As we did earlier with LSA, we will treat documents as *bags of words*, meaning that documents are represented by *sets of word tokens* or *frequency distributions over word tokens*, ignoring order altogether.
PAIRWISE MEASURES
JACCARD SIMILARITY

Cardinality (i.e., size) of the intersection of these two sets divided by the cardinality of their union:

\[ J(S, T) = \frac{|S \cap T|}{|S \cup T|} \]

[Source: Rajaraman & Ullman 2012]
TEXT SHINGLES

- A $k$-shingle is any $k$-length substring of a document.

- A stopword-$k$-shingle consists of a stopword token and the following $k - 1$ tokens in the document.

- We can compute Jaccard similarity using bags or sets of shingles, but this gets expensive quickly as $k$ and the document collection grows.
MANY HASH MINHASH

- Pick a good hash function $h$, and $n$ random numbers
- For each shingle $s$, generate the $n_i$-th hash as:

$$h_i = h(s) \ xor n_i$$

where $\ xor$ is the bitwise exclusive-or (XOR) operator.

Given documents $S$, $T$, let $c$ be the the number of hashes $h_i$ for which $\min h_i(S) = \min h_i(T)$. Then, $c/n$ is an estimate of $J(S, T)$.

[Source: Broder 1997]
TOPIC MODELING
LATENT SEMANTIC ANALYSIS DEMO
LATENT DIRICHLET ALLOCATION

• Simple generative (graphical) model:
  
  • every *document* is a mixture of various *topics*
  
  • every *topic* is a mixture of various words

• Estimation of this model from observed documents is not so simple, but it often produces interpretable “topics”:

  topic #0: 0.009*river + 0.008*lake + 0.006*island + 0.005*mountain + 0.004*area + 0.004*park...
  topic #1: 0.026*relay + 0.026*athletics + 0.025*metres + 0.023*freestyle + 0.022*hurdles...

[Source: http://radimrehurek.com/gensim/wiki.html#latent-dirichlet-allocation]
$w_{i,j}$: $j$-th word in document $i$
$z_{i,j}$: topic of word $w_{i,j}$
$\Theta_i$: topic distribution for document $i$
$\beta$: Dirichlet prior parameter on topic/word distribution
$\alpha$: Dirichlet prior parameter on document/topic distribution

[Source: Wikipedia]
A document can thus be summarized by the mixtures of topics it comprises, and this used for comparison, clustering, and retrieval.