# School of Computer Science Engineering and Technology

Course- BTech

Course Code- 301

Year-   2022

Date- 24-01-2022

Type- Core

Course Name-AIML

Semester- Even

Batch- 4th Sem (SPL)

## 2 - Lab Assignment No. 2.1

**Objective:  To learn data pre-processing operations involved in a machine learning tasks for text, audio and video data.**



1. **Working with Text Data**
   a) **Create a dataset:** Create one folder and add 5 text files (.txt) to it. Add appropriate data in the text file (You may use the Wikipedia data to create the text dataset).  (15)
   b) **Read** all the text documents from the folder and print the data. (15)
   c) **Standardize** the dataset using the following pre-processing techniques. (20)
       - Remove all the special characters, also remove the numbers, and keep only alphabetical data in the txt.
       - Convert uppercase letters into lowercase letters.
   d) Store the pre-processed data in another file (You may store it in .txt or .csv or xlsx)

2. **Read and play the video:** Read the sample video file in MP4 format and play it (10)
3. **Read and play the audio:** Read the sample audio file in MP3 format and play it (10)

**Suggested Platform:** Python: Azure Notebook/Google Colab Notebook, packages such as numpy, nltk, regular expression package re.

**Additional (**Not a part of evaluation**)**

*** Try to read the other formats of video and audio data. Also, you may explore the different operations involved in the video and audio.