Comp Sci 839 Stage 3

Chen Qian, Jiahan Li, Fangfei Lin

First, we selected 50 pairs from the candidate set and labeled them as 0 or 1. Next, we calculated the density, (number of actual positive matches/all the matches), which was 0%.

Since our density is lower than 0.2, we applied blocking rule to the dataset. We first blocked on the brands of the cosmetics, and applied Jaccard measure. Each brands were delimited by space. If the Jaccard measure is larger than 0.5, we kept the pair.

Second, we blocked on the product names, and applied Jaccard measure. We used 3-gram rule. Similarly, if the Jaccard measure is 0.5, we kept the pair. The size of our candidate set is originally 142,405, and after applying the blocking rule, there are 938 pairs left.

Moreover, we used the notebook and run the 'debug_blocker' module. We found that most pairs are not matching pairs. In the reduced candidate set (size of 938), we randomly selected 50 pairs, labeled then and calculated the corresponding density. The density is 30/50, which is greater than 0.2, thus we stopped the iteration.