Definizione 0.0.1 (Indice di similarità di insiemi di Jaccard) *Dati due insiemi A e B, la similarità dei due insiemi sarà definita come:*

$$SIM(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Es: due utenti sono definibili simile dall'insieme di oggetti che essi hanno acquistato.

Definizione 0.0.2 (K-gramma (Shingling)) *Stringa di k caratteri che appare consecutivamente in un documento. Se volessimo rappresentare un documento tramite il suo k-gramma, il suo indice di Jaggard del k-gramma misura la similarità testuale dei documenti.*

Definizione 0.0.3 (Stop word) *Parole comuni nel linguaggio naturale ma che non aggiungono particolare valore semantico ad un testo.*

In alcuni contesti vengono tolti gli spazi nei documenti per calcolare il k-gramma di un documento, ma questo in alcuni casi può far perdere informazioni sul documento (Es: "Touch down" nel contesto dell'atterraggio di un aereo o di una partita di rugby).

Definizione 0.0.4 (Matrice rappresentativa di un Insieme) *Le colonne della matrice corrispondono agli insiemi, mentre le righe corrispondono agli elementi del set universale da cui i set sono estratti. Viene posto un 1 nella cella sulla riga r e colonna c se l'elemento r è un membro del set c, altrimenti è 0.*

Elemento	S_1	S_2	S_3	S_4
a	1	0	0	1
b	0	0	1	0
c	0	1	0	1
d	1	0	1	1
e	0	0	1	0

Figure 1: Nella matrice rappresentativa troviamo $\Delta = \{a,b,c,d,e\}, S_1 = \{a,d\}, S_2 = \{c\}, S_3 = \{b,d,e\}, S_4 = \{a,c,d\}$.

0.1 Minhash

Definizione 0.1.1 (Minhash) *Il valore di minhash di una qualsiasi colonna è il primo numero nella prima colonna, nella data permutazione (le righe della matrice possono essere permutate) che ha come valore 1. Ogni calcolo minhash va a costruire la firma di un set, che è composta da un grande numero di questi calcoli.*

0.1.1 Connessione tra minhashing e indice di Jaccard

La probabilità che una funzione di minhash per una permutazione randomica di righe produca lo stesso valore per due insiemi **è uguale** alla similarità di Jaccard per questi due insiemi.