Fakultet Elektrotehnike i Računarstva

3. Domaća zadaća - ROVKP

Vinko Kolobara
21. svibnja 2017.

1 Zadatak: Indeksiranje tekstualne kolekcije

Koliko se zapisa nalazi u izlaznoj datoteci?
U izlaznoj datoteci se nalazi 11175 $\binom{100}{2}$ zapisa.

Koja šala je najsličnija šali s ID-jem 1?
Najsličnija je ona šala sa ID-jem 87.

Vidite li zašto su te dvije šale slične?
U obje šale se spominje doktor.

Što mislite, hoće li preporuka po sadržaju imati smisla u slučaju ovih šala?
Trebala bi imati smisla. Ljudima bi se trebale sviđati šale slične po sadržaju

2 Zadatak: Izgradnja i evaluacija centraliziranog preporučitelja

Kojih 10 preporuka je za korisnika s ID-jem 220 je izračunao prvi, a koje drugi preporučitelj?
Prvi preporučitelj: 141, 140, 44, 22, 36, 52, 86, 37, 129, 43
Drugi preporučitelj: 62, 68, 105, 66, 53, 104, 35, 114, 148, 106

Koje preporučitelj ima bolju kvalitetu?
Nešto bolju kvalitetu ima prvi preporučitelj.

Je li za ove ulazne podatke bolje koristiti mjeru log-likelihood ili Pearsonovu korelaciju u slučaju drugog preporučitelja?
Bolje je koristiti log-likelihood.

3 Zadatak: Pokretanje raspodijeljenog preporučitelja

Listing 1: Pokretanje raspodijeljenog preporučitelja

```
mahout recommenditembased \
--similarityClassname SIMILARITY_PEARSON_CORRELATION \
--input data/jester_ratings.csv \
--usersFile data/users.txt \
--output data/result \
--numRecommendations 10
```

Koliko ste MapReduce poslova izvršili u vašem kôdu?
9 MapReduce poslova je izvršeno.

Pogledajte izlazne datoteke i objasnite postoji li razlika u izračunatim preporukama u odnosu na preporučitelja iz 2. zadatka?
Postoje razlike, prvenstveno zato što je u 2. zadatku korištena druga mjera sličnosti koja ne daje iste rezultate kao Pearsonova korelacija.