CS 839 : Stage 3 Report

Abbinaya Kalyanaraman Bob Effinger Rajan Dalal

Team 9

Estimating Accuracy

We obtained a candidate set file 'Job_Movie_apply_rules_ds.csv' containing 770 tuples.

We computed the density by selecting the first 50 tuples from our candidate set.

We checked if each tuple containing A_id and B_id was a match or non match and computed our density based on that.

The code computing our density can be found in this file shared in the repository: density_check.ipynb

We obtain a density of 0.8, and hence did not have to write additional blocking rules since the density is much higher than 0.2.

We added 350 tuples to this reduced candidate set and manually labeled the 400 tuples to create the file labeled.csv.

After running the Jupyter notebook code on the above labeled csv file, we obtain the following results:

Recall = [1.0 - 1.0]

Precision = [0.972 - 0.987]