[LRL14] <http://infolab.stanford.edu/~ullman/mmds/book.pdf> (S.79)

Note LSH

(S.79)

(Shingling- k-Shingles: Substring mit Länge k) Ein Dokument muss repräsentiert werden als kleine String, die ins Dokumentauftauchen: Man muss auch die Leerzeichen wegmachen   
„good rule of thumb is to imagine that there are only 20 characters and estimate the number of

K -shingles as 20^k. For large documents, such as research articles, choice k= 9 is considered safe“

(S.88)

Dissimilar pairs that do hash to the same bucket are: **false positives**