ESTIMATING PRECISION and RECALL

Group-3(Lumen Science)

Aditya Rungta Arjun Bala Rohit Kumar Sharma

Size of the candidate set downloaded from cloud matcher = **2997152**
Size of the prediction set downloaded from cloud matcher = **753**

Due to a huge candidate set size relative to prediction set size, the density of matches in the candidate set was very low. So we applied a blocking rule to reduce the candidate set size by eliminating non-matches from the candidate set.

Blocking Rule:
In order to reduce the candidate set, we applied **Jaccard similarity measure** on movie names with a threshold of **0.3**. On applying this blocking rule, the candidate set size reduced to 2029.
On getting this reduced candidate set, we ran the run_debug_blocker function to make sure that the blocking rule did not eliminate a lot of matches.

Once we were sure about the reduced candidate set, we labeled a randomly drawn sample of 400 tuples from the candidate set. The density of actual positives for the first 50 tuples was almost 0.5(greater than 0.2).

We labeled everything and ran the estimate_PR function to get precision and recall.

Our precision and recall values:

```
Recall = [1.0 - 1.0]
Precision = [0.9597402500318443 - 0.9986098101849362]
```