**Documentation**

Pre-processed files:

* File binary\_doc\_vector.csv is the binary term incidence matrix.

doc vector generation.py:

* Check():

Takes a list as input and checks if a term is present in the list, if it is present returns 1 otherwise returns -1.

* train{} DS

It is a dictionary containing the original data with the key being the track ids and the corresponding element is the binary incidence vector.

min hash.py:

* euclidean():

This function take updated signature matrix and generates list of bucket dictionaries, where each dictionary is for each band. A hash function from Euclidean LSH family is used to map the sliced signatures into different buckets.

* query\_euclidean():

This is used to answer a query, and takes the signature of the query and the list of buckets as input. Finally returns the union of all the required buckets.

* hemming():

This function take updated signature matrix and generates list of bucket dictionaries, where each dictionary is for each band. A hash function from Hemming LSH family is used to map the sliced signatures into different buckets.

* query\_hemming():

This is used to answer a query, and takes the signature of the query and the list of buckets as input. Finally returns the union of all the required buckets.

* signum():

This function take updated signature matrix and generates list of bucket dictionaries, where each dictionary is for each band. A hash function from Euclidean LSH family is used to map the sliced signatures into different buckets.

* query\_signum():

This is used to answer a query, and takes the signature of the query and the list of buckets as input. Finally returns the union of all the required buckets.

* Give\_document\_number()

This function finds the first occurrence of 1 in a document and returns that index for further computation.

* Music.txt and words.txt contain the data which are handled using standard file handling procedures of python.