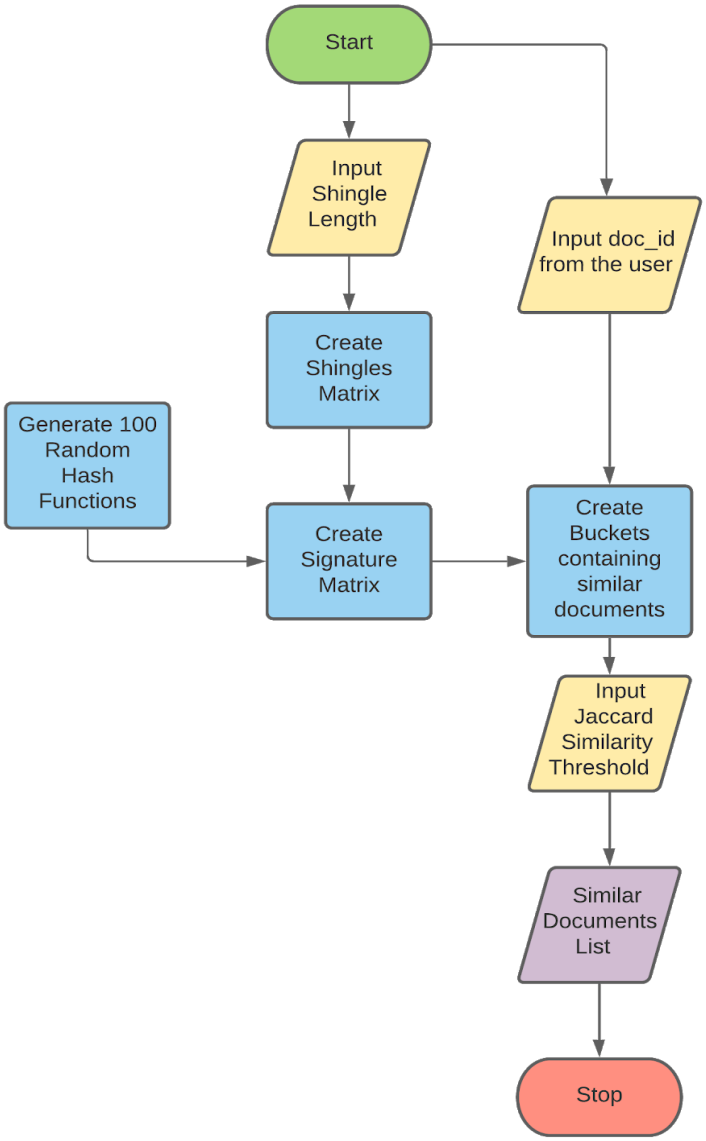
Design Documentation



DATA STRUCTURES USED:

1. unique_shingles
    - It is a list that contains all the unique possible shingles of the length given by the user. Example -
      [‘nma ’, ‘asdf’, ‘qwer’, ...] is a list of shingles of length 4.
2. shingle_matrix
    - It is 2-D numpy array having 0’s and 1’s with the dimension (no of unique shingles * no of docs)
3. signature_matrix
    - It is a 2-D numpy array with dimensions (number of hash functions * number of docs) which contains min hashed values of the shingle matrix.
4. band_dict
    - It is a python dictionary having keys as bucket ID and values as list of documents present in it.
5. answer
    - It is a python dictionary having keys as doc ID and values as the jaccard similarity between query doc and current doc.

Runtime for creating shingles matrix ~ 0.5 seconds
Runtime for creating signature matrix (4-shingle) ~ 10 minutes
Runtime for loading signature matrix if already saved ~ 0 seconds
Runtime for finding similar documents ~ 0.1 seconds