Documentation

# Group Members:

Nitin Chandra Surapur -2018A7PS0188H

Mohit Gudanti -2018AAPS0321H

Abhinava Maddha -2018A8PS0844H

# Aim

This assignment is aimed at designing a text based information retrieval system using Locality Sensitive Hashing

# Dataset Used:

The dataset used was obtained from Kaggle which is a DNA Dataset.

# Modules/ Frameworks used:

Itertools - This module implements a number of iterator building blocks inspired by constructs from APL, Haskell, and SML. Together, they form an “iterator algebra” making it possible to construct specialized tools succinctly and efficiently in pure **Python**

math - To perform mathematical operations like logarithmic operations and finding the square root.

Numpy -**NumPy** is a **python** library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Pandas - pandas is a software library written for the Python programming language for data manipulation and analysis.

# Functions

createDocShinglesList- creating a list of shingles for each doc takes shingle length and data as input

nextPrime- Function to return the smallest prime number greater than N

generateCoefficients - Generate list of k unique values mapped to each of the random hash functions

table\_rearr- Splitting array into query signature and remaining signatures

compare - comparing given query signature column with other columns in a band and returning a list of lists

LSH - LSH step takes signature matrix as input

Similarity- Printing similarity percentage of n neighbours asked

prec\_recall - precision and recall function for the documents retrieved