CS 839 – Data Science Project – Stage 3

Estimating Precision and Recall

Team Members:

- Madan Raj Hari – mhari2@wisc.edu
- Raghavan Vellore Muneeswaran – velloremunee@wisc.edu
- Shadana Subramanian – ssubramani23@wisc.edu

Size of the candidate set downloaded from cloud matcher = **93803**
Size of the prediction set downloaded from cloud matcher= **509**

We sampled 50 data and found very few matches, the density was very low. So we went with blocking rule to reduce the candidate size set.

Blocking Rule: We analyzed our two tables and found that overlap blocker is the best choice (even we experimented with other blockers but got better results with the overlap blocker).

Overlap blocker: The overlap blocker helped us find the overlapping strings as either words or q-grams. This helped us narrow down with overlaps between Name and removed the non-matching tuples from the dataset.

Name – overlap-size = **3** with overlap on words

Number of tuple pairs in the candidate set obtained after the blocking step: 1054
Number of tuple pairs in the sample G that you have labelled: 450

Estimating precision and recall: We used the provided Jupyter notebook to compute precision and recall from the following data: We obtained the following results:

Recall = [0.9493719362801628 - 0.9981037925246917]
Precision = [0.9271042891472905 - 0.9719258230923765]