Similar Documents

Euclidean Distance

Cosine Distance

Euclidean Distance

Hamming Distance

Distance functions

Minhashing

3 Word shingles

Shingling

Pre-processing

Read CSV file

Create signature matrix

Generate hash\_funcs

FLOW CHART

* The data structures Pandas DataFrame has been used to load the CSV file for performing further preprocessing.
* The list data structure has been used to store the tokenized words of each document and further each list formed is again stored as a list of lists.
* The set data structure has been used to add shingles without repetition.
* Distance Measures: We use Euclidean distance, Cosine distance, and Jaccard distance to measure the similarity.