**JACCARD SIMILARITY:**

**Overview:**

Jaccard(U, V) = (U dotProduct V) / (L22(U) + L22­(V) – (U dotProduct V))

Also note, jaccard of sets A and B = |A and B| / |A or B|

From now on, will use binary vectors and associated jaccard similarity

Suppose you have two documents Da and Db

Suppose also |Da| = n and |Db| = m

Jaccard(Da, Db) = |Da and Db| / |Da or Db|

Given two documents, compute their similarity = O(size of documents)

For a group of documents D1 … Dn:

For I and j, want to compute sim(Di, Dj)

Store each document in a hash table of size k

**Permutation:**

A permutation of set S is a function that is one-to-one and onto

**Random Permutation:**

Suppose we have a function from set A to set B

For all a in A, randomly pick b in B and set f(a) = b

The result is a random permutation

Suppose P is a random permutation on S = {1 … M)

Suppose i and j are elements of S

Prob(P(i) = j) Pr(P(1) != j and P(2) != j and … P(i-1) != j and P(i) = j)