README

Instructions for Locality sensitive hashing program:

The training files(corpus) of the dataset should be stored at the location(directory) 'D:\corpus\'. If you wish to change location(directory) of corpus, then edit the path at line no. 134 of the file 'code.py'.

The queries should be stored at the location(directory) 'D:\query\'. If you wish to change location of queries, then edit the path at line no. 143 of the file 'code.py'.

All training files(corpus) and queries should be stored in the '.txt' format. Other file formats are not supported.

Extract the files of the zip folder uploaded on CMS. File containing the python code is 'code.py'.

To run the python file:

In the command prompt, go to directory where file 'code.py' is stored and type

$ python code.py

After running the above code, the file will be executed and results will be displayed. Top similar training set documents will be displayed with respect to closeness from the test document in decreasing order along with similarity percentage.

Following python modules are required: Numpy, NLTK

To use a different dataset and query, replace the files in the 'corpus' and 'query' folder respectively.