Text Analysis

Text Analysis

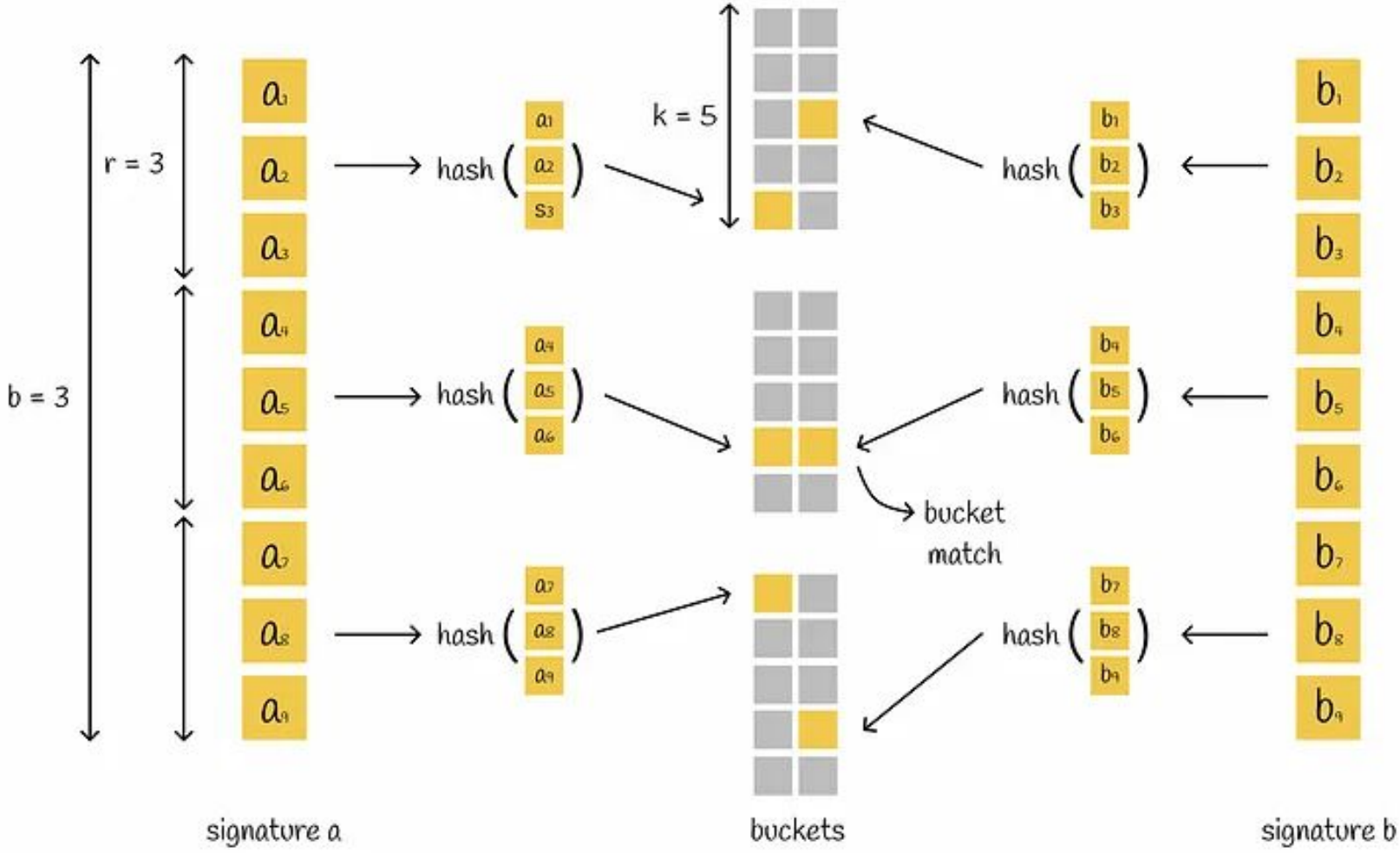
Data Flattening, Filtering and Chunking, Feature Scaling, Dimensionality Reduction

Reference: Book Feature Engineering for Machine Learning

LSH

LSH mechanism builds a hash table consisting of several parts which puts a pair of signatures into the same bucket if they have at least one corresponding part.

LSH takes a signature matrix and horizontally divides it into equal *b* parts called **bands** each containing *r* **rows**. Instead of plugging the whole signature into a single hash function, the signature is divided by *b* parts and each subsignature is processed independently by a hash function. As a consequence, each of the subsignatures falls into separate buckets.


r = 3
b = 3
k = 5
hash
hash
hash
hash
hash
hash
bucket match
signature a
buckets
signature b

LSH

Example of using LSH. Two signatures of length 9 are divided into b = 3 bands each containing r = 3 rows. Each subvector is hashed into one of k possible buckets. Since there is a match in the second band (both subvectors have the same hash value), we consider a pair of these signatures as candidates to be the nearest neighbours.

LSH

LSH mechanism builds a hash table consisting of several parts which puts a pair of signatures into the same bucket if they have at least one corresponding part.

LSH takes a signature matrix and horizontally divides it into equal b parts called bands each containing r rows. Instead of plugging the whole signature into a single hash function, the signature is divided by b parts and each subsignature is processed independently by a hash function. As a consequence, each of the subsignatures falls into separate buckets.

Collaborative Filtering

References:

Book:Mining Massive Datasets

https://www.youtube.com/watch?v=h9gpufJFF-0