Optimization

z5132086 Jingxuan Li

- Prefix filtering

Consider an ordering O of the token universe U and a set of records, each with tokens sorted in the order of O. Let the p-prefix of a record x be the first p tokens of x. If O(x,y) ≥ α, then the (|x|−α+1)-prefix of x and the (|y|−α+1)-prefix of y must share at least one token.

- Size filtering
  - J(x, y) ≥ t => t *|x| ≤ |y|
  - For example, it won't consider<x, w> as a candidate pair, as | w | < 4 (|x| = 5, t =0.8)