DESIGN DOCUMENT

INFORMATION RETRIEVAL (CS F469)

Assignment - 2

BUILDING AN INFORMATION RETRIEVAL SYSTEM USING LOCALITY SENSITIVE HASHING (LSH) ALGORITHM

Group Members-

SUMANASA SOMU	2017A7PS0114H
L SRIHARI	2017A7PS1670H
PRANAV ANAND P	2017AAPS0379H
ASHWIN KUMAR RAJA	2017B4A70599H

About the System

The main aim of the system is to group very similar documents and be used as recommendations for the user. The corpus used is a Dataset of songs with lyrics. It takes the input of Document id and returns the document set of similar document sets for different distance measures.

Dataset and Implementation details

The dataset for the search engine is obtained from the Kaggle. It consists of "Song name", "Artist", "year of production" and the "lyrics". The dataset is of the format CSV(Column seperated Vectors) with 5 columns - Doc-Id, Song, Artist, Year, Lyrics.

Number of Documents = 50
Number of Hash Functions = 200
Shingles size = 4
Threshold = 0.6
Optimized number of bands = 30

Data Structures Used

Dictionaries for storing bucket lists of hash functions as key and document list as the number of documents with that hash function.
NumPy Signature Matrix - Number of columns as no. Of documents and the number of rows as number of has functions.

Runtime for Different Distance Measures

Jaccard - 0.2393sec
Cosine -0.2094sec
Euclidean -0.0069sec

Total Runtime - 0.8sec

Distances Measures

1. **Euclidean distance-** The euclidean distance between any two points(vectors) (x1,y1) and (x2,y2) is defined as sqrt((x1-x2)^2+(y1-y2)^2)

2. **Cosine distance**- In order to remove discrepancies that could arise due to the variation in the size of the query and the document vectors themselves, we use a cosine distance measure. The higher the cosine coefficient between the two vectors, the lesser is the angle between them and the more similar they are. (Lesser distance)

3. **Jaccard coefficient measure**- It is a number between 0 and 1. It is defined as the number of elements in the intersection of two sets A and B divided by the number of elements in their union. The higher the coefficient, more is the similarity. (Lesser distance)

4. **Hamming distance -** Hamming distance between two vectors to be the number of components in which they differ. Most commonly, Hamming distance is used when the vectors are Boolean; they consist of 0's and 1's only. However, in principle, the vectors can have components from any set

Working Model/ Architecture:

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.