Advanced Database Systems (CS60113)

Assignment 3. Retrieval from Large Image and Document Databases (Due Date: September 30 2020)

NOTE: For this assignment, you may work in two groups together. A total of 3 applications will be acceptable. Some group may also work on its own. Marks for this assignment will be relative.

Develop an application as discussed in the class on September 11, 2020. You may go through the recorded video for a description of the assignment.

Essentially, it will be a combination on image and keyword based search. First you will create your image and document database by crawling sports or news or any other websites. Make sure you follow appropriate protocols (Honor exclusion protocols) while crawling. You should be able to keep a link between the webpages and the images contained in those. Represent each image by its histogram. For the documents, you may keep either only the nouns or all the words, whichever is convenient.

The application will filter the documents based on the keywords and select the images based on histogram similarity. Images should be kept in R tree as done in assignment 2. Document terms will be represented by their TF-IDF values. You may use standard libraries for creating list of nouns/generating TF-IDF, values, etc.

The result will be retrieved by specifying the number of nearest neighbors the user is interested in. It should be possible to go to the corresponding page by clicking on the image.

Record and share a video of the working of the application in your YouTube channel and share the link. You will also demonstrate your application from your machine and submit either a zip of the application files or a link to Google Drive where such files will be stored.