Definizione 0.0.1 (Indice di Jaccard) *Dati due insiemi $\mathbb{S}$ e $\mathbb{T}$, la similarità dei due insiemi sarà definita come:*

$$SIM(\mathbb{S},\mathbb{T}) = \frac{|\mathbb{S} \cap \mathbb{T}|}{|\mathbb{S} \cup \mathbb{T}|}$$

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

Definizione 0.0.2 (K-gramma) *Stringa di k caratteri che appare in un documento.*

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).