SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 826

SUSTAV ZA PREPORUKU FILMOVA

Nikolina Špehar

Zagreb, lipanj 2023.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 826

SUSTAV ZA PREPORUKU FILMOVA

Nikolina Špehar

Zagreb, lipanj 2023.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 10. ožujka 2023.

ZAVRŠNI ZADATAK br. 826

Pristupnica:	**Nikolina Špehar (0036526861)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentor:	doc. dr. sc. Marko Đurasević

Zadatak: **Sustav za preporuku filmova**

Opis zadatka:

Proučiti problem preporučivanja filmova korisnicima na temelju njihovih preferencija. Istražiti postojeću literaturu na temu tog problema te metode korištene za izradu preporuka. Pronaći prikladne skupove podataka s korisničkim preferencijama za pojedine filmove koji će se koristiti za učenje i ocjenu. Razviti sustav baziran na metodama strojnog učenja koji će na temelju korisničkih preferencija daje preporuku sličnih filmova na temelju preferencija drugih korisnika. Ocijeniti učinkovitost razvijenog sustava i istaknuti moguća poboljšanja. Radu priložiti izvorne tekstove programa, dobivene rezultate uz potrebna objašnjenja i korištenu literaturu.

Rok za predaju rada: 9. lipnja 2023.

SADRŽAJ

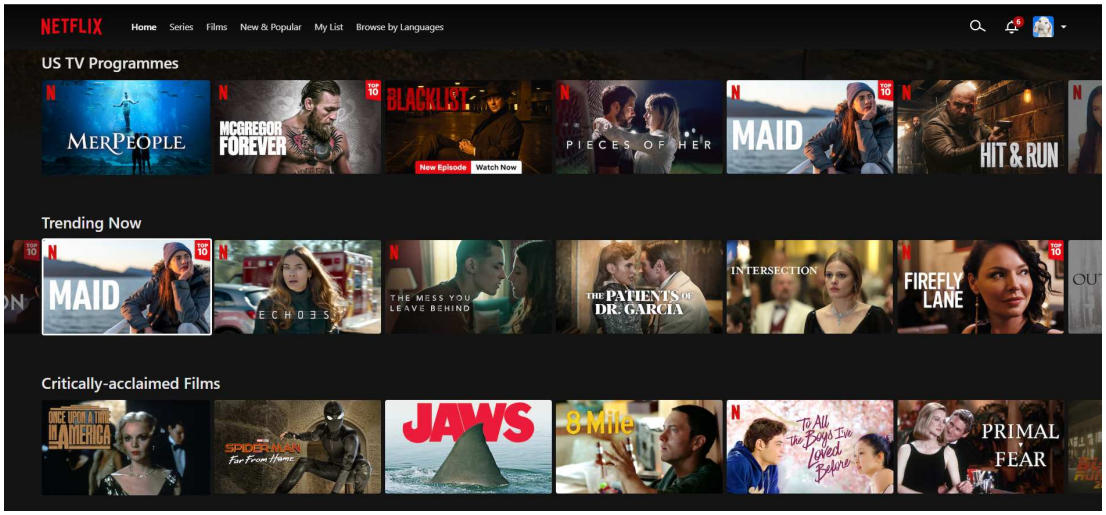
1. Uvod

Zbog napretka u tehnologiji, veće povezanosti, te zbog jednostavnosti prikupljanja, ali i objavljivanja informacija, jedan od sve većih problema današnjice je prevelika količina podataka kojoj smo izloženi. U toj ogromnoj količini podataka postaje teško pronaći baš ono što tražimo.

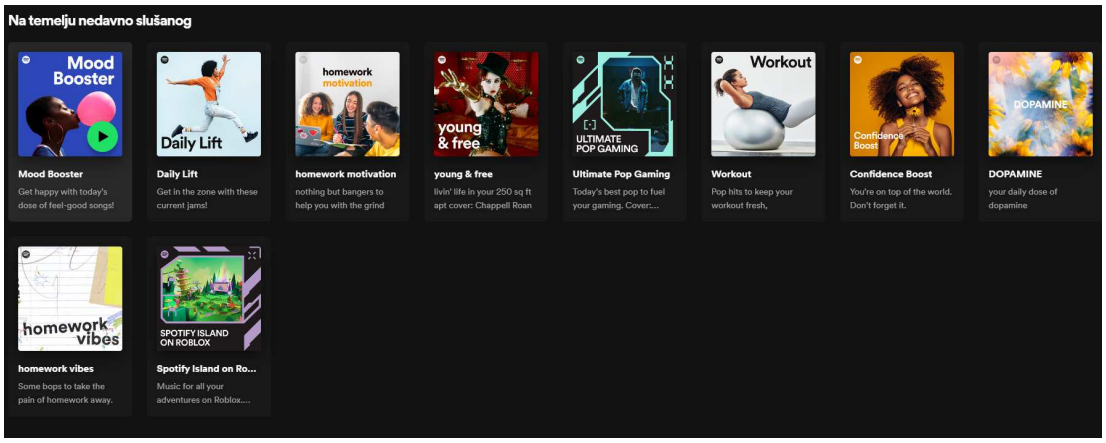
Tijekom zadnjih desetljeća, s usponom firmi kao što su Amazon, Netflix i YouTube sustavi za preporuku dobili su na važnosti. Sustavi za preporuku tu pomažu tako što korisnicima preporučuju baš ono što bi im se moglo svidjeti na temelju različitih izračuna. Primjere dva servisa koji uspješno koriste sustave za preporuku vidimo na slici 1.1: Netflix i na slici 1.2: Spotify. Pomoću sustava za preporuku Netflix neće morati jednom korisniku preporučiti sve svoje filmove na "home screenu", što zbog velike količine podataka zapravo nije ni moguće, nego će koristeći metode strojnog učenja preporučiti samo one filmove za koje s velikom vjerojatnošću pretpostavlja da će se svidjeti upravo tom korisniku [10]. Uz to što servisi ne znaju što preporučiti svojim korisnicima, tako ni korisnici, odnosno gledatelji, u nekim slučajevima jednostavno ne mogu izabrati ono što će gledati, zbog preplavljujuće količine izbora kojeg imaju.

Takvi problemi rješavaju se sustavima za preporuku. Sustavi za preporuku su oblik umjetne inteligencije, koji je najčešće povezan sa strojnim učenjem i s Big Data-om [6]. Ovaj rad bavit će se upravo problemom preporuke proizvoda korisniku, kada imamo ogromnu količinu proizvoda na izbor. Cilj rada je razviti sustav baziran na metodama strojnog učenja koji će na temelju korisničkih preferencija dati preporuku sličnih filmova na temelju preferencija drugih korisnika. Ocijeniti učinkovitost razvijenog sustava i istaknuti moguća poboljšanja.

Drugo poglavlje opisati će sustave za preporuku. Objasniti će se njihova podjela i pobliže će se opisat one vrste koje će biti korištene u sustavu za preporuku implementiranom u ovom radu. Treće poglavlje opisuje bazu podataka koja se koristila prilikom izrade i testiranja ove implementacije sustava. U četvrtom poglavlju se opisuje sustava za preporuku implementiran u ovom radu. U petom poglavlju opisani rezultati rada, a šesto poglavlje je zaključak.



Slika 1.1: Primjer sustava za preporuku Netflix [3]

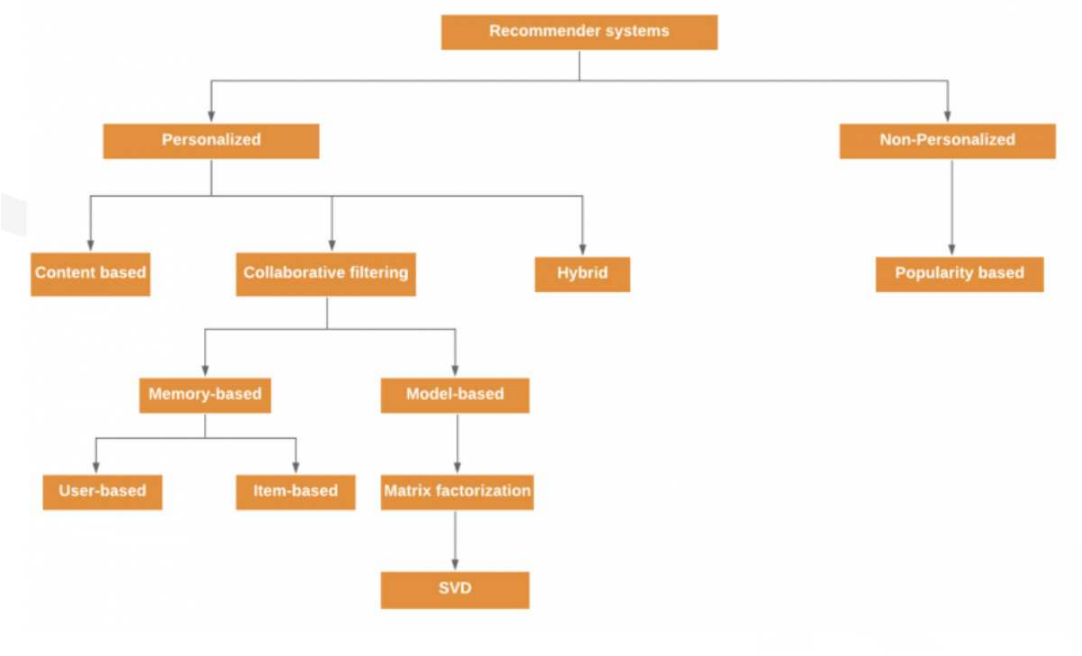


Slika 1.2: Primjer sustava za preporuku Spotify [4]

2. Sustavi za preporuku

Sustavi za preporuku su podvrsta sustava za filtriranje informacija koji za cilj imaju predvidjeti ocjenu koju bi korisnik dao nekom proizvodu ili korisnikovu preferencu kada su u pitanju neki proizvodi [12]. Sustav koristi koncepte i algoritme dubokog učenja. Duboko učenje je podskup strojnog učenja, koji koristi neuronske mreže s tri ili više slojeva. Koristeći neuronske mreže pokušava imitirati rad ljudskog mozga, s krajnjim ciljem da ta neuronska mreža "uči" pomoću velike količine podataka, kao što bi i ljudi učili [1]. Sustav za preporuku koristeći algoritme dubokog učenja pronalazi uzorke u ponašanju korisnika ili u ponašanju sličnih korisnika. Koristeći sustave za preporuku uvelike olakšavamo izbor i servisima ali i korisnicima. Zbog isplativosti korištenja sustava za preporuku tijekom posljednjeg desetljeća velike firme, kao što je Netflix, uložile su velike količine sredstava u razvijanje sustava za preporuku. To je jedan od razloga zašto je i došlo do velikog napretka u razvoju sustava za preporuku. Prednosti korištenja sustava za preporuku su povećana prodaja servisima koji ga koriste, smanjenje pritiska na sustav, te povećanje zadovoljstva i angažmana kod korisnika. Razlog toga je upravo dobra preporuka proizvoda koji bi se mogli svidjeti korisniku [9].

Sustavi za preporuku mogu se podijeliti na personalizirane i na nepersonalizirane sustave. Takvu podjelu vidimo na slici 2.1.



Slika 2.1: Podjela sustava za preporuku [12]

Jedna podvrsta nepersonaliziranog sustava za preporuku je preporuka po popularnosti. Takav će sustav svakom korisniku dati potpuno istu preporuku; 10 najboljih triler filmova, 10 najboljih detektivskih knjiga... Tu nam nije bitno koje su karakteristike korisnika, zbog toga što će svi dobiti istu preporuku. Preporučit će se oni proizvodi koji su popularni u općoj populaciji tijekom nekog vremenskog razdoblja. Prednosti ove vrste sustava su jednostavna implementacija i lakoća kojom se prikupljaju podaci. Nedostatak je baš to što su preporuke nepersonalizirane, i takve se mogu ne svidjeti svima. Također problem je i taj, što će popularne stvari postajati sve popularnije, dok se manje znani proizvodi neće niti preporučivati.

Druga vrsta sustava za preporuku su personalizirani sustavi. Personalizirani sustavi analiziraju korisničke podatke, kao što su prošle kupovine, ocijene koje su dane ostalim proizvodima i odnosi s drugim korisnicima. Korištenjem svih tih podataka u izračunu preporuke osigurava se to da svaki korisnik dobije različitu i personaliziranu preporuku. Dvije glavne vrste personaliziranih sustava za preporuku su sustavi koji koriste kolaborativno filtriranje i sustavi koji filtriraju na temelju sadržaja.

2.1. Kolaborativno filtriranje

Sustavi daju preporuke izričito na temelju prošlih interakcija između korisnika i proizvoda. Interakcije se spremaju u takozvane matrice interakcije korisnik-proizvod

(engl. *user-item interactions matrix*) [10]. Za preporuku algoritam koristi sličnosti u ponašanju dva ili više korisnika, pomoću prošlih interakcija između korisnika i proizvoda, sustav uči kako predvidjeti budućnost. Ovakvi sustavi model grade na pretpostavci; ako su dva korisnika imali slične odluke i slične preference u prošlosti, da onda postoji velika vjerojatnost da će se njihove odluke i preference podudarati i u budućnosti [6].

Na primjer sustav za preporuku filmova koji zna da korisnik A i korisnik B imaju sličan ukus za filmove. Sustav će korisniku A preporučiti filmove koji su se već svidjeli korisniku B. Glavna prednost ovakvog sustava za preporuku je ta što ne moramo imati nikakve informacije o proizvodu ili o korisniku, tako da jednom napravljen sustav može biti korišten u različitim situacijama. Također, s povećanjem vremena korištenja sustava povećava se i preciznost preporuka. Zbog toga što sustav tijekom tog vremena prikuplja dodatne informacije o interakcijama između korisnika i proizvoda.

Kako se ovakav sustav oslanja na prošle interakcije kako bi napravio preporuke, javlja se problem hladnog starta (engl. *cold start problem*). Problem nastaje kada nam u sustav dođe novi korisnik ili novi proizvod, za koje nemamo zabilježene prijašnje interakcije. Te je onda za korisnika nemoguće napraviti preporuku korištenjem kolaborativnog filtriranja, a proizvod ne možemo preporučiti korisnicima, jer ne postoji korisnik kojem se proizvod ranije svidio. Taj problem se u praksi najčešće rješava korištenjem drugog načina preporuke, kao što je preporuka nasumičnih proizvoda korisniku, to jest preporuka novog proizvoda nasumičnim korisnicima ili preporuka najpopularnijih proizvoda korisniku [10].

2.2. Filtriranje na temelju sadržaja

Sustavi koji preporučuju na ovaj način koriste različite značajke proizvoda i korisnika kako bi napravili pretpostavku. Uzevši za primjer sustav za preporuku filmova. Od korisnika će sustav dobiti različite informacije, koji filmovi su mu se ranije svidjeli, koliko korisnik ima godina, gdje živi, kojeg je spola... Dok će za filmove znati žanr, glavne glumce, koje godine je izašao film, tko je bio redatelj... Na temelju tih informacija sustav gradi model koji objašnjava uočenu interakciju između korisnika i filma. Proučavajući te modele možemo na primjer uočiti da mlade žene bolje ocjenjuju jedne filmove, dok mladi muškarci bolje ocjenjuju druge filmove. Davanje preporuke na temelju tih modela onda postaje lagano, zbog toga što samo moramo pogledati karakteristike korisnika koji traži preporuku, vidjeti u koji model se najbolje može smjestiti, te mu onda dati preporuku na temelju izabranog modela [10].

Postoje i jednostavniji oblici ovakvog sustava za preporuku, na primjer sustav može od korisnika dobiti samo informaciju koji filmovi su mu se ranije svidjeli. Sustav će onda pogledati različite značajki filmova, te će onda na temelju tih značajki pokušati preporučiti korisniku film koji je svojim značajkama što sličniji tim filmovima za koje zna da se sviđaju korisniku. Sustavi koji koriste algoritme ove vrste u puno manjoj mjeri pate od problema hladnog starta [6].

Zbog toga što se i novi korisnici, ali i novi proizvodi mogu opisati svojim značajkama, koje će imati uvijek bez obzira na to kada su dodani u sustav, te se zbog toga na temelju tih značajki uvijek može dobiti nekakva preporuka [10].

3. Korištena baza podataka

Za izradu ovog sustava za preporuku bilo je potrebno pronaći odgovarajući skup podataka s kojim se moglo raditi. Korišteni skup podataka je dobiven iz servisa za preporuku filmova Movie Lense. Skup podataka sadrži 25,000,095 ocjenjivanja i 1,093,360 komentara kroz 62,423 filma. Podatci su se na servisu Movie Lense skupljali između 9. siječnja 1995. i 21. studenog 2019. godine. Skup podataka generiran je 21. studenog 2019. Korisnici koji se nalaze u skupu izabrani su nasumično, unutar tablica predstavljeni su samo s posebnom id oznakom. Nikakvi demografski podatci, kao što su; spol, vjerska orijentacija, nacionalnost..., nisu uključeni u skup. Svaki korisnik koji je uključen u podatkovni skup morao je ocijeniti barem 20 filmova. Filmovi koji su uključeni u skup podataka morali su imati barem jednu ocjenu ili jedan komentar [8].

Tablica `ratings.csv` sadrži sve ocijene koje su uključene u skup podataka. Tablica je sortirana prvo po `userId`-u, a onda unutar svakog korisnika po `movieId`-u. Dio tablice vidljiv je na slici 3.1. Zaglavlje tablice sadrži:

- `userId` oznaka po kojoj identificiramo korisnike
- `movieId` oznaka po kojoj razlikujemo filmove
- `rating` ocjena koju je korisnik dodijelio filmu, u intervalu od 0.5 do 5.0 zvjezdica
- `timestamp` vrijeme kada je ocjena dodijeljena, prikazano u sekundama

userId	movieId	rating	timestamp
1	296	5	1.15E+09
1	306	3.5	1.15E+09
1	307	5	1.15E+09
1	665	5	1.15E+09
1	899	3.5	1.15E+09
1	1088	4	1.15E+09
1	1175	3.5	1.15E+09
1	1217	3.5	1.15E+09
1	1237	5	1.15E+09
1	1250	4	1.15E+09
1	1260	3.5	1.15E+09
1	1653	4	1.15E+09
1	2011	2.5	1.15E+09

Slika 3.1: Tablica ratings.csv

Tablica `movies.csv` sadrži informacije o filmovima, svaki redak tablice predstavlja jedan film. Dio tablice vidi se na slici 3.2. Zaglavlje tablice sadrži:

- `movieId` oznaka po kojoj razlikujemo filmove
- `title` naziv i godina filma
- `genres` popis žanrova kojima film pripada

Vrijednosti koje se mogu nalaziti u popisu žanrova su:

- `Action` = akcijski film
- `Adventure` = avanturistički film
- `Animation` = animirani film
- `Children's` = film za djecu
- `Comedy` = komedija
- `Crime` = kriminalistički film
- `Documentary` = dokumentarac
- `Drama` = drama
- `Fantasy` = fantazija
- `Film-Noir` = kriminalistički crno-bijeli film
- `Horror` = horor film
- `Musical` = mjuzikl

- `Mystery` = film misterije
- `Romance` = romantični film
- `Sci-Fi` = znanstvena fantastika
- `Thriller` = triler
- `War` = ratni film
- `Western` = kaubojski film
- `(no genres listed)` = nema žanrova na popisu

movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action|Crime|Thriller
7	Sabrina (1995)	Comedy|Romance
8	Tom and Huck (1995)	Adventure|Children
9	Sudden Death (1995)	Action
10	GoldenEye (1995)	Action|Adventure|Thriller

Slika 3.2: Tablica movies.csv

4. Implementacija

Zadatak ovog rada bio je izrada vlastitog sustava za preporuku filmova. U ovom poglavlju pobliže će se opisati kako je sustav za preporuku implementiran. Kao skup podataka koji je korišten za davanje preporuka koriste se tablice opisane u poglavlju 3. U ovom sustavu za preporuku filmova, implementirane su 3 različite personalizirane metode za preporuku; prva metoda je kolaborativno filtriranje, druga je filtriranje prema sadržaju, u ovom slučaju filtrirat ćemo na temelju žanrova. U trećoj metodi se spaja indekse sličnosti iz prve dvije metode.

4.1. Pretraga tablice

Prvi problem koji sam trebala riješiti, nakon pronalaska odgovarajućeg skupa podataka, bila je pretraga tog istog skupa. Zbog toga što se u tablici `movies.csv` nalazi 62,423 filma. Bio mi je potreban optimalan način za prolazak kroz cijelu tablicu `movies.csv` i pronalazak filma za kojeg korisnik želi dobiti preporuke.

Naslove filmova prvo je bilo potrebno "urediti", pomoću regularnog izraza iz njih smo maknuli sve ono što nisu bile velika ili mala slova ili brojevi. Također zbog jednostavnosti sva velika slova prebacili smo u mala slova.

Za pretragu tablice koristimo TF-IDF frekvencija izraza/inverzija frekvencije dokumenta (engl. *term frequency/inverse document frequency*). TF-IDF će vektorizirati, pretvoriti u vektor s kojim možemo računati, stupac tablice `movies.csv title`, odnosno `clean_title`; tamo smo spremili "uređene" naslove filmova. Iz originalnih naslova uklonjeni su svi znakovi osim brojki i slova. Nakon što korisnik upiše naslov filma za kojega želi preporuku, naslov tog filma će se vektorizirati istom metodom, te će se onda među vektorima tražiti najsličniji vektor. Kada se on pronađe, pronašli smo i korisnikov uneseni film [7]. TF-IDF metoda može se podijeliti u dva dijela: TF dio u kojem izračunavamo frekvenciju pojedinačnog izraza u odnosu na promatrani dokument, i na IDF dio u kojem gledamo koliko je neka riječ česta kada gledamo sve promatrane dokumente [5]. TF se računa kao broj ponavljanja izraza

unutar dokumenta podijeljen s ukupnim brojem riječi u dokumentu:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

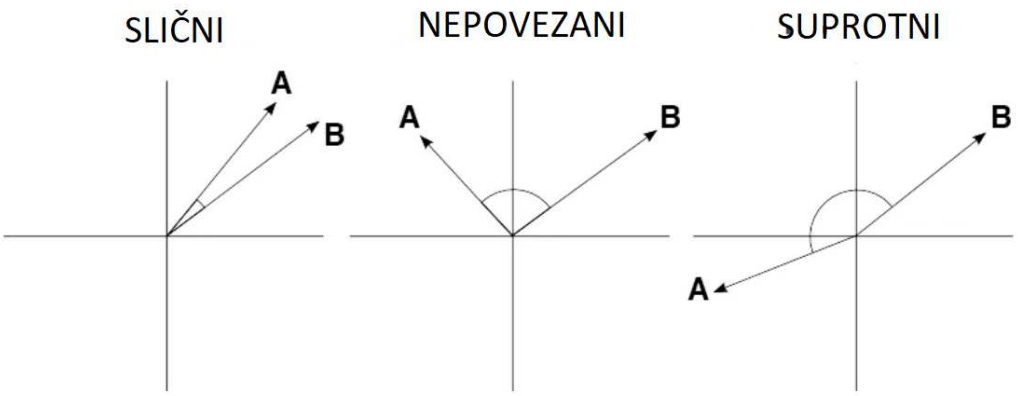
IDF dobijemo iz logaritma broja dokumenata podijeljenog sa brojem dokumenta u kojem se pojavljuje izraz za kojeg računamo TF-IDF. IDF određuje značajnost nekog izraza gledano za sve dokumente s kojima se radi:

$$idf(w) = \log(\frac{N}{df_t})$$

Na kraju kako bi dobili konačnu vrijednost za promatrani izraz, vrijednosti TF i IDF trebamo jednostavno pomnožiti:

$$w_{i,j} = tf_{i,j} x \log(\frac{N}{df_i})$$

Ovu metodu nije bilo potrebno samostalno implementirati, nego sam koristila biblioteku `Sklearn`. Naposljetku za izračun sličnosti između vektora naslova iz tablice `movies.csv` i vektora filma kojeg je korisnik unio koristi se kosinusna sličnost (engl. *cosine similarity*). Kosinusna sličnost mjeri sličnost između dva vektora u prostoru, tako da gleda kosinus kuta između dva vektora kako bi odredila pokazuju li ta dva vektora u istom smjeru, to možemo vidjeti na slici 4.1 [2].



Slika 4.1: Kosinusna sličnost - vektori [11]

4.2. Kolaborativno filtriranje - implementacija

Proces pronalaska sličnih filmova započinje, pronalaskom sličnih korisnika našem korisniku koji koristi aplikaciju. Tražimo one korisnike koji su traženi film ocijenili s ocjenom 4 ili više. Kada smo pronašli slične korisnike, onda pretpostavljamo, na temelju dobre ocijene traženog filma, da ti korisnici imaju ukus sličan našem korisniku koji traži preporuku. Zbog toga od sličnih korisnika tražimo ostale filmove koje su ocijenili s ocjenom 4 ili više. To su filmovi slični onom kojeg je korisnik unio.

Znamo da postoje filmovi, koje jednostavno svi ili barem velika većina ljudi voli. Kako svi znamo za te filmove i velika je vjerojatnost da ih je naš korisnik već pogledao. Ali i ne želimo da se naša preporuka sastoji samo od takvih popularnih filmova, nego želimo filmove koje su pozitivno ocijenili samo korisnici slični našem korisniku. Problem rješavamo tako da tražimo sve korisnike koji su s ocjenom 4 ili više ocijenili filmove slične korisnikovom filmu, tada iz našeg popisa sličnih filmova izbacujemo one koje je pozitivno ocijenilo više od 10% korisnika. Sada su iz našeg popisa izbačeni popularni filmovi za koje svi znaju, naša preporuka je time postala puno personalnija i preporučit će manje znane filmove.

4.3. Filtriranje žanrova

Implementacija metode filtriranja po sadržaju bila je jednostavnija od implementacije kolaborativnog filtriranja. Prvi korak bio je pronalazak žanrova filma kojega je unio korisnik. U idućem koraku prolazimo kroz tablicu `movies.csv` i uspoređujemo žanr našeg filma sa žanrovima filmova iz tablice, tu vrijednost upisujemo u tablicu i nakon prolaska kroz cijelu tablicu, sortiramo po njoj.

4.4. Povezano filtriranje

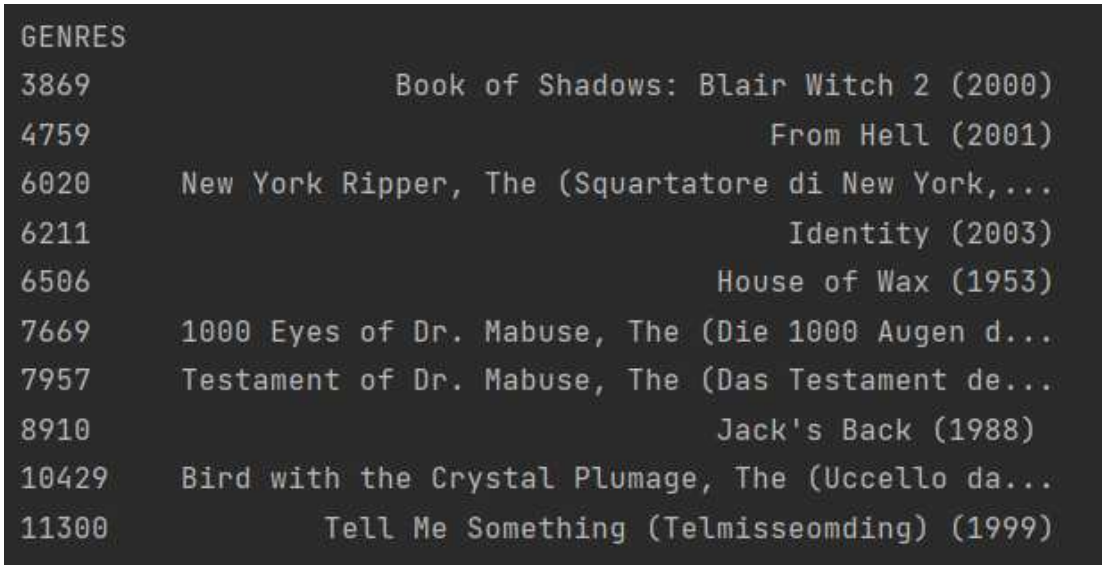
Povezano filtriranje implementirali smo tako da smo prvo izračunali sličnosti na temelju prve i druge metode, kolaborativnog filtriranja i filtriranja po žanrovima. Novu vrijednost sličnosti smo dobili tako da smo pronašli aritmetičku sredinu između sličnosti kolaborativnog filtriranja i filtriranja po žanrovima. Tablicu smo sortirali na temelju novih vrijednosti sličnosti.

5. Rezultati

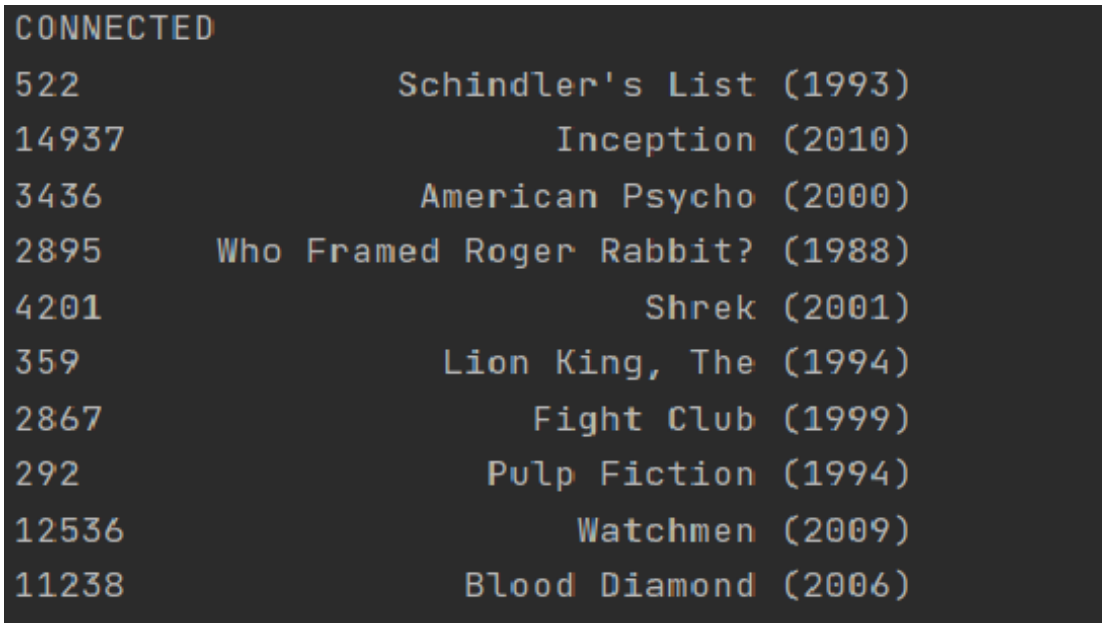
Promatrati ćemo rezultate sva tri sustava za preporuku. Preporuke kolaborativnog filtriranja i filtriranja ne temelju sadržaja, odnosno filtriranje po žanrovima daju donekle slične rezultate, koji se podudaraju u općem ozračju filmova, gledajući radnju i žanr. Filmovi dani preporukama su različiti. Neki od filmova su dosta poznati i znani su svima, dok su neki filmovi puno manje znani. Na primjer na slikama 5.1 i 5.2 vidimo filmove kao što su Kuća voštanih figura (engl. *House of wax*), Zodiac i Hanibal. To su neki opće poznati filmovi i postoji vrlo velika šansa da je korisnik te filmove gledao. Zbog će biti poželjnije da sustav za preporuku preporučuje one filmove koje je manje korisnika pogledalo. Korisnik tijekom traženja preporuka ne želi da mu se preporuče samo oni poznati filmovi za koje je već čuo i oni filmovi o kojima svi pričaju. Preporuka nepoznatijih filmova je dobra stvar iz tog razloga što je onda veća vjerojatnost da korisnik već ranije nije pogledao taj film. Tako će mu se više svidjeti sustav za preporuku, pa će ga zbog toga više i koristiti.

```
FILTERING
3436            American Psycho (2000)
11647               Death Proof (2007)
8235              The Machinist (2004)
10250    History of Violence, A (2005)
11509                Grindhouse (2007)
11435                    Zodiac (2007)
4044                   Hannibal (2001)
8199                     Primer (2004)
11830          Eastern Promises (2007)
4001             Evil Dead, The (1981)
```

Slika 5.1: American Psycho - kolaborativno filtriranje



Slika 5.2: American Psycho - filtriranje po žanrovima



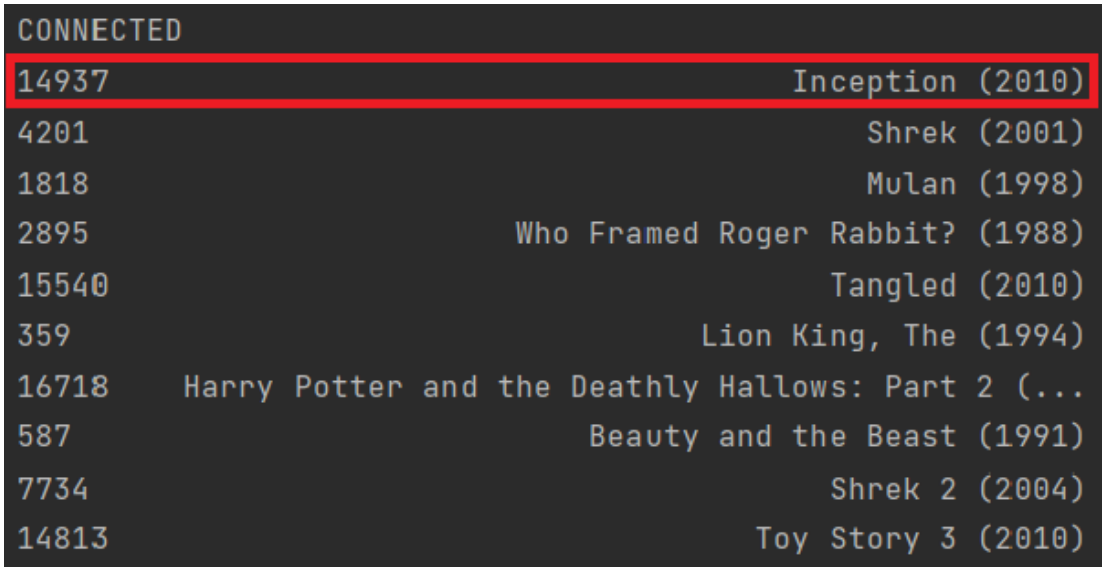
Slika 5.3: American Psycho - povezano

Treća vrsta sustava koji je implementiran u ovom radu je hibridni sustav. Hibridni sustav smo dobili tako da smo pomoću funkcije aritmetičke sredine povezali sličnosti iz prve dvije vrste filtriranja. Sličnosti kolaborativnog filtriranja i sličnost filtriranja po žanrovima, po kojima smo u tim vrstama sustava uspoređivali i davali preporuke, sada samo spojili kako bi dobili novi hibridni sustav. Kod ove vrste preporuke primjećujemo to da se neki filmovi dosta često preporučuju, iako na prvi pogled nisu povezani s filmom za kojeg tražimo preporuku. Na primjer film Početak (engl. *Inception*) često

je izlazio kao preporuka. Pretpostavka zašto se to događa je ta da se Početak ubraja u puno žanrova, pa zbog toga ima veliku sličnost s puno, zapravo nepovezanih filmova. Na slici 5.4 vidimo da je za jednu od preporuka, za film Priča o igračkama (engl. *Toy Story*), dan film Početak. Na slici 5.5 vidimo da je Početak preporučen i kao preporuka za jedan od filmova iz serijala o Harry Potter-u. Film Priča o igračkama je dječiji animirnani film, serijal Harry Potter su fantastičnog žanra i filmovi su koje mogu bez problema gledati i razumjeti djeca. Dok je Početak film poprilično mračnog i akcijskog ozračja, također je film koji jako dobro treba pratiti kako bi se shvatilo o čemu se radi u filmu. Iz tog razloga on objektivno nije film koji bi se preporučio kada se traži preporuka za filmove kao što su serijal Harry Potter i Priča o igračkama.

```
CONNECTED
3021                                    Toy Story 2 (1999)
0                                         Toy Story (1995)
4201                                          Shrek (2001)
2895                        Who Framed Roger Rabbit? (1988)
14937                                     Inception (2010)
359                                  Lion King, The (1994)
1818                                          Mulan (1998)
1171     Good, the Bad and the Ugly, The (Buono, il bru...
587                             Beauty and the Beast (1991)
4780                                  Monsters, Inc. (2001)
```

Slika 5.4: Inception kao preporuka za Toy Story - povezano



Slika 5.5: Inception kao preporuka za Harry Potter film - povezano

6. Zaključak

Zadatak ovog završnog rada bila je izrada vlastitog sustava za preporuku filmova. U radu su implementirane tri vrste sustava za preporuku. Učinkovitost sustava za preporuku općenito teško je ocijeniti samo na temelju rezultata, zbog toga što ne znamo što će se točno svidjeti nekom korisniku. Na primjer film po svim izračunima može biti savršen za korisnika, i na temelju kolaborativnog filtriranja i na temelju žanrova i na temelju glumačke postave, ali iz nekog potpuno subjektivnoga razloga korisnik može mrziti film. Sustav za preporuku neće moći predvidjeti takve stvari. Iz tog razloga teško je objektivno ocijeniti rad sustava za preporuku implementiranog u ovom radu.

Dobra strana kolaborativnog filtriranja i filtriranja po žanrovima implementiranih u ovom radu je ta da daju nepoznatije flimove kao preporuke. Što je, kako je već ranije objašnjeno, poželjana karakterisika sustava za preporuku. Funkcija kojom se povezuju sličnosti kolaborativnog filtriranja i filtriranja po žanrovima, kod implementacije sustav povezanih karakteristika te dvije vrste filtriranja, je obična funkcija aritmetičke sredine. Moguće poboljšanje sustava je korištenje kompliciranije funkcije za spajanje dviju sličnosti, također jedno poboljšanje može biti i dodavanje težina korištenima sličnostima. Tako bi mogli utjecati na važnost pojedine vrste filtriranja prilikom povezivanja više njih.

Poboljšanje kolaborativnog filtriranja mogli bi dobiti proširenjem skupa podataka. Tako bi naš sustav bio "pametniji", imao bi više informacija i o filmovima, ali i o korisnicima, te bi onda njegove preporuke bile puno točnije. Sustavu bi se mogla dodati i memorija s kojom bi unosili nove podatke u sustav, pamćenjem prošlih korisnikovih interakcija sa sustavom. Time bi osigurali da ne radimo s potpuno zastarjelim podacima.

LITERATURA

[1] What is deep learning? | IBM. URL https://www.ibm.com/topics/deep-learning.

[2] Cosine similarity - an overview | ScienceDirect topics, . URL https://www.sciencedirect.com/topics/computer-science/cosine-similarity.

[3] Netflix hrvatska – gledaj serije na mreži, gledaj filmove na mreži, . URL https://www.netflix.com/hr/.

[4] Registrirajte se - spotify, . URL https://www.spotify.com/hr-hr/signup?forward_url=https%3A%2F%2Fopen.spotify.com%2F.

[5] Understanding TF-IDF for machine learning, . URL https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/.

[6] What is a recommendation system?, . URL https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/.

[7] Mukesh Chaudhary. TF-IDF vectorizer scikit-learn. URL https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a.

[8] F. Maxwell Harper i Joseph A. Konstan. The MovieLens datasets: History and context. 5(4):1–19. ISSN 2160-6455, 2160-6463. doi: 10.1145/2827872. URL https://dl.acm.org/doi/10.1145/2827872.

[9] Khang Pham. What are recommendation systems? URL https://medium.com/@khang.pham.exxact/what-are-recommendation-systems-6bb5036042db.

[10] Baptiste Rocca. Introduction to recommender systems. URL https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada.

[11] Sindhu Seelam. Machine learning fundamentals: Cosine similarity and cosine distance. URL https://medium.com/geekculture/cosine-similarity-and-cosine-distance-48eed889a5c4.

[12] Valentina. Introduction to recommender systems. URL https://thingsolver.com/introduction-to-recommender-systems/.

Sažetak

Sustavi za preporuku su područje strojnog učenja, koje nam pomaže pri radu s ogromnim količinama podataka s kojom smo u današnje doba zatrpani. Ovaj rad bavi se implementacijom sustava za preporuku filmova. Implementirane su tri različite vrste sustava za preporuku; kolaborativno filtriranje, filtriranje po sadržaju i spoj te dvije vrste. Pokazano je da kolaborativno filtriranje i filtriranje po sadržaju daju relativno slične preporuke. Dok implementacija preporuke povezivanjem te dvije vrste daje nepovezane preporuke.

Ključne riječi: sustav za preporuku, strojno učenje, kolaborativno filtriranje, filtriranje po sadržaju

Movie recommendation system

Abstract

Recommendation systems are a field of machine learning, which helps us work with the huge amounts of data we are overwhelmed whit today. This paper deals with the implementation of a movie recommendation system. Three different types of recommendation systems have been implemented; collaborative filtering, content based filtering and a combination of the two types. It has been shown that collaborative filtering and content based filtering provide similar recommendations. While implementation of recommendation system, which is based on connecting collaborative and content based filtering, gives unrelated recommendations.

Keywords: recommendation system, machine learning, collaborative filtering, content based filtering