

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN



PROPOSAL

KHOA HỌC DỮ LIỆU ỨNG DỤNG:

HỆ THỐNG GỢI Ý PHIM

TEAM

PNHT

Thành phố Hồ Chí Minh - 2024

Mục lục

0	Quá trình làm việc (18/03 - 24/03)	3
1	Giới thiệu đề tài	4
1.1	Hệ thống gợi ý (Riêng)	4
1.2	Hệ thống gợi ý (Chung)	5
2	Thu thập dữ liệu	6
2.1	Ý tưởng Crawl Data	6
2.2	Thực hiện Crawl Data	8
2.2.1	Thu nhập những ratings của user cho movie:	8
2.2.2	Thu nhập các thông tin về movie và user:	12
3	Khám phá & Tiền xử lý dữ liệu.....	14
4	Trực quan hóa dữ liệu	15
5	Xây dựng mô hình.....	16
6	Đánh giá mô hình	17

Phần 0: Quá trình làm việc (18/03 - 24/03)

- 18/03: Sau buổi báo cáo đầu tiên thì nhóm đã quyết định chọn lại chủ đề, và chủ đề được chọn là "Hệ thống gợi ý phim"
- 19/03 - 23/03: Hình thành bước đặt câu hỏi và mô tả bài toán cùng với đó là thực hiện bước crawl data
- 24/03: Họp Meeting để thảo luận để thực hiện báo cáo

Phần 1: Giới thiệu đề tài

Thử nghiệm xây dựng một hệ thống gợi ý gồm nhiều hạng mục gợi ý khác nhau như:

- Có thể bạn sẽ thích
- Nội dung phim tương tự
- Phim Hot
- ...

⇒ Có thể thực hiện gợi ý cho các thể loại users khác nhau (Riêng cá nhân), cụ thể là

- **User cũ**: Thực hiện dự đoán Rating theo phương pháp Collaborative Filtering, hoặc dựa Content-based Filtering
 - Content-based Filtering: Nội dung phim tương tự
 - Collaborative Filtering: Có thể bạn sẽ thích
- **Guest**: Cho thực hiện Survey để khảo sát, từ đó đưa ra top film theo Survey của Guest
- **User mới**:
 - Nếu chưa có thông tin gì thì thực hiện như Guest
 - Nếu có ít thông tin: (Chưa rõ)
 - Nếu đã có nhiều thông tin thì thực hiện như User cũ

⇒ Ngoài ra, có thể xây dựng một hệ thống gợi ý (Chung) cho toàn bộ users. (Phim Hot)

[1]: Hệ thống gợi ý (Riêng)

Bài toán đặt ra:

- **Content-based Filtering (Lọc dựa trên nội dung)**: Tìm hiểu các thuộc tính (genres, actor, director,...) của bộ phim mà user cho rating cao và đưa ra bộ phim tương tự

⇒ Xây dựng model để dự đoán movie_id mà user yêu thích hoặc viết hàm để Filter

- **Collaborative Filtering (Lọc cộng tác)**: Dựa trên sự tương tác của user với item để thực hiện dự đoán rating của user.
 - Ví dụ như 2 user này cùng coi 1 bộ phim cho rating cao như nhau, mà user kia coi 1 bộ phim khác cũng cho rating cao thì gợi ý cho user còn lại bộ phim khác đó

⇒ Xây dựng model để dự đoán rating của user cho các movie và lấy top

- **Survey**: Cho user(Guest) chọn các thuộc tính (genres, actor, director,...) để đưa ra top film theo các thuộc tính đó

⇒ Viết hàm để Filter

[2]: Hệ thống gợi ý (Chung)

Bài toán đặt ra:

- Phân tích toàn bộ dữ liệu để tìm ra những bộ phim hot
→ Ví dụ như lấy phim có genre, actor, director, ... nằm trong top các phim có rating cao (Chưa rõ)

⇒ Phân tích + Trực quan hóa dữ liệu

Phần 2: Thu thập dữ liệu

[1]: Ý tưởng Crawl Data

Thu nhập những ratings của user cho movie:

- 1.) Đầu tiên thì thu thập 250 movie_id của top dựa trên trang chủ:
→ https://www.imdb.com/chart/top/?ref_=nv_mv_250
- 2.) Sau đó thì dùng movie_id để crawl toàn bộ ratings của toàn bộ user trong movie đó, và lưu lại thông tin user_id:
→ https://www.imdb.com/title/movie_id/reviews/?ref_=tt_ql_2
- 3.) Dùng user_id để crawl toàn bộ ratings của user đó cho toàn bộ movie mà user đã đánh giá, và lưu lại thông tin movie_id:
→ https://www.imdb.com/user/user_id/?ref_=tt_urv
⇒ 2 quá trình (2) và (3) được thực hiện xen lẫn và thực hiện vét cạn đến khi có đủ thông tin cần thiết

Thu nhập các thông tin về movie và user:

- Trang thông tin Movie:

→ Data Schema:

Keyword	Mô tả	Loại dữ liệu
movie_id	Id movie	String
title	Movie name	String
runtime	Tổng thời gian bộ phim	String
rating	Số lượng đánh giá và rating trung bình	Int & Float
award	Số chiến thắng và giải thưởng của phim	Int & Int
genre	Thể loại phim	List string
releaseDate	Ngày/tháng/năm ra mắt	Datetime
releaseLocation	Quốc gia sản xuất	String
actors	Các diễn viên chính	List String
director	Đạo diễn phim	List String

→ Đường Link:

https://www.imdb.com/title/movie_id/?ref_=chttp_t_1

- Trang thông tin User:

→ Data Schema:

Keyword	Mô tả	Loại dữ liệu
user_id	Id User	String
user_name	User name	String
joined_year	Năm tham gia	Int

→ Đường Link:


https://www.imdb.com/user/user_id/?ref_=tt_urv

[2]: Thực hiện Crawl Data

(1). Thu nhập những ratings của user cho movie:

- Đối với trang Movie_id:

Thu thập dữ liệu từ mỗi trang (Đối với mỗi movie_id) như thế này:



The Shawshank Redemption (1994)

User Reviews

[+ Review this title](#)

10,972 Reviews

☐ Hide Spoilers Filter by Rating: Show All Sort by: Featured ↓↑

★ 10/10

Some birds aren't meant to be caged.

[hitchcockthelegend](#) 24 July 2010

Warning: Spoilers

1,014 out of 1,095 found this helpful. Was this review helpful? Yes No

[Report this](#) | [Permalink](#)

★ 10/10

An incredible movie. One that lives with you.

[Sleepin_Dragon](#) 17 February 2021

It is no wonder that the film has such a high rating, it is quite literally breathtaking. What can I say that hasn't said before? Not much, it's the story, the acting, the premise, but most of all, this movie is about how it makes you feel. Sometimes you watch a film, and can't remember it days later, this film loves with you, once you've seen it, you don't forget.

The ultimate story of friendship, of hope, and of life, and overcoming adversity.

I understand why so many class this as the best film of all time, it isn't mine, but I get it. If you haven't seen it, or haven't seen it for some time, you need to watch it, it's amazing. 10/10.

288 out of 312 found this helpful. Was this review helpful? Yes No

[Report this](#) | [Permalink](#)

→ Ta sẽ thu được:

	user_id	movie_id	user_rating
0	ur83822756	tt0050083	10
1	ur0968789	tt0050083	10
2	ur1318549	tt0050083	
3	ur0643062	tt0050083	
4	ur0688559	tt0050083	
...
245	ur17816293	tt0468569	6
246	ur113320171	tt0468569	10
247	ur2488512	tt0468569	9
248	ur49526876	tt0468569	
249	ur3387663	tt0468569	8

- Đối với trang User_id:

Thu thập dữ liệu từ mỗi trang (Đối với mỗi user_id) như thế này:

hitchcockthelegend

Reviews

3,939 Reviews

☐ Hide Spoilers
 Filter by Rating: Show All
Sort by: Review Date

Requiem for a Gringo (1968)

★ 7/10

Logan's Run!
20 September 2020

Once sheared of twenty minutes, "Requiem for a Gringo" is now available to be seen in a full uncut version. Not that it's outrageously violent or sexually repugnant, it would appear some stiff backed suits back in the late 1960's had a bug where the sun doesn't shine.

This is a little treat for fans of Euro-Westerns of the 60's. Plot holds familiar traits, where a ruthless gang of scumbags terrorise locals and kill indiscriminately. Enter a lone stranger, Ross Logan/Django (Lang Jeffries), who after having been dealt a family mortal blow, sets about revenge - good job he is one seriously hard and smart dude!

6 out of 9 found this helpful. Was this review helpful?

Yes
No

[Report this](#) | [Permalink](#)

The Five Pennies (1959)

★ 8/10

Mr. Paradise, I play New Orleans style. You know, it's the newest thing. As a matter of fact I got an arrangement right here of the very number that you're doing.
20 September 2020

The Five Pennies is a musical biopic of jazz great Red Nichols, who is here played by Danny Kaye. As the famed Dixieland cornetist, he runs into opposition to his sound, but breaks through barriers to achieve success. Upon marrying an understanding patient woman (Barbara Bel Geddes) he begins to raise a family. But when tragedy strikes the family, "Red" puts down his horn to focus on matters of the heart.

Out of Paramount, The Five Pennies was released at a time when

2 out of 3 found this helpful. Was this review helpful?

Yes
No

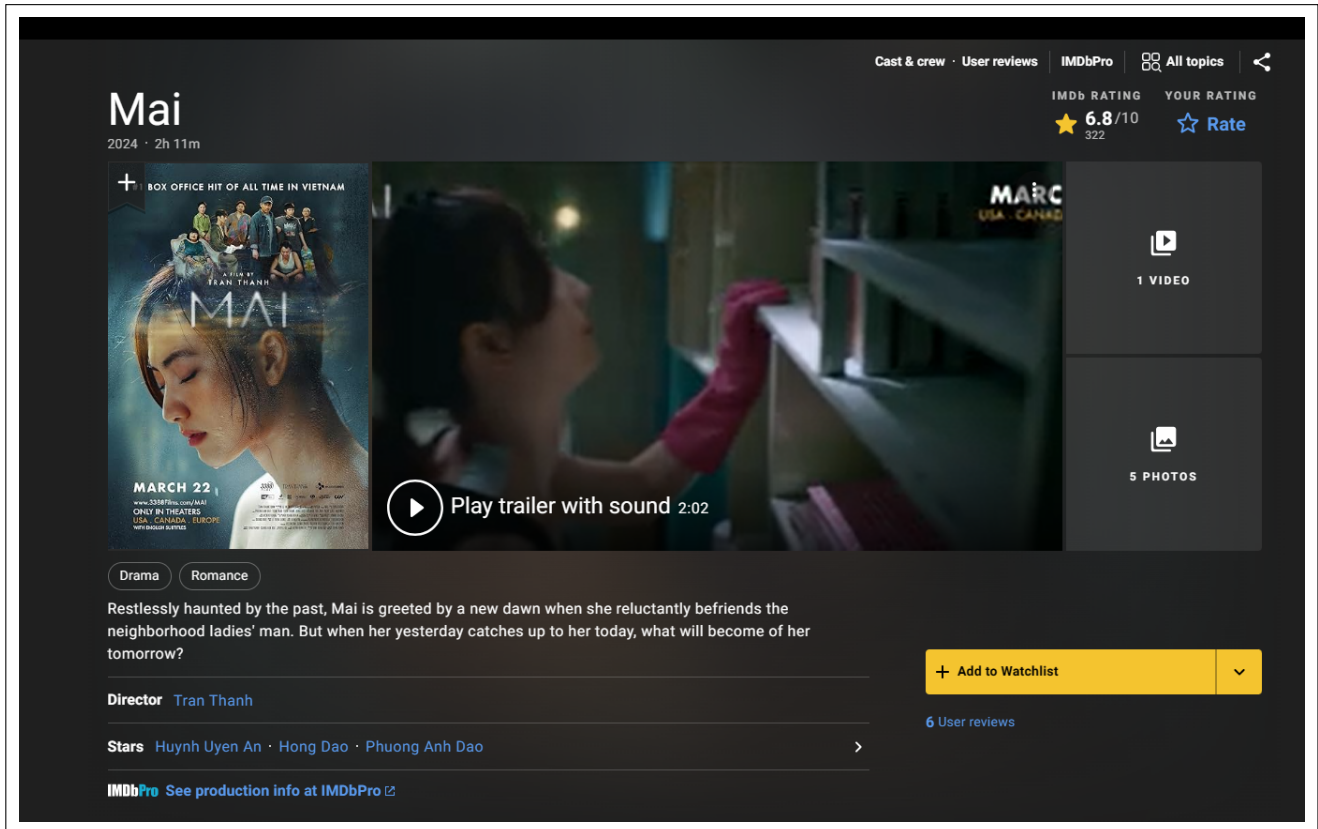
→ Ta sẽ thu được:

	user_id	movie_id	user_rating
0	ur0257957	tt0093145	
1	ur0257957	tt0145505	
2	ur0257957	tt0106017	
3	ur0257957	tt0071870	
4	ur0257957	tt0107756	
...
295	ur16161013	tt1392190	9
296	ur16161013	tt0237572	9
297	ur16161013	tt1392214	9
298	ur16161013	tt2328900	3
299	ur16161013	tt0049833	9

(2). Thu nhập các thông tin về movie và user:

- Trang thông tin Movie:

Thu thập dữ liệu từ api sau: https://imdb-api.huyvongmongmanh75.workers.dev/title/{movie_id}




→ Ta sẽ thu được:

```
{
  "movie_id": "tt31174028",
  "title": "Mai",
  "introduction": "Restlessly haunted by the past, Mai is greeted by a new dawn when she reluctantly befriends the neighborhood ladies' man. But when her yesterday catches up to her",
  "runtime": "2h 11m",
  "rating": {
    "count": 320,
    "star": 6.8
  },
  "award": {
    "wins": 0,
    "nominations": 0
  },
  "genre": [
    "Drama",
    "Romance"
  ],
  "releaseDate": "2024-02-10",
  "releaseLocation": "Vietnam",
  "actors": [
    "Huynh Uyen An",
    "Hong Dao",
    "Phuong Anh Dao"
  ],
  "directors": [
    "Tran Thanh"
  ]
}
```

Trang thông tin User:

Thu thập dữ liệu từ mỗi trang (Đối với mỗi user_id) như thế này:



hitchcockthelegend


IMDb member since July 2007


We all go a little crazy sometimes.


Doc says I'm better now, I'm free to roam in society again :-)


Me? Middle aged British punker who is heavily in love with anything punk related circa 1976 - 1982. Film fanatic who indulges in any genre of film but


[See more▼](#)

 Lifetime Total
2,500+

 Lifetime Plot
1+


 Top Reviewer

 Poll Taker
100x


 IMDb Member
16 years

Ratings

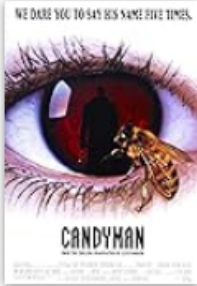
Most Recently Rated




The Invisible Man
★ 8




Ride with the D...
★ 9



Candyman
★ 7



Machine Gun Mc...
★ 7



Carry on Campi...
★ 8

[See all 4,715 ratings »](#)

Lists

→ Ta sẽ thu được: (Chưa thực hiện)

Phần 3: Khám phá & Tiền xử lý dữ liệu

Sử dụng pandas kết hợp với các thư viện khác để khám phá, hiểu sâu hơn về dữ liệu cũng như tiền xử lý dữ liệu trước khi bước vào các giai đoạn phân tích khác

Tìm hiểu các thông tin sơ lược như:

- Số lượng thuộc tính, số lượng mẫu.
- Ý nghĩa của từng thuộc tính (cột).
- Dữ liệu có bị trùng lặp hay không?

Đến các thông tin cụ thể và chuyên sâu hơn:

.Các thuộc tính nên có kiểu dữ liệu gì? Xử lý các thuộc tính có kiểu dữ liệu không phù hợp. Xử lý các mẫu, thuộc tính bị thiếu dữ liệu.

Cuối cùng, lưu dữ liệu đã được tiền xử lý để phục vụ cho giai đoạn trực quan cũng như huấn luyện mô hình học máy

Phần 4: Trực quan hóa dữ liệu

Phần 5: Xây dựng mô hình

Phần 6: Đánh giá mô hình