CS839 Project Stage 3 Blocking Rules

Pan Wu, Dongqiangzi Ye, Sek Cheong

Iteration 1

Size before blocking: 675218
Density: 0.02
Blocking rules:
In the set B, we found that many pairs of movie name are different. So we do Jaccard measure in the name attribute (discard stop words).
Size after blocking: 1336
The density then is greater than 0.2 then we stop.

Table L

We random sample 400 instances of the reduced candidate sets.