Min hash is a technique for quickly estimating how similar two sets are.

In [computer science](http://en.wikipedia.org/wiki/Computer_science), **MinHash** (or the **min-wise independent permutations** [locality sensitive hashing](http://en.wikipedia.org/wiki/Locality_sensitive_hashing) scheme) is a technique for quickly estimating how similar two sets are. The scheme was invented by [Andrei Broder](http://en.wikipedia.org/wiki/Andrei_Broder) ([1997](http://en.wikipedia.org/wiki/MinHash#CITEREFBroder1997)),[[1]](http://en.wikipedia.org/wiki/MinHash#cite_note-b97-0) and initially used in the [AltaVista](http://en.wikipedia.org/wiki/AltaVista) search engine to detect duplicate web pages and eliminate them from search results.[[2]](http://en.wikipedia.org/wiki/MinHash#cite_note-bcfm-1) It has also been applied in large-scale [clustering](http://en.wikipedia.org/wiki/Cluster_analysis) problems, such as clustering documents by the similarity of their sets of words.[[1]](http://en.wikipedia.org/wiki/MinHash#cite_note-b97-0)

Here the set distance can be estimated using the Jaccard method.

<http://roba.rushcj.com/?p=533&cpage=1#comment-5985>

and a book *Mining of Massive Datasets*

A is a set, B is a set too.

a doc and b doc can separated into discrete parts using the every k words. So it is reduce to the set A and set B, now their similarity can use the Jaccard

J(A,B) = ( A intersect b ) / ( A union B ) -----🡪 natural, makes sense

The [Jaccard similarity coefficient](http://en.wikipedia.org/wiki/Jaccard_index) of two sets *A* and *B* is defined to be[[3]](http://en.wikipedia.org/wiki/MinHash#cite_note-2)

 J(A,B) = {{|A \cap B|}\over{|A \cup B|}}.