Homework 4: Locality Sensitive Hashing

Problem Description

Locality-sensitive hashing (LSH) is an algorithmic technique that hashes similar input items into the same "buckets" with high probability. Given the following shingling matrix and permutations for some documents (d_1, d_2, d_3):

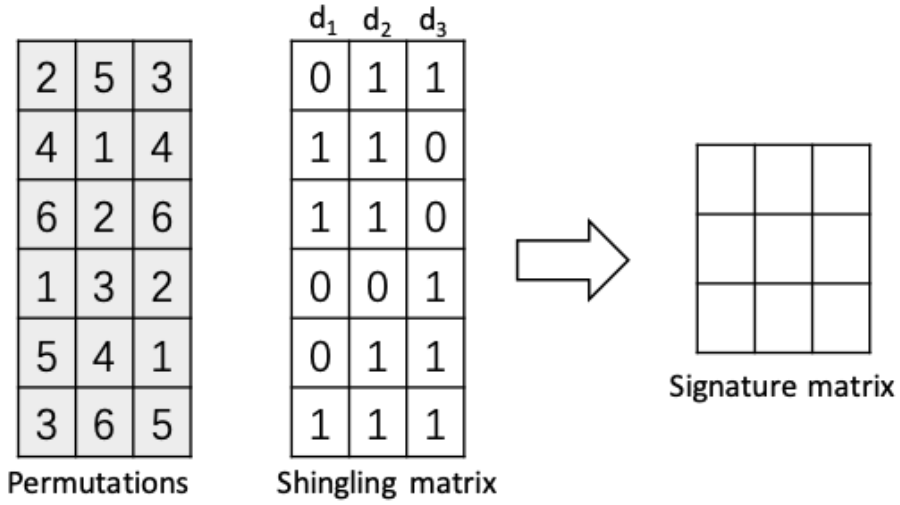

Figure 1: Shingling matrix and permutations

1) Complete the corresponding signature matrix by Mini-Hashing.

2) Compute the Jaccard similarities between documents.

3) If we have a bigger signature matrix, we divide the signature matrix into 10 bands of 4 rows. If two documents C_1 and C_2 have 0.7 similarity in original space, what is probability of C_1 and C_2 are truly similar after LSH?

Solution: Your answers and proofs go here.

Please submit the PDF version to assistant.