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**

**Đề tài: Hệ thống gợi ý phim dựa trên những bộ phim liên quan và sự đánh giá những bộ phim của người dùng.**

*Phần chữ màu vàng là phần bổ sung, sửa đổi của nhóm theo sau sự nhận xét của thầy*

1. ***Đánh giá tổng quan tình hình nghiên cứu***

***Ngoài nước*** (*Phân tích đánh giá được những công trình nghiên cứu có liên quan với những kết quả nghiên cứu mới nhất trong lĩnh vực nghiên cứu của nhiệm vụ; nêu được những bước tiến về trình độ KHCN của những kết quả nghiên cứu đó)*

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.

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.

* Netflix: Nền tảng xem film trực tuyến được thành lập ngày 29/08/1997, ở Scotts Valley, California, United States. ứng dụng gợi ý những bộ film phù hợp với người dùng dựa trên phương pháp collaborative filtering ứng dụng thư viện tensorflow.
* Youtube: Nền tảng cho phép chia sẻ video, xem video trực tuyến có trụ sở ở San Bruno, California, United States và được công bố ngày 14/02/2005 bởi Steve Chen, Chad Hurley, và Jawed Karim. Ứng dụng gợi ý video cho người dùng dựa trên mạng học sâu DNN.

***Trong nước*** *(Phân tích, đánh giá tình hình nghiên cứu trong nước thuộc lĩnh vực nghiên cứu của nhiệm vụ, đặc biệt phải nêu cụ thể được những kết quả KHCN liên quan đến nhiệm vụ mà các cán bộ tham gia nhiệm vụ đã thực hiện)*

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 do ít công ty phát triển lĩnh vực giải trí này. Điển hình như là FPT Play (đây là ứng dụng dịch vụ truyền hình Internet đa dạng nội dung trong đó có cung cấp những bộ phim), VIEON.

***Luận giải về việc đặt ra mục tiêu và những nội dung cần nghiên cứu của nhiệm vụ***

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.

Mục tiêu cụ thể chúng ta đề ra trong bài toán toán:

* Nghiên cứu xây dựng thiết kế một hệ thống hoàn chỉnh cho phép người dùng tương tác với hệ thống như là tìm kiếm phim, đánh giá phim và lấy các gợi ý về các bộ phim dựa trên thông tin đánh giá của người dùng đối với các bộ phim khác và thông tin đánh giá của các người dùng khác để tăng tối đa trải ngiệm của người dùng.
* Hệ thống này sẽ được viết dựa trên nền website và bao gồm các phần như giao diện người dùng, hệ thống lưu trữ và xử lý dữ liệu.
* Hệ thống sẽ được tích hợp các module dùng để đưa ra các bộ phim phù hợp nhất với người dùng

***Liệt kê danh mục các công trình nghiên cứu, tài liệu có liên quan đến nhiệm vụ đã trích dẫn khi đánh giá tổng quan***

Các tài liệu danh mục 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>

1. ***Nội dung nghiên cứu khoa học và triển khai thực nghiệm của nhiệm vụ và phương án thực hiện***

***Các nội dung chính cần thực hiện của đề tài.***

Để thực hiện các nhiệm vụ của đề tài, những nọi dung nghiên cứu và triển khai thực nghiệm phù hợp cần thực hiện như sau:

* 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.
* Nội dung 3: Nghiên cứu xây dựng, thiết kế mô hình gợi ý phim dựa trên bộ dữ liệu ở trên.
* 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).
* 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).
* Nội dung 6: Thử nghiệm đánh giá chất lượng hệ thống.

***Chi tiết cụ thê các nội dung cần thực hiện sẽ được làm rõ dưới 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***

Để xây dựng được một thống gợi ý phim hiệu quả, phù hợp cho người dùng thì chúng ta cần từng bước nghiên cứu, xây dựng, thực hiện một số công việc cốt lõi sau đây: 1) hệ thống lưu trữ dữ liệu có trước (data ban đầu) dưới dạng các bảng .csv (Cố định và chưa thể scale data) giúp quản lý lượng lớn dữ liệu và trích xuất xử lý nhanh chóng trên tập dữ liệu này. 2) Hệ thống lưu trữ thông tin của người khi đăng nhập vào hệ thống giúp lưu trữ và xử lý, phân tích thông tin người dùng để đưa ra gợi ý. 3) hệ thống giúp xử lý, phân tích các loại dữ liệu sử dụng các mô hình, kỹ thuật ứng dụng trí tuệ nhân tạo để đưa ra kết quả là các bộ phim phù hợp nhất với người dùng. Ngoài ra việc tích hợp kết hợp giữa các hệ thống riêng biệt để đưa ra được một hệ thống hợp nhất cũng là một công việc cần được nghiên cứu thực hiện chuyên sau. Các công việc cần thiết cần được thực hiện trong nội dung này là:

* ***Công việc 1.1:*** Nghiên cứu tổng quan hệ thống, cách lưu trữ, quản lý truy xuất và cách xử lý các bảng dữ liệu ban đầu (một lượng lớn data có sẵn và chưa thể scale data)
* ***Công việc 1.2:*** Nghiên cứu tổng quan các hệ thống lưu trữ dữ liệu để lưu trữ, truy xuất, xử lý dữ liệu của người dùng.
* ***Công việc 1.3:*** Nghiên cứu tổng quan về các module, mô hình ký thuật gợi ý recommend có thể sử dụng được trong hệ thống.
* ***Công việc 1.4:*** Nghiên cứu tổng quan về việctích hợp các hệ thống xử lý, lưu trữ dữ liệu với các mô hình recommend. Cách tạo và liên kết dữ liệu với giao diện người dùng.
* ***Nội dung 2: Tìm kiếm, đánh giá và phân tích bộ dữ liệu để phục vụ cho hệ thống gợi ý phim.***

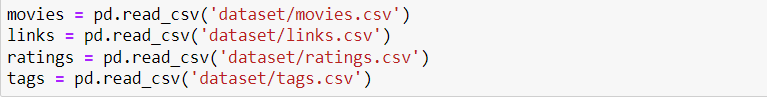
Để có thể hoàn thiện được một mô hình, hệ thống gợi phim hiệu quả, phù hợp cho người dùng trước hết chúng ta cần phải có 1 bộ dữ liệu phù hợp cho hệ thống.

* ***Công việc 2.1:*** Tìm kiếm dữ liệu phù hợp cho hệ thống

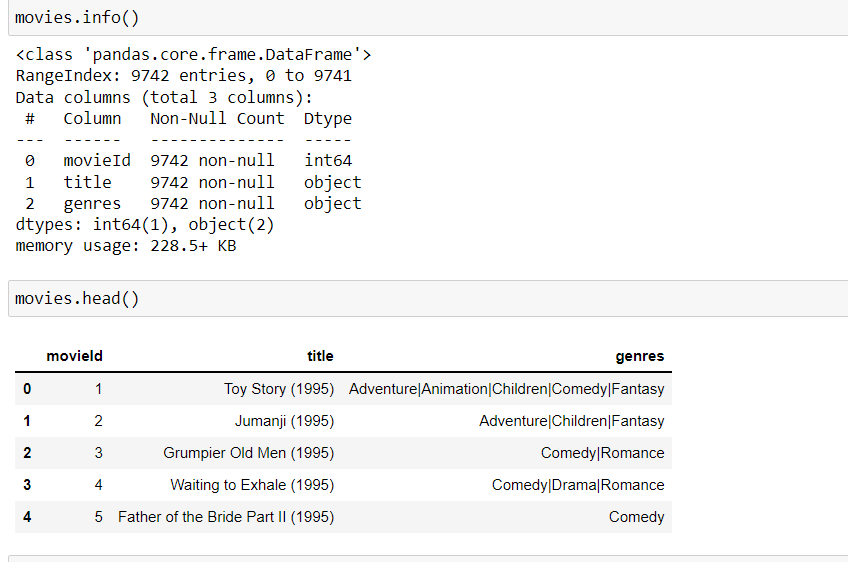
Chúng ta đã tìm được 1 bộ dữ liệu phù hợp cho hệ thống gợi ý phim

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

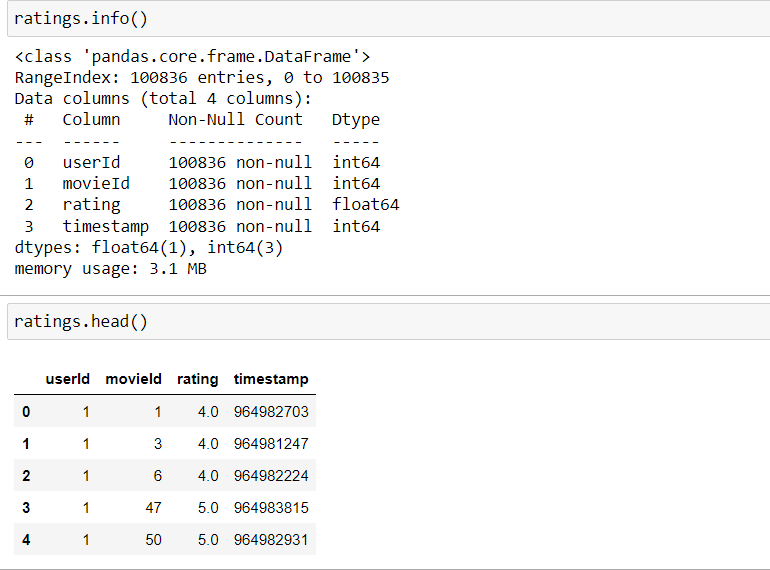
Đẻ phát triền hệ thống gợi ý phim này chùng ta lựa chọn sử dụng bộ dữ liệu về hơn 9000 bộ phim của movieLens cùng với lượng lớn đánh giá của các người dùng thử nghiệm khác mà họ đã thu thập.



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



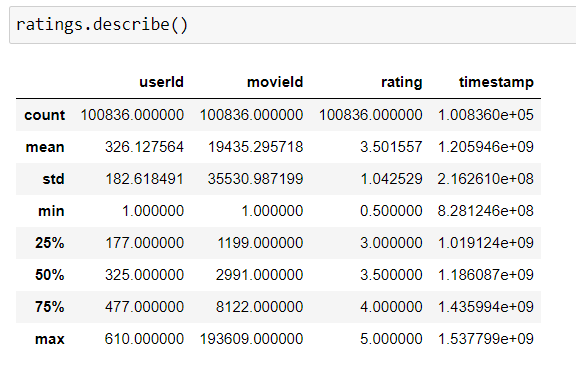
File movies.csv gồm 9742 bản ghi chứa thông tin của các bộ film như là id, tên, mô tả.



File ratings.csv chứa thông tin về các lượt đánh giá của các người dùng trước cho các bộ phim.

Kết quả của công việc 2.1 là bộ dữ liệu thu thập được

* ***Công việc 2.2:*** đánh giá, phân tích bộ dữ liệu.

Đánh giá và phân tích bộ dữ liệu: 

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.

Kết quả của công việc 2.2 là có cái nhìn tổng quan về bộ dữ liệu này để từ đó đưa ra những quyết định tiếp theo.

*Kết quả của nội dung 2: 1 báo cáo công việc, 1 tập dữ liệu đã được làm sạch*

* ***Nội dung 3: Nghiên cứu xây dựng, thiết kế mô hình gợi ý phim dựa trên bộ dữ liệu ở trên.***
* ***Công việc 3.1:*** Lựa chọn các đặc trưng thích hợp của các bộ dữ liệu để đưa vào mô hình triển khai.

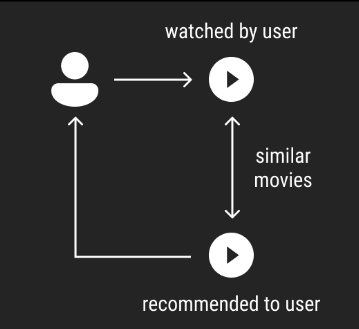
Từ những bộ dữ liệu đã thu thập được ở trện cùng với những đánh giá phân tích thì chúng ta cần lựa chọn ra những thuộc tính đặc trưng phù hợp với hệ thống để cho ra được 1 mô hình gợi ý phim hiệu quả.

VD: như đối với tập dữ liệu trên thì chúng ta có thể lấy ra được các cột các thuộc tính các đặc trưng cần thiết như là title, genres của phim, số đánh giá của người dùng về bộ phim, giá trị mà người dùng đánh giá về bộ phim. Kết quả của công việc này là bộ các thuộc tính cần để đưa vào hẻ thống.

* ***Công việc 3.2:*** Lựa chọn xây dựng mô hình kỹ thuật gợi ý phim cho người dùng.

Hiện nay có rất nhiều phương pháp dùng để gợi ý đề xuất phim đang được sử dụng dựa trên ững dụng các thuật toán học máy và phổ biến nhất là 2 phương pháp Content-based filtering và Collaborative-filtering.

***Phương pháp Content-based filtering****:*



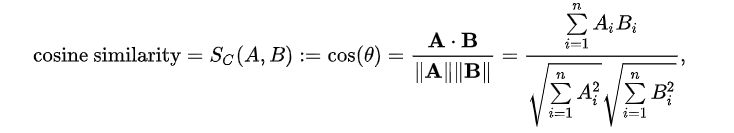
Đối với phương pháp này thì chúng ta chỉ đánh giá các đặc tính của các items ở đây là các bộ phim và so sánh sự tương dồng giữa các bộ phim với nhau. Khi 1 người xem 1 bộ phim nào đấy thì sẽ đưa ra kết quả là các bộ phim tương đồng nhất với những bộ phim mà người này đã xem.

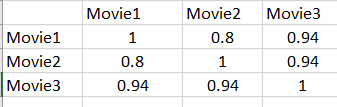
Hệ thống sẽ tiến hành tạo ra các profile cho bộ phim là các feature vectors đặc trưng cho các bộ phim đó sau đó tinh độ tương đồng của các bộ phim.

Ví dụ có ba bộ phim là Movie1 [‘Hi Hello Hi’], Movie2 [‘Hello Hi Hello’], Movie3 [‘Hi Hello’]

Ta có feature của 3 bộ phim lần lượt là Movie1 ( 2 1), Movie2 (1 2), Movie3 ( 1 1). Là số lần xuất hiện các từ trong mỗi bộ phim.

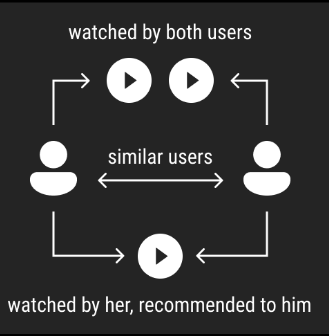
Sử dụng hàm cosine similarity để tính khoảng cách giữa các vector đặc trưng của các bộ phim, khoảng cách càng lớn thì độ tương đồng càng lớn.



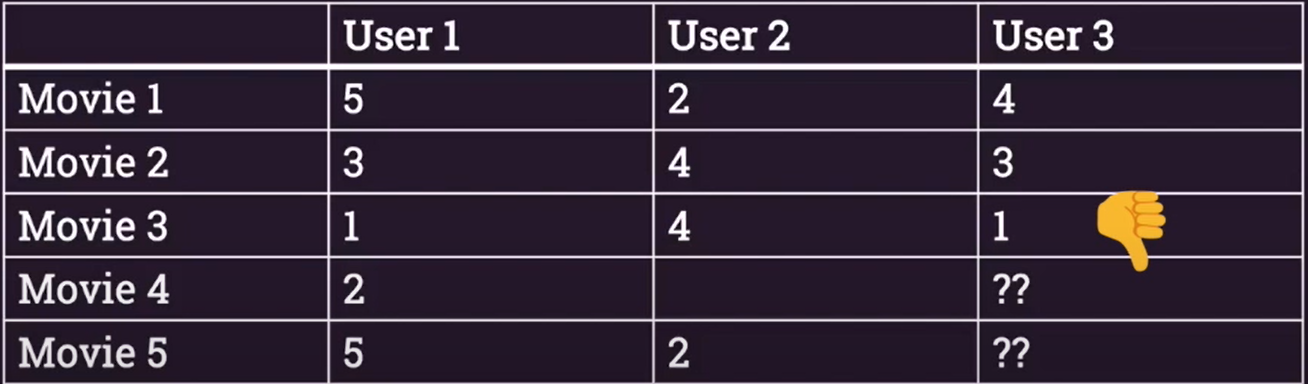


Đầu ra, kết quả của phương pháp này là các bộ phim có sự tương đồng lớn nhất về nội dung như là tên, thể loại với bộ phim mà người dùng đã tương tác tìm kiếm.

**Phương pháp Collaborative filtering:**

****

Đối với phương pháp này chúng ta cần thông tin đánh giá của các người dùng khác với các bộ phim trong hệ thống.



Đối với các ô trống là những bộ phim chưa được đánh giá của các user và chúng ta cần điền vào để đánh giá xem bộ phim đấy có phù hợp với người dùng đấy không.

Có 2 cách tiếp cận đối với phương pháp này là User-based collaborarive filtering tìm các user tương đồng với user cần gợi ý và gợi ý những bộ phim mà người dùng trước đấy đã xem hoặc đã thích và Item-based collaborative filtering tìm kiếm các bộ phim tương tự mà người dùng nhắm mục tiêu đến

Kết quả thu được trong công việc này là module gợi ý phim tương ứng với tương tác của người dùng.

Đầu ra, kết quả của phương pháp này theo cả 2 cách tiếp cận là những bộ phim được đánh giá cao bởi những người dùng khác có sự tương đồng với người dùng cần gợi ý hoặc là những bộ phim có sự tương đồng với bộ phim được người dùng đánh giá cao. Người dùng tương đồng là những người dùng có sự đánh giá gần giống nhau đối với 1 số bộ phim nhất định,

*Kết quả của nội dung 3: 1 báo cáo và 1 phần mềm là module đã thiết kế được*

* ***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)***
* ***Công việc 4.1:*** Nghiên cứu các cách, các phần mềm quản lý dữ liệu người dùng và quản lý dữ liệu dataset của hệ thống và trả về dữ liệu cho phần hiển thị.

Phần này được xem là trung tâm kết nối của hệ thống, nó bao gồm các chức năng chính như sau:

* Lữu trữ dữ liệu của người dùng
* Liên kết module gợi ý vào hệ thống
* Kết nối với phần giao diện người dùng để lấy dữ liệu và xử lý trả về kết quả cho người dùng

Cần lựa chọn 1 phần mềm có thể lưu trữ được lượng lớn đữ liệu, truy nhập và xử lý dữ liệu 1 cách nhanh chóng.

* ***Công việc 4.2:*** Nghiên cứu tích hợp module vào hệ thống.

Trong phần này chúng ta cần tích hợp module gợi ý đã xây dựng ở trên vào hệ thống xử lý chung để có thể đưa ra được gợi ý về các bộ phim cho người dùng.

Để có thế tích hợp mô hình vào hệ thống 1 cách dễ dàng chúng ta cần lựa chọn các ngôn ngữ lập trình thích hợp. Trong dự án này chúng ta sẽ sử dụng ngôn ngữ lập trình python cùng với các thư viện như là sklearn.. và framework để tạo ứng dụng là django.

*Kết quả của nội dung 4: 1 báo cáo, 1 phần mềm hoàn chỉnh đã được tích hợp module gợi ý phim.*

* ***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).***
* ***Công việc 5.1:*** Nghiên cứu xây dựng hệ thống phần mềm để người dùng tương tác với hệ thống và hiển thị kết quả cho người dùng.

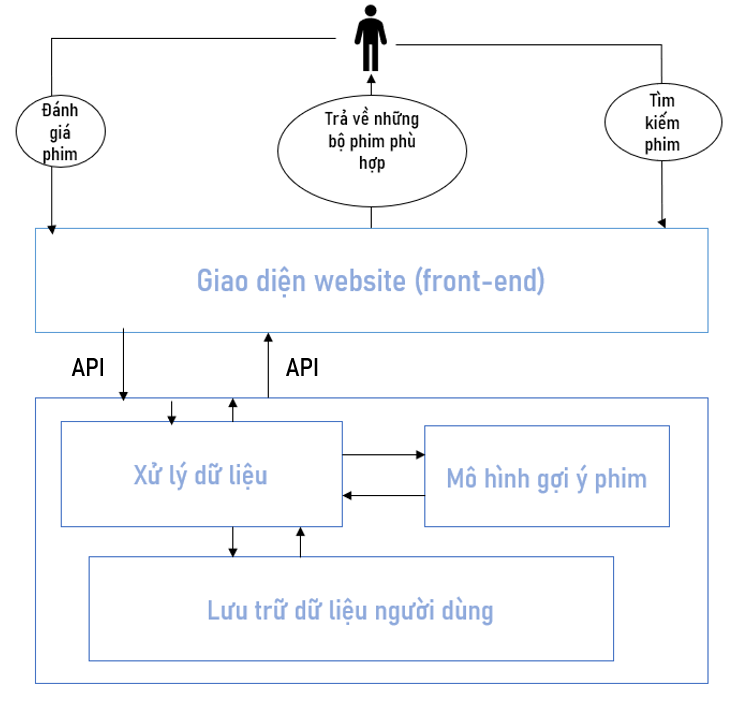
Định hướng chúng ta sẽ xây dựng hệ thống trên phương diện website nên cần thiết kế phần xử lý dữ liệu và phần giao diện riêng. Phần này sẽ cho phép người dùng tương tác và hiển thị dữ liệu cho người dùng.

Phần này cần những thành phần chính sau đây (Chúng ta chưa làm mà mới là ý tưởng nên sẽ liệt kê những thành phần cần có ở đây):

* Phần đăng nhập, đăng ký của người dùng
* Một ô tìm kiếm để người dùng thực hiện tìm kiếm phim mà người dùng muốn xem
* Một phần hiển thị danh sách các bộ phim cho người dùng
* Phần đánh giá của người dùng dối với mỗi bộ phim.
* ***Công việc 5.2:*** Nghiên cứukết nối phần giao diện với phần hệ thống xử lý dữ liệu ở trên.

Như chúng ta đã nói ở trên thì chúng ta sẽ xây dựng ứng dụng trên nền tảng web và có 2 phần riêng biệt là phần xử lý dữ liệu (Back-end) và phần giao diện người dùng (Front-end). Vì vậy chúng ta cần nghiên cứu xây dựng kết nối giữa 2 phần này lại với nhau để tạo ra được 1 trang web hoàn chỉnh.

Và phương pháp gần như là duy nhất ở đây là sử dụng kết nối thông qua API 1 cổng kết nối chung. Khi cần tương tác giữa người dùng và hệ thống có liên quan đến dữ liệu 2 phần này sẽ trao đổi dữ liệu với nhau bằng việc gọi dến API này.

*Kết quả của nội dung 5: 1 báo cáo và 1 hệ thống đã hoàn chỉnh cả phần giao diện và phần xử lý dữ liệu*******

* ***Nội dung 6: Thử nghiệm đánh giá chất lượng hệ thống.***
* ***Công việc 6.1:*** Đánh giá độ chính xác của hệ thống

Để đánh giá sự chính xác của hệ thống chúng ta sẽ dùng công thức sau:



Chuẩn hóa đánh giá của người dùng về giá trị từ -2.5 đến 2.5

True positive: Số bộ phim gợi ý đúng với sự yêu thích của người dùng là những bộ phim dược gợi ý có tổng giá trị của sự tương đồng với các bộ phim mà người dùng đã đánh giá cao là > 0

False positive: Số bộ phim gợi ý không đúng với sự yêu thích của người dùng là những bộ phim đã được gợi ý nhưng có tổng giá trị của sự tương đồng với các bộ phim mà người dùng đã đánh giá cao bởi người dùng < 0

False Negative: Số bộ phim đã được gợi ý nhưng có tổng giá trị của sự tương đồng với các bộ phim mà người dùng đã đánh giá thấp

bởi người dùng > 0

Với độ chính xác Precision mục tiêu là > 80%

Quả trình này cần được thực hiện nhiều lần với nhiều user khác nhau để có đánh giá khách quan nhất. Và quá trình này cũng cần được thực hiện song song với nội dung 3 xây dưng mô hình.

* ***Công việc 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 nhiều user 1 lúc.

Nếu có lỗi thì sẽ tiến hành đánh giá lại từ đầu

1. ***Phân công nhiệm vụ***

***Bảng phân công nhiệm vụ chi tiết cho các thành viên trong nhóm.***

*Các nội dung công việc đã được sắp xếp theo thứ tự lần lượt từ trước đến sau*

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Nội dung | Công việc | Người thực hiện |
| 1 | Nội dung 1: khảo sát các hệ thống | Công việc 1.1 | Cả nhóm |
| Công việc 1.2 | Cả nhóm |
| Công việc 1.3 | Cả nhóm |
| Công việc 1.4 | Lê Khắc Nam |
| 2 | Nội dung 2: Tìm, xử lý dữ liệu | Công việc 2.1 | Phạm Anh Khiêm |
| Công việc 2.2 |
| 3 | Nội dung 3: Nghiên cứu xây dụng mô hình gợi ý phim | Công việc 3.1 | Cả nhóm |
| Công việc 3.2 |
| 4 | Nội dung 4: Nghiên cứu xây dụng hệ thống lưu trữ xử lý dữ liệu | Công việc 4.1 | Lê Khắc Nam |
| Công việc 4.2 |
| 5 | Nội dung 5: nghiên cứu, xây dụng giao diện và kết nối với phần hệ thống xử lý | Công việc 5.1 | Nguyễn Khả Thiết |
| Công việc 5.2 |
| 6 | Nội dung 6: Thử nghiệm đánh giá hệ thống | Công việc 6.1 | Cả nhóm |
| Công việc 6.2 |