Priprava podatkov, osnovne statistike, vizualizacija

Amon Stopinšek (63150273)

19. marec 2017

1 Uvod

V nalogi smo se spoznali s podatkovno zbirko MovieLens 1995-2016. Cilj naloge je bila uporaba osnovnih veščin podatkovnega rudarjenja: branje podatkov iz datoteke, priprava in obdelava podatkov ter vizualiacija.

2 Podatki

Pri nalogi smo uporabili podatkovno zbirko Movie Lens. Za potrebe naloge smo uporabili le podatke iz datotek movies.csv, ratings.csv in cast.csv.

Datoteka movies.csv vsebuje podatke o 9125 filmov, ki so opisani z id-jem filma, naslovom in žanri.

V datoteki ratings.csv je 104 000 ocen filmov. Vsaka ocena je predstavljena z id-jem uporabnika, id-jem filma, oceno in časovno značko.

Cast.csv vsebuje igralsko zasedbo posameznega filma. Podatki so opisani z id-jem filma in imeni ter priimki igralcev, ki so med seboj ločen z '|'.

3 Metode

Za branje podatkov smo uporabil knjižico csv in razred DictReader. Pri branju podatkov smo le te shranili v slovar. Kot ključ smo najpogosteje uporabili id filma, za vrednost pa seznam s terkami.

```
reader = DictReader(open('data/ratings.csv', 'rt', encoding='utf-8'))
movieRatings = dict()

for row in reader:
    user = row['userId']
    movie = row['movieId']
    rating = row['rating']
    timestamp = int(row['timestamp'])

    if movie not in movieRatings.keys():
        movieRatings[movie] = []

    movieRatings[movie] = movieRatings[movie] + [(timestamp, float(rating))]
```

Za sortiranju podatkov smo uporabili metodi sort() in sorted(), odvisno od oblike podatkov, ki smo jih želeli urediti.

```
genres = sorted(genres.items(), key=lambda x: x[1])[::-1]
```

Pri vizualizaciji smo uporabili knjižico matplotlib. Podatke smo predstavili z navadnim grafom ali histogramom.

```
plt.figure(figsize=(20, 15))
plt.bar(x, [n for genre, n in genres])
plt.xlim(-0.5, len(genres) - 0.5)
plt.xticks(x)
plt.gca().set_xticklabels([genre for genre, n in genres], rotation=90)
plt.ylabel('Stevilo filmov')
plt.show()
plt.savefig('genres dist.png')
```

4 Rezultati

4.1 Najbolje povprečno ocenjeni filmi

Tabela 1: Najbolje povprečno ocenjeni filmi.

film	povprečna ocena
Godfather, The (1972)	4.488
Shawshank Redemption, The (1994)	4.487
Maltese Falcon, The (1941)	4.387
Godfather: Part II, The (1974)	4.385
Usual Suspects, The (1995)	4.371
Chinatown (1974)	4.336
Rear Window (1954)	4.315
12 Angry Men (1957)	4.304
Schindler's List (1993)	4.303
City of God (Cidade de Deus) (2002)	4.297

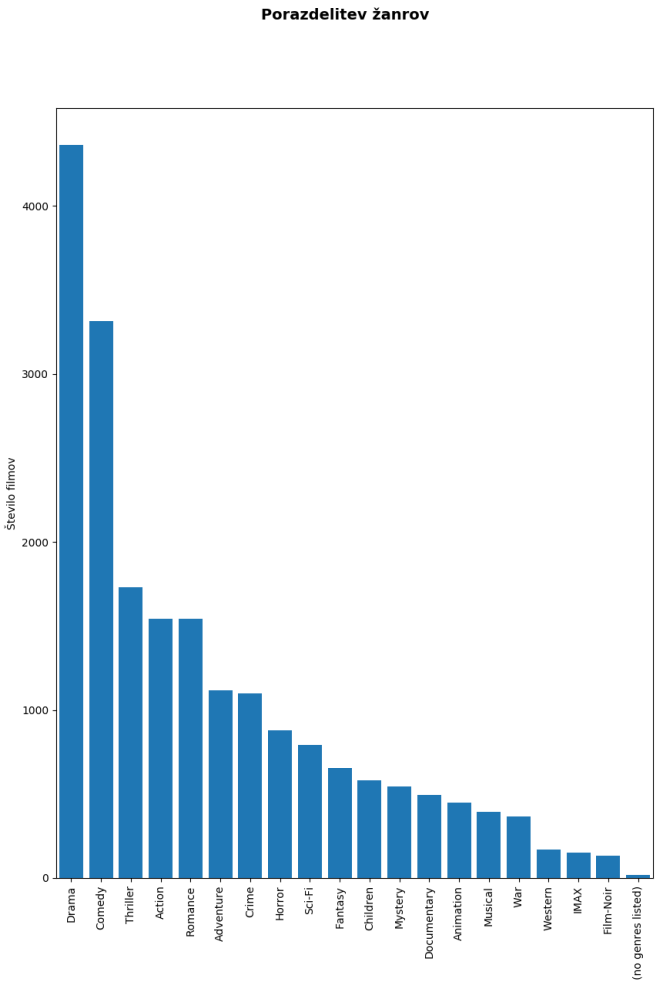
Pri razvrstitvi filmov po povprečni oceni (prvi poskus 3) se je pojavil problem, da so bili najvišje na tej lestvici filmi s samo eno oceno. Problem smo rešili tako, da smo na lestvico 1 uvrstil le filme z več kot 50 ocenami.

4.2 Žanri

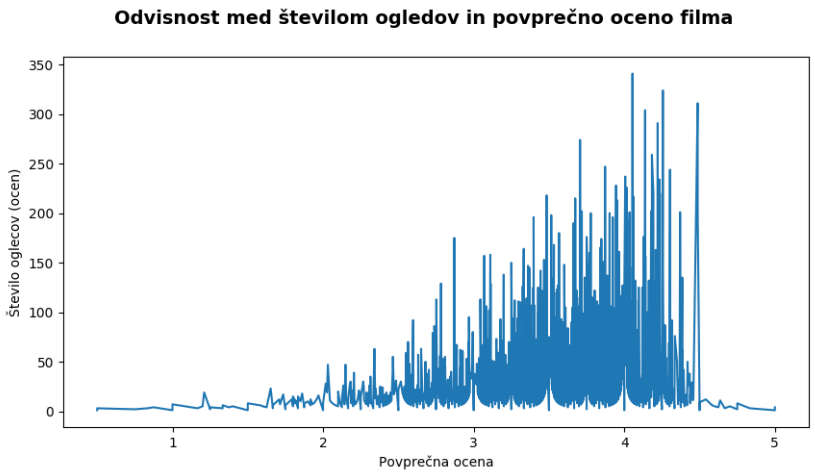
Filmi pripadajo 19 oz. 20 (če kot žaner štejemo tudi 'no genres listed') žanrom. Iz histograma 1 lahko vidimo, da je najbolj pogost žanr filma drama.

4.3 Povezava med gledanostjo in povprečno oceno filma

Vizualizacija 2 ni najbolj posrečena, še vseeno pa lahko iz nje vidimo, da imajo povprečno zelo slabi oz. zelo dobri filmi zelo majhno število ogledov, običajno pod 25. Če od daleč pogledamo



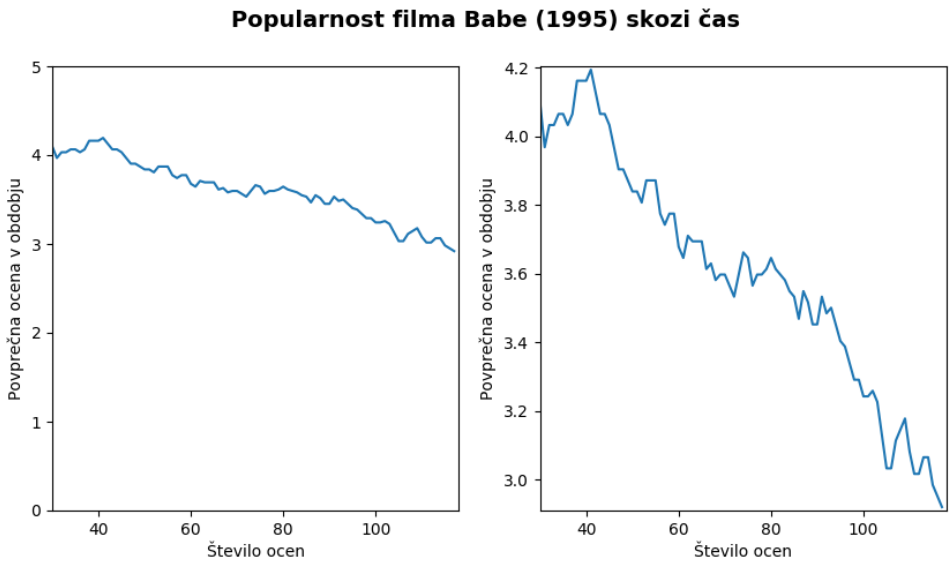
Slika 1: Porazdelitev žanrov po številu filmov



Slika 2: Povezava med povprečno oceno filma in številom ogledov

graf lahko vidimo, da imajo filmi z več ogledi boljšo povprečno oceno (okoli 4), tisti z manj ogledi pa slabšo (med 2 in 3). Izjema so že omenjeni filmi z malo ogledi in izredno slabo ali dobro povprečno oceno.

4.4 Popularnost filmov skozi čas



Slika 3: Popularnost filma Babe skozi čas

Večina filmov nima velikih skokov v popularnosti skozi čas, če izvzamemo prvih nekaj ocen. Histogram 4 prikazuje, razliko v povprečni oceni med dvema zaporednima časovnima obdobjema (za dolžino obdobja 30 ocen). Zanimiva primerka filmov, ki izstopata iz povprečja sta filma Babe 3 (povprečna ocena skozi čas pada iz 4,2 do 2,8) in Beauty and The Beast 5 (povprečna ocena za kratko časovno obdobje skoči za 0,8 in nato pade za več kot 1).

4.5 Popularnost posameznih igralcev

Za izračun popularnosti posameznih igralcev 2 smo izbrali preprosto formulo. Za nastop v filmu smo igralcu k oceni popularnosti prišteli povprečno oceno filma. Ocena popularnosti igralca je tako sestavljena iz vsote povprečnih ocen filmov v katerih je nastopal.

4.6 Najljubši film

Moj najljubši film je V for Vendetta Všeč mi je zaradi sporočila ki ga nosi - če se ljudje v dovolj velikem številu za nekaj združimo in zavzamemo lahko spremenimo svet na bolje.

5 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.

Tabela 2: Popularni igralci

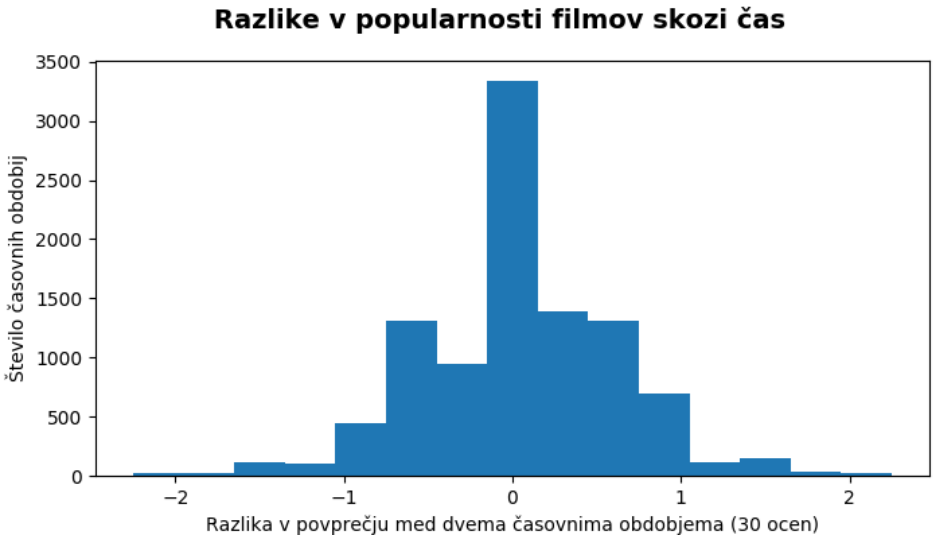
igralec	ocena popularnosti
Harrison Ford	9041.5
Tom Hanks	8426.0
Bruce Willis	6693.5
Robert De Niro	6159.0
Morgan Freeman	6063.0
Brad Pitt	5516.0
Kevin Spacey	5172.5
Tom Cruise	5149.5
Bill Murray	4903.5

Priloge

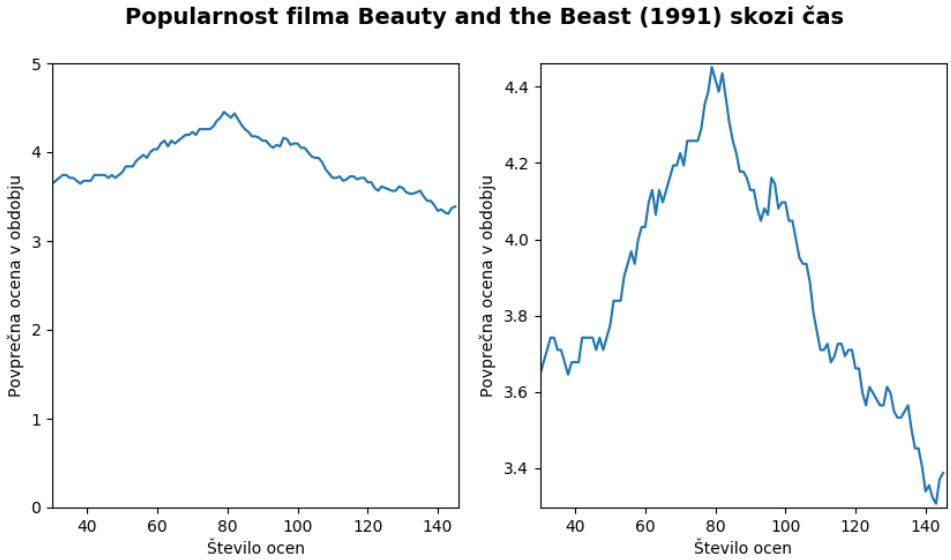
A Podrobni rezultati poskusov

Tabela 3: Najbolje povprečno ocenjeni filmi.

film	povprečna ocena
Zerophilia (2005)	5.0
Zelary (2003)	5.0
Z Channel: A Magnificent Obsession (2004)	5.0
Yossi (Ha-Sippur Shel Yossi) (2012)	5.0
Wrong Cops (2013)	5.0
Wrong (2012)	5.0
World of Tomorrow (2015)	5.0
Woman on the Beach (Haebyeonui yeoin) (2006)	5.0
Woman on Top (2000)	5.0
Wolf Children (Okami kodomo no ame to yuki) (2012)	5.0



Slika 4: Sprememba v popularnosti med dvema zaporednima časovnima obdobjema



Slika 5: Popularnost filma Beauty and the Beast skozi čas