Yangzhou University

Software Data Analysis 2022

鲁卡娅+ MH20055

Recommender System (content based filtering)

dataset resource : https://grouplens.org/datasets/movielens/100k/

In [1]:

```
import numpy as np
import pandas as pd
```

In [10]:

```
#### datasets
unames = ['user_id','age','gender','occupation','zipcode']
users = pd.read_csv('ml-100k/u.user',sep='|',names=unames)
rnames=['user_id','item id','rating','timestamp']
ratings = pd.read_csv('ml-100k/u.data',sep='\t',names=rnames)
```

In [12]:

```
ratings.head(5)
```

Out[12]:

	user_id	item id	rating	timestamp
0	196	242	3	881250949
1	186	302	3	891717742
2	22	377	1	878887116
3	244	51	2	880606923
4	166	346	1	886397596

In [19]:

```
user_df = users.loc[:,['user_id','gender']]
ratings_df=ratings.loc[:,['user_id','rating']]
```

In [20]:

```
rating_df=ratings_df.merge(user_df)
```

In [24]:

```
rating_df
```

Out[24]:

	user_id	rating	gender
0	196	3	M
1	196	4	M
2	196	4	M
3	196	3	M
4	196	5	M
...	...	...	...
99995	941	5	M
99996	941	3	M
99997	941	5	M
99998	941	4	M
99999	941	4	M

100000 rows × 3 columns

In [25]:

```
rating_df.groupby('gender').rating.std()
```

Out[25]:

```
gender
F    1.170951
M    1.109556
Name: rating, dtype: float64
```

In [27]:

```
rating_df.groupby('gender').rating.apply(pd.Series.std)
```

Out[27]:

```
gender
F    1.170951
M    1.109556
Name: rating, dtype: float64
```

In [31]:

```
df1=rating_df.groupby(['user_id','gender']).apply(np.mean)
```

```
C:\Users\Rukaia\AppData\Local\Programs\Python\Python39\lib\site-packages\numpy\core\fromnumeric.py:3438: FutureWarning: Dropping of nuisance columns in DataFr
ame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError.  Select only valid columns before calling the reductio
n.
  return mean(axis=axis, dtype=dtype, out=out, **kwargs)
```

In [32]:

```
df1.groupby('gender').rating.std()
```

Out[32]:

```
gender
F    0.481241
M    0.430076
Name: rating, dtype: float64
```

In [36]:

```
pd.pivot_table(df1, values = 'rating', index = 'gender', aggfunc = pd.Series.std)
```

Out[36]:

	rating
gender	
F	0.481241
M	0.430076

In [38]:

```
pd.pivot_table(rating_df, index = ['user_id','gender'], values = 'rating')
```

Out[38]:

		rating
user_id	**gender**	
1	**M**	3.610294
2	**F**	3.709677
3	**M**	2.796296
4	**M**	4.333333
5	**F**	2.874286
...	**...**	...
939	**F**	4.265306
940	**M**	3.457944
941	**M**	4.045455
942	**F**	4.265823
943	**M**	3.410714

943 rows × 1 columns

In [41]:

```
t = pd.pivot_table(rating_df, index = ['user_id','gender'], values = 'rating')
female = t.query("gender == ['F']")
pd.Series.std(female)
```

Out[41]:

```
rating    0.481241
dtype: float64
```

In [43]:

```
t = pd.pivot_table(rating_df, index = ['user_id','gender'], values = 'rating')
male = t.query("gender == ['M']")
pd.Series.std(male)
```

Out[43]:

```
rating    0.430076
dtype: float64
```

In [ ]: