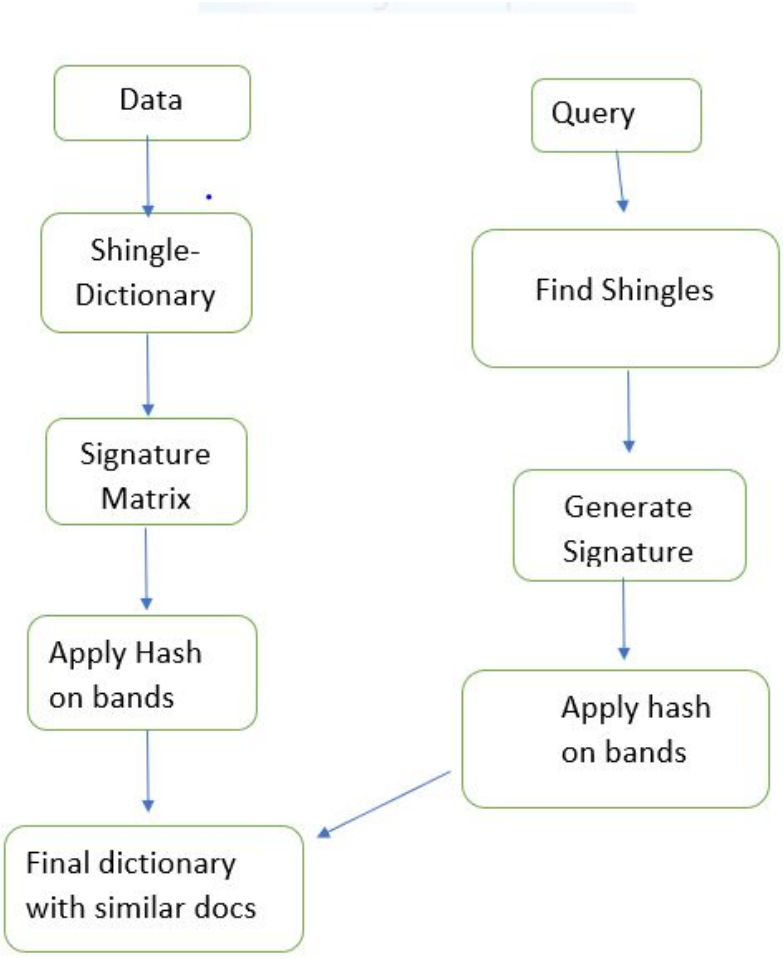
LSH based similar document Retrieval

This is a retrieval system made using Locality Sensitive Hashing (LSH) algorithm to find similar/duplicate DNA sequences trained over the Human DNA dataset containing **4380** sequences.

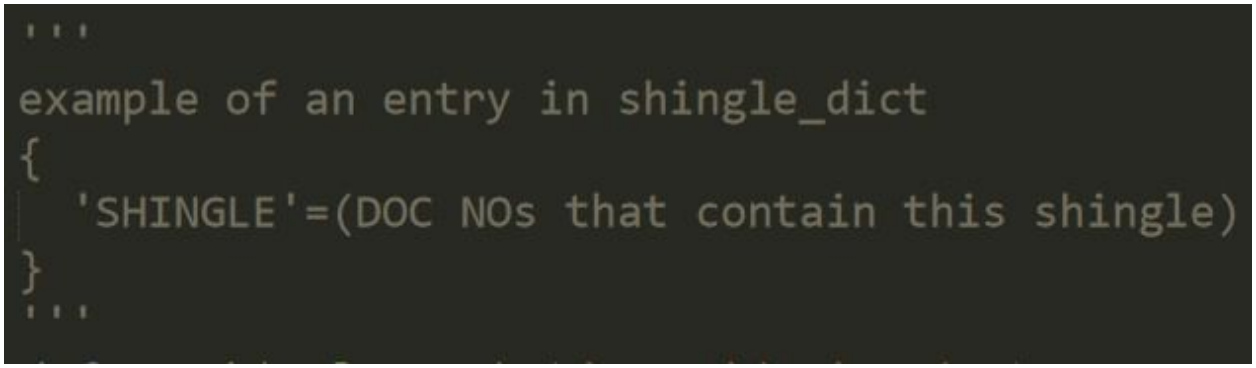
Architecture - Flow Diagram



Shingling

In this step, shingles present in our dataset are extracted and the shingle matrix is constructed. Instead of a boolean Shingle Matrix, we have constructed a dictionary of shingles with all the sequences/documents in which shingle is present.

We have taken shingle size to be ‘6’.



Min-Hashing

We are using Jaccard Similarity as a similarity measure. And the Hash Family that is suitable to be used is the Permutations of the shingles. Here we randomly generated 100 hash functions and also seed them to later generate the same random behavior while querying. Using these 100 hash functions the signature matrix is generated. The hash function is created using random.permutation of the numeric arrangement of shingles.

LSH

We have chosen the values of no. of bands(b) and no. of rows in a band(r) based on the threshold value of similarity(t) using the following formula.

$t = (1/b)^{1/r}$

We have taken three cases

If t=0.9, b=5, r=20

If t=0.8, b=10, r=10

If t=0.55, b=20, r=5

The Signature matrix is split into bands and in every band, a column is hashed (using python’s in-built hash function) and the

documents that get the same hash value are candidate pairs and are checked for Jaccard similarity.

Document Retrieval

We can take a DNA sequence as an input or pick a DNA sequence randomly from the dataset and retrieve the similar/duplicate DNA sequences by following the same procedure -

i) Generating shingles
ii) Creating the signature
iii) Applying hash on bands

After the above steps, Jaccard similarity is checked for the documents which are hashed to the same bucket as the query in each band. Based on the Jaccard similarity value, false positives are eliminated.

Group Members

Adesh Kumar Pradhan – 2017B3A70960H
Pranavi Marripudi – 2018A7PS0507H
Mereddy Aishwwarya Reddi – 2018A7PS0276H