Exploratory Data Analysis and Text Classification for Urdu Speech Dataset

---

Objective:

- To Understand Preprocessing steps for Speech Datasets
- To Understand Exploratory Data Analysis process for Speech Datasets

Dataset:

Download Urdu stopwords and Font files for the tasks from the github repository (*both datasets are public datasets*).
Download the Urdu Speech dataset from the given GitHub link.
https://github.com/siddiquelatif/URDU-Dataset/tree/master

Assignment 1. Data Preprocessing

- Merge all Speeches of the same speaker into one file.
- Convert the all merged audio files into text.

Assignment 2. Exploratory Data Analysis

- Identify the top 10 frequent words for each emotion in the transcribed text and plot them using *Bar Graph*
- Visualize the most frequent word for each emotion in the transcribed text using *Word Cloud Analysis*
- Calculate the average length of transcribed text for each emotion and plot them using *Bar Graph*
- Analyze the sentimental polarity of transcribed text for each emotion using *Vader Sentiment Analysis* and plot the results using *Line Plot.*
- Identify the top 10 frequent *Bigrams* for each emotion in the transcribed text and plot them using *Bar Graph*
- Identify the top 10 frequent *Trigrams* for each emotion in the transcribed text and plot them using *Bar Graph*

Assignment 3. Text Classification

- Perform Text Classification on the generated transcription using *scikit-learn library*