Big Data Analytics

Assignment # 1 Due Date: 19th February 2023

Assignments are to be done in the groups assigned. No late assignments will be accepted.

HONOR POLICY

This assignment is a learning opportunity that will be evaluated based on your ability to think, work through a problem in a logical manner. You may however discuss verbally or via email the assignment with your classmates or the course instructor, and use the Internet to do your research, but the written work should be your own. Plagiarized reports or code will get a zero. If indoubt, ask the course instructor.

Objectives:

To implement Locality Sensitive Hashing (LSH) on audio data and evaluate its performance for duplicate detection.

Tasks:

You are given an audio dataset. Download it from the classroom.

Part-I

Q1:

- First you must perform Pre-Processing of the given audios.
    1. Extract the MFCCs* from the audio files.
- Create LSH hash tables using the right number of tables and hash size.
- Create the feature vector (MFCC) and run a query on the LSH tables for a recently new audio.
- Compare the feature vector of the audio that was recorded with the matches returned in the previous step. (Depending on the components of the feature vector, the metrics for comparison can be L2 distance, cosine similarity, or Jaccard similarity.)
- Return the result that has lowest/highest metric value (depending on the chosen metric) as the match.

Q2:

- As you have studied LSH in the class. Now find something Interesting from LSH of audio, except for finding duplicates audios.

Part-II

Now for the second part you just must create a responsive flask application. Your application must perform the following functions.

- Upload Audio file (Mp4).
- Shows the audio similarity if any.
- Implementation of Q2.

Submission

One of the group members is required to submit a report with response to each Part-I and Part-II with supporting screenshots with time stamps. Each member should also explain their findings in the report. Py script for Audio Processing, implementing LSH and zip folder containing your Flask application.

NOTE: No need to submit the audio files. If submitted will result in marks Reduction.

MFCC:

MFCC stands for Mel-Frequency Cepstral Coefficients. MFCCs are widely used in speech recognition, music information retrieval, and audio classification tasks, among others, due to their ability to capture important spectral information while reducing the dimensionality of the data and ignoring unimportant information such as phase.