

Entity Matching

Anshu Verma, Srujana, Arpit Jain

{averma27,srujana,ajain74}@wisc.edu

DataSet:

Table A - IMDb_Dataset[[link](#)]

Size : 3475

Table B - RottenTomatoes_DataSet[[link](#)]

Size : 3060

Candidate Set [[link](#)]

Size : 867840

Prediction List [[link](#)]

Size: 1289

Candidate Set after applying blocking rule [[link](#)]

Size: 1250

Blocking Rule

We have applied several blocking rule to reduce our candidat set from 867840 to 1250 as listed below and also find the code here[[link](#)].

1. Applied jaccard on the two tables from candidate set on 'Release Year' with tolerance ± 1 .
 - [Candidate set](#) reduced to 66443 from 867840.
2. We apply blocking rule 2 on the output of blocking rule 1. In this rule, we check the following.
 - Jaccard measure of the 3 grams on the Movie Name ≥ 0.6 .
 - Jaccard measure of the 3 grams on the Movie Name is between (0.3, 0.6).
 - Release year should have an absolute difference of 1 year.
 - Jaccard measure between Director Name > 0.6 .
 - The second part of this rule is applied to retrieve the following kind of entity matches:
 - MISSION: IMPOSSIBLE II v/s MISSION: IMPOSSIBLE 2

- LES QUATRE CENTS COUPS v/s THE 400 BLOWS (LES QUATRE CENTS COUPS)
- [Candidate set](#) reduced to 1250 from 66443.
3. Drop duplicates (remained the same)

Iteration:

1. We sampled 50 tuples from our final candidate set and manually label them. We got 2 FP out of 50, thus the density is $48/50 = 0.96$.

We sampled 400 tuples from our final candidate set and manually label them. We got 5 FP out of 400, thus the density is $395/400 = 0.9875$

Sampled Candidate Set (400) [[link](#)]

Labeled Candidate Set (400) [[link](#)]

Estimating Precision and Recall on our Sampled Candidate Set:

Recall : [1.0 - 1.0]

Precision : [0.945 - 0.976]