

Sistemi preporuke filmova

Djordje Karišić

Februar 2022

1 Uvod

Potreba za pružanjem izvrsne preporuke korisnicima internet platformi nikad nije bila veća. Internet platforme za trgovinu, društvene mreže, sisteme za plaćanje, prodavnice aplikacija itd. moraju imati besprekoran sistem preporuka kako bi uštedeli korisniku vreme, i privukli ga primamljivim ponudama po njegovom ukusu. Od takvih sistema zavisi dosta korisničko iskustvo. Dakle, takvi sistemi moraju biti jako precizni. Naravno, kako bismo obezbedili preciznost jednog sistema, potrebno je koristiti odgovarajuću strategiju, i imati dovoljnu količinu relevantnih podataka nad kojima možemo primeniti strategiju.

Kako bismo preporučili korisniku odgovarajuće stavke, možemo koristiti neku od sledeće tri strategije:

1. Saradnički sistemi za preporuku
2. Sistem preporuka zasnovan na sadržaju
3. Hibridni sistem preporuka

Govorićemo o sistemima preporuka zasnovanim na sadržaju.

2 Sistem preporuke zasnovan na sadržaju

Cilj ovog sistema je, da na osnovu sličnosti izmeu elemenata i korisničkog ukusa, zaključi koji element je preporučljiv. Na primer, ako korisnik oceni filmove X,Y sa najvećom ocenom, znamo da, film Z koji je sličan po npr. žanrovima filmovima X i Y, može biti dobar predlog. Što više korisnik reaguje sa elementima, pretpostavka će biti preciznija. Važno je da, pored aktivnog korisnika, imamo dosta atributa za jedan element. Na primer, ako imamo situaciju kao u prošlom primeru, ali ako znamo da postoji film M koji je dovoljno približan po žanrovima filmovima X i Y, i koji ima zajedničkog režisera sa tim filmovima, možemo doneti razumnu pretpostavku da će taj film odgovarati korisniku. Dakle, atributi igraju dosta veliku ulogu u povećanju preciznosti ovih sistema. Kako korisnik nastavlja interakciju sa programom, možemo primetiti i neke šablone, npr. korisnik gleda filmove u kojima isključivo glumi glumac Q, ili samo one koje je režirao režiser P, pa čak i samo one koji su na R jeziku.

3 Implementacija sistema preporuka zasnovanog na sadržaju nad bazom filmovima

3.1 Baza podataka

Zaključak iz prethodne sekcije je da, kako bismo imali dobar sistem, moramo imati adekvatnu bazu podataka, sa dovoljno atributa nad kojima možemo izvršiti određene operacije i grupisanje.

Dobra baza podataka, u slučaju preporuke filmova, treba imati dovoljno filmova sa različitim žanrovima, režiserima, glumcima, kao i ostalih elemenata.

Baza koja je upotrebljena u ovom projektu je MovieLens(20M), koja sadrži oko 27000 filmova, 20 miliona ocena korisnika i oko 38 hiljada korisničkih oznaka.

<https://www.kaggle.com/grouplens/movielens-20m-dataset>

3.2 Žanrovi

Postoji više metoda kako bi se odredila sličnost izmeu filmova na osnovu žanrova. Razlika u preciznosti tih metoda nije velika na skupu podataka koji koristimo. Ideja algoritma koji je upotrebljen u ovom projektu je sledeća:

Figure 1: Prvi korak je doći od ovoga

```
1  "movieId","title","genres"
2  1,"Toy Story (1995)","Adventure|Animation|Children|Comedy|Fantasy"
3  2,"Jumanji (1995)","Adventure|Children|Fantasy"
4  3,"Grumpier Old Men (1995)","Comedy|Romance"
5  4,"Waiting to Exhale (1995)","Comedy|Drama|Romance"
6  5,"Father of the Bride Part II (1995)","Comedy"
7  6,"Heat (1995)","Action|Crime|Thriller"
8  7,"Sabrina (1995)","Comedy|Romance"
9  8,"Tom and Huck (1995)","Adventure|Children"
10 9,"Sudden Death (1995)","Action"
11 10,"GoldenEye (1995)","Action|Adventure|Thriller"
12 11,"American President, The (1995)","Comedy|Drama|Romance"
13 12,"Dracula: Dead and Loving It (1995)","Comedy|Horror"
14 13,"Balto (1995)","Adventure|Animation|Children"
15 14,"Nixon (1995)","Drama"
16 15,"Cutthroat Island (1995)","Action|Adventure|Romance"
17 16,"Casino (1995)","Crime|Drama"
18 17,"Sense and Sensibility (1995)","Drama|Romance"
19 18,"Four Rooms (1995)","Comedy"
20 19,"Ace Ventura: When Nature Calls (1995)","Comedy"
21 20,"Money Train (1995)","Action|Comedy|Crime|Drama|Thriller"
22 21,"Get Shorty (1995)","Comedy|Crime|Thriller"
23 22,"Copycat (1995)","Crime|Drama|Horror|Mystery|Thriller"
24 23,"Assassins (1995)","Action|Crime|Thriller"
25 24,"Powder (1995)","Drama|Sci-Fi"
26 25,"Leaving Las Vegas (1995)","Drama|Romance"
27 26,"Othello (1995)","Drama"
28 27,"Now and Then (1995)","Children|Drama"
29 28,"Persuasion (1995)","Drama|Romance"
30 29,"City of Lost Children, The (Cité des enfants perdus, La) (1995)","Adventure|Drama|Fantasy|Mystery|Sci-Fi"
31 30,"Shanghai Triad (Yao a yao yao dao walpo qiao) (1995)","Crime|Drama"
32 31,"Dangerous Minds (1995)","Drama"
```

Figure 2: Do ovoga

movieId	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	Drama	Fantasy	Film-Noir	Horror
1	0	1	1	1	0	0	0	1	0	0	0
2	0	1	0	1	0	0	0	1	0	0	0
3	0	0	0	1	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	1	0	0	0	0	0	0	0
6	0	0	0	1	0	0	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0	0
8	0	0	0	1	0	0	0	0	0	0	0
9	0	1	0	1	0	0	0	0	0	0	0
10	0	1	0	0	0	0	0	0	0	0	0
11	0	1	0	0	0	0	0	0	0	0	0
12	0	0	0	1	0	0	0	0	0	0	0
13	0	0	0	1	0	0	0	0	0	0	0
14	0	0	0	1	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0

Kao što se vidi sa druge slike, svakom filmu smo dodelili kolone za svaki od mogućih žanrova. Ako dati film sadrži neki žanr, vrednost u koloni tog žanra na indeksu tog filma će biti 1, inače 0. Kada se završi dodela svih žanrova filmovima imaćemo, umesto String žanra, jasnu Boolean vrednost žanrova za svaki od filmova, što nam olakšava posao.

Sledeći korak bi podrazumevao nalaženje sličnosti svakog filma sa svakim drugim. Ovo možemo lako ostvariti sa novom matricom M , i to tako što ćemo primeniti sledeću formulu:

$$M = M \cdot M^T \quad (1)$$

Kada izvršimo ovu operaciju nad matricom M , dobićemo odnos filmova.

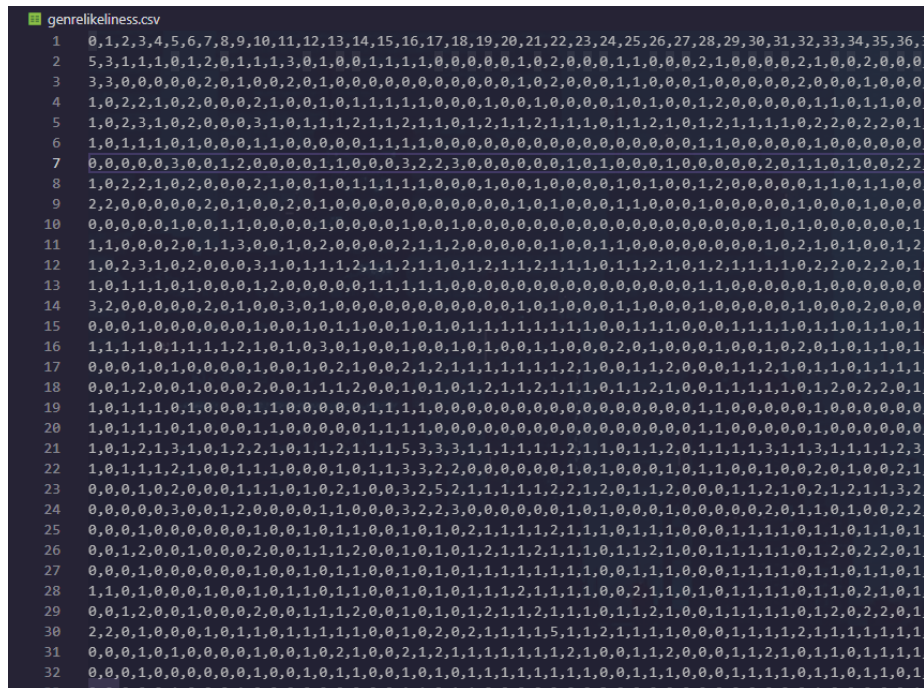


Figure 3: Kolone i redovi su obeleženi indeksima filmova

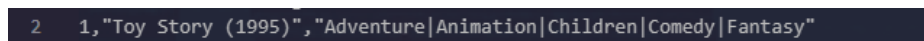


Figure 4: Film sa indeksom 0

Sa slike 3 vidimo da film sa indeksom 0 ima 5 sličnih žanrova sa samim sobom, slika 4 nam to potvrđuje.

Sada kada imamo matricu sličnosti, ostalo je još samo da korisnik unese ocene o filmovima kako bismo našli sličan film po žanrovima.

3.3 Modifikacija ocene filmova

Ocene jednog filma su glavni atribut kada je u pitanju sortiranje filmova. Meutim, ako je neki film ocenilo samo nekoliko korisnika, taj film je nepouzdan za predlog. Kako rešiti ovaj problem?

Ponderisani rejting filmova

Iza ove metode stoji IMDB i koristi se aktivno u kalkulaciji rejtinga filmova sa njihovog sajta.

Ideja iza ove formule je pronaći način kako oceniti dati film, na osnovu popularnosti i ocena, i dodeliti težinu svakom od tih atributa.

```
def IMDBeq(x):  
    v = x['popularityScore']  
    R = x['ratingScore']  
    m = x['popularityScore'].quantile(0.90)  
    C = x['ratingScore'].mean()  
  
    return (v/(v+m) * R) + (m/(m+v) * C)
```

Figure 5: Formula

Potrebno je sledeće:

1. popularityScore-Broj ljudi koji je ocenio film
2. ratingScore-Prosečna ocena filma
3. popularityScore.quantile(0.90)-Vrednost popularnosti u tačno 90 procenata od maksimalne vrednosti
4. ratingScore.mean()-Prosečna ocena svih filma

```
popularity['frequency']=(ratingsData['movieId'].value_counts())
```

Figure 6: Kako naći broj korisnika koji je ocenio neki film, brojimo ponavljanja Id-jeva svakog filma u bazi recenzija

```
means['rating'] = ratingsData.groupby(['movieId'])['rating'].mean()
```

Figure 7: Kako naći prosečnu ocenu filma, grupišemo po Id-jevima, i računamo srednju vrednost ocene za svaki Id

Ostale dve stavke se mogu izračunati kao što je dato u slici 5.
Kada se ova funkcija primeni nad bazom podataka, dobiće se sledeće:

```

5
Унети број филмова које желите да вам се прикажу...100
movieId Action Adventure ... ratingScore title IMDBScore
315 318 0 0 ... 4.446990 Shawshank Redemption, The (1994) 4.260132
843 858 0 0 ... 4.364732 Godfather, The (1972) 4.125101
49 50 0 0 ... 4.334372 Usual Suspects, The (1995) 4.123911
523 527 0 0 ... 4.310175 Schindler's List (1993) 4.115059
293 296 0 0 ... 4.174231 Pulp Fiction (1994) 4.044189
587 593 0 0 ... 4.177057 Silence of the Lambs, The (1991) 4.039772
257 260 1 1 ... 4.190672 Star Wars: Episode IV - A New Hope (1977) 4.033004
1173 1198 1 1 ... 4.219009 Raiders of the Lost Ark (Indiana Jones and the... 4.023626
2486 2571 1 0 ... 4.187186 Matrix, The (1999) 4.022667
2873 2959 1 0 ... 4.227123 Fight Club (1999) 4.018154
1171 1196 1 1 ... 4.188202 Star Wars: Episode V - The Empire Strikes Back... 4.006822
1195 1221 0 0 ... 4.275641 Godfather: Part II, The (1974) 3.986345
1169 1193 0 0 ... 4.248879 One Flew Over the Cuckoo's Nest (1975) 3.984267
2772 2858 0 0 ... 4.155934 American Beauty (1999) 3.981338
895 912 0 0 ... 4.258327 Casablanca (1942) 3.952612
1113 1136 0 1 ... 4.174146 Monty Python and the Holy Grail (1975) 3.949909
1172 1197 1 1 ... 4.176732 Princess Bride, The (1987) 3.949612
602 608 0 0 ... 4.112359 Fargo (1996) 3.943383
4897 4993 0 1 ... 4.137925 Lord of the Rings: The Fellowship of the Ring,... 3.943262
4132 4226 0 0 ... 4.178547 Memento (2000) 3.939967
737 750 0 0 ... 4.247287 Dr. Strangelove or: How I Learned to Stop Worr... 3.936579
352 356 0 0 ... 4.029000 Forrest Gump (1994) 3.922922
7041 7153 1 1 ... 4.142382 Lord of the Rings: The Return of the King, The... 3.920971
1187 1213 0 0 ... 4.183633 Goodfellas (1990) 3.919604
108 110 1 0 ... 4.042534 Braveheart (1995) 3.914284
4877 4973 0 0 ... 4.197072 Amelie (Fabuleux destin d'Amélie Poulain, Le) ... 3.914018
537 541 1 0 ... 4.133706 Blade Runner (1982) 3.909871
5853 5952 0 1 ... 4.107521 Lord of the Rings: The Two Towers, The (2002) 3.907318
46 47 0 0 ... 4.053493 Seven (a.k.a. Se7en) (1995) 3.899075
12525 58559 1 0 ... 4.220129 Dark Knight, The (2008) 3.895715
887 904 0 0 ... 4.271334 Rear Window (1954) 3.893314
1944 2028 1 0 ... 4.064417 Saving Private Ryan (1998) 3.888520
2676 2762 0 0 ... 4.054999 Sixth Sense, The (1999) 3.887954
2244 2329 0 0 ... 4.151208 American History X (1998) 3.877763
1184 1210 1 1 ... 4.004622 Star Wars: Episode VI - Return of the Jedi (1983) 3.870700
1067 1089 0 0 ... 4.089361 Reservoir Dogs (1992) 3.865214
453 457 0 0 ... 3.985690 Fugitive, The (1993) 3.861754
109 111 0 0 ... 4.110576 Taxi Driver (1976) 3.860376
1564 1617 0 0 ... 4.083377 L.A. Confidential (1997) 3.856955
891 908 1 1 ... 4.233538 North by Northwest (1959) 3.850560
1182 1208 1 0 ... 4.108900 Apocalypse Now (1979) 3.850032
1188 1214 0 0 ... 4.041784 Alien (1979) 3.849019
2239 2324 0 0 ... 4.175837 Life Is Beautiful (La Vita è bella) (1997) 3.848123
7249 7361 0 0 ... 4.105628 Eternal Sunshine of the Spotless Mind (2004) 3.842824
1645 1704 0 0 ... 4.022517 Good Will Hunting (1997) 3.831061
209 293 1 0 ... 4.050574 Léon: The Professional (a.k.a. The Professional... 3.830020
1224 1252 0 0 ... 4.190673 Chinatown (1974) 3.828947
1263 1291 1 1 ... 4.007593 Indiana Jones and the Last Crusade (1980) 3.827021
583 589 1 0 ... 3.931954 Terminator 2: Judgment Day (1991) 3.824524
906 923 0 0 ... 4.130443 Citizen Kane (1941) 3.819062

```

Figure 8: Kada se primeni prethodna formula, dobijaju se identične ocene kao u IMDB bazi podataka, ova lista liči na IMDB listu top 100 filmova

3.4 Prikaz najpopularnijih filmova

Nekad je najjednostavnije prikazati najpopularnije filmove korisniku, jer postoji dovoljno velika šansa da će se bar jedan njemu svideti.

```
def getMostPopular(): #vraca na
    pd.set_option("display.max_colwidth", None)
    output=pd.read_csv('ratingsFiltered.csv',index_col=0)
    genres=pd.read_csv('movie.csv')
    output['genres']=genres['genres']
    output['rating']=round(output['rating'],2)
    x=int(input('Колико најгледанијих филмова желите да видите?'))
    print(output.head(x))
```

Figure 9: Jednostavan kod, sortira bazu i ispisuje u odnosu na korisnički unos

```
Колико најгледанијих филмова желите да видите?10
frequency  movieId  rating  title
293      67310      296   4.17   Pulp Fiction (1994)
352      66172      356   4.03   Forrest Gump (1994)
315      63366      318   4.45   Shawshank Redemption, The (1994)
587      63299      593   4.18   Silence of the Lambs, The (1991)
476      59715      480   3.66   Jurassic Park (1993)
257      54502      260   4.19   Star Wars: Episode IV - A New Hope (1977)
108      53769      110   4.04   Braveheart (1995)
583      52244      589   3.93   Terminator 2: Judgment Day (1991)
2486     51334     2571   4.19   Matrix, The (1999)
523      50054      527   4.31   Schindler's List (1993)
```

Figure 10: Ovo je izlaz

3.5 Sličnost na osnovu korisničkih oznaka

Prethodno je napisano, da se meu bazama koje se koriste u ovom programu nalazi i baza sa korisničkim oznakama, tj. tagovima. Tagovi su poput kratkih korisničkih komentara, sadrže uglavnom reč ili dve koje ih podsećaju na taj film, npr. Pulp Fiction ima tagove Quentin Tarantino, cult film, violence... Što nam daje dosta materijala na osnovu kojeg možemo doneti pretpostavku. Recimo, da korisnik oceni filmove i da naemo filmove slične po tagovima filmovima koje je on najvećom ocenom ocenio, ili da dozvolimo korisniku da unese neku reč ili par, pa na osnovu toga da naemo filmove sa tim tagom, ili oba. Ovo je adekvatno rešenje problema baze, manjak kolona za režisere i glumce.

Ideja za rešenje ovog problema je ta da nakon korisničkog unosa, izdvojimo Id-jeve filmova ocenjenih najvišom ocenom, i pretražimo bazu sa tagovima, kako bismo našli film sa najviše istih tagova kao filmovi koje je smo izdvojili. Moguće koristiti fuzzywuzzy biblioteku za bolju pretragu tagova. Ako želite da pogledate kod kliknite na ovaj broj 15.

```
Читам матрицу релевантности...
(1-5 | било шта друго ако нисте гледали овај филм)
Pulp Fiction (1994)
5
(1-5 | било шта друго ако нисте гледали овај филм)
Forrest Gump (1994)
4
(1-5 | било шта друго ако нисте гледали овај филм)
Shawshank Redemption, The (1994)
3
(1-5 | било шта друго ако нисте гледали овај филм)
Silence of the Lambs, The (1991)
5
(1-5 | било шта друго ако нисте гледали овај филм)
Jurassic Park (1993)
3
Да ли желите оцените још филмова? (Y)/(било шта друго)
z
```

Figure 11: Na osnovu ovih ocena...


```

Колико сличних филмова по ознака желите да видите?...10

Филмови налик онима који Вам се свиђају, на основу корисничких ознака
=====
movieId      title      ...      rating \
3      2959      Fight Club (1999) ... 4.227123
1      1089      Reservoir Dogs (1992) ... 4.089361
4      4226      Memento (2000) ... 4.178547
8      32587      Sin City (2005) ... 3.874201
5      4878      Donnie Darko (2001) ... 4.015990
0      47      Seven (a.k.a. Se7en) (1995) ... 4.053493
2      1206      Clockwork Orange, A (1971) ... 3.995303
6      6874      Kill Bill: Vol. 1 (2003) ... 3.901338
7      7361      Eternal Sunshine of the Spotless Mind (2004) ... 4.105628
9      79132      Inception (2010) ... 4.156172

popularity
3      40106
1      27635
4      30443
8      15481
5      18731
0      43249
2      24694
6      21746
7      22352

```

Figure 12: ...Izlaz je ovakav.

3.6 Sličnost na osnovu korisničkog unosa oznake

Kao što je prethodno napisano, moguće je korisniku dati slobodu da unese neki termin, preko kojeg se može pretražiti baza podataka tagova i naći film(ove) sa takvom oznakom.

	movieId	title	genres	tagScore
2	296	Pulp Fiction (1994)	Comedy Crime Drama Thriller	217
13	68157	Inglourious Basterds (2009)	Action Drama War	108
4	1089	Reservoir Dogs (1992)	Crime Mystery Thriller	92
7	6874	Kill Bill: Vol. 1 (2003)	Action Crime Thriller	89
8	7438	Kill Bill: Vol. 2 (2004)	Action Drama Thriller	73
9	32587	Sin City (2005)	Action Crime Film-Noir Mystery Thriller	60
14	99114	Django Unchained (2012)	Action Drama Western	52
11	53519	Death Proof (2007)	Action Adventure Crime Horror Thriller	31
10	52281	Grindhouse (2007)	Action Crime Horror Sci-Fi Thriller	28
1	70	From Dusk Till Dawn (1996)	Action Comedy Horror Thriller	27
5	1729	Jackie Brown (1997)	Crime Drama Thriller	25
0	18	Four Rooms (1995)	Comedy	14
3	555	True Romance (1993)	Crime Thriller	8
6	3979	Little Nicky (2000)	Comedy	4
12	61073	Hell Ride (2008)	Action Drama Thriller	2

Figure 13: Za uneseni termin 'Quentin Tarantino' imamo ovaj izlaz

Nakon što korisnik unese termin, potrebno je da taj termin prebacimo u lowercase, kako bismo preciznije našli tagove koji odgovaraju. Takoe, tagove iz baze tagova moramo prebaciti u lowercase.

Nakon toga, sortiramo matricu po Id-jevima, i brojimo koliko se tagova koji odgovaraju nalaze za svaki Id. Ta vrednost će predstavljati 'Tag Score' tj. vrednost koja će govoriti o tome koliko je uneti termin relevantan za svaki od filmova koje sadrže taj termin. Uz pomoć te vrednosti možemo sortirati finalnu matricu. Linije koda koje se nalaze u programu pored ovih navedenih su tu radi reorganizacije indeksa matrice kako bi ta matrica mogla da se uporedi sa drugom. Ako želite da pogledate kod kliknite na ovaj broj 14

```

402
403 def getMoviesByTagInput():
404     print("Унесите неку карактеристичну реч за филм-име, глумца, режисера, мотив филма...")
405     q=str(input(" нпр: 007, marvel, scary, horror, plot twist, Quentin Tarantino, Leonardo DiCaprio..."))
406
407     tagDf=pd.read_csv('tag.csv')
408     moviesDf=pd.read_csv('movie.csv')
409     #tagDf['worthy'] = tagDf.apply(lambda x: stringIsSimilar(tagDf,q), axis=1)
410     tagDf=tagDf.loc[tagDf['tag'].str.lower()==q.lower()]
411     tagDf=tagDf.sort_values(by='movieId',ascending=True)
412     tagDf=tagDf.groupby(['movieId'])['tag'].count()
413     #tagDf=tagDf.sort_values(ascending=False)
414     print("Нађено ",len(tagDf.index)," филмова са датом ознаком")
415     x=int(input("Колико таквих филмова желите да прикажете?..."))
416
417     printable=moviesDf.loc[moviesDf['movieId'].isin(tagDf.index)]
418     printable.index=printable['movieId']
419     printable['tagScore']=tagDf.values
420     printable=printable.reset_index(drop=True)
421     printable=printable.sort_values('tagScore',ascending=False)
422     print(printable.head(x))
423

```

Figure 14: Kod

410. U Series-u tagDf tražimo vrednosti koje su, kada se prebace u lowercase, jednake terminu koji je korisnik uneo. tagDf sada sadrži samo tagove koji odgovaraju terminu.
411. Sortiramo vrednosti po Id-ju.
412. Grupišemo po Id-ju i brojimo koliko svaki Id ima tagova.
417. U liniji 408 smo napravili DataFrame koji sadrži neobraene podatke o filmovima. Ovo će biti DataFrame koji ćemo ispisati nakon modifikacije. Pravimo novi DataFrame printable koji će sadržati samo filmove čiji Id se nalazi u novom tagDf Series-u. tagDF je Series, što znači da vrednost njegovog indeksa odgovara Id-ju filмова (linija 412)
418. Menjamo indeks printable DataFrame-a kako bismo 'nalepili' vrednosti iz tagDf
419. tagDf je series, što znači da tagDf.values sadrži tagScore
420. Vraćamo indeks
421. Sortiramo printable po tagScore parametru

3.7 Kod za sekciju 3.5

```

343 def getSimilarByTag():                                     #ova funkcija je za nalazenje slicnih filmova po
344     pd.set_option('display.max_rows', 500)
345     tags=pd.read_csv('tag.csv')
346     tags=tags.sort_values(by=['movieId'],ascending=True)
347
348     getUserPref()
349     user=pd.read_csv('userPreferred.csv')
350
351     idm=user.loc[user['rating']==5,'movieId']
352     idm=idm.to_numpy()                                     #izdvajamo filmove ocenjene sa 5...
353     #novo
354     tags['Status']=tags['movieId'].isin(idm)               #status oznacava filmove koji su ocenjeni sa 5
355
356     tagged=tags.loc[tags['Status'],'tag']
357     taggedArray=tagged.to_numpy()                          #niz tagova kod kojih je status==True, tj. filmova koji
358     #novo
359     tags['isSimilar']=tags['tag'].isin(taggedArray)        #True ako se tag filma poklapa sa bilo kojim tagom prot
360     final=tags.groupby(['movieId'])['isSimilar'].sum()    #sabiramo koliko slicnih tagova svaki film ima sa -||-
361     final=final.sort_values(ascending=False)              #sortiramo seriju tako da je film sa najvise tagova pri
362
363     final=final.drop(final.loc[final.index.isin(idm)].index) #posto ce svaki prethodno ocenjeni da ima najvise slicnih
364
365     x=int(input("Кликните лични филм на знак жelite да видите?..."))
366     recommend=final.head(x)
367     rec=pd.DataFrame()                                     #series--->dataframe
368     rec['movieId']=recommend.index
369     rec['value']=recommend.values
370     stuff=pd.read_csv('movie.csv')
371
372     stuff['recommendable']=stuff['movieId'].isin(rec['movieId']) #ako movieId postoji u rec znaci da se treba
373     #stuff['recommendableScore']=rec.loc[rec['movieId'].isin(stuff['movieId']),'value'] #recommendable score je zapravo br
374     stuff=stuff.loc[stuff['recommendable']==True]            #uzmemo samo filmove koji su za
375     new_df = stuff.merge(rec[['movieId',"value"]], on="movieId", how="left") #spojimo matricu preporucljivih
376     new_df["recommendableScore"] = new_df["value"].fillna(0) #ako nema ni jedan slican tag sa
377     new_df=new_df.drop(columns=['value'])                   #value kolona nam je sada nepotrebna
378
379     movies=pd.read_csv('ratingsFiltered.csv',index_col=0)
380
381     movies.index=movies['movieId']                         #menjamo indeks kako bismo lakse stavili vredno
382     new_df.index=new_df['movieId']
383
384     new_df['rating']=movies['rating']                       #Lakse uz merge, ali smo vec zamenili indekse i
385     new_df['popularity']=movies['frequency']
386     new_df=new_df.reset_index(drop=True)
387
388     new_df=new_df.sort_values(by=['recommendableScore'],ascending=False) #od prvog najpreporucljivijeg do 'x'-tog
389     print("\n\nФилмви налик аниме који Вам се свиђају, на основу корисничких оцена")
390     print("=====")
391     print(new_df)

```

Figure 15: Kod

- 351. Uzimamo indekse filmova koje je korisnik ocenio sa ocenom 5.
- 354. Označavamo sve filmove u tags čiji Id se nalazi u listi idm
- 356. U DataFrame tagged pakujemo sve tagove elemenata iz tags čiji je status True
- 359. Pravimo novu kolonu u u DataFrame tags i punimo je Boolean-ima koji ukazuju da li se dati tag iz DF tags nalazi u nizu tagova koje je korisnik ocenio.
- 363. Odbacujemo sve vrste Series-a čiji se indeks nalazi u gornje navedenom nizu

368. i 369. Dodajemo u rec DataFrame vrednosti x najrelevantnijih vrsti.
370. i 372. Pravimo novi DataFrame koji će podatke uzeti iz movie.csv, koji ćemo obraditi i štampati na izlazu. Nova kolona recommendable govori o tome da li vrsta sadrži Id filma koji je relevantan
375., 376. i 377. Spajamo novi DataFrame sa DataFrame-om koji sadrži skor svakog filma, i stavljamo u recommendableScore kolonu