COMP9313 Project4 Optimization

Z5045582

Yunhe Zhang

1. In stage 2 finding “similar” id pairs, partition using prefixes. In map part, calculate prefix length n, then get first n tokens of record, at last emit the pairs (each token as key, total record as value)
2. In reducer part, calculating similarity of records with same key, but for every two records, they will be calculated only once.
3. For calculating similarity, the records can be regard as an ordered number list from small to large since a smaller ID always means that it appears less frequently, thus the union set and intersection set can be calculated with time complexity O(m + n) where m and n are length of two records.