CS 839 Project Stage 3: Blocking Rule Description

Team Members:

Shuoxuan Dong sdong34@wisc.edu
Ziyun Zeng ziyun.zeng@wisc.edu
Chushi Shi cshi46@wisc.edu

Iteration 1

Size of Candidate set : 26, 262
Density : 0.04 (done by sampling 50 pairs)
Blocking rule added :
Discard tuple pairs as significant non-matches in **time** and **title**
Details : From the output of the sample (50 paires), we found that most of negative samples have significant difference in **time** (released date) and **title**. Because of existing errors, on the one hand, we just match the year but not month and day under a certain tolerance (difference less than 1 in year). For example, we got last two numbers of **time** (eg. `23-June-2015` → `15`) and if the last two numbers of its pair is 16 or 14, they are matched. On the other hand, we use **jaccard** (more than 0.15) to classify **title**. And then we ran the run_debug_blocker function given in jupyter notebook to verify the effectiveness of our blocking rule.

Iteration 2

Size of Candidate Set : 461
Density : 0.71 (done by sampling 50 pairs)
As our density was above 0.2, we moved on to estimating our precision and recall using the estimate_precision_recall module from the jupyter notebook.