Sustavi preporuke: predikcija korisničkih ocjena

Tehnička dokumentacija

Tomislav Bjelčić

28. siječnja 2021.

OPIS ZADATKA

Cilj ovog projekta je bio upoznavanje sa sustavima preporuke: kako funkcioniraju, gdje se primjenjuju te koji algoritmi se koriste kako bi učinkovito radili. Zadatak je bio razviti jednostavnu početnu verziju jednog takvog sustava koji će za specificiranog korisnika i predmet u domeni odrediti ocjenu kojom bi taj korisnik ocijenio navedeni predmet.

SKUP PODATAKA

Skup podataka (dataset) koji određuje domenu ovog sustava preporuke je podskup Netflixove baze podataka koji je objavljen 2006. godine za natjecatelje koji žele sudjelovati u natjecanju Netflix Prize. Skup podataka se sastoji od 17770 filmova i za svaki film je dan popis svih korisnika koji su ocijenili taj film, zajedno sa datumom i samom ocjenom od 1 do 5. Svaki film i svaki korisnik ima svoj jedinstveni identifikator (ID) koji se za filmove kreće od 1 do 17770, a za korisnike od 1 do 2649429, doduše iz tog raspona postoji samo 480189 korisnika. Ocjena postoji oko 100 milijuna, no u sklopu ovog projekta, radi jednostavnosti, u obzir je uzet samo manji dio ocjena.

KORIŠTENE METODE

Ovaj preporučiteljski sustav koristi item-item collaborative filtering pristup kako bi, kao prvi korak rada, za film *f* odredio skup njemu najsličnijih filmova. Kao funkciju sličnosti između filmova *a* i *b* korištena je sljedeća formula:

$$sim(a,b) = \frac{\sum_{u \in U_{ab}} (r_{ua} - \overline{r_u})(r_{ub} - \overline{r_u})}{\sqrt{\sum_{u \in U_{ab}} (r_{ua} - \overline{r_u})^2} \sqrt{\sum_{u \in U_{ab}} (r_{ub} - \overline{r_u})^2}}$$

gdje je U_{ab} skup korisnika koji su ocijenili film *a* i film *b*, r_{ui} označava ocjenu koju je dao korisnik *u* filmu *i*, a $\overline{r_u}$ označava prosječnu ocjenu korisnika *u*. U slučaju da korisnik *u* nije ocijenio niti jedan film, kao prosjek se uzima 0. Ocjene su korigirane sa prosječnom ocjenom korisnika kako bi se uzela u obzir činjenica koja ocjena zapravo predstavlja visoku ocjenu za nekog korisnika. Ako korisnik većinom daje ocjene 5, njegova ocjena 5 je kao kad neki drugi korisnik, koji u prosjeku daje ocjene 3, ocijeni taj isti film sa ocjenom 3. Skup najsličnijih filmova S_f nekom filmu *f* je određen na sljedeći način:

$$S_f = \{skup\ svih\ filmova\ g \neq f\ za\ koje\ vrijedi: sim(f,g) \geq C\}$$

gdje je C neka predefinirana konstanta. Za ovaj projekt vrijednost te konstante je $C = 0.2$. U Nakon što se odredi skup najsličnijih filmova S_f, predikcija ocjene r_{uf} korisnika *u* za film *f* se tada računa po sljedećoj formuli:

$$r_{uf} = \frac{\sum_{g \in F} sim(f,g)\, r_{ug}}{\sum_{g \in F} sim(f,g)}$$

gdje je F presjek skupa S_f i skupa filmova koji su ocijenjeni od strane korisnika *u*. U slučaju da je skup F prazan, kao predikcija se koristi prosječna ocjena korisnika *u*. Ako korisnik *u* nije ocijenio niti jedan film, kao predikcija se koristi prosječna ocjena filma *f*. Ako uz to za film *f* ne postoji niti jedna ocjena, onda je predikcija ocjene 3.0.

IMPLEMENTACIJA I KORIŠTENE BIBLIOTEKE

Ovaj jednostavan sustav je implementiran u programskom jeziku Java. Uz Javine standardne biblioteke, korištene su i dvije vanjske biblioteke:

- Apache Commons Math 3.6.1 – korištena podrška za rijetke matrice za spremanje podataka o ocjenama.
- Apache Commons Collections 4.4 – korištena podrška za dvosmjerne mape (bijekciju).

REZULTATI

Kao rezultat projekta razvijena je jednostavna aplikacija sa grafičkim korisničkim sučeljem u koju se unosi brojevi koliko filmova i koliko korisnika se želi učitati te putanju do korijenskog direktorija sa datotekama dataseta. Nakon što se dataset učita u memoriju, onda se može unijeti ID korisnika te ID filma, a aplikacija će izračunati i ispisati predikciju za unesenog korisnika i uneseni film.