Topic 12

1. **Title:** ***Data Redundancy Removal Algorithm***
2. **Description**: Bhuvan is constantly flooded with different types of data from multiple sources like the crowd sourced information and other data generated inhouse. This large inflow of the unstructured data leads to cumbersome management and data storing challenges. It is required to implement an efficient data management technique to eliminate redundant datasets. A robust algorithm using AI/ML techniques is required to detect/identify and display on a user interface for decision making to delete/retain the duplicate datasets.
3. **Objectives**:
    a. Data pre-processing/Data cleaning.
    b. De-duplication Algorithm to identify redundant/repetitive datasets.
    c. Provide in a user interface for decision making to delete/choose the latest dataset.
4. **Expected Outcomes:**
    a. Data pre-processing/Data cleaning.
    b. De-duplication Algorithm to identify redundant/repetitive datasets.
    c. Provide in a user interface for decision making to delete/choose the latest dataset.
5. **Relevant data and steps to get the data from Bhuvan/ other sources:**

    Test Datasets are provided in the repository (link details in the web page)

6. **Steps to be followed for achieving the objectives:**
    a. Data pre-processing.
    b. Develop an ML algorithm to read and store the dataset information, metadata info etc.
    c. Train the model with sample datasets.
    d. Run to detect repetitive/redundant datasets.
    e. Develop a user interface displaying the list of duplicate files in a specified folder/directory to end user for decision making.

7. **Evaluation**
    a. The algorithm will be run on sample unstructured data to detect the redundant datasets.
    b. The efficiency/accuracy of the model is evaluated based to ratio of actual duplicates to the detected duplicates.