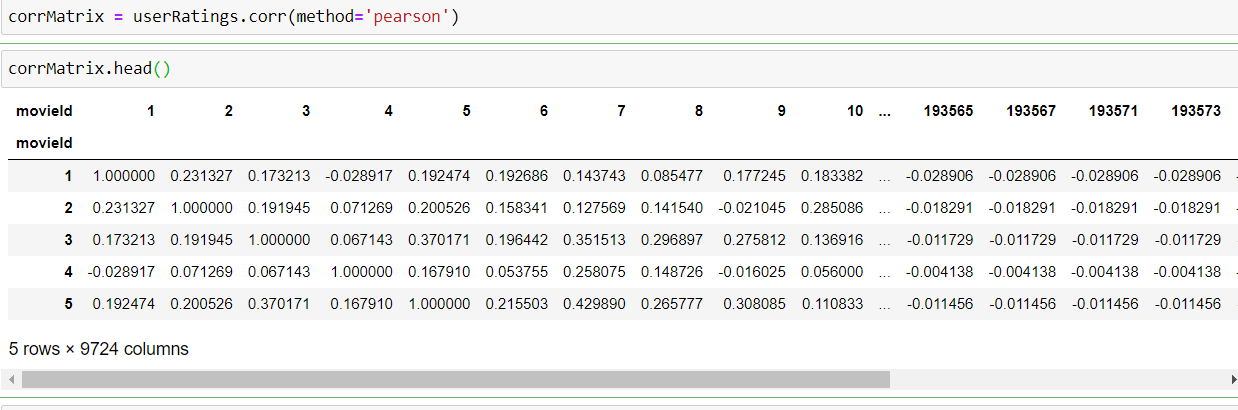
Ta thấy bảng trên có rất nhiều bộ phim chưa được đánh giá bởi tất cả người dùng nên sẽ có giá trị NaN. Chúng ta cần thay thế các giá trị này bằng các giá trị có thể tính toán được đó là 0

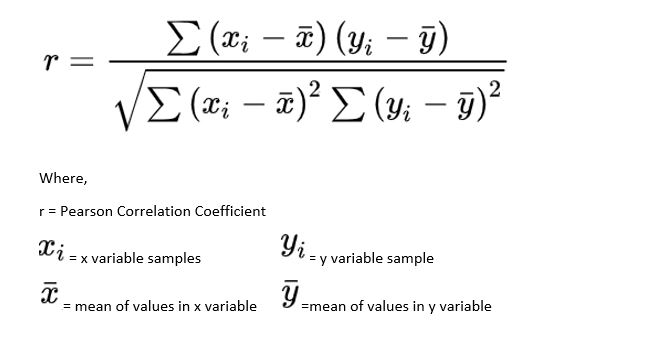


* Bước 2: Xây dựng matrix tương quan giữa các bộ phim đối với các dữ liệu đánh giá của người dùng

Sử dụng hàm corr của thư viện pandas và hàm hệ số tương quan pearson để tính độ tương đồng, mối tương quan correlation giữa các cột dữ liệu của các các bảng trong dataframe tạo ra 1 matrix các mối tương đồng của các bộ phim dựa trên đánh giá của tất cả người dùng.

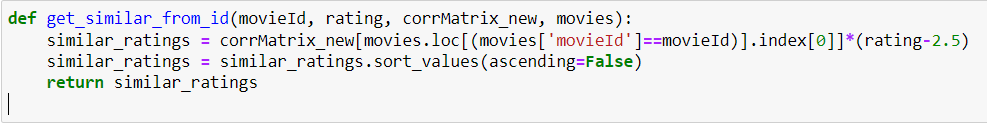
Giá trị về độ tương quan của các bộ phim khi sử dụng hàm này cũng sẽ có giá trị là từ [0:1] và giá trị này càng cao đồng nghĩa với việc 2 bộ phim càng tương đồng





***Hình 20: hàm pearson tính độ tương quan giữa các item profile với nhau***

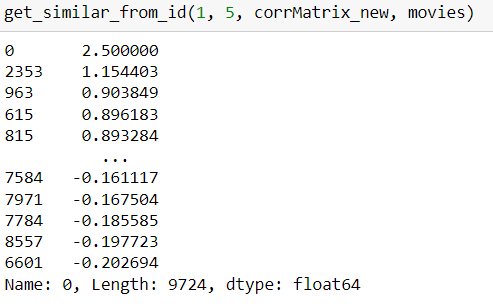
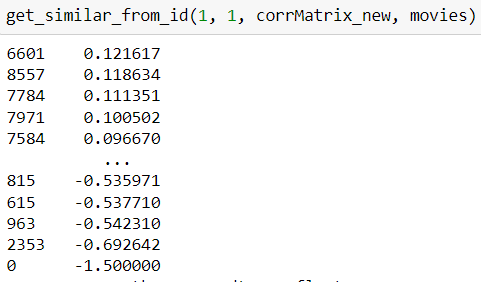
* Bước 3: Xây dựng hàm gợi ý phim cho người dùng dựa vào matrix đã tính toán được ở trên



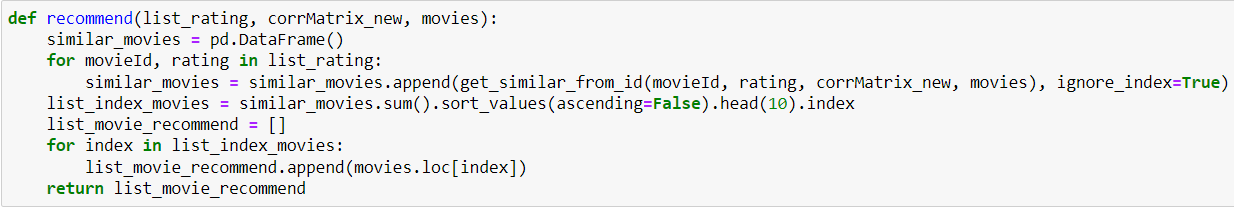
Hàm get\_similarity\_from\_id(): Nhận 4 tham số đầu vào bao gồm moiveId: Id của bộ phim trong bảng dữ liệu ban đầu, rating: giá trị mà người dùng đã đánh giá cho bộ phim này, corrMatrix\_new: Matrix về mối tương quan giữa các bộ phim dựa trên đánh giá của người dùng đã tính toán được ở trên. movies dữ liệu thực của các bộ phim.

Chúng ta sẽ lấy ra được vị trí của bộ phim trong Matrix bằng cách lấy ra index của bộ phim trong bảng movies bằng movieId được cung cấp.

Tiếp đến chúng ta sẽ nhân tất cả các hàng của vector bộ phim đầu vào với (giá trị rating của người dùng – 2.5). Giá trị rating sẽ là từ [0:5] vì vậy chúng ta sẽ nhân chúng với giá trị rating – 2.5 để nếu bộ phim đấy được rating dưới 2.5 thì các bộ phim có sự tương quan cao với bộ phim đầu vào sẽ có giá trị thấp hơn so với các bộ phim khác và ngược lại

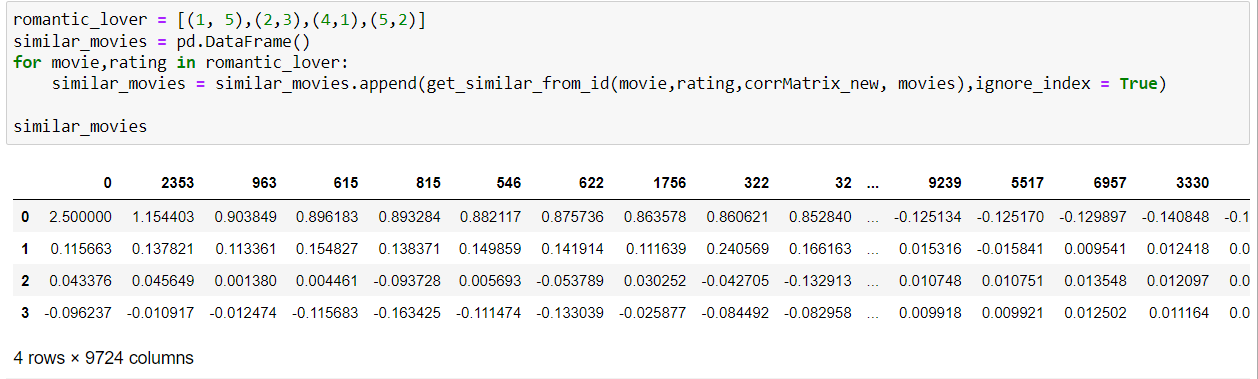
Sự khác nhau về giá trị của các bộ phim so với sự khác nhau giữa các lượt rating khác nhau của người dùng đối với 1 bộ phim



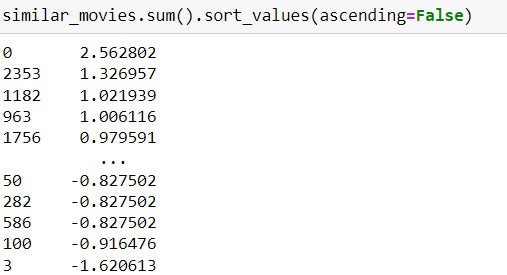
Hàm gợi ý phim cho người dùng recommend(): nhận tham số đầu vào là list\_rating thông tin về các lượt đánh giá người dùng đến các bộ phim

Đầu tiên chúng ta sẽ lặp trên list\_rating để lấy được từng rating của người dùng

Sau đấy chúng ta sẽ đưa từng cái rating của người dùng vào hàm get\_similarity\_from\_id(): đã xây dựng trước đó và nhận về kết quả là 1 mảng chứa giá trị tương đồng đến từng lượt rating.



Sau đấy chúng ta cộng tất cả các giả trị về mức độ tương đồng với các đánh giá của người dùng và tiến hành sắp xếp lại để được các index của các bộ phim có mức độ tương quan lớn nhất với các đánh giá của người dùng với các đánh giá của người dùng khác lên các bộ phim.



Từ index của các bộ phim chúng ta sẽ lấy được dữ liệu thực của các bộ phim dựa trên bảng movies và đây là kết quả cuối cùng.



***Hình 21. Kết quả thử nghiệm của phương pháp Collaborative Filtering***

# **Chương 4**

# **Nghiên cứu xây dựng hệ thống để lưu trữ dữ liệu người dùng, dữ liệu hệ thống và tích hợp mô hình đã xây dựng ở trên (phần Back-end).**

## *4.1 Nghiên cứu các framework , ngôn ngữ lập trình, thư viện giúp lưu trữ dữ liệu có thể tích hợp mô hình, thuật toán và cách tích hợp.*

### ***4.1.1 Framework, ngôn ngữ lập trình, thư viện để cài đặt hệ thống***.

Hiện nay với sự phát triển mạnh mẽ của ngành công nghệ thông tin vì vậy có rất nhiều ngôn ngữ lập trình website backend khác nhau với vô vàn các framework khác nhau cho chúng ta lựa chọn.

Có thể kể đến các framework sau:

[Laravel](https://laravel.com/) là một Framework back end dựa trên PHP, nó có cú pháp đẹp, khả năng phục vụ các team lớn với nhiều chức năng và công cụ hiện đại.  
  
  
Laravel tuân theo mô hình kiến ​​trúc MVC và được xây dựng để tạo điều kiện phát triển dự án sâu, rộng.

Laravel cũng cung cấp hệ thống di chuyển cơ sở dữ liệu của riêng mình và có một hệ sinh thái mạnh mẽ.  
Tính năng chính của Laravel:

* Công cụ định tuyến đơn giản và nhanh chóng
* Đi kèm với CLI của riêng mình
* Hệ thống template mạnh mẽ (Blade)
* Tài liệu tốt

[Ruby on Rails](https://rubyonrails.org/) là một Web Framework phía máy chủ được viết bằng ngôn ngữ lập trình Ruby.  
Nó cung cấp một thiết kế và triết lý tương tự như Django, tuy nhiên, nó có thiết lập quen thuộc hơn nhiều cho các lập trình viên Ruby.  
Ruby khuyến khích sử dụng các mẫu thiết kế (design pattern) như MVC ([MVC là gì](https://niithanoi.edu.vn/mo-hinh-mvc-la-gi.html)?) và DRY (Đừng lặp lại chính mình).  
Một vài ví dụ về các trang web lớn được xây dựng trên Ruby on Rails bao gồm: Shopify, SoundCloud, Basecamp, GitHub.  
Tính năng chính của Ruby on Rails:

* Thư viện lớn plugin có sẵn
* Ruby cung cấp cú pháp rất rõ ràng
* Cộng đồng lớn
* Dự án dễ phát triển và quản lý

[Flask](https://palletsprojects.com/p/flask/) là một Framework khác dựa trên Python. Tuy nhiên, không giống như Django, Flask nhẹ và phù hợp hơn cho việc phát triển các dự án nhỏ hơn.  
Flask cung cấp hỗ trợ cho những thứ như tạo Jinja2 templating, secure cookie, unit testing và RESTful requets dispatching.  
Nó cũng cung cấp tài liệu phong phú và là một giải pháp tuyệt vời cho các lập trình viên mới tự học Python.  
Nếu bạn muốn nhiều thứ hơn, hãy tiến lên Django. Chắc chắn sẽ rất dễ dàng.  
Tính năng chính của Flask:

* Linh động
* Nhỏ nhẹ, phù hợp với dự án nhỏ
* Tài liệu tốt
* Khả năng xây dựng nguyên mẫu nhanh chóng

[Spring Boot](https://spring.io/projects/spring-boot) là một web framework dựa trên Java.  
Nó được tạo ra bởi Spring với mục đích dễ sử dụng và cung cấp một ứng dụng dựa trên Spring ở cấp độ sản xuất mà bạn có thể cần chạy.  
Có thể nói, Spring boot là bản rút gọn của Spring. Vì thế, bạn cũng   
Tính năng chính của Spring Boot:

* Tạo các ứng dụng Spring độc lập
* Khả năng mở rộng cao
* Tài liệu tuyệt vời
* Được xây dựng cho các ứng dụng quy mô lớn sử dụng cách tiếp cận đám mây
* Hệ sinh thái mở rộng

Tuy nhiên để có thể tích hợp được mô hình, thuật toán mà chung ta đã xây dựng ở trên thì ở đây chúng tôi sẽ lựa chọn framework django của python đã dễ dàng cài đặt tích hợp hệ thống.

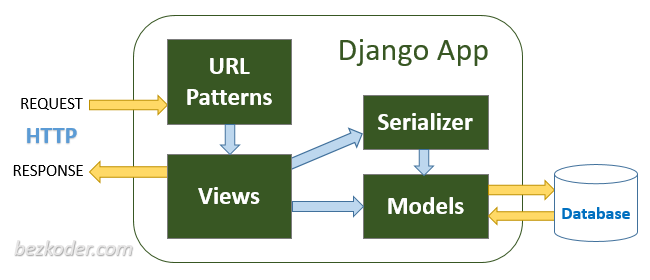
Django ( JANG -goh ; được cách điệu thành django ) là một nguồn mở và miễn phí dựa trên Python khung web, theo mô hình kiến ​​trúc mô hình-khung nhìn mẫu (MVT). Nó được duy trì bởi Django Software Foundation (DSF), một tổ chức độc lập được thành lập dưới dạng phi lợi nhuận 501

Mục tiêu chính của Django là dễ dàng tạo ra các trang web dựa trên cơ sở dữ liệu phức tạp. Khung này nhấn mạnh khả năng sử dụng lại và &quot;khả năng cắm&quot; của các thành phần, ít mã hơn, khớp nối thấp, phát triển nhanh và nguyên tắc không lặp lại chính mình.  Python được sử dụng xuyên suốt, ngay cả đối với các tệp cài đặt và mô hình dữ liệu.

Django cũng cung cấp giao diện tạo, đọc, cập nhật và xóa quản trị tùy chọn được tạo động thông qua hướng nội và được định cấu hình qua các mô hình quản trị.

Một số trang web nổi tiếng sử dụng Django bao gồm Dịch vụ phát thanh công cộng, Instagram,  Mozilla, The Washington Times  Disqus,  Bitbucket, và Nextdoor. Nó đã được sử dụng trên Pinterest, nhưng sau đó trang web đã chuyển sang một khung được xây dựng trên Flask.

Ngoài ra chủng tôi cũng sử dụng thư viện django-rest-framework để tạo api gửi và nhận dữ liệu cho phần server qua dạng json.



***Hình 22: Workflow của framwork Django và thư viện rest framework***

### ***4.1.2 Nghiên cứu cách tích hợp mô hình thuật toán vào hệ thống.***

Vì ở đây cả 2 phương pháp, thuật toán mà chúng ta đã thiết kế ở trên đều chỉ sử dụng các vector, matrix để lưu thông tin các bộ phim vì vậy cách đơn giản nhất để tích hợp mô hình, thuật toán là lưu các matrix, vector này vào các file( ở đây chúng tôi chọn file . npz là định dạng file nén giảm dung lượng của file) sau đấy khi cần tính toán xử lý chúng ta sẽ load dữ liệu từ các file này lên hệ thống và tính toán.

## *4.2 Xây dựng mô hình lưu trữ dữ liệu và tích hợp mô hình thuật toán gơi ý phim (backend)*

Như chúng ta đã nói ở trên, chúng ta sẽ lựa chọn ngôn ngữ lập trình python cùng với framework django của nó để xây dựng phần lưu trữ dữ liệu cho người dùng và hệ thống cùng với tích hợp thuật toán gợi ý phim. Cùng với đó là framework rest framwork để xây dựng api kết nối với phía người dùng của hệ thống.

### ***4.2.1 Thiết kế mô hình lưu trữ dữ liệu cho hệ thống***

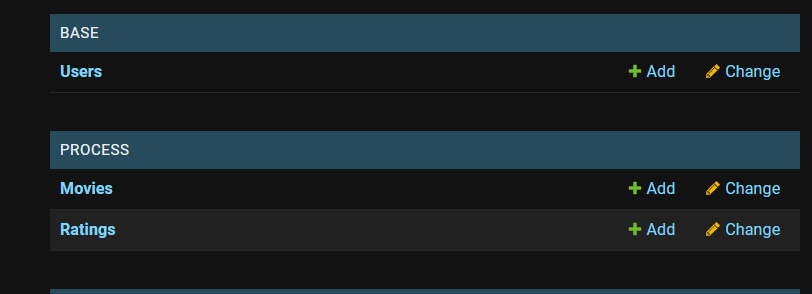
Django là 1 framework với mã nguồn mở giúp tạo lưu trữ và truy xuất dữ liệu 1 cách dễ dàng mà không cần phải viết các câu lệnh như trong các hệ quản trị cơ sở dữ liệu khác



***Hình 23: database của hệ thống***

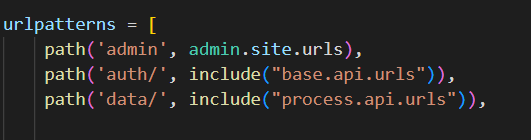
******

***Câu lệnh để tạo bảng lưu dữ liệu***

******

***Các bảng và dữ liệu được lưu trong hệ thống***

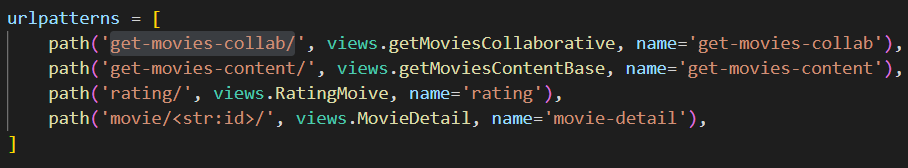
### ***4.2.2 Cấu hình router cho các đường dẫn của trang web***

******

Đường dẫn admin sẽ giúp chúng ta đến trang admin của phần backend này, nơi lưu trữ toàn bộ dữ liệu được lưu của hệ thống

Đường dẫn auth/ sẽ giúp chúng ta đến với các chức năng của trang website như là đăng nhập, đăng ký tài khoản, verify email, reset password cho các user.

Đường dấn data/ sẽ đưa chủng ta đến với các chứa năng chính của trang web như là lấy recommend cho người dùng



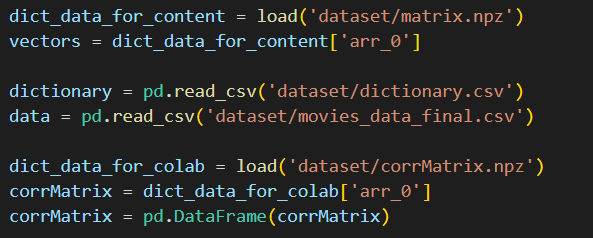
Đường dẫn get-movies-collab/ đưa chúng ta đến với hàm chức năng lấy gợi ý cho người người dùng dựa trên thuật toán collaborative – filtering

Đường dẫn get-movies-content/ đưa chúng ta đến với hàm chức năng lấy gợi ý cho người người dùng dựa trên thuật toán content-base filtering

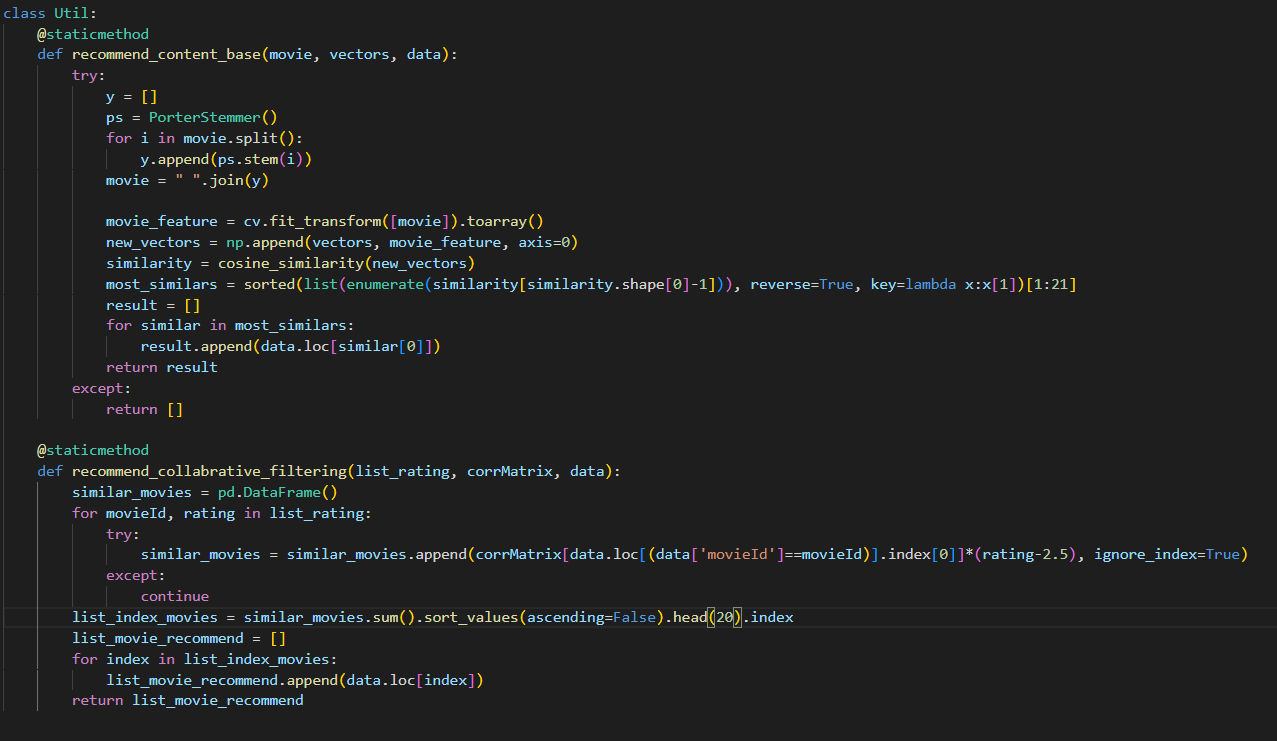
Đường dẫn rating/ đưa chúng ta đến với hàm chức năng đánh giá phim

Đường dẫn movie/<str:id>/ đưa chúng ta đến với hàm chứa năng lấy thông tin của 1 bộ phim cụ thể nào đấy

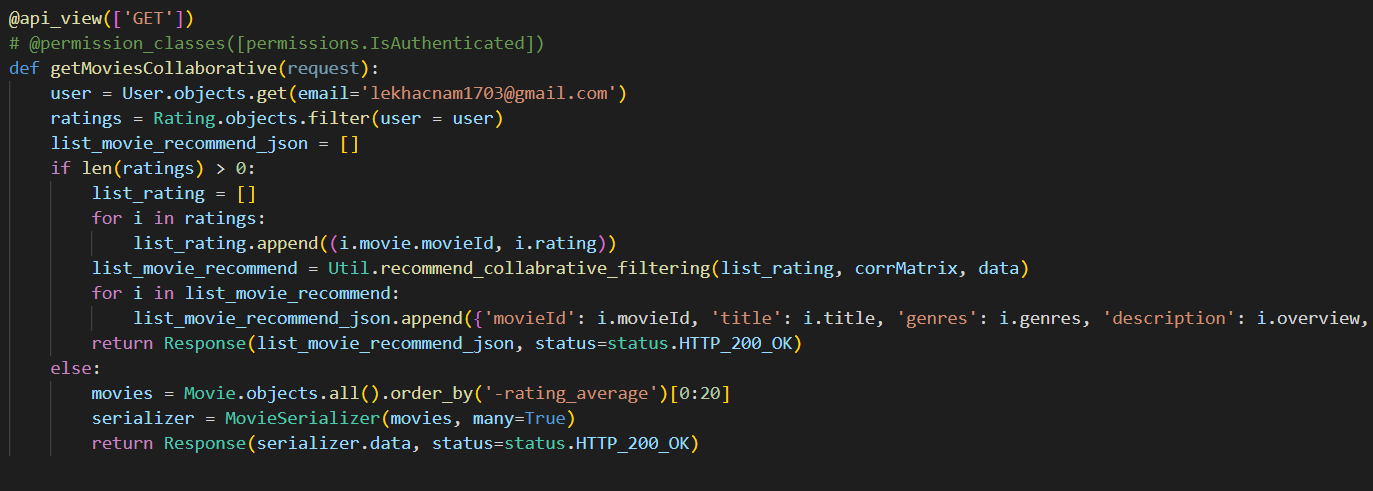
### ***4.2.3 Các hàm chức năng trong hệ thống***

******

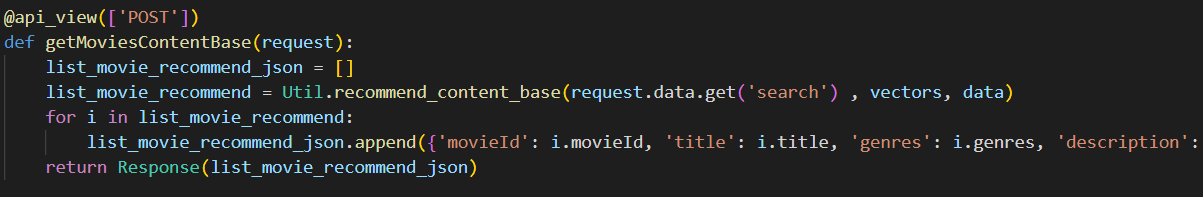
Load dữ liệu đã được lưu vào các file nén .npz vào hệ thống



2 hàm chức năng chính đưa ra các bộ phim theo nhu cầu của người dùng theo 2 phương pháp là content-based filtering và collaborative- filtering

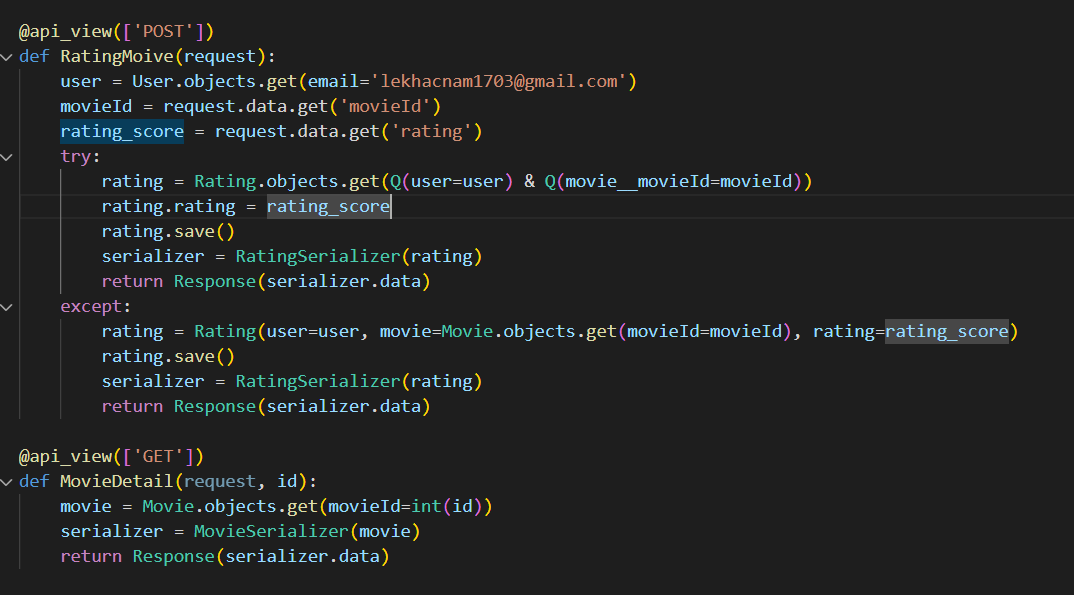
******

Hàm giúp lấy gợi ý các bộ phim cho người dùng bằng phương pháp collaborative filtering bằng cách gọi đến hàm của lớp Util và trả về dữ liệu cho phía client



Hàm giúp lấy gợi ý các bộ phim cho người dùng bằng phương pháp content-based filtering bằng cách gọi đến hàm của lớp Util và trả về dữ liệu cho phía client

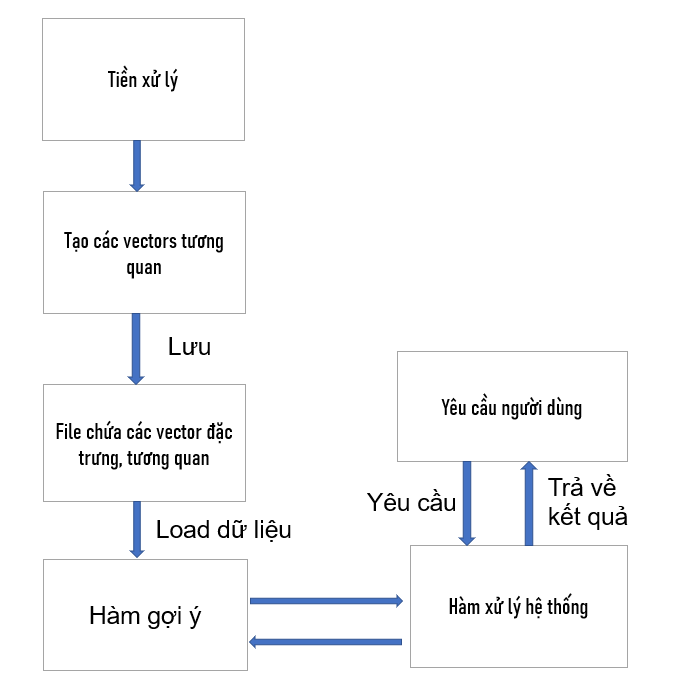
Cả 2 hàm này đều sẽ trả về dữ liệu bằng các file dữ liệu json thông qua framework django rest framework.

******

Hàm xử lý để lấy dữ liệu của 1 bộ phim cụ thế và hàm xác nhận đánh giá phim của người dùng.

******

Kết quả nhận được khi gọi đến hàm get movie collaborative đã trả về dữ liệu dạng json cho người dùng thông qua framework django rest framework



***Hình 23: workflow của hệ thống backend***

Tóm tắt quá trình làm việc của hệ thống backend sau khi đã tích hợp được mô hình gợi ý phim:

Bước 1: tiền xử lý dữ liệu

Bước 2: tạo các vector đặc trưng, tương quan cho các bộ phim

Bước 3: Lưu các vector đặc trưng, tương quan vào các file nén

Bước 4: load dữ liệu từ các file nén lên hệ thống để xây dựng các hàm gợi ý

Bước 5: Khi có yêu cầu từ người dùng đến server (hàm xử lý trung tâm) Hàm xử lý trung tâm sẽ gọi đến hàm gợi ý để lấy dữ liệu

Bước 6: Hàm gợi ý trả về dữ liệu cho hàm xử lý chung của hệ thống và hàm này sẽ trả về dữ liệu cho người dùng

# **Chương 5:**

# **Nghiên cứu xây dựng giao diện để người dùng tương tác với hệ thống và hiển thị kết quả cho người dùng (phần Front-end) và kết nối với phần xử lý dữ liệu (Back-end).**

## *5.1 Nghiên cứu các ngôn ngữ lập trình, framework, thư viện giúp tạo giao diện hiển thị cho người dùng*

Cũng giống như backend, frontend cũng là 1 phần giúp phát triển website và hiện nay nghành lập trình frontend cũng đang rất phát triển vì vậy có rất nhiều ngôn ngữ lập trình cũng như là framework có thể viết được giao diện người dùng và 1 trong số đó không thể không nhắc đến javascript 1 ngôn ngữ lập trình được coi là phổ biến nhất hiện nay.

Cùng với sự phát triển như vũ bão của cộng đồng JavaScript, chúng ta có cảm giác như mỗi tuần lại có thêm các framework, library JavaScript mới được ra mắt. Thật khó có thể để liệt kê hết ra đây các framework, library, tool hiện có trong cộng đồng. Bài viết này chỉ giới thiệu với các bạn những framework, library, tool nổi tiếng nhất và có ảnh hưởng nhất trong cộng đồng JavaScript, cũng như khi nào nên áp dụng các framework, library hay tool đó.

Một số framework nối bật của JS

AngularJS là framework phổ biến nhất, được nhiều developer sử dụng để xây dựng phần front-end cho những web application có phần front-end phức tạp. Nó là một opensource được hỗ trợ bởi Google.

AngularJS là một MVC framework. Nó cung cấp two-way data binding giữa model và view. Cách binding dữ liệu như thế này cho phép tự động update ở cả hai phía bất kể khi nào dữ liệu bị thay đổi. Vì thế hỗ trợ các bạn xây dựng các thành phần view có thể tái sử dụng ở nhiều chỗ. Ngoài ra nó cũng cung cấp cho các bạn một services framework để hỗ trợ tối đa giao tiếp giữa back-end và front-end.

ReactJS Đây có lẽ là thứ tốn giấy mực nhiều nhất trong thời gian gần đây. Tất cả mọi người ai ai cũng nói về ReactJS. ReactJS là một opensource được phát triển bới Facebook. Bản thân nó là một JavaScript library dùng để xây dựng user interface.

React cung cấp một phương thức dễ dàng để xây dựng giao diện thông qua việc phát triển các component và kết hợp chúng lại với nhau. Nó cũng cung cấp một phương thức trừa tượng hoá DOM để tối ưu việc rendering cũng như cho phép có thể render thông qua Node.js. Một điểm nữa đó là nó implement one-way reactive data flow, giúp cho chúng ta có một cái nhìn dễ dàng hơn.

Trong một số trường hợp, người ta thường kết hợp React để làm view layer trong một số MVC framework như Angular, Ember.

BackboneJS nổi tiếng là một framework đơn giản gói gọn trong một file JavaScript. Nó cực kì phổ biến đối với các web application nhỏ.

BackboneJS cung cấp một MVC framework đầy đủ với cả phần routing. Phần model trong BackboneJS cho phép key-value binding và sử dụng event để handle khi dữ liệu thay đổi. Phần view đảm nhận việc handle các event và router dùng để handle URL cũng như trạng thái trong application. Quá đầy đủ cho bạn để xây dựng một Single Page Application mà không cần phải apply quá nhiều thứ phức tạp.

Vue.js là một framework JavaScript nhỏ gọn mới xuất hiện. Nó là một [framework JavaScript phổ biến](https://topdev.vn/blog/top-7-javascript-frameworks/) trên GitHub tính theo lượng sao trên GitHub. Vue là một framework không quá cứng nhắc và do đó giúp nhà phát triển dễ dàng nắm bắt được nó. Các mẫu cú pháp HTML của Vue gắn kết phần DOM đã render với giá trị của dữ liệu.

Framework này cung cấp trải nghiệm giống React với những Virtual DOM của nó và các component có thể tái sử dụng giúp bạn tạo ra cả widgets và toàn bộ ứng dụng web. Hơn nữa, bạn cũng có thể dùng cú pháp JSX để viết phần chức năng render trực tiếp.

Khi trạng thái thay đổi, Vue.js sẽ mở một hệ thống phản ứng để xác định rằng điều gì đã thay đổi và render số lượng nhỏ nhất các component.

Vue.js cũng hỗ trợ tích hợp những thư viện khác vào framework mà không cần tốn công sức nhiều.

Và trong bài lần này chúng tôi sẽ sử dụng framework Reactjs của ngôn ngữ JS cùng với các ngôn ngữ thuần khác như là HTML, CSS và thư viện Bootstrap để viết nên giao diện cho website của chúng tôi.

## *5.2 Nghiên cứu các hệ thống, các cách kết nối phần giao diện người dùng với phần lưu trữ dữ liệu hệ thống*

Như chúng tôi đã đề cập ở trên thì ứng dụng web của chúng tôi sẽ chia thanh 2 phần riêng biệt là backend (đã nói ở trên) và frontend vì vậy chúng tôi cần phải sử dụng 1 hệ thống giúp kết nối 2 phân riêng biệt lại với nhau.

Ở phần xây dụng hệ thống lưu trữ xử lý dữ liệu chúng tôi đã đề cập đến việc sử dùng api và phần backend đấy đã sử dụng 1 api để cho phép kết nối từ bên ngoài server để lấy và gửi dữ liệu

Phần việc còn lại là chúng ta cần sử dụng 1 công cụ, 1 hệ thống có thế giúp chúng ta kết nối được đến các cổng mà backend đã mở ra.

Hiện nay có 3 công cụ để gọi API phổ biến nhất đó là:

XMLHttpRequest (XHR) là một API ở dạng một đối tượng có các phương thức truyền dữ liệu giữa trình duyệt web và máy chủ web. Đối tượng được cung cấp bởi môi trường JavaScript của trình duyệt. Đặc biệt, việc truy xuất dữ liệu từ XHR với mục đích liên tục sửa đổi một trang web đã tải là khái niệm cơ bản của thiết kế Ajax. Bất chấp tên gọi, XHR có thể được sử dụng với các giao thức khác ngoài HTTP và dữ liệu có thể ở dạng không chỉ XML, mà còn cả JSON, HTML hoặc văn bản thuần túy.

Fetch API: cung cấp một giao diện để tìm nạp tài nguyên (bao gồm trên toàn mạng). Nó sẽ có vẻ quen thuộc với bất kỳ ai đã sử dụng XMLHttpRequest, nhưng API mới cung cấp một bộ tính năng mạnh mẽ và linh hoạt hơn có sẵn trong ngôn ngữ lập trình JS.

Axios là một HTTP Client dựa trên base-promise cho node.js và trình duyệt. Nó là isomorphic (= nó có thể chạy trong trình duyệt và nodejs với cùng một cơ sở mã). Ở phía máy chủ, nó sử dụng mô-đun http node.js gốc, trong khi trên máy khách (trình duyệt), nó sử dụng XMLHttpRequests.

Tính năng của axios

Tạo XMLHttpRequests từ trình duyệt

Thực hiện các yêu cầu http từ node.js

Hỗ trợ API Promise

Chặn yêu cầu và phản hồi

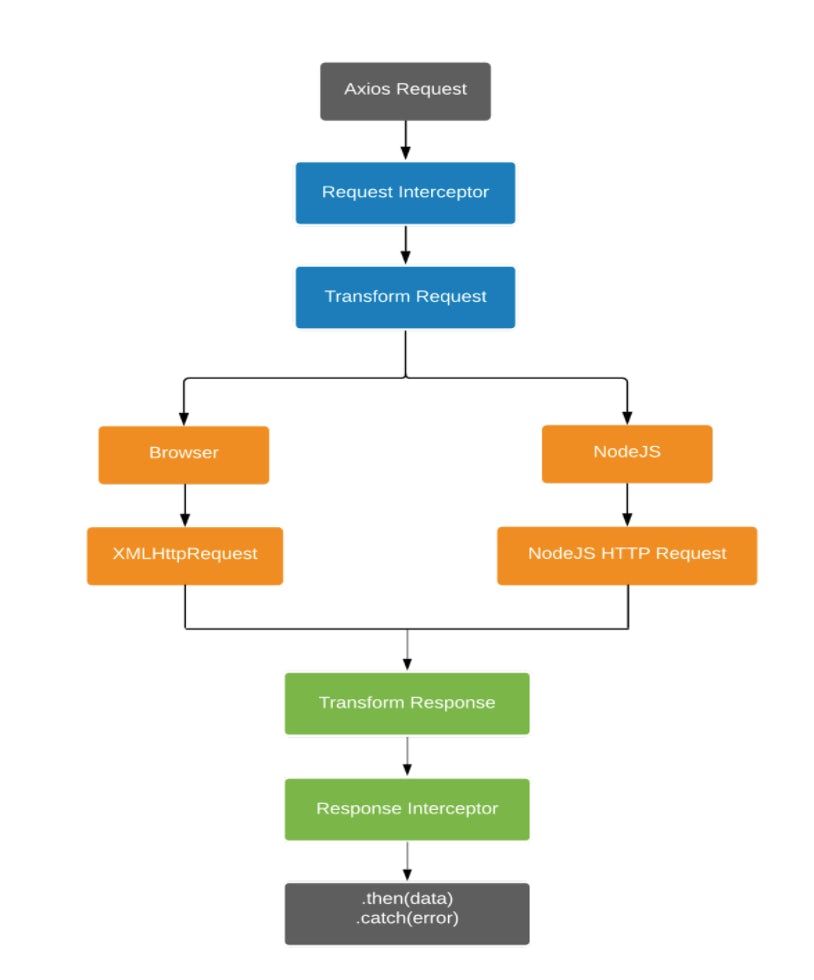
Chuyển đổi dữ liệu yêu cầu và phản hồi

Hủy yêu cầu

Chuyển đổi tự động cho dữ liệu JSON

Hỗ trợ từ phía khách hàng để bảo vệ chống lại XSRF

Và trong bài lần này chúng ta sẽ sử dụng thư viện axios để làm việc với api kết nối đến máy chủ để lấy và gửi dữ liệu

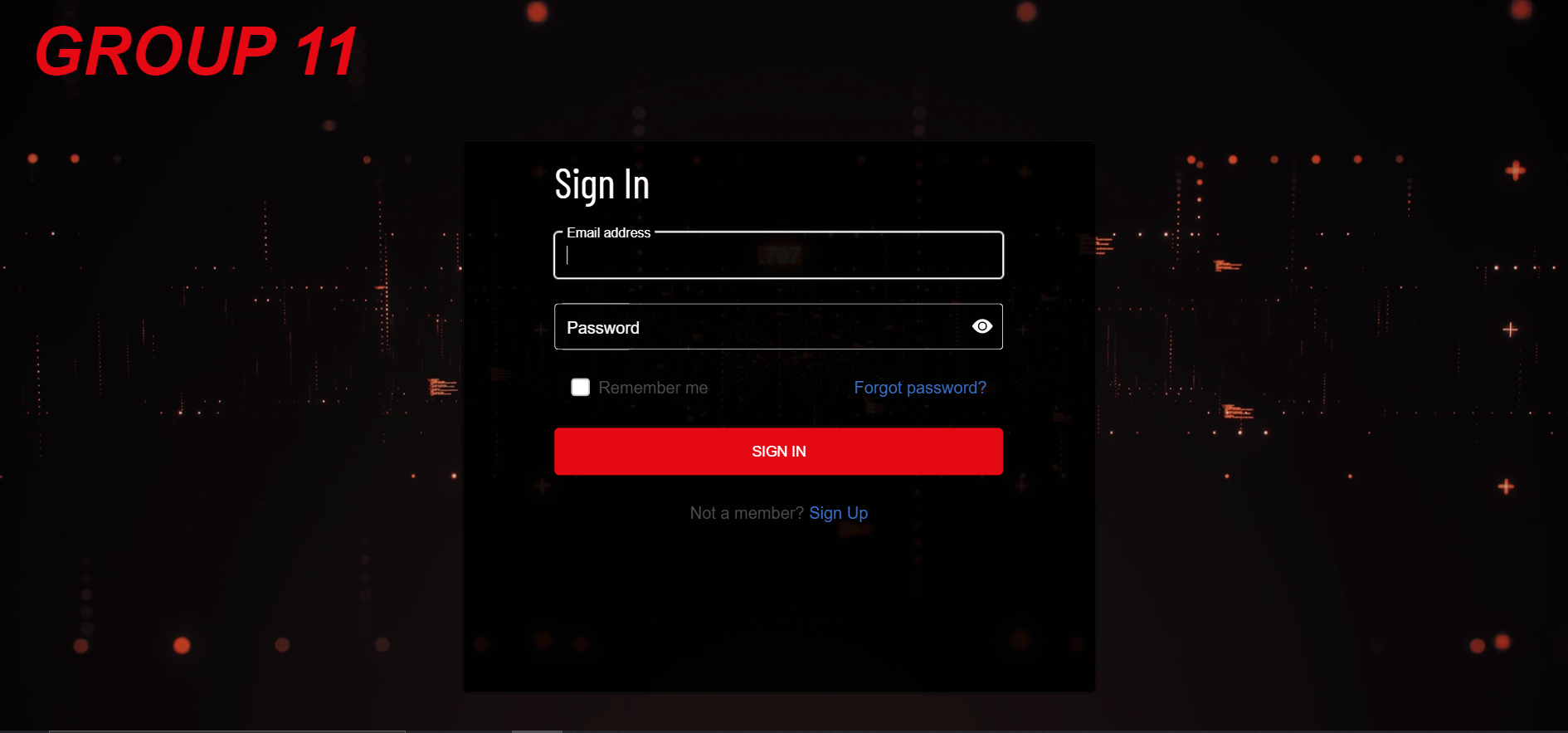


***Hình 24. Vòng đời của axios khi thực hiện request call api***

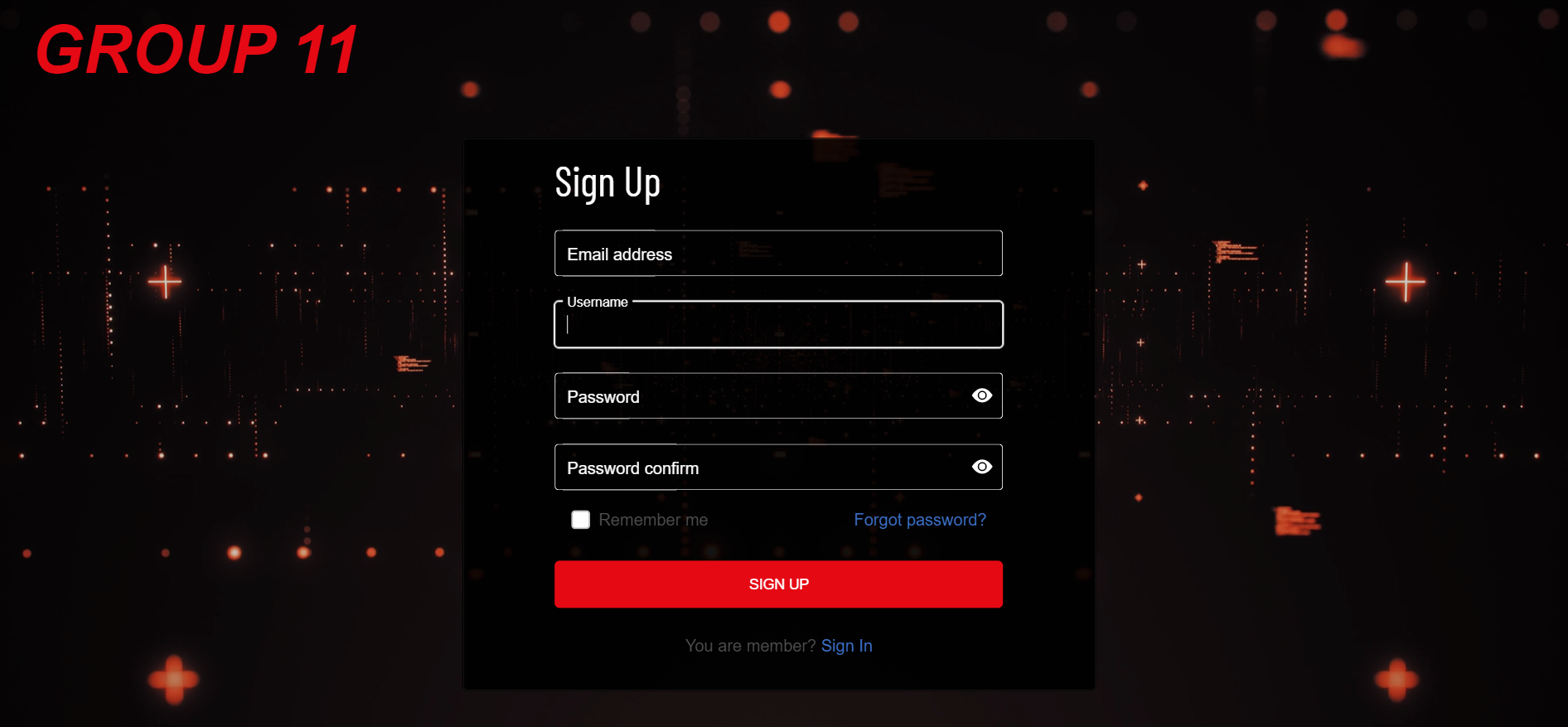
## *5.3: Xây dựng hệ thống giao diện của hệ thống và kết hợp kết nối hệ thống giao diện với hệ thống xử lý dữ liệu*

Trong phần này chúng ta sẽ sử dụng thư viện reactJs của ngôn ngữ JS cùng với HTML, CSS, Bootsrap để xây dựng hệ thống giao diện cho người dùng và sử dụng thư viện axios để kết nối với phần xử lý dữ liệu

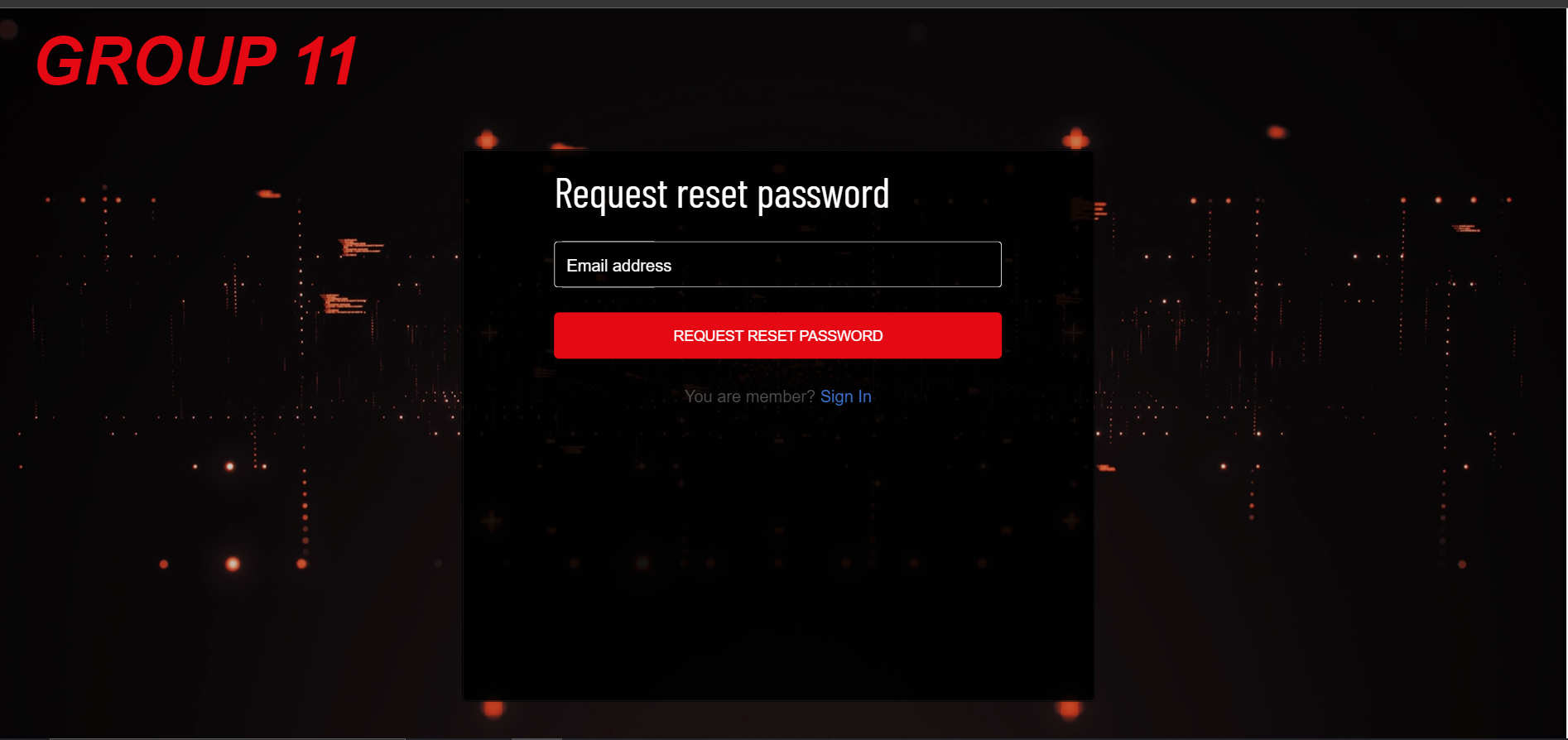
Hệ thống giao diện của hệ thống:



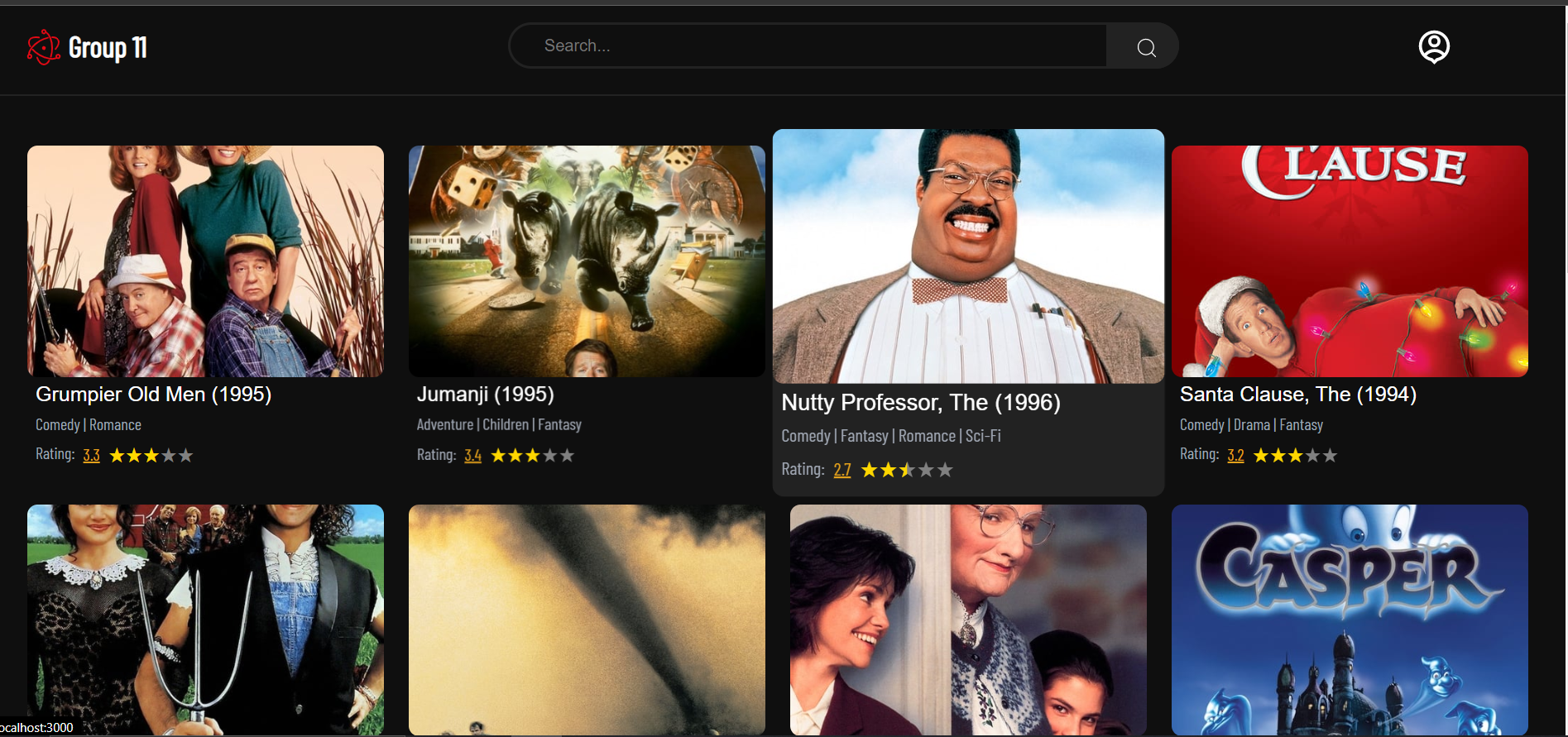
Trang Login của hệ thống cho phép người dùng đăng nhập vào hệ thống sử dụng email và password.



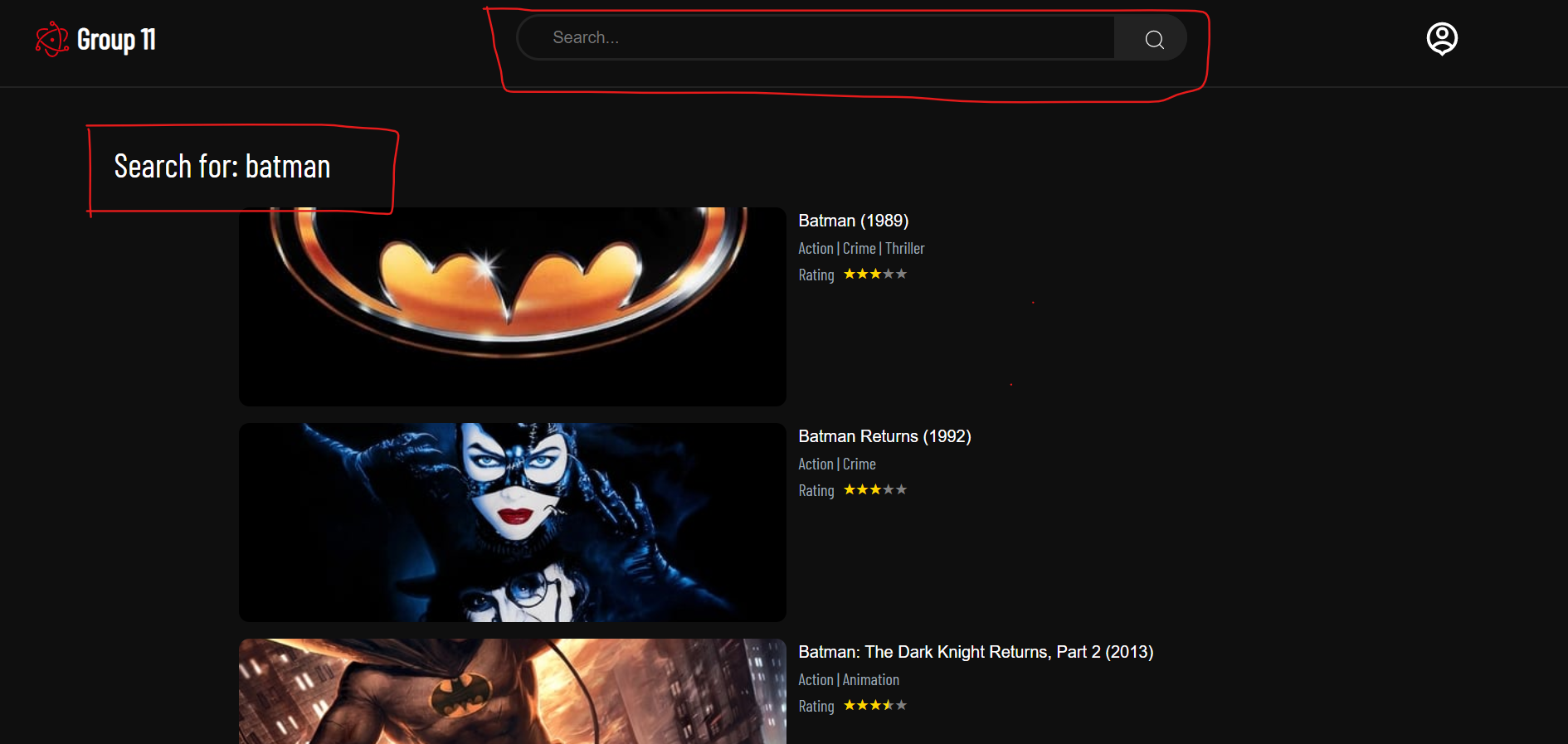
Trang sign up cho phép người dùng đăng ký 1 tài khoản mới cho sử dụng email.



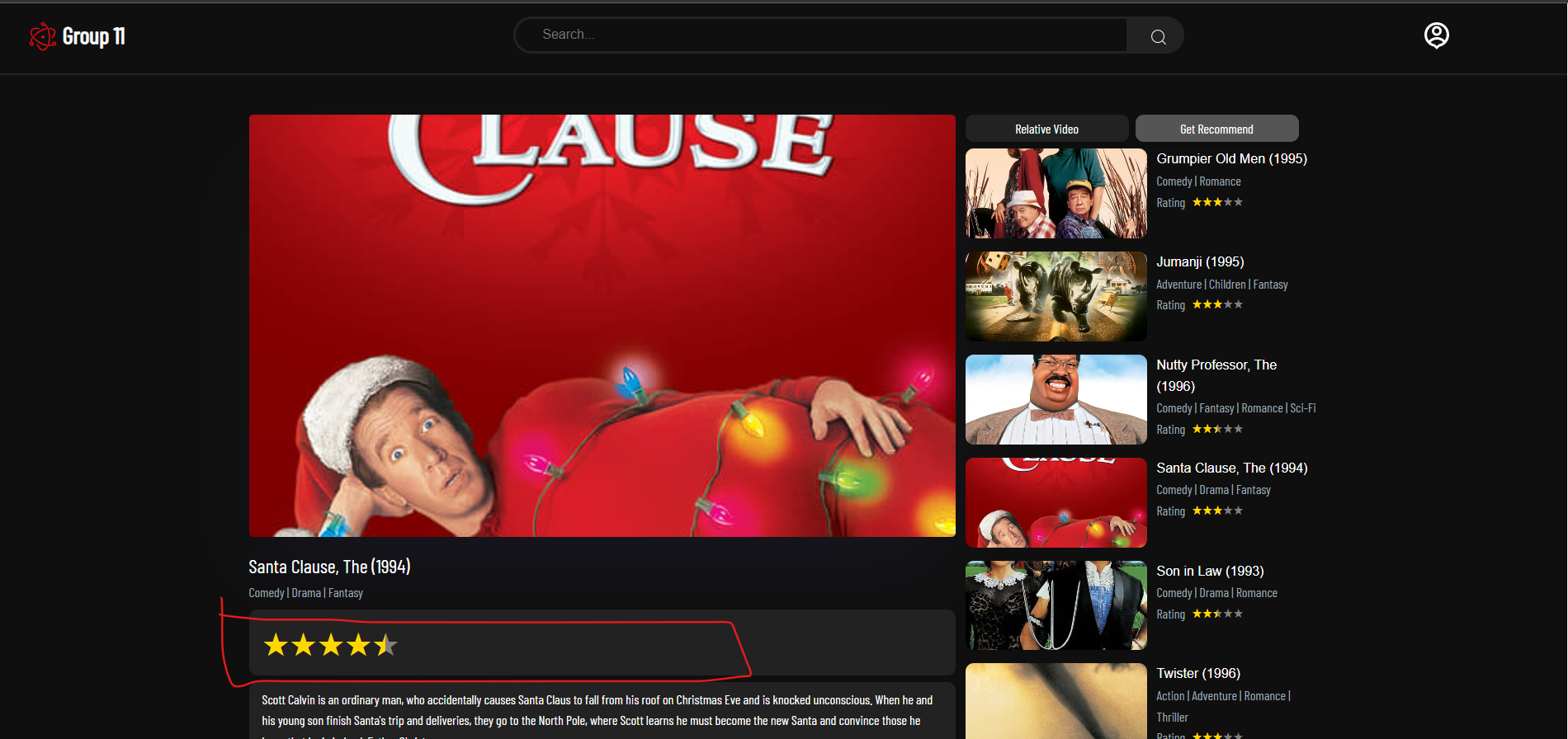
Trang Request reset password cho phép người dùng yêu cầu đổi mật khẩu



Giao diện sau khi login: sau khi login nếu người dùng chưa từng đánh giá 1 bộ phim nào hoặc là người mới thì hệ thống sẽ đưa ra những bộ phim có tổng rating cao nhất, Còn nếu người dùng đã rating 1 số bộ phim thì sẽ đưa ra recommend những bộ phim khác cho người dùng.



Giao diện tìm kiếm phim cho người dùng. Khi người dùng nhập vào 1 yêu cầu nào đó vào ô search hệ thống sẽ đưa ra những bộ phim tương đồng cho người dùng.

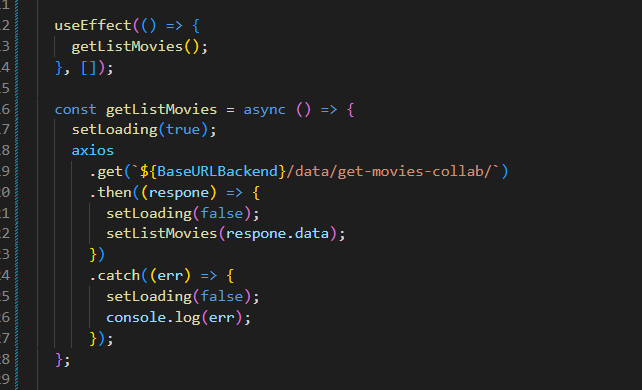
******

Giao diện đánh giá phim của người dùng sau khi người dùng đánh giá phim xong thì có thể lấy được các bộ phim khác do hệ thống gợi ý bằng 2 tab relative video và get recommend ở bên cạnh

Kết nối phần giao diện với phần quản lý xử lý dữ liệu thông qua API và thư viện axios

$ npm install axios

Chúng ta có thể sử dụng câu lệnh này đê cài đặt thư viện axios

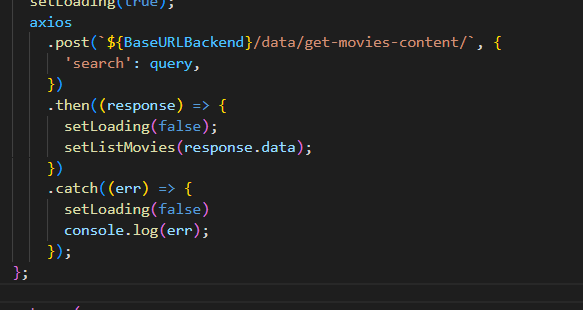
******

Trong đoạn code này chúng ta đã gọi đến api của phần backend có đường dẫn là BaseURLBackEnd/data/get-movies-collab/ với phương thức .get mà chúng ta đã cài đặt ở phần backend lúc năy sau khi gọi đến api với đường dẫn này chủng ta sẽ lấy được dữ liệu về những bộ phim được gợi ý bằng phương pháp collaborative – filering mà chúng ta đã trả về lúc nãy tại server.

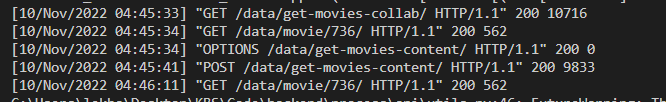
******

Và chúng ta sẽ lấy dữ liệu này và load lên phần giao diện của mình.

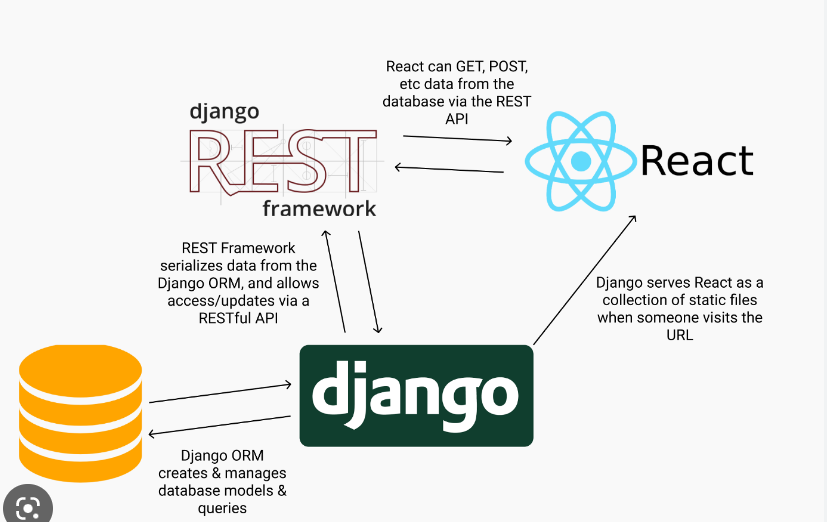
Tương tự với các phần khác. Đây là phần lấy dữ liệu từ server về client. Còn người lại khi chúng ta muốn gửi dữ liệu lên server để xử lý thì chúng ta cần làm như sau

******

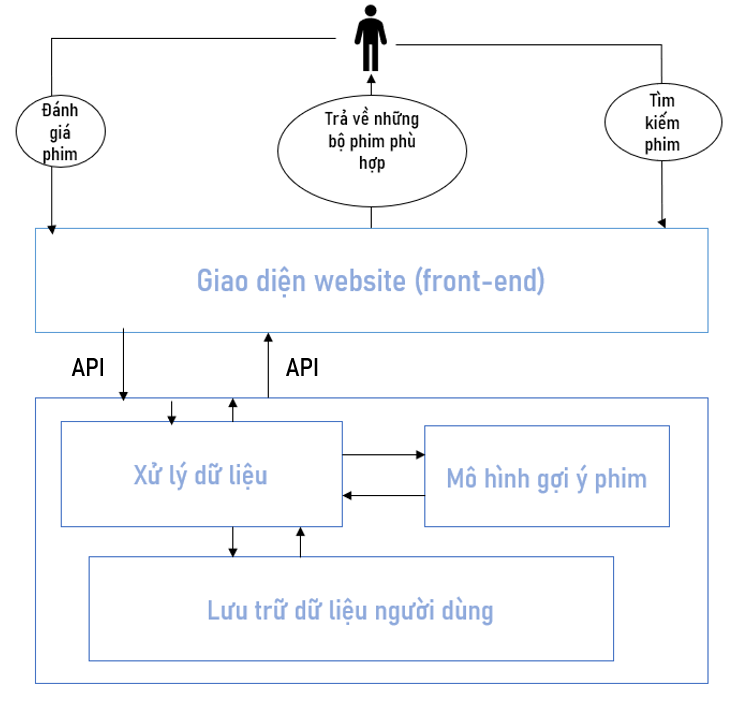
Khi mà chúng ta muốn gửi dữ liệu từ client lên server chúng ta chỉ cần set phương thức cho thư viện axios là .post cùng với paramater mà chúng ta muốn gửi đi.

******

Khi front end gửi đi 1 request thì phía server cũng ngay lập tức nhận được 1 request với nội dung tương ững mà phía front đã gừi.



***Hình 25: workflow của hệ thống sử dụng Django REST Framework và ReactJS***

******

***Hình 26: Sơ dồ hoạt động của toàn bộ hệ thống***

# ***Chương 6***

# **Thử nghiệm đánh giá chất lượng hệ thống.**

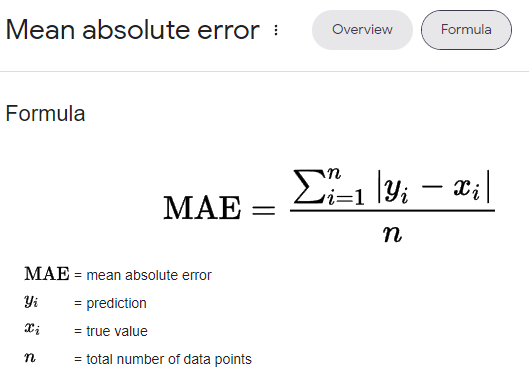
## 6.1 Đánh giá sự chính xác của hệ thống

### ***6.1.1 Đánh giá độ mất mát của hệ thống***

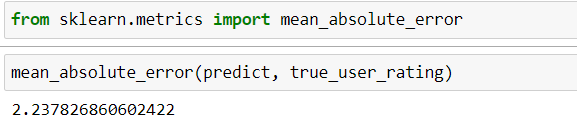
Đối với cả 2 phương pháp content-base filtering và colaborative filtering đểu là những phương pháp lọc nên rất khó đưa ra sự đánh giá về độ chính xác của thuật toán bởi vì cả 2 đều đưa ra nhiều kết quả khác nhau và không có chuẩn chung để đánh giá là đúng hay sai.

Đối với phương pháp collaborative thì chúng ta đưa ra 1 hàm đánh giá sự mất mát đối với các bộ phim như sau:

Sử dụng hàm MEA: Mean absolute error để đánh giá sự mất mát của ma trận



***Hình 27: công thức đánh giá sự mất mát của hệ thống***



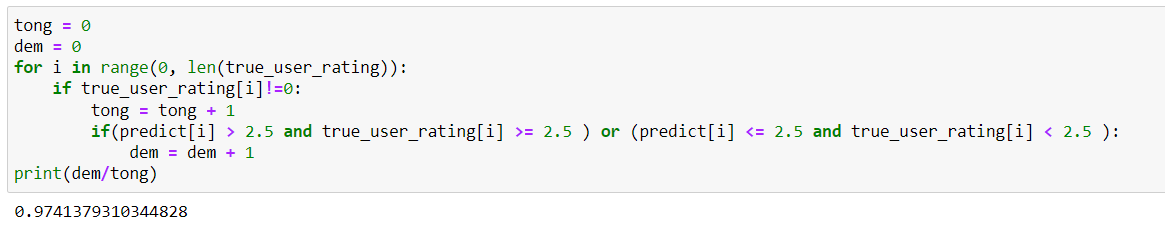
*Predict là các rating mà hệ thống đã đánh giá cho các user*

*True\_user\_rating là các đánh giá thực của các user*

### **6.1.2 Đánh giá độ chính xác của hệ thống collaborative filtering**

Vì bài toán là bài toán gợi ý dựa trên việc lọc nên chúng ta sẽ viết 1 hàm kiểm tra độ chính xác của hệ thống.

Chúng ta quy định với những bộ phim mà người dùng đánh giá < 2.5 là người dùng không thích bộ phim và ngược lại nếu >= 2.5 thì người dùng thích bộ phim chúng ta đưa bài toán và bài toán nhị phân thích và không thích và xây dựng hàm tính độ chính xác



Với cách tính này thì độ chính xác của hệ thống tính được lên đến 97%

Công thức tính độ chính xác như sau:

Precison =

Trong đó

True positive là trường hợp bộ phim được người dùng đánh giá trên 2.5 sao và hệ thống cũng dự đoán đúng như vậy

True Negative là trường hợp bộ phim được người dùng đánh giá dưới 2.5 sao và hệ thống cũng đã dự đoán đúng

False positive là trường hợp bộ phim được người dùng đánh giá dưới 2.5 sao và hệ thống dự đoán người dùng sẽ đánh giá trên 2.5 sao

False Negative là trường hợp bộ phim được người dùng đánh giá trên 2.5 sao và hệ thống dự đoán người dùng sẽ đánh giá dưới sao

## 6.2 Đánh giá sự ổn định của hệ thống

Đẻ đánh giá sự ổn định của hệ thống chúng ta cho hệ thống chạy và tương tác với đồng thời 20 user 1 lúc.

Hệ thống sau khi chạy với 20 user cùng 1 lúc và sau 5 lần thử nghiệm hệ thống đảm bảo chạy ổn định trong cả 5 lần với 20 user cùng 1 lúc không xảy ra bất ký sự cố nào không mong muốn.

# **Kết luận và Định hướng nghiên cứu trong tương lai**

Trong những năm gần đây xảy ra đại dịch lớn covid trên toàn thế giới cũng kiến nhu cầu giải trí của con người dần chuyển dịch hướng online internet thay vì ra rạp xem phim thì mọi người lựa chọn xem phim ở nhà thông qua các ứng dụng, website và đây cũng là 1 cơ hội lớn để các app và các website giải trí phát triển đồng thời cũng là 1 sự cạnh tranh giữa các ứng dụng đòi hỏi phải đưa ra được trải nghiệm tốt nhất cho người dùng, với sự phát triển của công nghệ, lượng dữ liệu về phim được thu thập trong nhiều năm, vì vậy nghiên cứu tập trung vào việc xây dựng các hệ thống gợi ý phim là 1 vấn đề tiên quyết để tăng tối ưu trải nghiệm và thu hút người dùng.

Việc xây dựng hệ thống gợi ý phim giúp cải thiện thời gian tìm kiếm phim của người dùng, tăng trải nghiệm sử dụng dịch vụ.

Đồng thời đây cũng là yếu tố giúp các công ty, tổ chức cung cấp dịch vụ phim ảnh có thể dựa vào hệ thống có thể giữ chân, thu hút khách hàng mới và cạnh tranh với các công ty, tổ chức trong cùng lĩnh vực.

Qua đây chúng ta thấy việc tạo ra 1 hệ thống gợi ý sản phẩm đang là một nhu cầu gần như là thiết yếu đối với mọi ngành nghề, lĩnh vực không riêng gì với ngành phim ảnh.

Kết quả thực nghiệm cho thấy nhóm chỉ đưa ra 1 thuật toán đơn giản tuy nhiên vẫn cho thấy độ chính xác cao cùng với sự ổn định tốt quan nhiều lần thử nghiệm Nhóm thấy rằng có thể đưa vào thực tế sử dụng.

Hệ thống của nhóm phát triển vẫn còn những khuyết điểm cần phải cải thiện thêm. Đầu tiên là thời gian phản hồi vẫn còn chậm, Thuật toán gợi ý còn đơn giản ngoài ra để có được kết quả như vậy cần có 1 bộ dữ liệu lớn và điều này không phải lúc nào cũng dễ dàng. Và nhóm sẽ có thể tiếp tục phát triển thêm ững dụng này trong thời gian sắp thời nhằm tăng khả năng phản hồi của ứng dụng

Chú thích

***Dịch vụ OTT*:**  ([tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh): **Over-the-top**, nghĩa đen: *lên trên đầu*) là thuật ngữ để chỉ dịch vụ [truyền thông](https://vi.wikipedia.org/wiki/Truy%E1%BB%81n_th%C3%B4ng) được cung cấp [trực tiếp đến người xem](https://vi.wikipedia.org/wiki/Doanh_nghi%E1%BB%87p_v%E1%BB%9Bi_Kh%C3%A1ch_h%C3%A0ng) thông qua [Internet](https://vi.wikipedia.org/wiki/Internet). OTT vượt qua các nền tảng [truyền hình cáp](https://vi.wikipedia.org/wiki/Truy%E1%BB%81n_h%C3%ACnh_c%C3%A1p), [mặt đất](https://vi.wikipedia.org/wiki/Truy%E1%BB%81n_h%C3%ACnh_m%E1%BA%B7t_%C4%91%E1%BA%A5t) và [vệ tinh](https://vi.wikipedia.org/wiki/Truy%E1%BB%81n_h%C3%ACnh_v%E1%BB%87_tinh), những loại hình công ty vồn đóng vai trò là nhà kiểm soát hoặc nhà phân phối nội dung.*OTT* cũng từng được dùng để chỉ những chiếc điện thoại không có nhà cung cấp dịch vụ, nơi mà tất cả giao tiếp liên lạc được tính phí dưới dạng dữ liệu, giúp phòng tránh [sự cạnh tranh độc quyền](https://vi.wikipedia.org/wiki/C%E1%BA%A1nh_tranh_%C4%91%E1%BB%99c_quy%E1%BB%81n), hay những ứng dụng điện thoại truyền dữ liệu theo cách này, bao gồm cả những ứng dụng thay thế phương thức gọi khácvà những ứng dụng cập nhật phần mềm

API, viết tắt của Application Programming Interfaces, là giao diện phần mềm với phần mềm. Có nghĩa là, chúng cho phép các ứng dụng khác nhau nói chuyện với nhau và trao đổi thông tin hoặc chức năng. Điều này cho phép các doanh nghiệp truy cập vào dữ liệu, đoạn mã, phần mềm hoặc dịch vụ của một doanh nghiệp khác để mở rộng chức năng của sản phẩm của chính họ - tất cả đều tiết kiệm thời gian và tiền bạc.

Tài liệu tham khảo:

<https://www.youtube.com/watch?v=1xtrIEwY_zY&list=LL&index=5&t=7355s&ab_channel=CampusX>

<https://www.youtube.com/watch?v=3ecNC-So0r4&list=LL&index=6&ab_channel=CodeHeroku>

<https://labelyourdata.com/articles/movie-recommendation-with-machine-learning#:~:text=A%20movie%20recommendation%20system%2C%20or,their%20past%20choices%20and%20behavior>.

<https://developers.google.com/machine-learning/recommendation/labs/movie-rec-programming-exercise>