数据观察

In [1]:
```
import pandas as pd
```

In [2]:
```
movies_df = pd.read_csv('../data/ml-latest-small/movies.csv')
movies_df.head()
```

Out[2]:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

In [3]:
```
links_df = pd.read_csv('../data/ml-latest-small/links.csv')
links_df.head()
```

Out[3]:

	movieId	imdbId	tmdbId
0	1	114709	862.0
1	2	113497	8844.0
2	3	113228	15602.0
3	4	114885	31357.0
4	5	113041	11862.0

In [4]:
```
ratings_df = pd.read_csv('../data/ml-latest-small/ratings.csv')
ratings_df.head()
```

Out[4]:

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931

In [5]:
```
tags_df = pd.read_csv('../data/ml-latest-small/tags.csv')
tags_df.head()
```

Out[5]:

	userId	movieId	tag	timestamp
0	2	60756	funny	1445714994
1	2	60756	Highly quotable	1445714996
2	2	60756	will ferrell	1445714992
3	2	89774	Boxing story	1445715207
4	2	89774	MMA	1445715200

数据合并

目的是给定一个用户id，找出用户可能喜欢的电影名。
但是两个文件电影信息和用户评分信息是分开的，所以需要合并。

In [6]:
```
data = pd.merge(movies_df, ratings_df, on='movieId')  # 通过两数据框之间的movieId连接
data[['userId', 'rating', 'movieId',
      'title']].sort_values('userId').to_csv('./output/merged.csv', index=False)
```