**Hash Functions** A hash function h takes a hash-key value as an argument and produces a bucket number as a result. The bucket number is an integer, normally in the range 0 to B − 1, where B is the number of buckets.

**Jaccard Similarity** Similar Items To define similar items we will use the Jaccard Similarity. The Jaccard similarity of sets S and T is |S ∩ T |/|S ∪ T |, that is, the ratio of the size of the intersection of S and T to the size of their union. We shall denote the Jaccard similarity of S and T by SIM(S, T ).

**Shingling of Documents (Sentences)**

The most effective way to represent documents as sets, for the purpose of identifying lexically similar documents is to construct from the document the set of short strings that appear within it. In our particular use-case we have referred to words as shingles.

*Similarity-Preserving Summaries of Sets*

Sets of shingles are large. Even if we hash them to four bytes each, the space needed to store a set is still roughly four times the space taken by the document. Our goal is to replace large sets by much smaller representations called "signatures." The important property we need for signatures is that we can compare the signatures of two sets and estimate the Jaccard similarity of the underlying sets from the signatures alone.

*Matrix Representation of Sets*

It is helpful to visualize a collection of sets as their characteristic matrix. The columns of the matrix correspond to the sets, and the rows correspond to elements of the universal set from which elements of the sets are drawn.

Each of the following examples only uses two sentences to show how to solve the similarity problem. However, the efficiency of the algorithm can be seen when we talk about a large data corpus. We use the algorithm to find similarities in peta-bytes of data

**Minhashing**

The signatures we desire to construct for sets are composed of the results of a large number of calculations, say several hundred, each of which is a "minhash" of the characteristic matrix. To minhash a set represented by a column of the characteristic matrix, pick a permutation of the rows. The minhash value of any column is the number of the first row, in the permuted order, in which the column has a 1.

*Minhashing and Jaccard Similarity*

There is a remarkable connection between minhashing and Jaccard similarity of the sets that are minhashed.

The probability that the minhash function for a random permutation of rows produces the same value for two sets equals the Jaccard similarity of those sets.

*Minhash Signatures*

Again, think of a collection of sets represented by their characteristic matrix M. To represent sets, we pick at random some number n of permutations of the rows of M.

Perhaps 100 permutations or several hundred permutations will do. Call the minhash functions determined by these permutations h1, h2, . . . , hn. From the column representing set S, construct the minhash signature for S, the vector

**References:**

1. <https://www.todaysoftmag.com/article/1553/finding-similar-entities-in-bigdata-models>
2. <https://www.slideshare.net/microlife/3-finding-similar-items>