1. domača naloga pri predmetu Podatkovno rudarjenje

Priprava podatkov, osnovne statistike, vizualizacija

6. marec 2017

1 Uvod

Neizogiben del vsakega projekta na področju podatkovnega rudarjenja je iskanje, urejanje in priprava podatkov. V tej nalogi boste spoznali primer podatkovne zbirke, pretvorbo podatkov v ustrezno obliko ter pregled in prikaz osnovnih statistik.

2 Podatki

V nalogi boste pregledali in pripravili podatke gledanosti Hollywoodskih filmov v obdobju 1995-2016, MovieLens. Podatke naložite s spletne učilnice. Iste podatke boste uporabili v vseh domačih nalogah, zato jih dodobra spoznajte.

Gre za podatkovno zbirko za vrednotenje priporočilnih sistemov, ki vsebuje gledalce ter njihove ocene za posamezni film na lestvici 1 do 5. Poleg osnovne matrike uporabnikov in ocen vsebuje še dodatne podatke o filmih (npr. žanr, datum, oznake, igralci).

Podatkovna zbirka vsebuje naslednje datoteke:

- `ratings.csv` podatki o uporabnikih in ocenah,
- `movies.csv` podatki o žanrih filmov,
- `cast.csv` podatki o igralcih,
- `tags.csv` podatki o oznakah (ang. *tags*),
- `links.csv` povezave na sorodne podatkovne zbirke.

Pred pričetkom reševanja naloge si dobro oglejte podatke in datoteko `README.txt`. Pripravite metode za nalaganje podatkov v ustrezne podatkovne strukture. Te vam bodo prišle prav tudi pri nadaljnih nalogah. Bodite pozorni na velikost podatkov.

3 Vprašanja

Glavni namen podatkovnega rudarjenja je *odkrivanje znanj iz podatkov*, torej odgovarjanje na vprašanja z uporabo računskih postopkov.

Z uporabo principov, ki ste jih spoznali na vajah in predavanjih, odgovorite na spodnja vprašanja. Pri vsakem vprašanju dobro premislite, na kakšen način boste najbolje podali, prikazali oz. utemeljili odgovor. Bistven del so odgovori na vprašanja in ne implementacija vaše rešitve.

1. (15 %) Kateri filmi so v povprečju najbolje ocenjeni? Pripravite seznam filmov ter njihovih povprečnih ocen in izpišite po 10 filmov z vrha seznama. Opazite pri takem ocenjevanju kakšno težavo? Kako bi jo lahko rešili? Kakšni so rezultati tedaj?

2. (15 %) Posamezni film pripada enemu ali več žanrom. Koliko je vseh žanrov? Prikaži porazdelitev žanrov z uporabo ustrezne vizualizacije.

3. (20 %) Število ocen (ogledov) se za posamezni film razlikuje. Ali obstaja povezava med gledanostjo in povprečno oceno filma? Opišite postopek, ki ste ga uporabili pri odgovarjanju na vprašanje.

4. (30 %) Vsaka ocena je bila vnešena na določen datum (stolpec *timestamp*). Ali se popularnost posameznih filmov s časom spreminja? Problem reši tako, da za dani film ocene razporediš po času ter v vsaku časovni točki izračunaš povprečje za zadnjih 30, 50, ali 100 ocen. Nariši graf, kako se ocena spreminja in ga prikaži za dva zanimiva primera filmov.

5. (20 %) Kako bi ocenili popularnost posameznih igralcev? Opišite postopek ocenitve ter izpišite 10 najbolj popularnih igralcev.

6. (Bonus 1%) Kateri je tvoj najljubši film? Zakaj?

4 Zapiski

Za nalaganje podatkov lahko uporabite vgrajeno knjižnico `csv`.

```
from csv import DictReader

reader = DictReader(open("ratings.csv", "rt", encoding="utf-8"))
for row in reader:
    user = row["userId"]
    movie = row["movieId"]
    rating = row["rating"]
    timestamp = row["timestamp"]
```

Pretvorba časovnega formata *Unix time*.

```
from datetime import datetime

t = 1217897793 # Unix-time
date = datetime.fromtimestamp(t).strftime('%Y-%m-%d')
```

5 Oddaja poročila

Oddaja vključuje datoteko `vpisnast_priimek_ime.zip` z naslednjo vsebino:

- Poročilo z odgovori na vprašanja. Oddajte tako datoteko `.tex` kot `.pdf`. **Pomembno: oddaje, ki ne bodo vsebovale poročil, ne bodo ocenjene.** Vzorec poročila najdete na **spletni učilnici predmeta**.
- morebitne slike, ki jih vsebuje poročilo,
- vso izvorno kodo za pridobitev rezultatov.