COMP9313 Report

1. Method for the project

(1) Prefix filtering: minimize the number of items emitted from the mappers.

$$P = |record| - \lceil |record| * t \rceil + 1$$

(2) Compute the length of shared tokens: minimize the number of computing items emitted from method 1.

Step 1: calculated the Jaccard Similarity:

$sim(r, s) = |r \cap s| / |r \cup s|$

Step 2: If $sim(r, s) \geq t, I = |r \cap s| \geq |r \cup s| * t \geq \max(|r|, |s|) * t$

Step 3: Given a record r, we can compute the prefix length as $P = |r| - I + 1$,

Step 4: r and s is a candidate pair, they must share at least one token in the first $(|r| - I + 1)$ tokens.

Step 5: If the record r=(A,B,C,D) and P=2, the mapper emits (A,r) and (B,r).

2. Outcome on AWS