User Preference
on Two Popular Movie Review Platforms
IMDB and Douban

Team 6	Haoyun Hong
	Siyuan Wu
	Yaxin Shi
	Yuxin Zhang
	Shizhuo Cheng

Data

- Douban Top 250 movie data and corresponding user reviews

- IMDB Top 250 movie data and corresponding user reviews

Research Question List

- **Analyse the data we crawled and mine the movie preferences of IMDB and Douban users from different dimensions:**

  - Flim genres/Geographical distribution/Release year/Language/Runtime Statistic.

  - What's the characteristics of movie rate and what's the rating level of different genres?

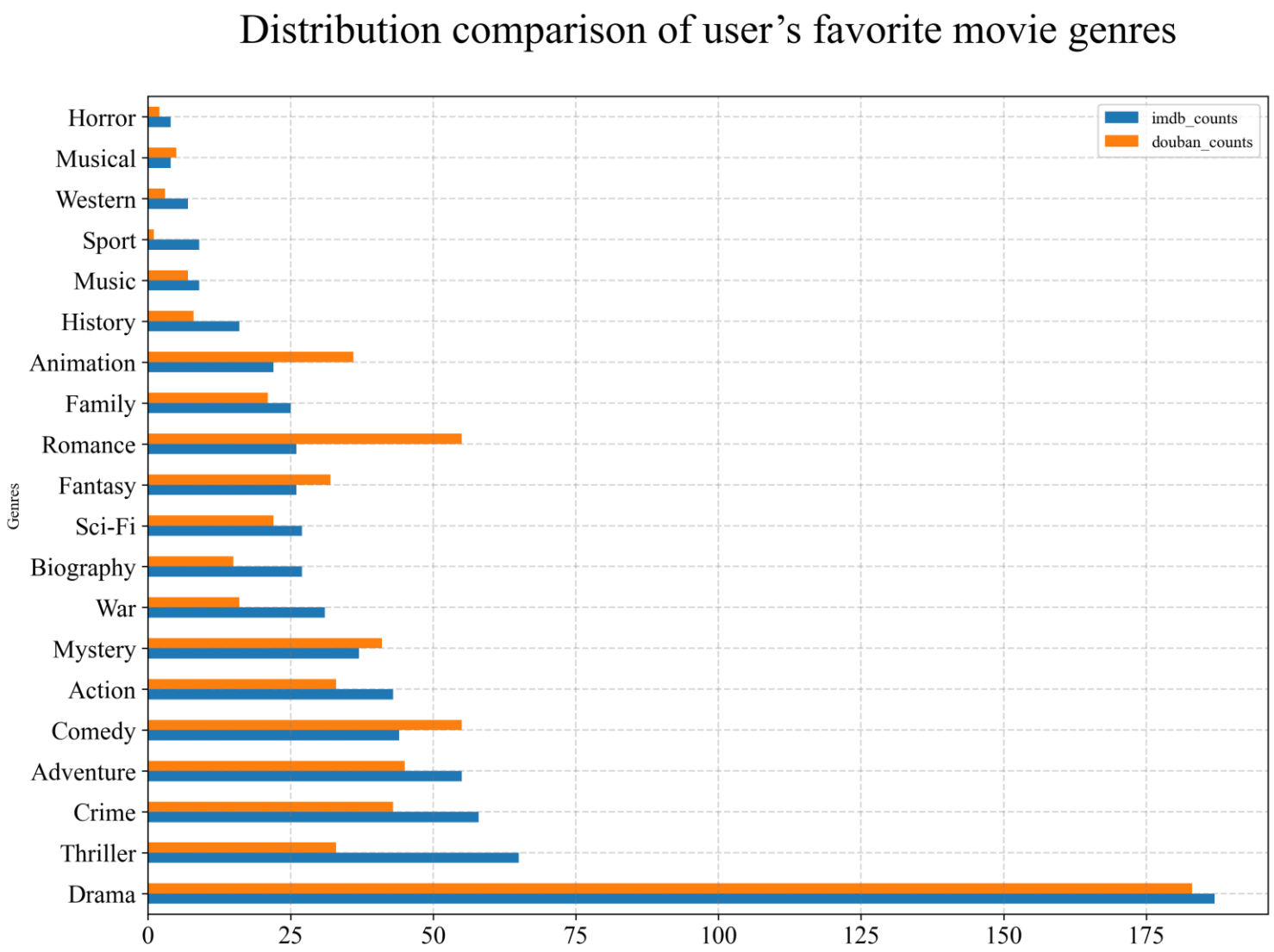
  - What words do users on imdb and Douban frequently say? What does this reveal ?

- **Predict the results from various factors (user factors and movie factors) on different platforms:**

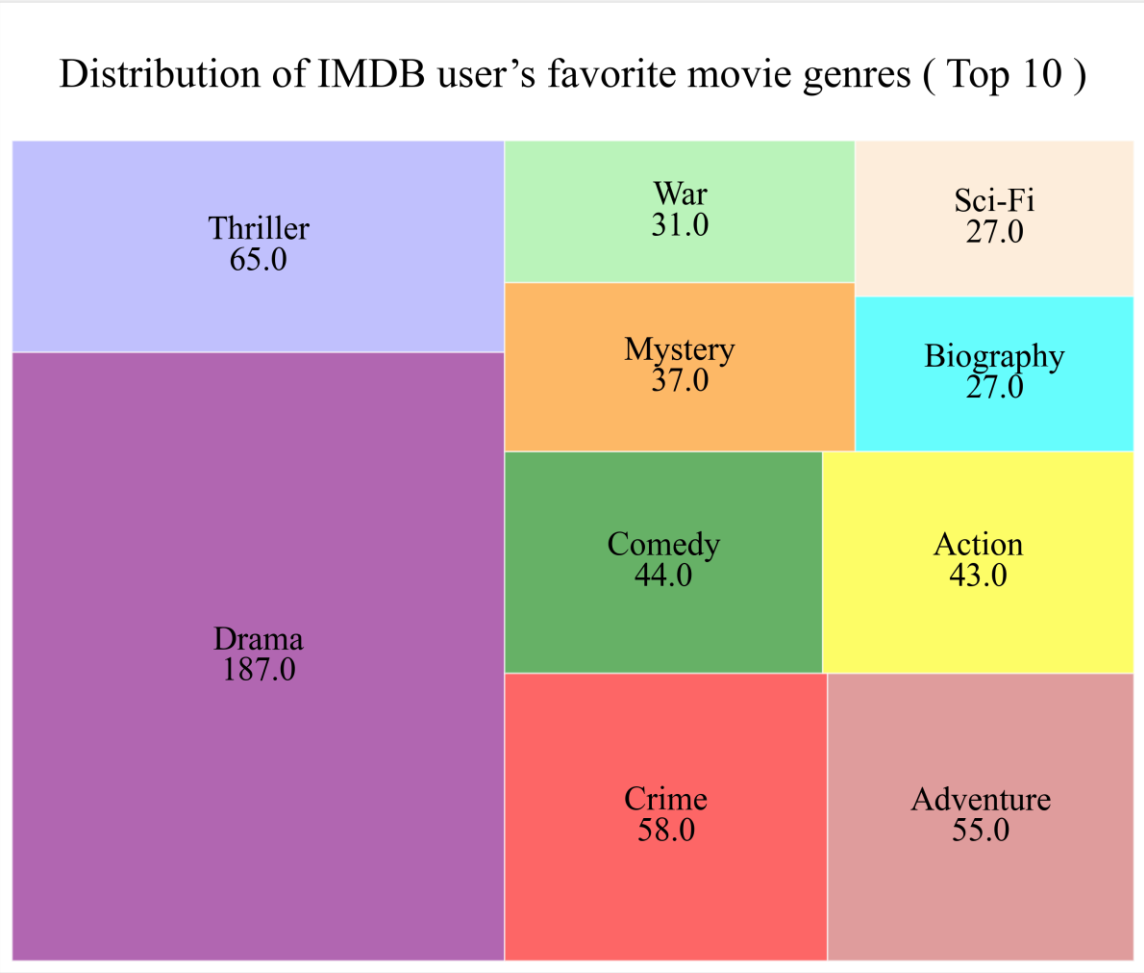
  - Finding features from user reviews to predict movie genres.

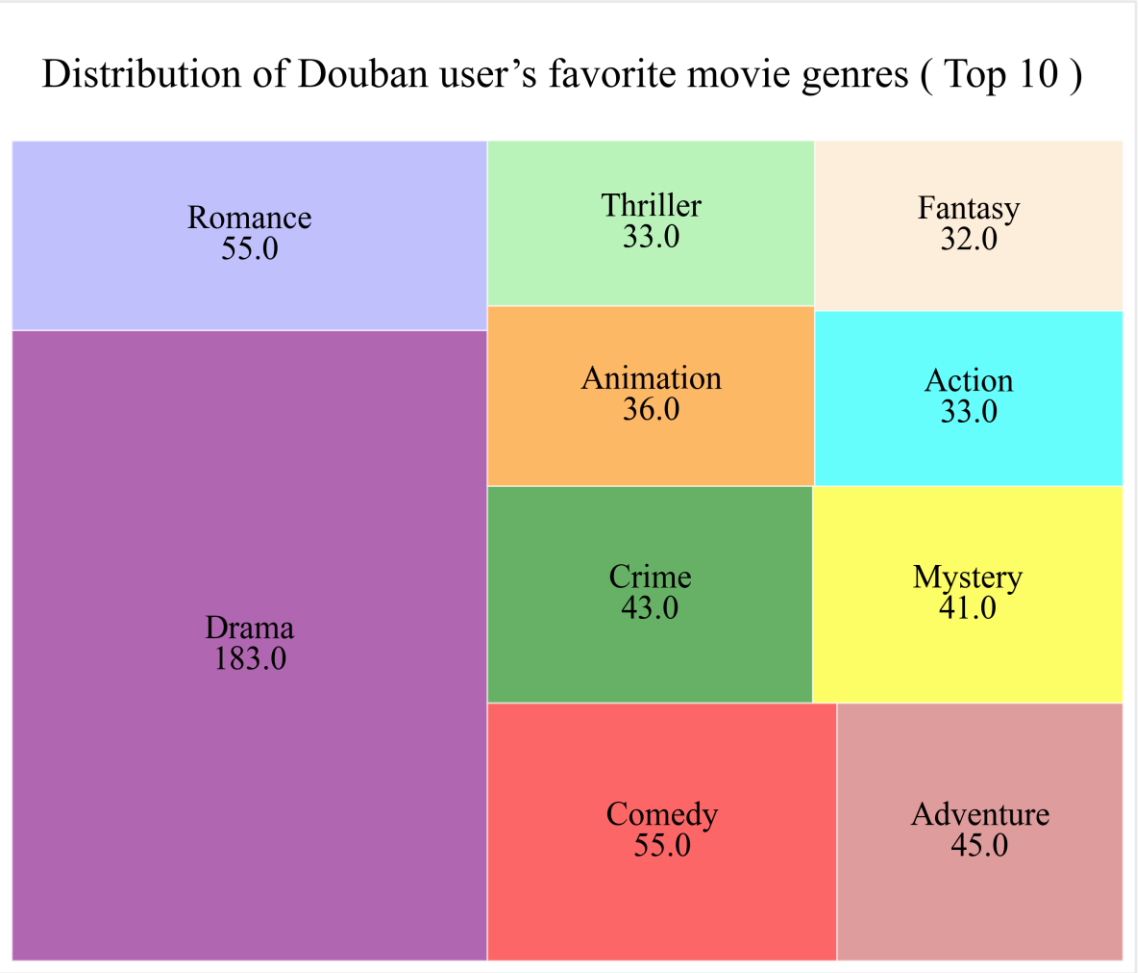
Distribution of common genres in two sites

	Genres	imdb_counts	douban_counts
0	Drama	187.0	183.0
1	Thriller	65.0	33.0
2	Crime	58.0	43.0
3	Adventure	55.0	45.0
4	Comedy	44.0	55.0
5	Action	43.0	33.0
6	Mystery	37.0	41.0
7	War	31.0	16.0
8	Biography	27.0	15.0
9	Sci-Fi	27.0	22.0
10	Fantasy	26.0	32.0
11	Romance	26.0	55.0
12	Family	25.0	21.0
13	Animation	22.0	36.0
14	History	16.0	8.0
15	Music	9.0	7.0
16	Sport	9.0	1.0
17	Western	7.0	3.0
18	Musical	4.0	5.0
19	Horror	4.0	2.0

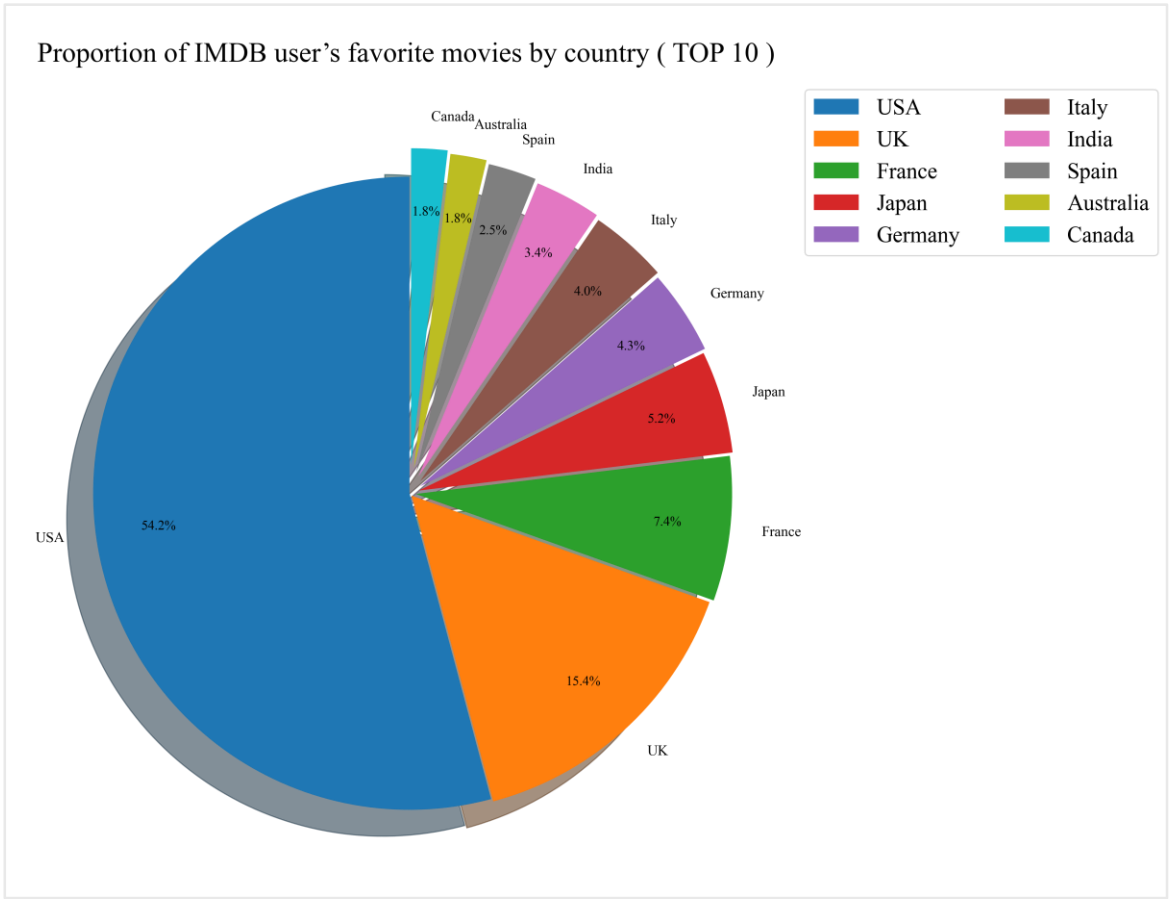


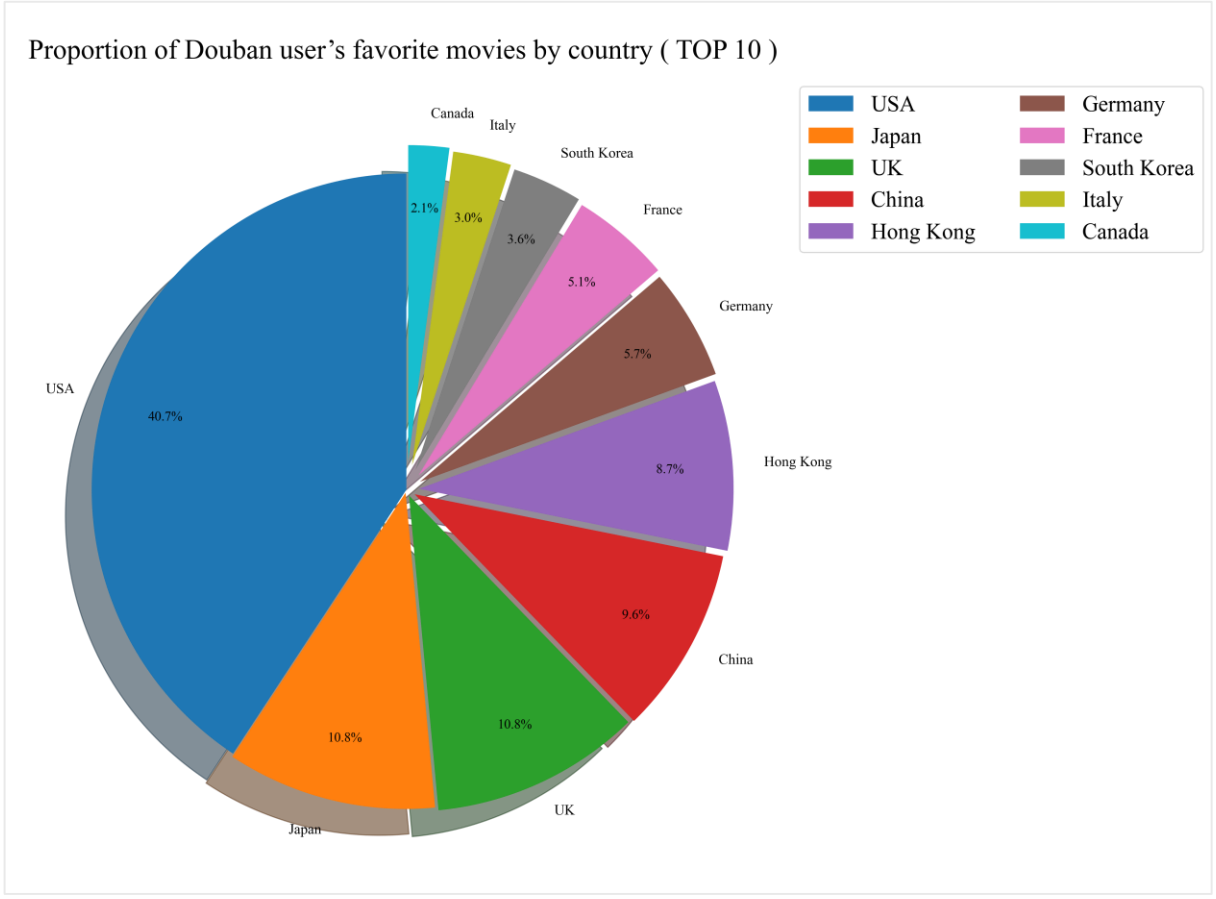
Genres distribution of movies (TOP10)



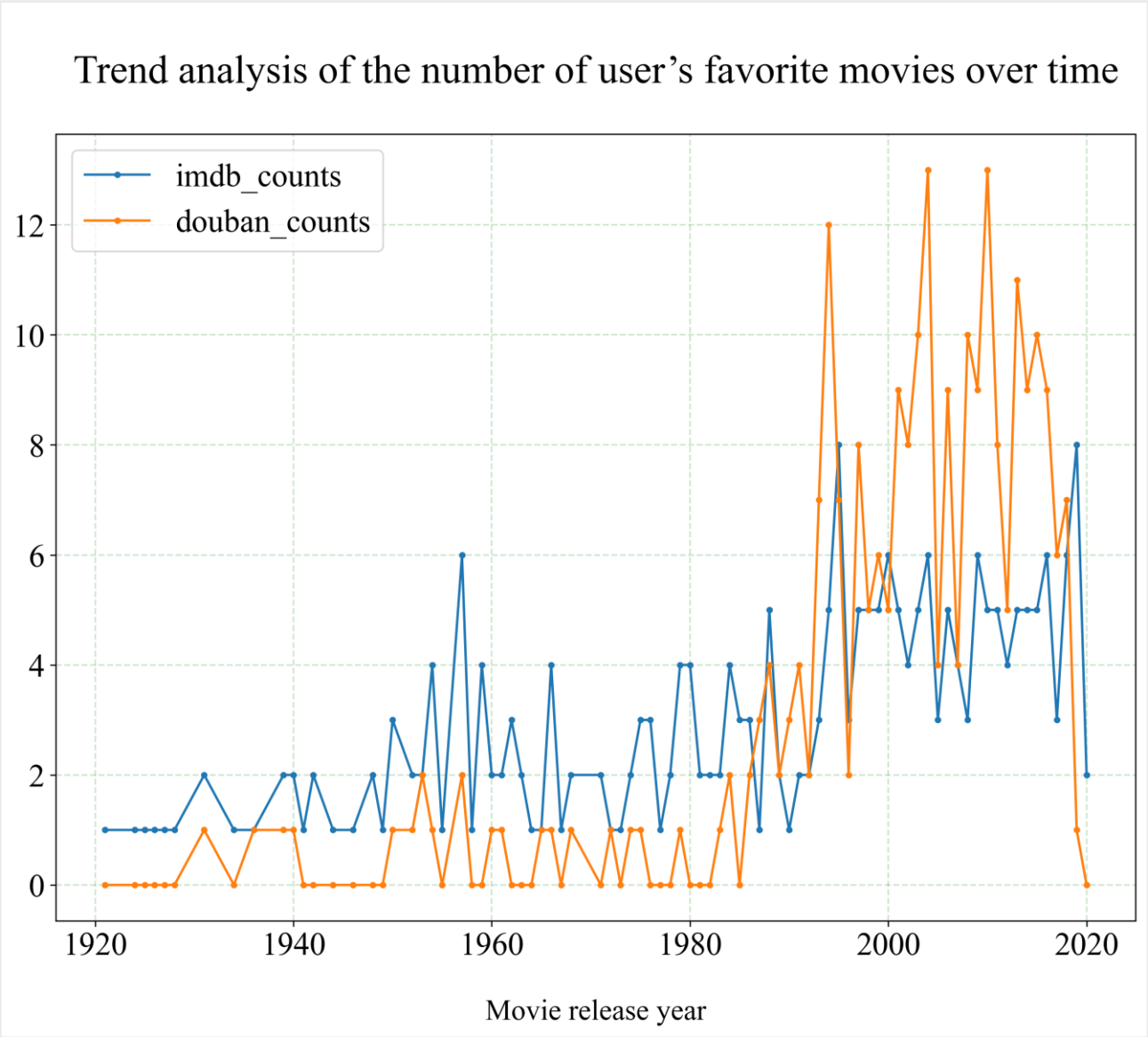


Geographical distribution of movies





Release Year distribution of movies



	Year	imdb_counts	douban_counts
63	2000	6	5.0
64	2001	5	9.0
65	2002	4	8.0
66	2003	5	10.0
67	2004	6	13.0
68	2005	3	4.0
69	2006	5	9.0
70	2007	4	4.0
71	2008	3	10.0
72	2009	6	9.0
73	2010	5	13.0
74	2011	5	8.0
75	2012	4	5.0
76	2013	5	11.0
77	2014	5	9.0
78	2015	5	10.0
79	2016	6	9.0
80	2017	3	6.0
81	2018	6	7.0
82	2019	8	1.0
83	2020	2	0.0

Languages

Languages most commonly used in the movies

Douban

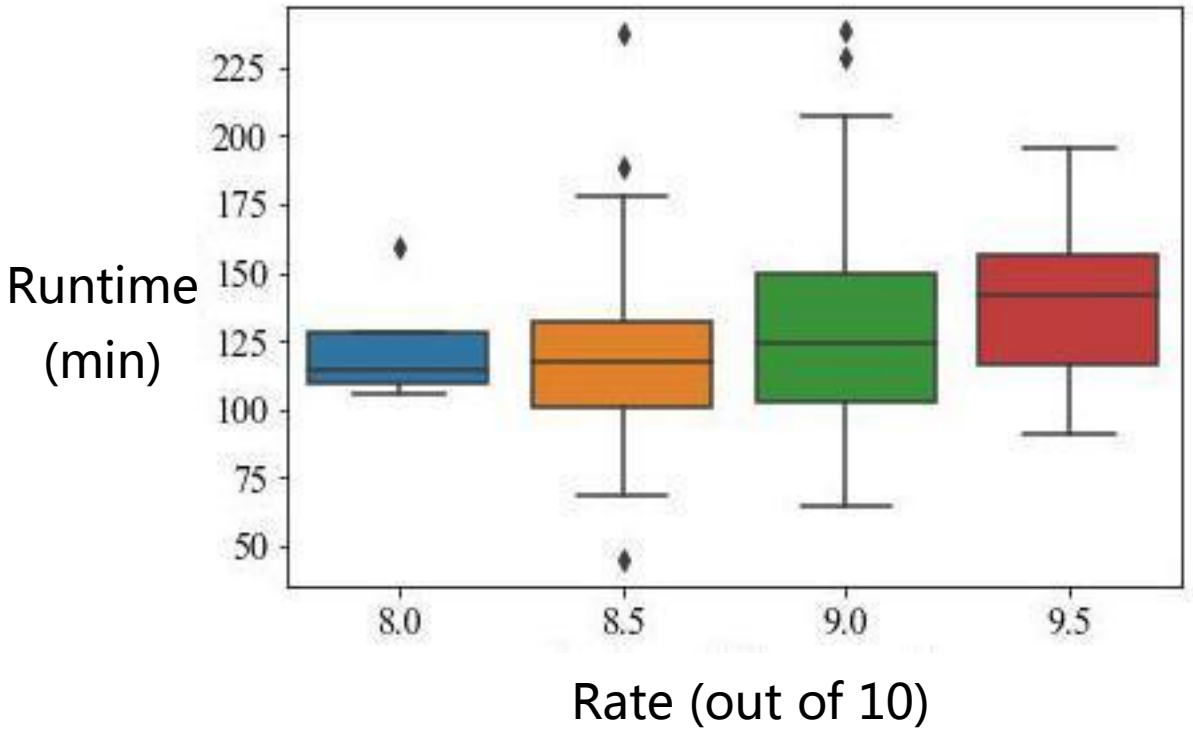
id	count	language
1	164	English
2	46	Japanese
3	42	Mandarin
4	39	French
5	27	German
6	25	Cantonese
7	19	Italian
8	15	Spanish
9	12	Russian
10	11	Korean

imdb

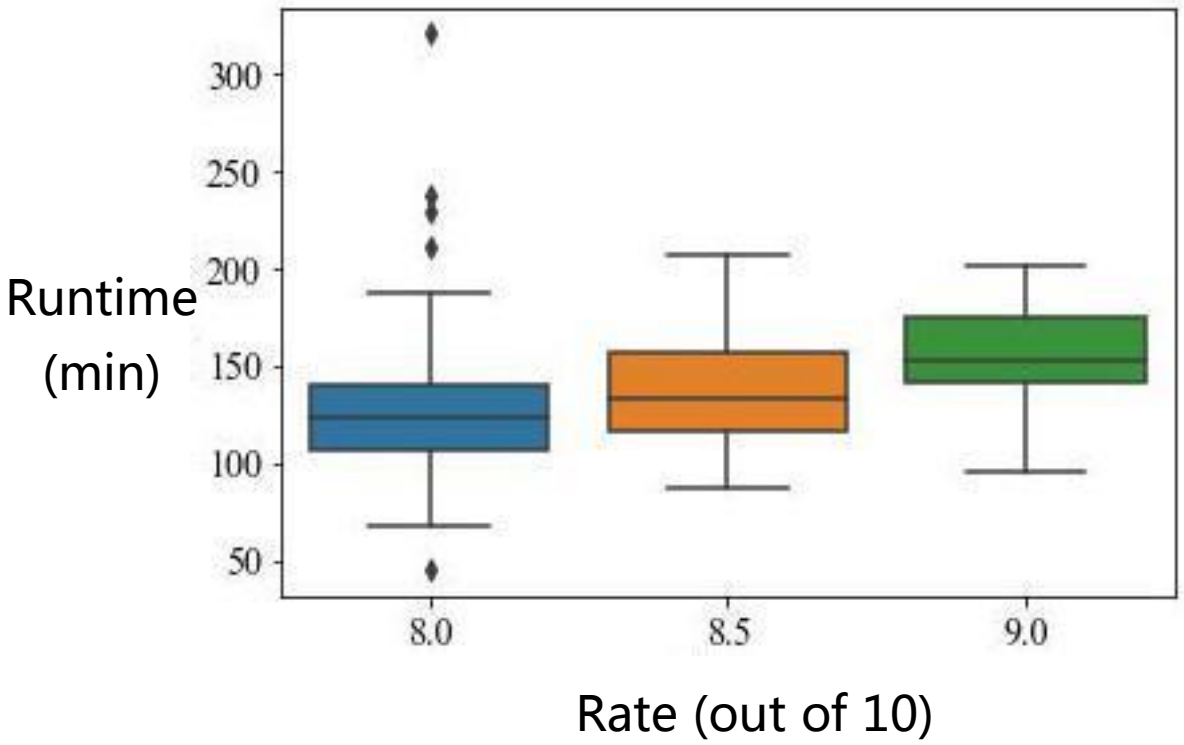
id	count	language
1	200	English
2	43	French
3	36	German
4	30	Spanish
5	26	Japanese
6	26	Italian
7	15	Russian
8	13	Latin
9	8	Hindi
10	8	Arabic

Runtime - general distribution

Douban

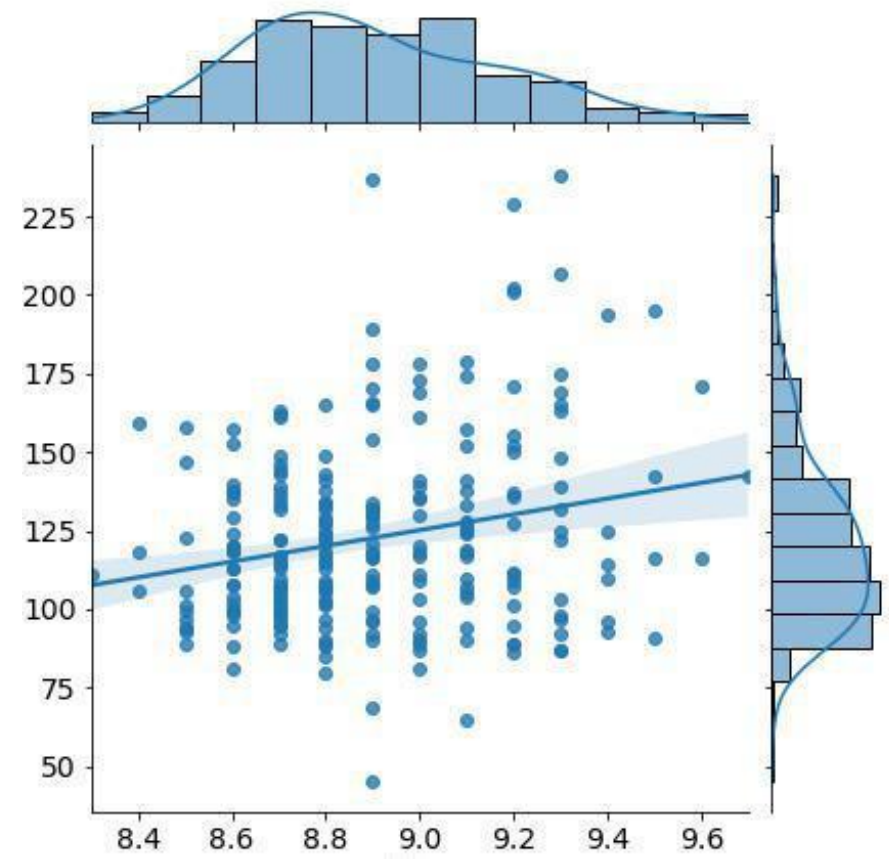


imdb

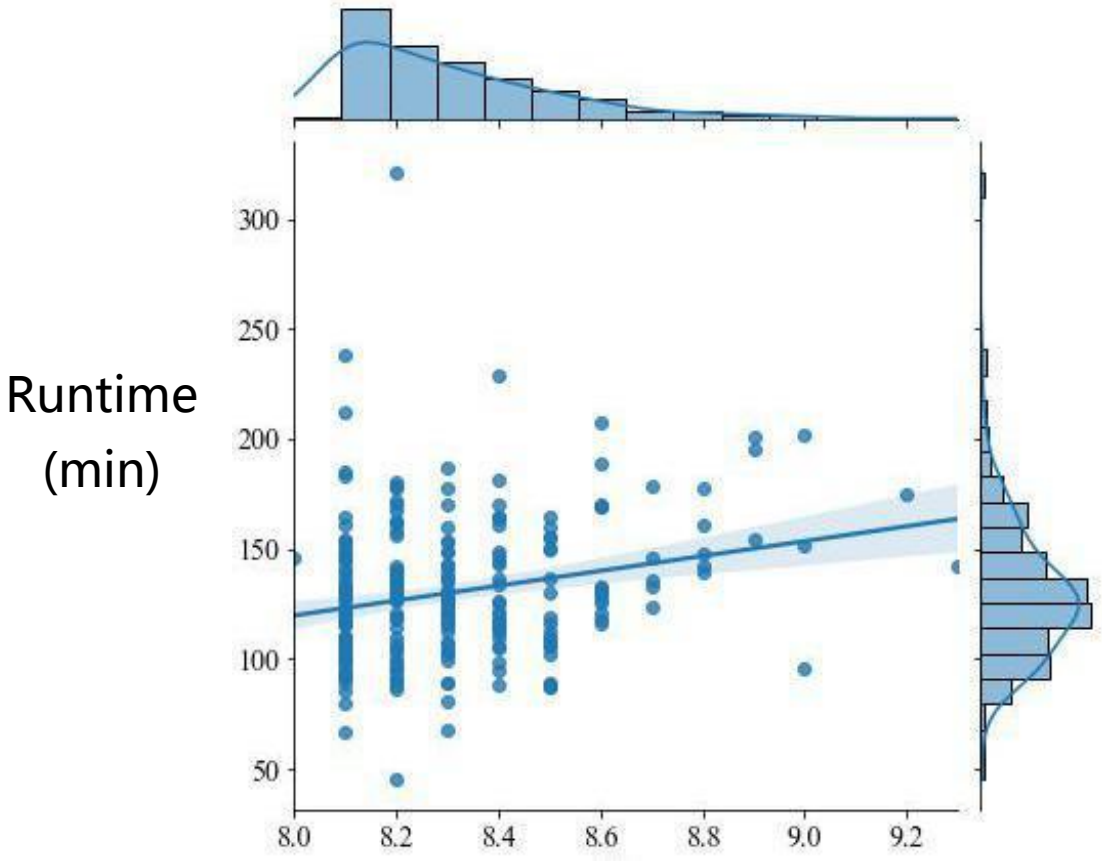


Runtime - regression analysis

Douban



imdb



Movie Rate Features in two sites

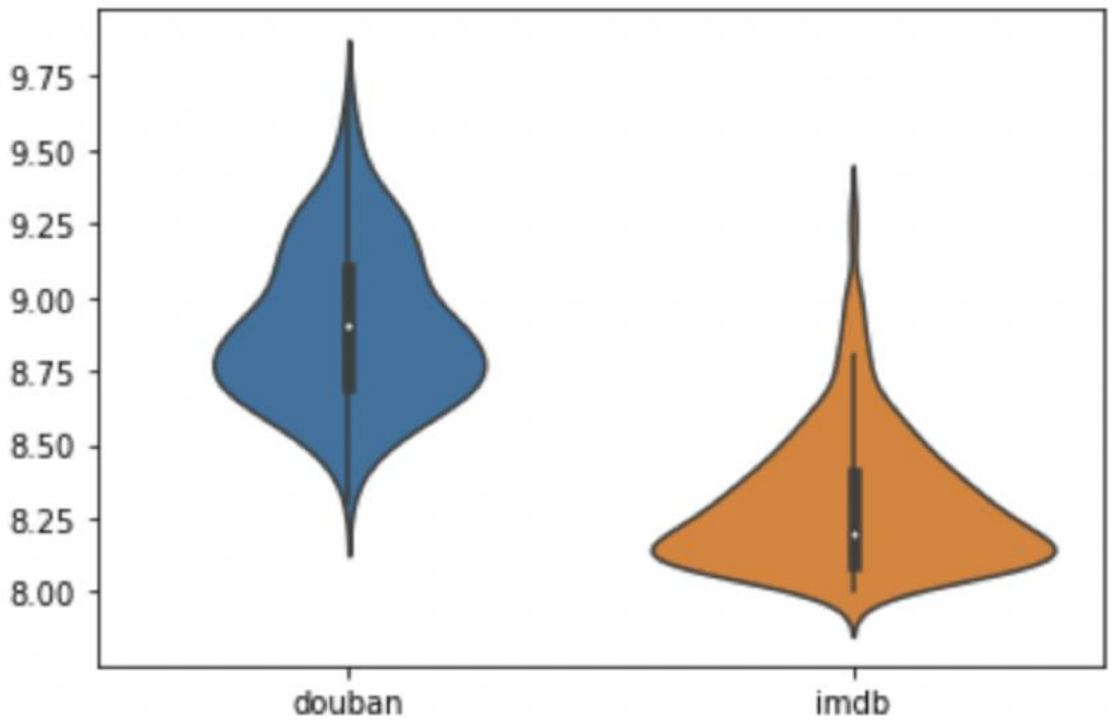
Statistical Analysis of 2 Sites' TOP250 Rate Distribution

	douban	imdb
count	250.000000	250.000000
mean	8.900800	8.303600
std	0.263433	0.223218
min	8.300000	8.000000
25%	8.700000	8.100000
50%	8.900000	8.200000
75%	9.100000	8.400000
max	9.700000	9.300000

Douban users tend to give higher ratings than IMDB users.

Movie Rate Features in two sites

Violin Plot of 2 Sites' TOP250 Movies Rate



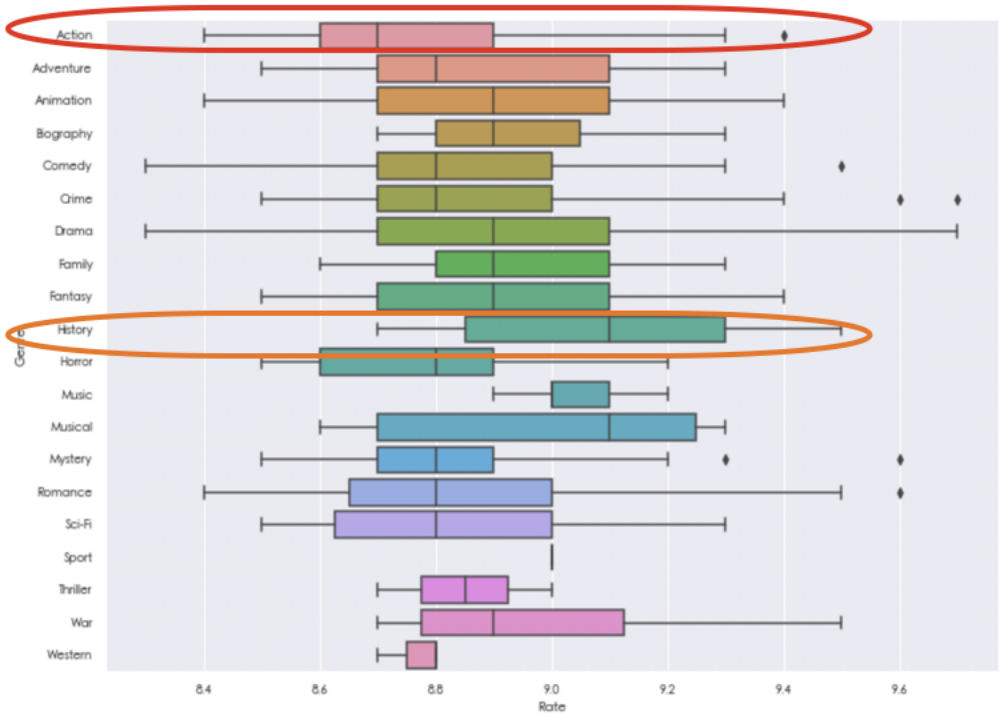
- IMDB: higher peak and thin tail
- Douban: more evenly distributed

Tips for IMDB user:
If you are looking up for movie recommendations in Douban, it may be helpful to do some conversion of rate first (such as subtracting 0.4 point, from 8.8 to 8.4) to acquire reasonable inference of the movie quality.
(vice versa for Douban users.)

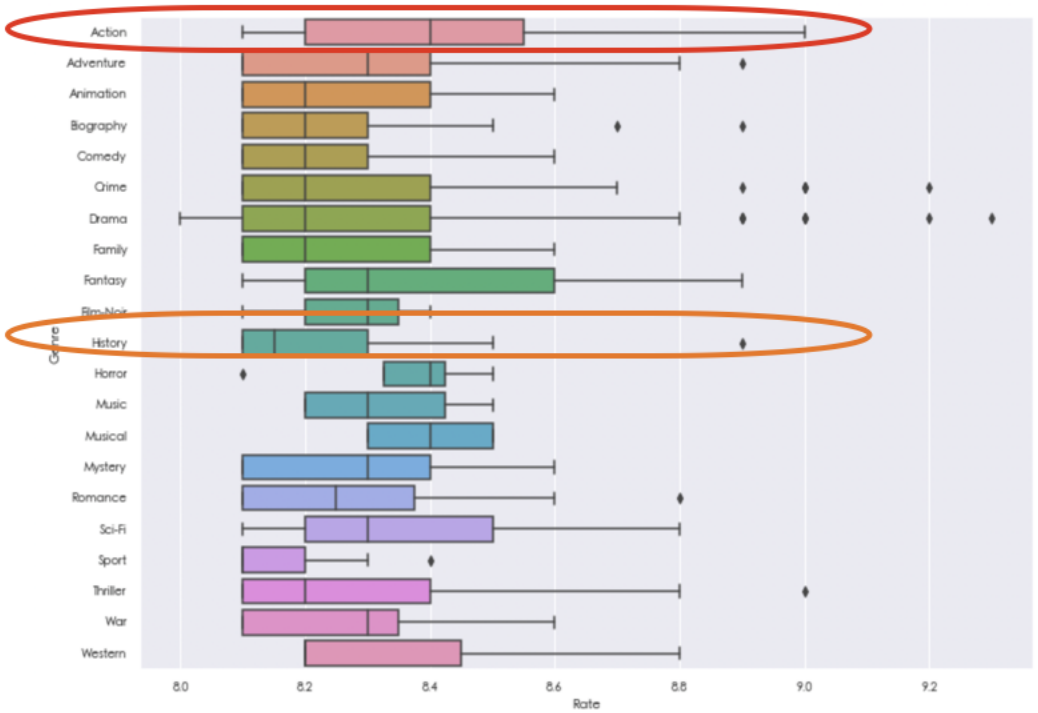
Douban is a community with more forgiving score givers.

Rate Distribution of Different Genres

Box Plot of 20 Genres' Rate in Douban & IMDB



Douban Rate Boxplot



IMDB Rate Boxplot

Rate Distribution of Different Genres

User Behavior Difference and Inspiring Strategy

	IMDB	Douban
Action	Highest	Lowest
History	Lowest	Highest

As douban is mainly composed of users from China, a country with long and attractive history, they love and respect traditions a lot and have more calm and stable personality.

Inspiration for a movie selection strategy

If you are an *action movie lover*, when IMDB gives a movie of action genre high score and Douban gives it low score, it's probably sensible to *believe in IMDB and give it a try*.

When checking the movie rate online, looking up for two sites and **give different weights of sites' rating considering movie's genre** can help us find movies suit to our taste best.

The culture background influences user behavior and leads to platform traits.

WordClouds for top-rated 100 reviews

WordClouds for top-rated 100 reviews

- douban:
- 'red': '电影', '导演', '故事', '剧情', '配乐', '剧本', '表演','角色','镜头', '音乐','主角','观众','片子'
- meaning: movie, director, story, plot, soundtrack, script, performance, character, shot, music, main character, audience, film(another)
-
- 'green': '真的', '感觉','精彩','感动','喜欢','特别','人生', '世界', '生活','人性','经典','现实'
- meaning: really, feel, excellent, touching, like, special(particularly), life, world, living(daily), humanity, classic, reality
-
- imdb:
- 'red': 'movie', 'film', 'character', 'performance', 'story', 'shot','actor','scene', 'director','plot','acting'
- 'green': 'life', 'people','good','like','bad','love', 'great', 'feel','world','excellent','perfect','real'

Predict Movie Genres via User Reviews

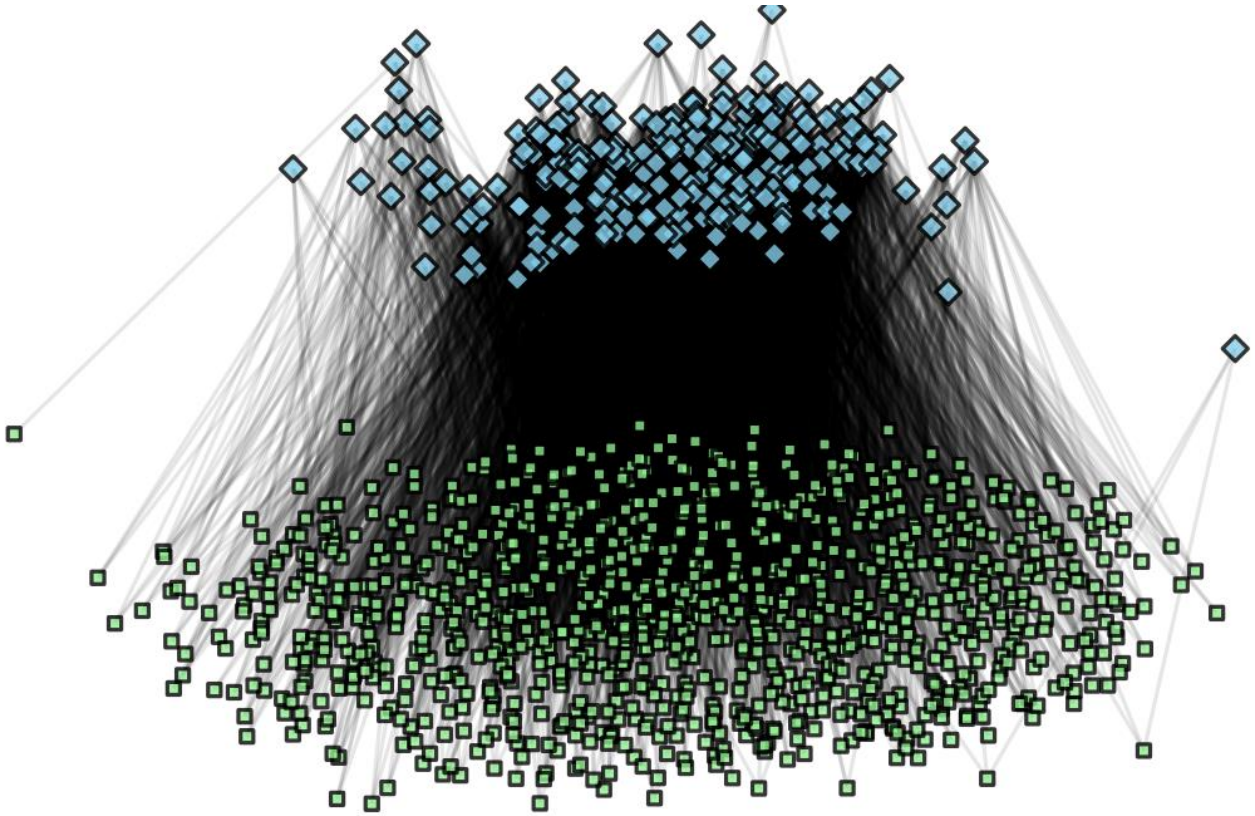
Is the prediction feasible? Let's do EDA first!

Assumption:

- **The movies of the same kind tend to have similar reviews**
- **The same user tends to use similar words in his or her movie reviews**
- **The movies of the same type may have similar audience groups**
- **The users may be more interested in some genres of movies**

User: audience who review on at least one of the top 250 movies on IMDB

Network Analysis



Blue nodes: movies

Green nodes: users

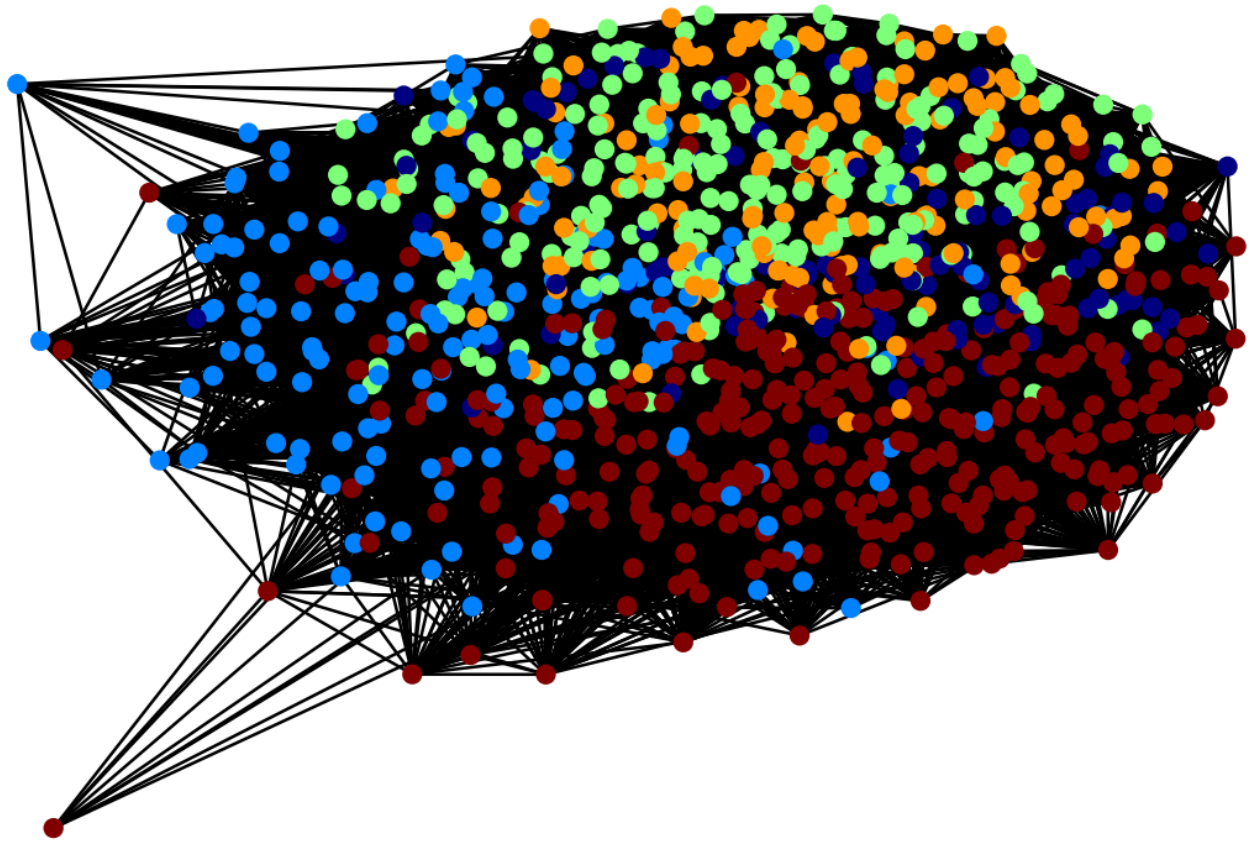
Edge: the user reviewed the movie

Number of nodes: 1284

Number of edges: 7763

Average degree: 12.0919

Community Detection



Node: user

Edge: review the same movie

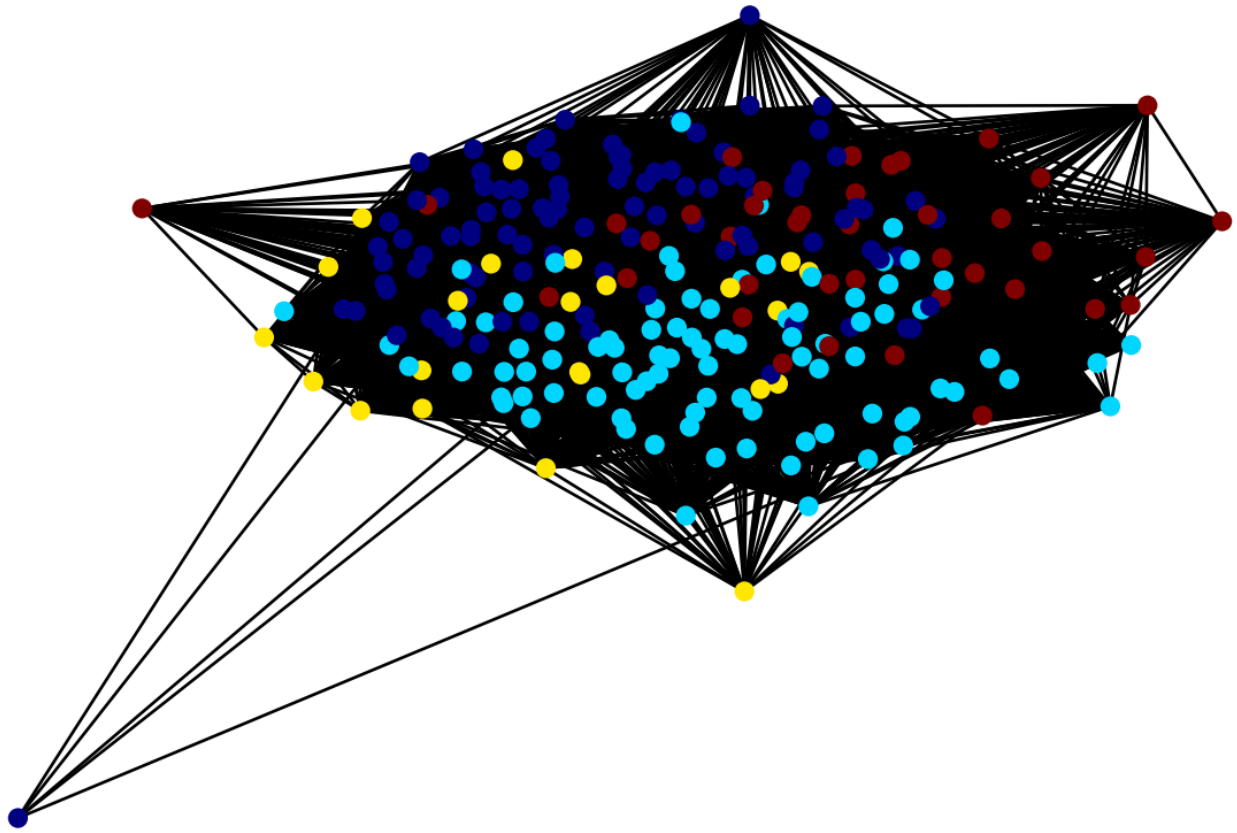
Modularity: 0.133

Number of nodes: 1034

Number of edges: 92486

Average degree: 178.8897

Community Detection



Node: movie

Edge: reviewed by the same user

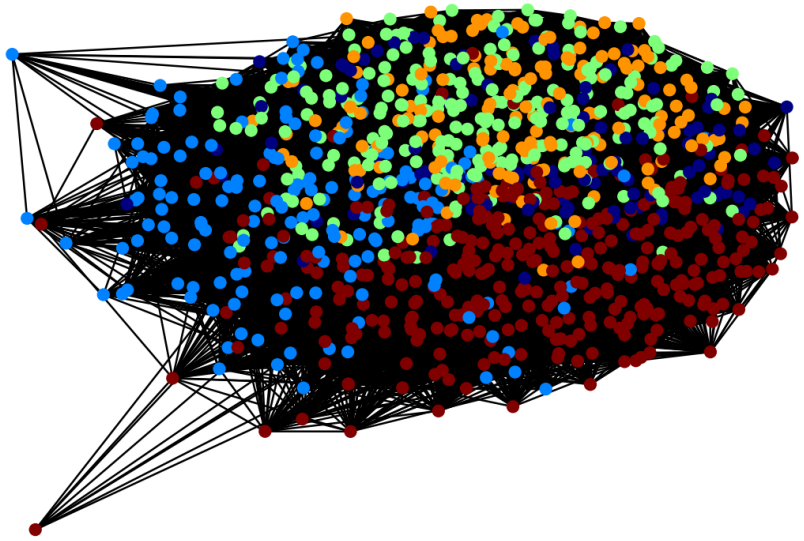
Modularity: 0.042

Number of nodes: 250

Number of edges: 23933

Average degree: 191.4640

User-User

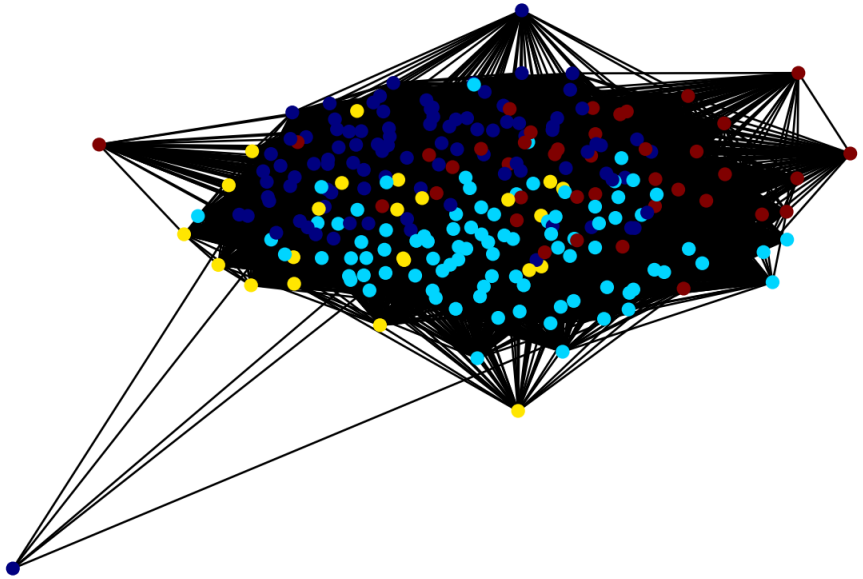


Number of nodes: 1034

Number of edges: 92486

Average degree: 178.8897

Movie-Movie



Number of nodes: 250

Number of edges: 23933

Average degree: 191.4640

Prediction is likely to be feasible

Predict Movie Genres via User Reviews

1. Reviews we crawled are the top 100 reviews sort by helpfulness on both sites
2. Use all the reviews of every movie to obtain TF-IDF scores
3. The indices are the unique names of the movies
4. The data set of each site is randomly divided into training set and test set according to 8:2

Predict Movie Genres via User Reviews

1. For every movie M in the test set, we choose the top 10 movies in the training set that have the highest TF-IDF similarity with M
2. Choose the most common genre of the top 10 training samples to be the predicted genre of M
3. Calculate the Hit@1 of the prediction results
4. Some movies fall into multiple genres. We think that as long as the predicted genre is one of movie's true genres, the predicted result is correct

IMDB

Hit@1 = 0.9

```
In [30]: hit1 = 0
         for name in test_names:
             print("Test movie: ", name)
             res = get_top_similar(name)
             hit1 += res
             print("**********************************************************************")

Test movie:  The Matrix
ground truth:  Action/Sci-Fi
predicte:  Sci-Fi  (with confidence of  0.7 )
**********************************************************************
Test movie:  Gisaengchung
ground truth:  Comedy/Drama/Thriller
predicte:  Drama  (with confidence of  0.7 )
**********************************************************************
Test movie:  Seppuku
ground truth:  Action/Drama/Mystery
predicte:  Drama  (with confidence of  0.8 )
**********************************************************************
Test movie:  The Usual Suspects
ground truth:  Crime/Mystery/Thriller
predicte:  Drama  (with confidence of  0.9 )
**********************************************************************
Test movie:  American History X
ground truth:  Drama
predicte:  Drama  (with confidence of  0.8 )
**********************************************************************

In [31]: hit1 = hit1/len(test_names)
         print("predict accuracy is: ", hit1)

predict accuracy is:  0.9
```

Douban

Hit@1 = 0.74

```
In [13]: hit1 = 0
         for name in test_names:
             print("Test movie: ", name)
             res = get_top_similar(name)
             hit1 += res
             print("**********************************************************************")

predicte:  剧情  (with confidence of  0.6 )
**********************************************************************
Test movie:  何以为家
ground truth:  剧情
predicte:  剧情  (with confidence of  0.8 )
**********************************************************************
Test movie:  指环王3：王者无敌
ground truth:  剧情/动作/奇幻/冒险
predicte:  剧情  (with confidence of  0.5 )
**********************************************************************
Test movie:  少年派的奇幻漂流
ground truth:  剧情/奇幻/冒险
predicte:  剧情  (with confidence of  0.6 )
**********************************************************************
Test movie:  天空之城
ground truth:  动画/奇幻/冒险
predicte:  动画  (with confidence of  0.5 )
**********************************************************************
Test movie:  猫鼠游戏

In [14]: hit1 = hit1/len(test_names)
         print("predict accuracy is: ", hit1)

predict accuracy is:  0.74
```

Github repository

https://github.com/HaoyunHong/Oxford-Group-Project.git

Data

https://cloud.tsinghua.edu.cn/d/13faa4a5a6304ad28385/

Thanks for listening !

March 7, 2021