|  |  |  |
| --- | --- | --- |
|  | UNIVERZITET U NIŠU  ELEKTRONSKI  FAKULTET |  |

**IMPLEMENTACIJA SISTEMA ZA PREPORUKU U PROGRAMSKOM JEZIKU PYTHON**

**SEMINARSKI RAD**

Predmet: Web mining

|  |  |
| --- | --- |
| Student: | Mentor: |
|  |  |
| Tina Radenković , br. ind. 1128 | Doc. dr Miloš Bogdanović |

Niš, februar 2021. god.

SADRŽAJ

[1. Uvod 3](#_Toc64987065)

[2. Kolaborativno filtriranje podataka 4](#_Toc64987066)

[3. Analiza projekta 5](#_Toc64987067)

[3.1. Analiza skupa podataka 5](#_Toc64987068)

[3.2. Implementacija sistema za preporuku filmova 7](#_Toc64987069)

[4. Analiza rezultata 9](#_Toc64987070)

[5. Zaključak 11](#_Toc64987071)

[6. Reference 12](#_Toc64987072)

# Uvod

Količina informacija u savremenom životu je postala toliko velika da je teško i zahtveno obraditi čak i njen mali deo. Sa razvojem Interneta, sajtova za onlajn kupovinu i društvenih mreža, drastično je porastao dnevni protok podataka i komunikacija sa provajderima mnogih usluga. Ti veliki sistemi teže da korisnicima ponude širok dijapazon proizvoda čime bi privukli još veći broj kupaca i svoj sadržaj plasirali na adekvatan način. Iz tog razloga, veliki sistemi sve više ulažu u implementaciju automatizovanih sistema za preporuke (eng. recommender systems).

Sistemi za preporuke koriste različite izvore podataka kako bi pronašli sadržaje od interesa. Entiteti kojima se daje preporuka nazivaju se korisnicima (eng. users), dok je predmet preporuke označen kao objekat (eng. item). U osnovi implementacije sistema za preporuku stoji činjenica da su ranija interesovanja i sklonosti korisnika dobar pokazatelj budućeg ponašanja.

Prošle interakcije korisnika sa objektima u sistemu mogu se analizirati i posmatrati kroz dva aspekta. Jedna od mogućnosti je korišćenje samih objekata, proizvoda kompanije, gde se na osnovu njihovih osobina formira preporuka. Ovakav sistem je definisan kao tzv. content-based sistem. Druga mogućnost je formiranje preporuke na osnovu implicitnih ili eksplicitnih ocena drugih korisnika u sistemu. Pošto se zasniva na postojanju zajednice korisnika, ovaj sistem označen je kao kolaborativni sistem preporuke. Takođe, postoje i sistemi koji u jednom kombinuju oba navedena pristupa, označeni kao hibridni (eng. hybrid) sistemi.

U ovom radu biće najpre ukratko objašnjene teorijske osnove kolaborativnih sistema preporuke, a zatim će biti opisana implementacija sistema za preporuku filmova u programskom jeziku Python.

# Kolaborativno filtriranje podataka

Sistemi za kolaborativno filtriranje nastoje da predvide korisnost određenog objekta datom korisnika, a na osnovu ocene tog proizvoda od strane ostalih korisnika u sistemu. Kod ovih sistema ocene koje su korisnici dodelili objektima posmatraju se kao aproksimacija njhovih interesovanja i merilo korisnosti objekta. Sličnost u ukusu između dvoje korisnika računa se na osnovu njihovih dosadašnjih interesovanja, ogledano u sličnosti ocena koje dodeljuju.

U okviru kolaborativnih sistema postoje dva načina za preciznije određivanje sličnosti između korisnika, a na osnovu kojih je izvršena i podela sistema. Razlikuju se:

* kolaborativno filtriranje zasnovano na korisnicima (eng. User-Based) i
* kolaborativno filtriranje zasnovano na proizvodima (eng. Item-Based).

U slučaju kolaborativnog filtriranja zasnovanog na korisnicima podrazumevano je da se sličnost između korisnika određuje na osnovu sličnosti u ocenama korisnika. Na ovaj način  određuje se grupa korisnika sa sličnim interesovanjima, a zatim se vrši preporuka onih proizvoda koji su najbolje ocenjeni.

Slično tome, u slučaju kolaborativnog filtriranja zasnovanog na proizvodima, preporuka se formira na osnovu ranije pregledanih, kupljenih ili ocenjenih proizvoda. Sličnost između ranijih objekata i objekta koji se preporučuje određuje se upoređivanjem ocene ciljnog korisnika i ostalih korisnika u sistemu.

I kod jednog i kod drugog tipa kolaborativnih sistema za preporuku koriste se ocene korisnika, a ne karakteristike objekata čime se daje mogućnost preporuke objekata sa kojima ciljni korisnik prethodno nije imao iskustva. Ipak, ovaj pristup poseduje i izvesna ograničenja. U obimnim sistemima često dolazi do toga da veliki broj objekata bude ocenjen od strane malog broja korisnika, što znači da na ovaj način vrednovani objekti nisu pogodni za korišćenje u algoritmima preporuke.

Dodatno, u ovako velikim sistemima javlja se i problem tzv. hladnog starta (eng. Cold-Start). Ukoliko na početku rada sistema nema ocenjenih sadržaja, nije moguće generisati odgovarajuće preporuke koristeći kolaborativno filtriranje. Na isti način i nove proizvode nije moguće preporučiti ukoliko nisu dovoljno puta ocenjeni. U novije vreme, javio se još jedan problem u korišćenju kolaborativnog filtriranja. Naime, konkurenti mogu namerno loše ocenjivati tuđe, a visoko ocenjivati svoje objekte i na taj način drastično uticati na rezultat preporuke.

# Analiza projekta

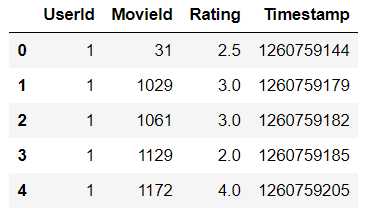
U ovom projektu izvršena je implementacija kolaborativnih sistema za preporuku filmova u programskom jeziku Python. Korišćene su biblioteke NumPy, Pandas, Matplotlib i Surprise dodatak.

* NumPy bibilioteka se koristi za rad sa nizovima i matricama, rad u domenu linearne algebre i Furijeove transformacije.
* Pandas biblioteka omogućava manipulaciju podacima i njihovu analizu, nudi strukture podataka i operacije nad numeričkim tabelama i vremenskim serijama.
* Matplotlib je sveobuhvatna biblioteka za kreiranje statičkih, animiranih i interaktivnih vizuelizacija u Python-u.
* Surprise je Python dodatak (eng. scikit) za sisteme za preporuku.

## Analiza skupa podataka

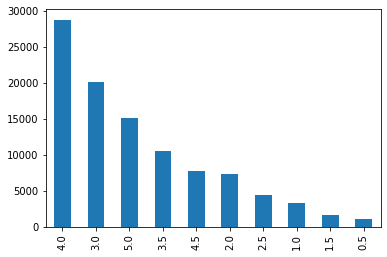
U ovom projektu korišćen je javno dostupan set podataka preuzet iz Kaggle baze[[1]](#footnote-1) koji sadrži informacije o ocenama koje su dobili filmovi uključeni u ovaj skup podataka. Celokupni podaci nalaze se u dokumentu *ratingsMovie.csv* i uključuju četiri relevantna atributa:

* UserId - predstavlja jedinstveni identifikator korisnika koji je ostavio ocenu, pri čemu jedan korisnik može da oceni veći broj filmova;
* MovieId - predstavlja jedinstveni identifikator filma koji se ocenjuje;
* Rating - predstavlja brojčanu ocenu koju korisnik dodeljuje filmu i može imati vrednosti u opsegu od 1 do 5, gde je 1 najlošija, a 5 najbolja ocena;
* Timestamp - predstavlja vremensku oznaku dodeljivanja ocene.



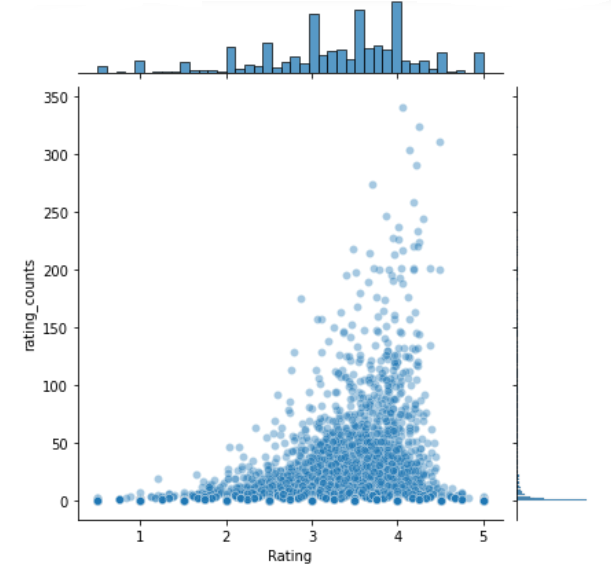
Slika 1. Isečak korišćenog skupa podataka

Svi podaci sadržani u dataset-u su numeričkog tipa. Nakon inicijalne analize utvrđeno je da nije bilo dupliranih podataka kao ni nedostajućih vrednosti. Distribucija dodeljenih ocena nije pratila normalnu raspodelu već su najučestalije dodeljene ocene bile 4, 3 i 5, redom (Slika 2).



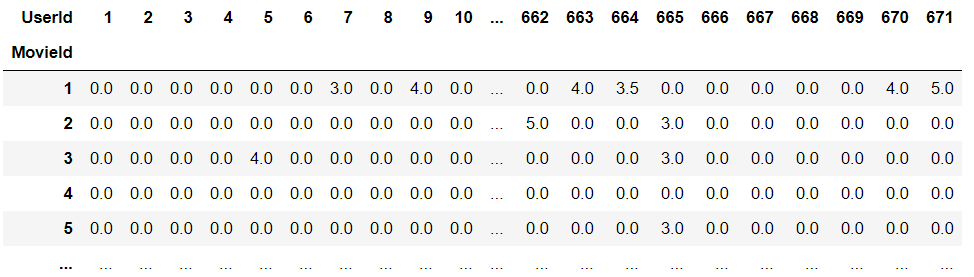
Slika 2. Distribucija ocena filmova

Da bi se utvrdila pouzdanost i validnost ukupne ocene filma, izvršena je analiza korelacije broja dodeljenih ocena i prosečne ocene za svaki film. Ocena se može smatrati relevantnom u slučaju da je dati film više puta ocenjen na isti ili sličan način. U suprotnom se ocena o filmu donosi na osnovu veoma malog uzorka, čime se stvara nepouzdana ocena koja se kao takva ne može koristiti u sistemima za preporuku. Na Slici 3. dat je rezultat ove analize, iz koje se vidi pozitivna korelacija između navedenih veličina: filmovi koji imaju veću prosečnu ocenu su u proseku i veći broj puta ocenjeni. Jedno od mogućih objašnjenja je to da se kvalitetniji filmovi više dopadaju publici, više preporučuju i bolje ocenjuju, a što dovodi do daljeg povećanja broja korisnika koji su ocenili film.



Slika 3. Odnos broj ocena i vrednosti ocena

Radi jednostavnije manipulacije podacima početni dataset je transformisan u matricu koja u kolonama sadrži jedinstvene identifikatore korisnika, a u redovima jedinstvene identifikatore filmova. U preseku vrste i kolone nalazi se vrednost ocene koju je korisnik sa identifikatorom određenim vrednošću kolone dodelio filmu čiji je identifikator određen vrednošću reda (Slika 4).



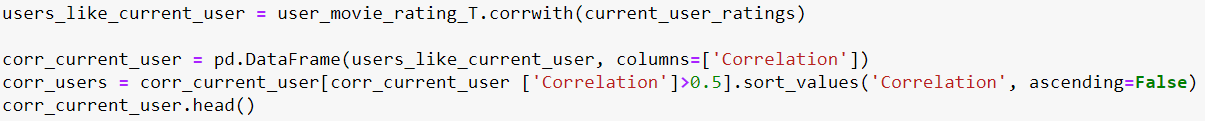
Slika 4. Novokreirani dataset

## Implementacija sistema za preporuku filmova

U ovom radu korišćena su dva pristupa u implementaciji sistema za preporuku. Prvi pristup podrazumeva preporučivanje filmova na osnovu ocena korisnika sa sličnim interesovanjima kao ciljni korisnik. U drugom pristupu nad skupom podataka su primenjeni algoritmi prediktivne klasifikacije u cilju kreiranja modela preporuke. Rezultati oba pristupa su objedinjeni i izvršena je uporedna analiza.

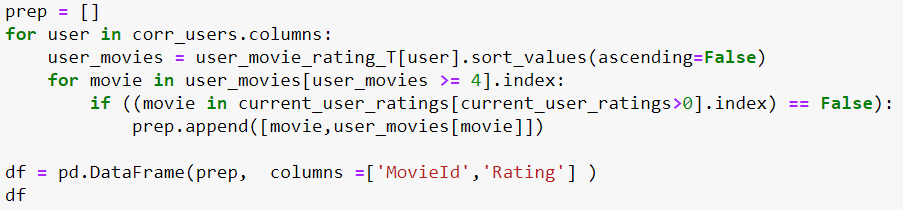
*Prvi pristup*

Radi izdvajanja korisnika koji dele interesovanja sa ciljnim korisnikom korišćena je *corrwith* funkcija Pandas biblioteke (Slika 5). Ova funkcija vrši analizu korelacije između ocena filmova ciljnog korisnika sa jedne i ocena svih ostalih korisnika iz dataseta sa druge strane. Kako bi se izdvojili najrelevantniji korisnici, postavljena je i pražna vrednost korelacionog koeficijenta na 0.5. Ova vrednost je odabrana da bi se uključile jake korelacije (vrednost Pirsonovog koeficijnta je veća od 0.7) kao i srednje do umereno jake korelacije (vrednost Pirsonovog koeficijnta u opsegu od 0.5 do 0.7) [1].



Slika 5. Korišćenje funkcije *corrwith* za pronalaženje sličnih korisnika ciljnom korisniku *current\_user\_ratings*

Nakon dobijanja podskupa korisnika sa najsličnijim interesovanjima, izvršeno je filtriranje najbolje ocenjenih filmova datog podskupa korisnika koje ciljni korisnik nije ocenio i njihovo smeštanje u listu filmova za preporuku (Slika 6).

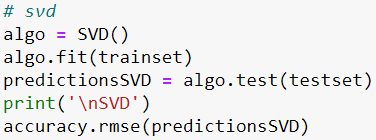


Slika 6. Izdvajanje filmova za preporuku

*Drugi pristup*

Nad početnim skupom podataka primenjena su tri algoritma (Slika 7) za izvršavanje prediktivne klasifikacije sistema za preporuku i to:

* SVD (Singular Value Decomposition) - algoritam koji vrši dekompoziciju matrice na njene sastavne delove radi pojednostavljenja proračuna nad njenim elementima [2].
* NMF (Non-negative matrix factorization) - algoritam koji vrši faktorizaciju matice na dve matrice sa osobinom da su elementi svih matrica pozitivni [3].
* KNN (K-Nearest Neighbors) - algoritam koji projektuje podatke na N dimenzija određenih dataset-om i traži k najbližih suseda na osnovu udaljenosti tačaka u N-dimenzionalnom prostoru [4].

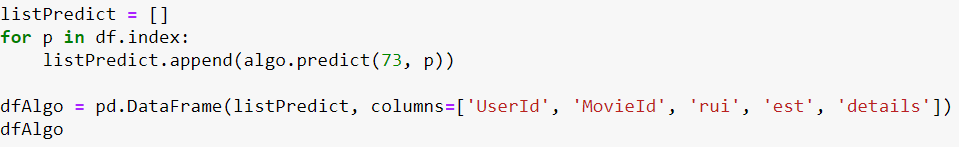


Slika 7. Primena SVD algoritma

Kako su sistemi za preporuke prediktivni sistemi, tačnost modela se može utvrditi korišćenjem evaluacionih metrika kao što su koren srednje kvadratne greške (RMSE) ili srednja apsolutna greška (MAE) [5]. U projektu je iskorišćena RMSE metrika koja za svaki od primenjenih algoritama iznosi:

* SVD – 0.90;
* NMF – 0.95;
* KNN – 0.97.

Na osnovu ovih vrednosti se vidi da je vrednost RMSE najveća za KNN algoritam, a najmanja za SVD algoritam. Kako manja vrednost RMSE ukazuje na manju grešku, a time i bolje formiranje modela, za dalju analizu korišćen je model dobijen SVD algoritmom. Na ciljnom korisniku se primenjuje dobijeni model i upotrebom funkcije *predict* iz celokupnog seta filmova određuju se oni koji su potencijalno zanimljivi ciljnom korisniku (Slika 8).

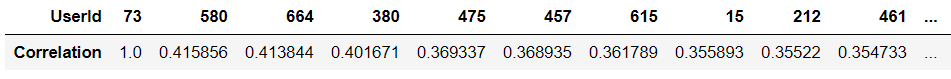


Slika 8. Korišćenje funkcije *predict* za odeđivanje filmova od interesa

# Analiza rezultata

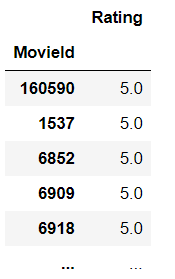
Dobijeni rezultati biće predstavljeni na primeru korisnika sa jedinstvenim identifikatorom 73.

Primenom prvog pristupa i korišćenjem funkcije *corrwith*, izdvojeni su korisnici sa sličnim interesovanjima u odnosu na ciljnog korisnika 73. Na slici 9 prikazani su korelacioni koeficijenti između korisnika 73 i ostalih korisnika.



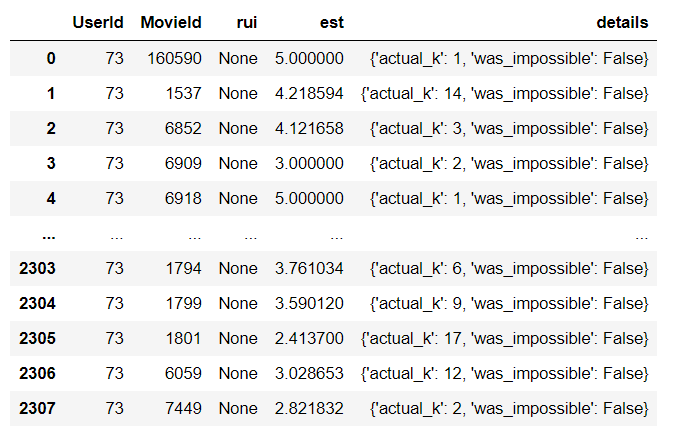
Slika 9. Prikaz dela rezultata korelacione analize između korisnika 73 i ostalih korisnika

Filtriranjem korisnika od interesa sa koeficijentom korelacije većim od 0.5 i izdvajanjem visokoocenjenih filmova, dobijena je lista predloga filmova koja se prosleđuje korisniku 73 (Slika 10).



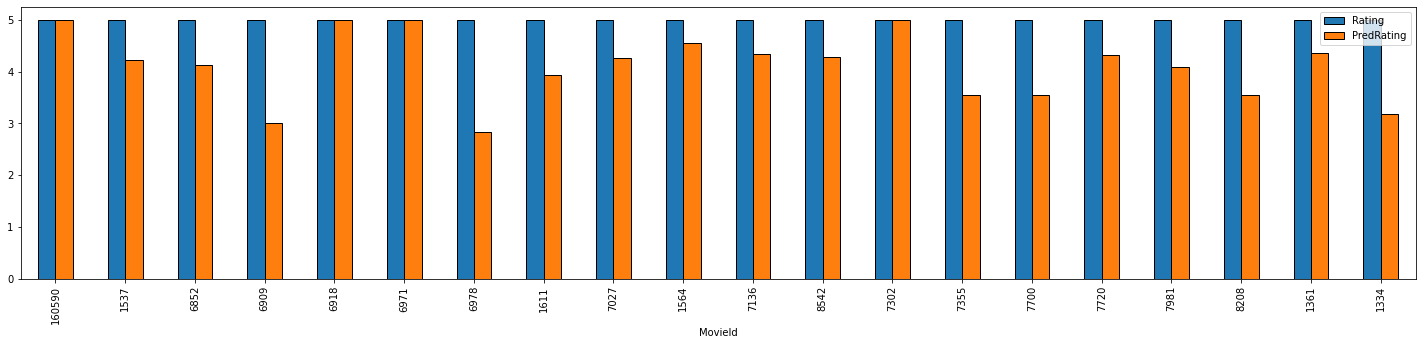
Slika 10. Isečak preporučenih filmova

Po drugom pristupu, nad dobijenim filmovima od interesa primenjen je SVD model čime je dobijena prediktivna ocena svakog od prosleđenih filmova (kolona est na slici 11). Sumarni rezultati oba pristupa prikazani su na slici 11.

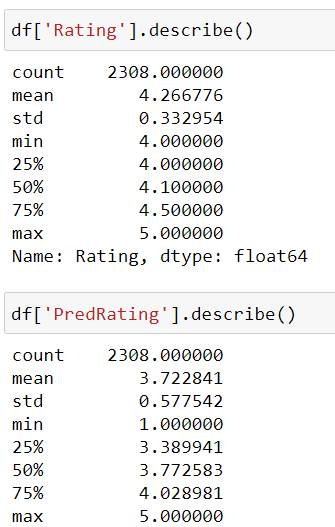


Slika 11. Sumarni rezultati prvog i drugog pristupa implementacije sistema za preporuke

Rezultati dobijeni uporednom primenom prve i druge metode ukazuju na razliku u ocenjivanju filmova između korisnika 73 i grupe korisnika sa sličnim interesovanjima (Slika 12). Prosečna ocena grupe korisnika je 4.26 sa najmanjom ocenom 4, dok prosečna ocena korisnika 73 iznosi 3.72 sa najmanjom ocenom 1 (Slika 13). Kako je razlika u prosečnoj oceni 0.54, svega polovina ocene, može se zaključiti da je preporuka dobijena implementacijom ovog sistema zaista relevantna.



Slika 12. Isečak uporednog prikaza ocena grupe korisnika i prediktivnih ocena korisnika 73



Slika 13. Uporedni prikaz sumarnih rezultata ocenjivanja grupe korisnika i prediktivnih ocena korisnika 73

# Zaključak

Sistemi za preporuke su efikasni alati koji se koriste u velikim sistemima za plasiranje adekvatnog sadržaja potencijalnim korisnicima. Jedna od čestih implementacija je ona koja koristi tzv. kolaborativno filtriranja podataka. Sistemi sa kolaborativnim filtriranjem nastoje da predvide korisnost određenog proizvoda za datog korisnika na osnovu vrednovanja istog proizvoda od strane ostalih korisnika u sistemu. U ovom radu implementaciji sistema za preporuke pristupljeno je na dva načina i to na primeru preporuke za gledanje filmova: prvi pristup podrazumeva preporučivanje na osnovu ocena korisnika sa sličnim interesovanjima kao ciljni korisnik, a u drugom pristupu su primenjeni algoritmi prediktivne klasifikacije nad skupom podataka u cilju kreiranja modela preporuke. Dobijeni rezultati su sumirani i zaključeno je da obe implementacije daju slične rezultate iz čega proizilazi da je konačna preporuka validna i relevantna ciljnom korisniku.

# Reference

[1] Kirch W. Pearson’s Correlation Coefficient. Springer, Dordrecht. 2008.

[2] Hogben L. Handbook of linear algebra. CRC. 2007

[3] Huang K., Sidiropoulos N.D., Swami A. Non-Negative Matrix Factorization Revisited: Uniqueness And Algorithm For Symmetric Decomposition. Tranasactions for signal processing. 2014

[4] Harrington P. Machine Learning in Action. Manning Publications. 2012

[5] Chai T., Draxler R.R. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geoscientific Model Development. 2014

1. Kaggle. The Movies Dataset. Dostupno na: <https://www.kaggle.com/rounakbanik/the-movies-dataset> [↑](#footnote-ref-1)