Ime, priimek:

LV07: Priporočilni sistem, podobnost vsebin

Namen vaje:

- Spoznati podatke o ocenah filmov
- Algoritem izračuna podobnosti vsebin (filmov) na osnovi ocen
- Generiranje priporočila podobnih filmov za izbrani film

1 Analiza podatkov

1.1 Podatkovni set MovieLens

```
MovieLens data sets were collected by the GroupLens Research Project
at the University of Minnesota.

This data set consists of:
        * 100,000 ratings (1-5) from 943 users on 1682 movies.
        * Each user has rated at least 20 movies.
        * Simple demographic info for the users (age, gender, occupation, zip)

The data was collected through the MovieLens web site
(movielens.umn.edu) during the seven-month period from September 19th,
1997 through April 22nd, 1998. This data has been cleaned up - users
who had less than 20 ratings or did not have complete demographic
information were removed from this data set. Detailed descriptions of
the data file can be found at the end of this file.
```

1.2 Nalaganje podatkov

Uvozimo podatke iz datoteke v data frame

```
import numpy as np
import pandas as pd
import sklearn.cross_decomposition
# Uvozi za vizualizacijo
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('white')
%matplotlib inline
```

```
column_names = ['user_id', 'item_id', 'rating', 'timestamp']
df = pd.read_csv('u.data', sep='\t', names=column_names)
```

Preglej in izpiši nekaj vrstic podatkov:

```
   user_id  item_id  rating  timestamp
0        0       50       5  881250949
1        0      172       5  881250949
2        0      133       1  881250949
3      196      242       3  881250949
4      186      302       3  891717742
```

Preberemo naslove filmov, in jih dodamo v podatkovni objekt

```
movie_titles = pd.read_csv("Movie_Id_Titles")
movie_titles.head()
df = pd.merge(df,movie_titles,on='item_id')
print(df.shape)
```

Izpiši nekaj vrstic podatkov:

```
(100003, 5)
   item_id              title
0        1   Toy Story (1995)
1        2   GoldenEye (1995)
2        3  Four Rooms (1995)
3        4  Get Shorty (1995)
4        5     Copycat (1995)
```

1.3 Vprašanje 1

Ugotovi, koliko uporabnikov in koliko filmov obsegajo podatki. Uporabi metodo nunique().

Koda in rezultat:

```python
# Število unikatnih uporabnikov
unique_users = df['user_id'].nunique()

# Število unikatnih filmov
unique_movies = df['item_id'].nunique()

print(f"Število uporabnikov: {unique_users}")
print(f"Število filmov: {unique_movies}")
```

✓ 0.0s

```
Število uporabnikov: 944
Število filmov: 1682
```

1.4 Ogled podatkov

```python
df.groupby('title').head()
```

	user_id	item_id	rating	timestamp	title
0	0	50	5	881250949	Star Wars (1977)
1	0	172	5	881250949	Empire Strikes Back, The (1980)
2	0	133	1	881250949	Gone with the Wind (1939)
3	196	242	3	881250949	Kolya (1996)
4	186	302	3	891717742	L.A. Confidential (1997)
...	...	...	...	...	...
98958	655	1641	3	887427810	Dadetown (1995)
99180	416	1594	5	893212484	Everest (1998)
99617	450	1490	3	882396929	Fausto (1993)
99752	399	1542	2	882348592	Scarlet Letter, The (1926)
99956	655	913	4	891817521	Love and Death on Long Island (1997)

Preskusi kodo in komentiraj rezultate:

```
title
'Til There Was You (1997)                  9
1-900 (1994)                               5
101 Dalmatians (1996)                    109
12 Angry Men (1957)                      125
187 (1997)                                41
                                         ...
Young Guns II (1990)                      44
Young Poisoner's Handbook, The (1995)     41
Zeus and Roxanne (1997)                    6
unknown                                    9
Á köldum klaka (Cold Fever) (1994)         1
Name: rating, Length: 1664, dtype: int64
```

```
title
Star Wars (1977)             584
Contact (1997)               509
Fargo (1996)                 508
Return of the Jedi (1983)    507
Liar Liar (1997)             485
                            ...
War at Home, The (1996)        1
Mirage (1995)                  1
Modern Affair, A (1995)        1
Dadetown (1995)                1
Yankee Zulu (1994)             1
Name: rating, Length: 1664, dtype: int64
```

```
title
'Til There Was You (1997)                 2.333333
1-900 (1994)                              2.600000
101 Dalmatians (1996)                     2.908257
12 Angry Men (1957)                       4.344000
187 (1997)                                3.024390
                                             ...
Young Guns II (1990)                      2.772727
Young Poisoner's Handbook, The (1995)     3.341463
Zeus and Roxanne (1997)                   2.166667
unknown                                   3.444444
Á köldum klaka (Cold Fever) (1994)        3.000000
Name: rating, Length: 1664, dtype: float64
```

```
title
Yankee Zulu (1994)                          1.0
Mighty, The (1998)                          1.0
Mille bolle blu (1993)                      1.0
Modern Affair, A (1995)                     1.0
New Age, The (1994)                         1.0
                                           ...
Great Day in Harlem, A (1994)               5.0
Saint of Fort Washington, The (1993)        5.0
Prefontaine (1997)                          5.0
They Made Me a Criminal (1939)              5.0
Marlene Dietrich: Shadow and Light (1996)   5.0
Name: rating, Length: 1664, dtype: float64
```

Star wars ima najvec ocen, ker je najbolj popularen.

Pogost proble, je da če ne sortiramo po popularnosti in oceni hkrati … Temveč samo po oceni, potem dobimo film »ki ga je ocenil samo 1 človek, a z oceno 5«

```
## Kaj naredimo, preizkusi ?

# Tabela filmov
#df.groupby('title')['rating'].count()

#df.groupby('title')['rating'].count().sort_values(ascending=False)

#
#df.groupby('title')['rating'].mean()

#
#df.groupby('title')['rating'].mean().sort_values()
```

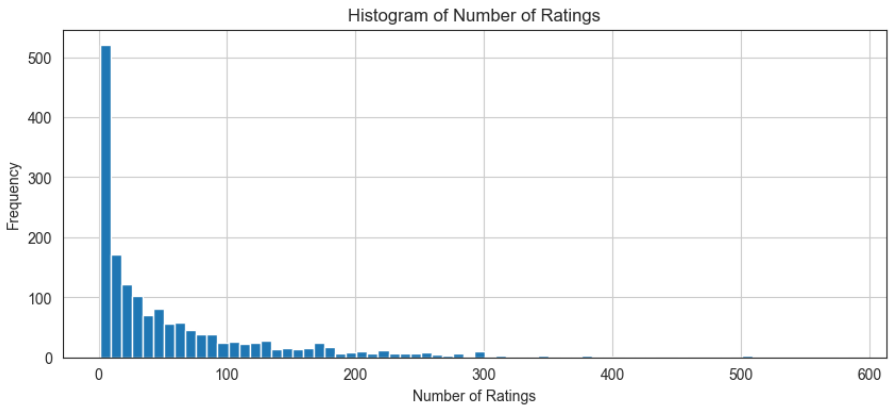
1.5 Podatki o filmih

```
# tabela povpr ocen filmov
ratings = pd.DataFrame(df.groupby('title')['rating'].mean())
ratings.head()
ratings['num of ratings'] = pd.DataFrame(df.groupby('title')['rating'].count())
ratings.head()
```

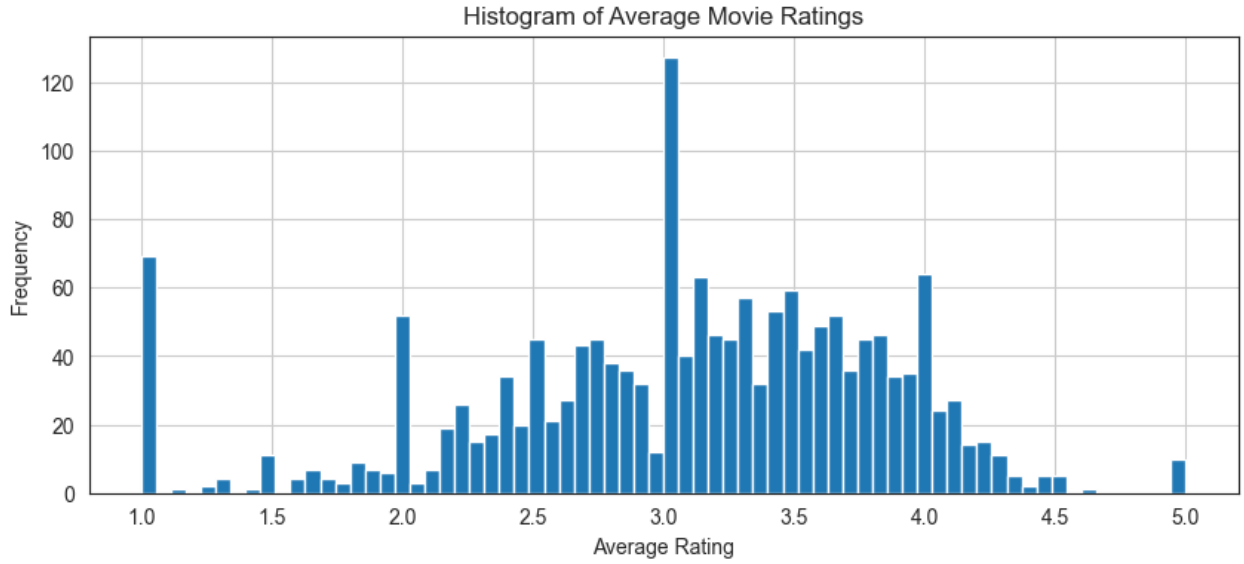
title	rating	num of ratings
'Til There Was You (1997)	2.333333	9
1-900 (1994)	2.600000	5
101 Dalmatians (1996)	2.908257	109
12 Angry Men (1957)	4.344000	125
187 (1997)	3.024390	41

1.6 Naloga : Izriši histograme

Izriši histogram števila ocen:



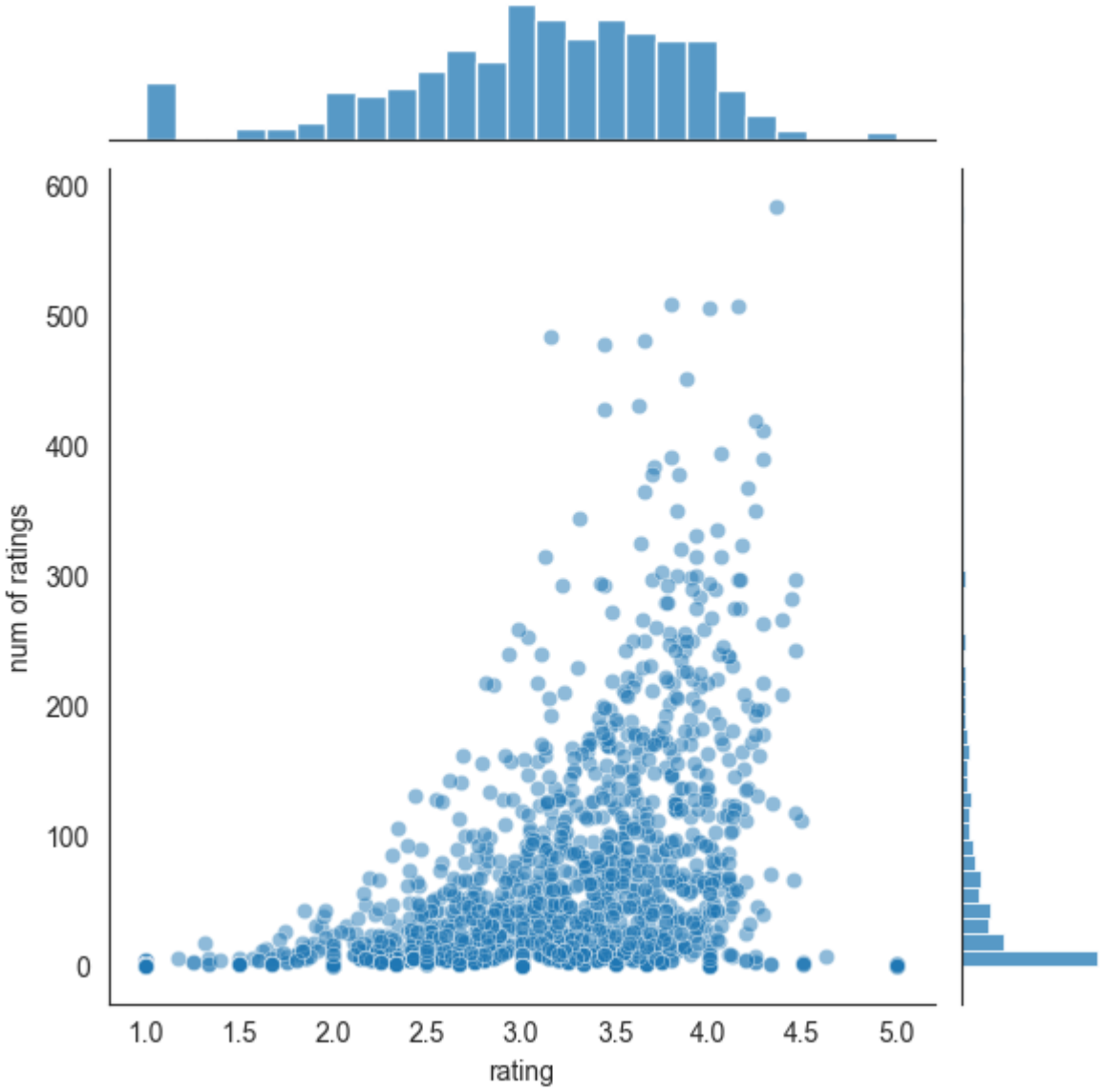
Izriši histogram povprečne ocene filmov (rating)



Izriši skupni histogram obeh:

```
sns.jointplot(x='rating',y='num of ratings',data=ratings,alpha=0.5)
```

Kaj nam skupni histogram pove o povprečnih ocenah filmov ?



2 Priporočanje podobnih filmov

```
moviemat = df.pivot_table(index='user_id',columns='title',values='rating')
moviemat.head()
```

Kaj so podatki v tej tabeli ?

title	rating	num of ratings
Star Wars (1977)	4.359589	584
Contact (1997)	3.803536	509
Fargo (1996)	4.155512	508
Return of the Jedi (1983)	4.007890	507
Liar Liar (1997)	3.156701	485
English Patient, The (1996)	3.656965	481
Scream (1996)	3.441423	478
Toy Story (1995)	3.878319	452
Air Force One (1997)	3.631090	431
Independence Day (ID4) (1996)	3.438228	429

```
ratings.sort_values('num of ratings',ascending=False).head(10)
```

Kaj naredi koda, kaj predstavlja rezultat?

Rezultat bo natisnjen rezultat 10 najboljših filmov z najvišjim številom ocen v podatkovnem okviru ocen. To pomaga prepoznati najbolj ocenjene filme v zbirki podatkov

2.1 Podobnost filmov

```
starwars_user_ratings = moviemat['Star Wars (1977)']
starwars_user_ratings.head()
```

```
similar_to_starwars = moviemat.corrwith(starwars_user_ratings)

similar_to_starwars
```

Kaj smo dobili, kaj vrne metoda corrwith, in kakšen algoritem uporablja za izračun?

Pojasnilo: metoda corrwith vrne parno korelacijo vrstic ali stolpcev dveh objektov DataFrame.

Privzeto uporablja Pearsonov korelacijski koeficient.

```
corr_starwars = pd.DataFrame(similar_to_starwars,columns=['Correlation'])
corr_starwars.dropna(inplace=True)
corr_starwars.head()
```

```
corr_starwars.sort_values('Correlation',ascending=False).head(100)
```

Izpis: kaj smo dobili ?

```
user_id
0    5.0
1    5.0
2    5.0
3    NaN
4    5.0
Name: Star Wars (1977), dtype: float64
title
101 Dalmatians (1996)                    0.211132
12 Angry Men (1957)                      0.184289
187 (1997)                               0.027398
2 Days in the Valley (1996)              0.066654
20,000 Leagues Under the Sea (1954)      0.289768
                                             ...
Wyatt Earp (1994)                        0.059560
Young Frankenstein (1974)                0.192589
Young Guns (1988)                        0.186377
Young Guns II (1990)                     0.228615
Young Poisoner's Handbook, The (1995)   -0.007374
Length: 1144, dtype: float64
                                    Correlation
title
101 Dalmatians (1996)                  0.211132
12 Angry Men (1957)                    0.184289
187 (1997)                             0.027398
...
E.T. the Extra-Terrestrial (1982)      0.303619
```

2.2 Obdelamo podatke o podobnosti filmov

```
corr_starwars = corr_starwars.join(ratings['num of ratings'])
corr_starwars.head()
```

```
index = corr_starwars['num of ratings']>10
index.head()
```

```
corr_starwars[index].sort_values('Correlation',ascending=False).head()
```

Vstavi izpis:

```
                                Correlation  num of ratings
title
101 Dalmatians (1996)              0.211132             109
12 Angry Men (1957)                0.184289             125
187 (1997)                         0.027398              41
2 Days in the Valley (1996)        0.066654              93
20,000 Leagues Under the Sea (1954)  0.289768            72
title
101 Dalmatians (1996)               True
12 Angry Men (1957)                 True
187 (1997)                          True
2 Days in the Valley (1996)         True
20,000 Leagues Under the Sea (1954)  True
Name: num of ratings, dtype: bool
                            Correlation  num of ratings
title
Star Wars (1977)               1.000000             584
That Old Feeling (1997)        0.750000              11
Empire Strikes Back, The (1980)  0.748353            368
American Buffalo (1996)        0.722592              11
Return of the Jedi (1983)      0.672556             507
```

Komentiraj, kaj smo naredili, in kaj smo dobili kot rezultat ?

Stolpec 'število ocen' smo združili z ratings DataFrame v corr_starwars DataFrame.

Nato smo ustvarili logični indeks za filtriranje filmov z več kot 10 ocenami. Končno smo razvrstili filtrirani DataFrame po stolpcu 'Korelacija' v padajočem vrstnem redu in prikazali najboljše rezultate.

2.3 Naloga

Ponovi postopek generiranja podobnih filmov za svoj izbrani film. Kopiraj ustrezno kodo iz primera.

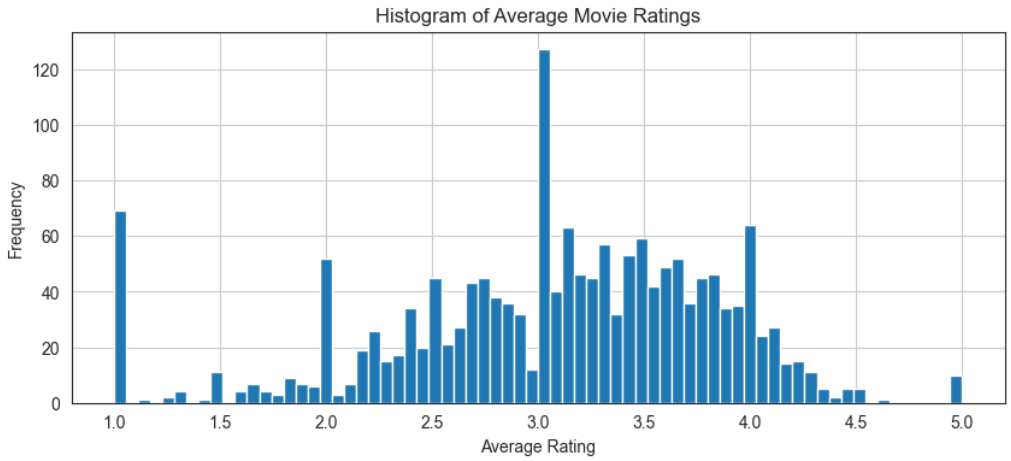
- Izberi si svoj film (z najmanj 50 ocenami)
- Izriši histogram njegovih ocen
- Izpiši vsaj 10 najbolj podobnih filmov.

Koda:

```python
#2.1
# DO NOT Get user ratings for Star Wars (1977)
starwars_user_ratings = moviemat['Toy Story (1995)']
print(starwars_user_ratings.head())

# Filter out movies with less than 10 ratings
movie_stats = ratings['num of ratings']
popular_movies = movie_stats[movie_stats >= 10].index
filtered_moviemat = moviemat[popular_movies]
```

Končni rezultat:




```
                                Correlation  num of ratings
title
101 Dalmatians (1996)              0.232118             109
12 Angry Men (1957)                0.334943             125
187 (1997)                         0.651857              41
2 Days in the Valley (1996)        0.162728              93
20,000 Leagues Under the Sea (1954)    0.328472          72
title
101 Dalmatians (1996)               True
12 Angry Men (1957)                 True
187 (1997)                          True
2 Days in the Valley (1996)         True
20,000 Leagues Under the Sea (1954)    True
Name: num of ratings, dtype: bool
                                Correlation  num of ratings
title
Phantoms (1998)                    1.000000             13
Critical Care (1997)               1.000000             11
Toy Story (1995)                   1.000000            452
Daytrippers, The (1996)            0.883883             15
Stalingrad (1993)                  0.867110             12
Gone Fishin' (1997)                0.818182             11
Afterglow (1997)                   0.816497             18
Perez Family, The (1995)           0.761165             14
Transformers: The Movie, The (1986)    0.753673         32
Unzipped (1995)                    0.739940             11
```

Komentar: Na čem temelji izračun podobnosti, in ali so rezultati smiselni glede na vsebino (naslov) filma, .. ?

Izračun podobnosti temelji na edinstvenem številu uporabnikov in filmov ter oceni filmov. Če bi imeli več podatkov bi naredili boljši sistem npr dodali še žaner itd. (genre). Ob pomanjkanju podatkov smo primorani delati preprostejše modele.