IR_Assignment_2

Locality Sensitive Hashing

Language Used: Python

Working Model:

1. The program can be started by running main.py.
2. The entire corpus is preprocessed (tokenizing and removing stop words, forming shingles of size 4, Hashing the shingles).
3. The matrix mapping document to list of shingles in the document is generated
4. Using the matrix mentioned above and 200 hash functions, the signature matrix is generated.
5. Signature matrix is divided into b bands each of r rows (n=b*r)
6. Each document from the bands are hashed to a bucket.
7. The user is asked to specify the Document to check the similarity.
8. The documents present in all the buckts where the search document is present are collected.
9. The similarity between the Search document and the collection formed are calculated using various distance measures like Jaccard Similarity, Cosine distance, etc.
10. All the relevant documents with similarity greater than or equal to the threshold are displayed to the user.

Requirements/Installation:

To run the following code, nltk have to be readily installed.

ntlk can be installed using ‘ntlk.download()’ in a python shell or in the program.

Libraries used are:

nltk, numpy, sympy, zlib, itertools, binascii, etc

To run the code in Python environment:

Change directory to the folder having the data set and type the command

“ cd /path/to/project/folder/ “

To run the program,

“ python3 main.py ”