CS-839 - Stage3 - PDF file describing blocking rules

Our candidate set had over 33000 entries and predicted matches were 245, so the initial density was negligible(almost zero). So, we decided to try a blocking rule. The rule we chose was to apply a Jaccard similarity measure on the names of the books in the 2 tables. We chose 2 thresholds, 0.5 and 0.2. At 0.5, from 33000 entries we were left with only 324 entries and some true positives were dropped. So, we decided the threshold to be 0.2, which gave 1022 entries in the candidate set. Then, using the debug_blocker, we found that only 1 true match out of 200 was missed out. Providing a lower threshold allows both matches and non-matches in the reduced candidate set. It also allows to match some books that are not heavily dependent on just the name, but also the author and year. Some book names referring to the same book in the 2 tables have completely different ordering of words and some book names have abbreviations in too. Hence, a lower threshold of 0.2 to filter out majority of false matching books served well for us. After this, we randomly sampled 50 entries and labeled them manually to find 18 actual positives, thus giving the density of 18/50, ie. 0.36. Since this was greater than the minimum density requirement (0.2) of Project Stage 3, we stopped making any other blocking rules and used this reduced candidate set of 1022 entries and randomly labeled 400 entries for the next steps in this Stage.