DOCUMENTATION

frun.py

NAME

frun

FUNCTIONS

frint()

Function for printing out the tuples in result set.

DATA

lines = []

resultset=[]

ish.py

NAME

lsh

FUNCTIONS

setGenerator()

Here a set is created from the data after it undergoing shingling, vectorizing and performing lsh.

DATA

lines = []

hashmap

arr = []

similarityMatrix

commonset = set()

similarityMatrix = []

hashtable = {}

minimumHashing.py

NAME

minimumHashing

FUNCTIONS

minHashing(arr)

This function reduces the dimensions of a matrix to form a similarity matrix having the same properties as of previous one.

shuffle(size)

shuffle is used to generate a random shuffling of an array of elements 0 to size-1.

shuffleList(hashedList)

shuffleList is used to generate a random shuffling of an array of elements 0 to size-1 but the input is different than the other function

DATA

hashedList = {}

signatureMatrix = []

hashMatrix = []

shingling.py

NAME

shingling

FUNCTIONS

createShingles(doc, k)

Creates a dictionary with key as the dna string and the set of shingles as the paired value.

DATA

lines = []

hashmap = {}

shingles = set()

vectorize.py

NAME

vectorize

FUNCTIONS

jaccardSimilarity(vec1, vec2)

Creates a numpy array with shingles as rows and documents as columns, the array has 0 or 1 value based on whether the shingle is present in document or not.

DATA

mainSet = set()

shinglesDictionary = {}

Dataset.txt