DESIGN DOCUMENT

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

### **SUBJECT- INFORMATION RETRIEVAL (CS F469)**

### **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.

### **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**

Number of Documents = 50 , Number of Hash Functions = 200 , Shingles - 4 shingles

Threshold - 0.6

Optimized number of bands - 30

Jaccard - 0.2393sec

Cosine -0.2094sec

Euclidean -0.0069sec

Total Runtime - 0.8sec

### **Distances Used**

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)