**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**---🙠**🕮**🙢---**



**BÁO CÁO ĐỒ ÁN MÔN HỌC**

**Nhập môn Học máy và Khai phá dữ liệu**

**Đề tài: Movie Recommedation System**

Giảng viên hướng dẫn: *PGS. Thân Quang Khoát*

Mã lớp : *131404*

Nhóm sinh viên thực hiện*: Cao Như Đạt – 20194013*

*Nguyễn Huy Hoàng – 20194058*

*Lê Huy Hoàng – 20190053*

MỤC LỤC

[Lời nói đầu 2](#_Toc12840)

[1. Giới thiệu bài toán: Movie Recommendation System 3](#_Toc26055)

[2. Tiến hành trên bộ dữ liệu TMDB 5](#_Toc25507)

[2.1 Tổng quan về bộ dữ liệu TMDB 5](#_Toc8021)

[2.2 Tiền xử lý dữ liệu 5](#_Toc3615)

[2.3 Thuật toán 5](#_Toc22284)

[2.4 Một số đánh giá và kết luận 5](#_Toc2968)

[3. Tiến hành trên bộ dữ liệu MovieLens 5](#_Toc4718)

[4. Kết luận 5](#_Toc22829)

# Lời nói đầu

Hiện nay, hầu khắp trên các ứng dụng hay phương tiện truyền thông chúng ta đều nhận thấy sự xuất hiện của Recommendation System (Hệ thống gợi ý). Khi ta xem một video, YouTube tự động gợi ý các video có nội dung liên quan hoặc có thể tự chuyển tiếp khi ta xem xong. Tương tự với Neflix hay Spotify, những bộ phim hay bản nhạc mà họ gợi ý có ý nghĩa rất lớn trong sự phát triển mạnh mẽ của họ trong thời gian gần đây.

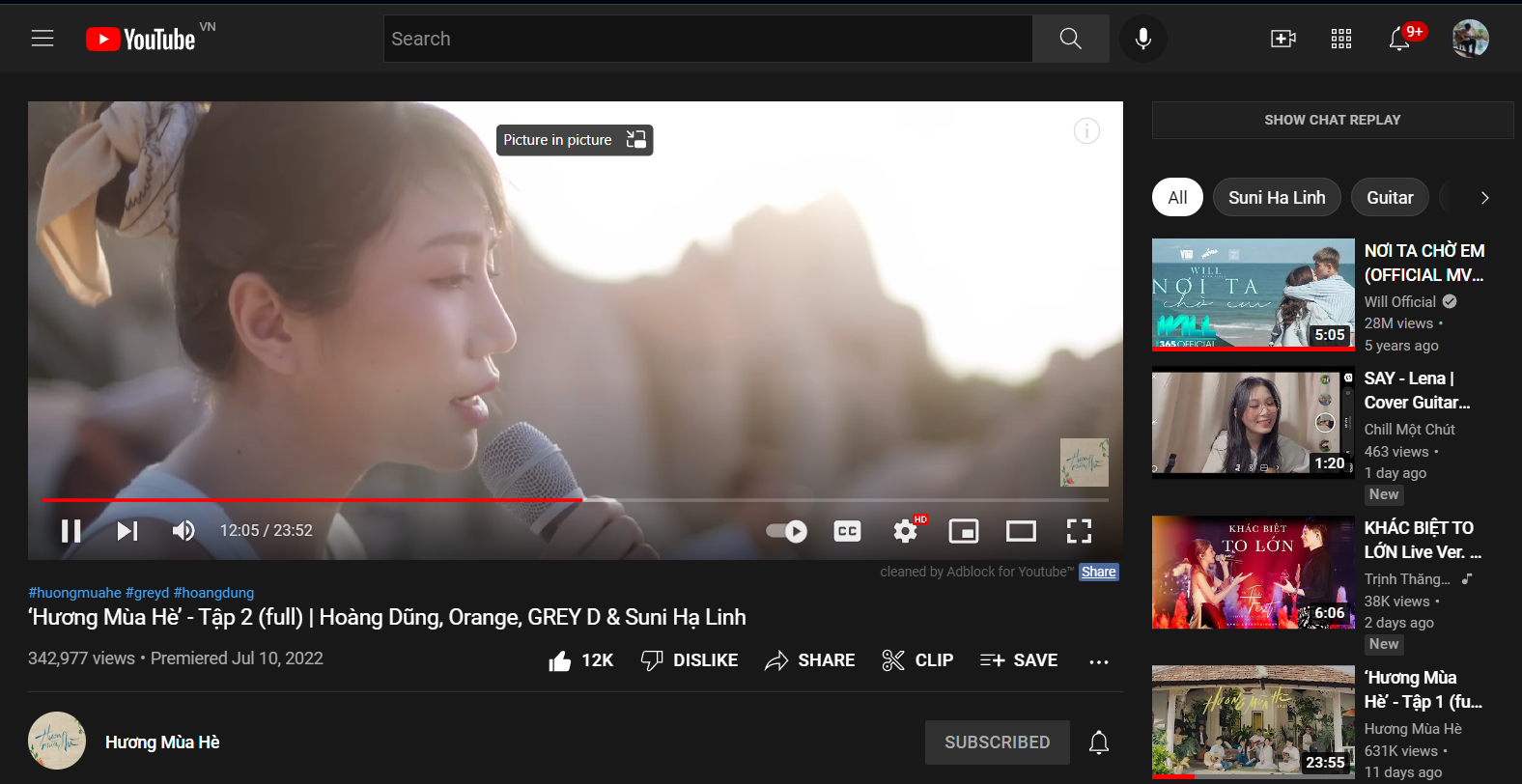
Nhận thấy đây là một đề tài thú vị và có nhiều ý nghĩa trong thực tiễn, nhóm chúng em quyết định lựa chọn đề tài Movie Recommendation System (Hệ thống gợi ý phim) làm đề tài báo cáo của môn học Nhập môn Học máy và Khai phá dữ liệu. Quá trình thực hiện đề tài không thể tránh khỏi thiếu sót, chúng em rất mong nhận được những sự góp ý của thầy!

# Giới thiệu bài toán: Movie Recommendation System

Một *hệ thống gợi ý* (hay cũng có thể gọi là *hệ thống khuyên dùng*, [tiếng anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh" \o "Tiếng 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" \o "Hệ thống lọc thông tin (trang không tồn tại))*, tìm kiếm dự đoán "đánh giá" hoặc "ưa thích" của người dùng với một sản phẩm hoặc đối tượng nào đó.

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.

Youtube tự động đề xuất các video khác liên quan khi người dùng sử dụng:



Hệ thống gợi ý (Recommendation system) dần trở thành một thành phần không thể thiếu của các sản phẩm điện tử có nhiều người dùng. Các sản phẩm cá nhân hóa điện tử ngày càng phổ biến với mục đích mang sản phẩm phù hợp tới người dùng hoặc giúp người dùng có các trải nghiệm tốt hơn. Nếu quảng cáo sản phẩm tới đúng người dùng, khả năng các món hàng được mua nhiều hơn. Nếu gợi ý một video mà người dùng nhiều khả năng thích hoặc gợi ý kết bạn đến đúng đối tượng, họ sẽ ở lại trên nền tảng của bạn lâu hơn. Khi họ ở trên nền tảng của bạn lâu hơn, họ sẽ nhìn thấy nhiều quảng cáo hơn và lợi nhuận từ quảng cáo sẽ lại càng nhiều hơn.

Quảng cáo điện tử ngoài việc giúp các doanh nghiệp bán được nhiều hàng còn giúp họ tiết kiệm được chi phí kho bãi. Họ sẽ không cần các cửa hàng ở vị trí thuận lợi để thu hút khách hàng hay phải trưng ra mọi mặt hàng ở vị trí đắc địa nhất trong cửa hàng. Mọi thứ có thể được cá nhân hóa sao cho mỗi người dùng nhìn thấy những sản phẩm khác nhau phù hợp với nhu cầu và sở thích của họ.

Quảng cáo trên Internet đã chiếm thị phần ngày càng cao so với quảng cáo truyền hình nhờ sự đa dạng và cá nhân hóa một cách tối đa. Một người dùng 20-30 tuổi thường xuyên nghe nhạc rap ít có khả năng thích nhạc Bolero. Một người dùng tìm kiếm các thông tin về xe hơi nhiều khả năng sắp mua xe và quan tâm tới những dịch vụ sửa và rửa xe. Một người dùng thường xuyên xe các video về làm vườn nhiều khả năng sẽ quan tâm tới việc mua bán hạt giống. Từ những thông tin thu thập được từ hành vi người dùng, hệ thống có thể gợi ý ra những lựa chọn phù hợp để đạt được hiệu quả cao nhất.

Các hệ thống gợi ý thường được chia thành ba nhóm chính:

**Lọc Nhân khẩu học (**Demographic Filtering):**** Đưa ra các đề xuất tổng quát cho mọi người dùng, dựa trên mức độ phổ biến và / hoặc thể loại phim. Hệ thống đề xuất những bộ phim tương tự cho những người dùng có các đặc điểm nhân khẩu học tương tự. Vì mỗi người dùng khác nhau nên cách làm này được coi là quá đơn giản. Ý tưởng cơ bản đằng sau hệ thống này là những bộ phim nổi tiếng hơn và được giới phê bình đánh giá cao hơn sẽ có xác suất được khán giả bình thường thích cao hơn.

****Lọc dựa trên nội dung (Content-based Filtering)****: nhóm thuật toán này gợi ý cho người dùng những sản phẩm tương tự như những sản phẩm mà người dùng đã có phản hồi tích cực. Hệ thống này cần xây dựng đặc trưng cho các sản phẩm sao cho những sản phẩm tương tự nhau có khoảng cách tới nhau nhỏ. Việc này khá tương tự như việc xây dựng các embedding cho các sản phẩm. Việc dự đoán cho mỗi người dùng hoàn toàn chỉ dựa trên lịch sử thông tin của người dùng đó.

****Lọc cộng tác (Collaborative Filtering)****: nhóm thuật toán này không chỉ dựa trên thông tin về sản phẩm tương tự mà còn dựa trên hành vi của những người dùng tương tự. Ví dụ: người dùng A, B, C đều thích các bài hát của Noo Phước Thịnh. Ngoài ra, hệ thống biết rằng B, C cũng thích các bài hát của Bích Phương nhưng chưa có thông tin về việc liệu user A có thích Bích Phương hay không. Dựa trên thông tin của những người dùng tương tự là B và C, hệ thống có thể dự đoán rằng A cũng thích Bích Phương và gợi ý các bài hát của ca sĩ này tới A.

Trong phạm vi nội dung của bài báo cáo, nhóm chúng em chỉ tập trung vào hệ thống gợi ý sử dụng phương pháp Lọc dựa trên nội dung (Content-based Filtering).

# Tiến hành trên bộ dữ liệu TMDB

## Tổng quan về bộ dữ liệu TMDB

## Tiền xử lý dữ liệu

## Thuật toán

## Một số đánh giá và kết luận

# Tiến hành trên bộ dữ liệu MovieLens

# Kết luận