2. domača naloga pri predmetu Podatkovno rudarjenje

# Osnovne vizualizacije in iskanje osamelcev

16. marec 2015

## 1 Uvod

V pričujoči domači nalogi bomo uporabili osnovne vizualizacije podatkov in porazdelitev ter metode za iskanje osamelcev.

## 2 Podatki

V nalogi uporabite podatke gledanosti Hollywoodskih filmov v obdobju 1996-1998, MovieLens, ki ste jih pripravili v prvi domači nalogi. Pri vseh vprašanjih upoštevajte le:

- filme z najmanj 20 ogledi,
- igralce, ki so igrali v najmanj 5 filmih,
- vse žanre razen 'unknown'.

## 3 Vprašanja

1. (20 %) Kakšne so *porazdelitve povprečnih ocen* po posameznem žanru? Ali se porazdelitve bistveno razlikujejo med žanri? Kakšna je oblika porazdelitve? Se v čem razlikuje od normalne porazdelitve? Nalogo rešite tako, da izberete ustrezno vizualizacijo za prikaz porazdelitev in slike shranite v datoteke z imenom `n1_vpisnastevilka_zanr.png`. Slike vključite v poročilo.
2. (20 %) Ali je število ogledov povezano s povprečno oceno filma? Nalogo rešite tako, da izberete ustrezno vizualizacijo za prikaz porazdelitev in rezultat shranite v datoteko z imenom `n2_vpisnastevilka.png`. Sliko vključite v poročilo.
3. (30 %) Poiščite pare najboljše ocenjenih filmov. Za dani par filmov, upoštevajte (izberite) le ocene od uporabnikov, ki so si ogledali oba filma. Iz izbranih ocen izračunajte povprečni oceni filmov v paru, ter to prikažite z ustrezno vizualizacijo (namig: rišete porazdelitev dveh spremenljivk). Sliko shranite v datoteko z imenom `n3_vpisnastevilka.png`. Izpišite

tudi pet v povprečju najbolje ocenjenih parov filmov (izračunajtve povprečje povprečnih ocen obeh filmov v paru).

4. (30 %) Kako posamezni uporabniki ocenjujejo žanre? Pretvorite vsakega uporabnika v vektor dolžine 19, ki predstavlja povprečje njegovih ocen po posameznem žanru (vseh žanrov brez 'unknown' je 19). V primeru, ko uporabnik pri posameznem žanru ni ocenil nobenega filma, namesto ničle vstavite povprečno oceno žanra. Na tako zgrajenih podatkih poiščite pet osamelcev (glede na Evklidsko razdaljo [1]), ki najbolj izstopajo (imajo *pozitivno* z-vrednost), ter izpišite pripadajoče vektorje s povprečji. Seznam osamelcev, njihove z-vrednosti in vektorje povprečnih ocen vključite v poročilo v obliki tabele LaTeX. V čem izstopajo ti posamezniki?

5. (Bonus 15%) Za vsakega igralca izračunajte povprečje ocen filmov, v katerih igralec nastopa. Narišite *porazdelitev povprečnih ocen*. Sliko porazdelitve shranite v datoteko z imenom `nb_vpisnastevilka.png`. Katerih pet igralcev je v povprečju najbolje ocenjenih? Seznam igralcev in povprečne ocene filmov, v katerih nastopajo, vključite v poročilo v obliki tabele LaTeX.

# 4 Rezultati

Rezultat naloge so:

- slike `n1_vpisnastevilka_zanr.png`,
- slika `n2_vpisnastevilka.png`,
- slika `n3_vpisnastevilka.png`,
- slika `nb_vpisnastevilka.png`,
- seznam najbolje ocenjenih parov filmov in igralcev (tabele v LaTeX-u),
- seznam osamelcev (uporabnikov, ki izstopajo), z-vrednosti in pripadajočih vektorjev,
- izvorna koda programov (datoteke .py, ...),
- poročilo z odgovori na vprašanja. Oddajte datoteko `.tex` in `.pdf` poročila.

**Pomembno:** oddaje, ki ne bodo vsebovale poročil, ne bodo ocenjene. Vzorec poročila najdete na **spletni učilnici predmeta.**

**Pomembno:** rezultate, ki nastopajo v obliki tabel, podajte kot tabele v LaTeX-u (in ne slike zaslona).

Vse datoteke oddajte v eni mapi - arhivu, brez poddirektorijev,
poimenovanem `priimek_ime_vpisna.zip`.

[1] Pomagajte si z http://orange.biolab.si/docs/latest/reference/rst/Orange.data.outliers.html