Project Report

Methods for optimization

1. Prefix filtering

I implemented this method to minimize the number of items emitted from the mappers.

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

2. Computing the length of shared tokens

I implemented this method to minimize the number of items emitted from method 1.

In order to minimize the number of items, I need to calculated the Jaccard Similarity:

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

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

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

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

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

Outcome on AWS