Stage 3 final report

Zhendong Zhou Yunwen Jiang Yixin Chen

In this part, we tried to calculate the precision and recall of entity matching.

The size of candidate set derived from cloud matcher is 92795, which is larger than 500, we took a sample with size 50 to calculate the density, and the output is 0.04, smaller than 0.2, so we decide to do some blocking.

We finally decide to block with the difference of released year smaller than 2 years.

After the blocking, the candidate size became 6696, we also took 50 samples and calculate the density, this time it became 0.2 exactly.

Then we chose 400 samples randomly from it, labeled them and calculate the density, this time it was 0.265, we fed it back to the function that we used to calculate the precision and recall, the final output was:

Precision: (0.9507454786427749, 1.0051845278381064)

Recall: (0.899774426900109, 0.9870180259300797)

The only one problem is that the upper bound of precision is slightly larger than 1, but it does make sense due to the formula we used to calculate it, and at the same time, the output itself seem to meet our need, so we decide to use it as our final consequence.