**Exploratory Data Analysis and NLP Downstream Tasks on Urdu Speech Dataset**  
  
  
 This assignment aims to perform Exploratory Data Analysis (EDA)  
on the Urdu Speech dataset and apply Natural Language Processing (NLP)  
downstream tasks on the generated transcription of the merged speeches using  
OpenAI Whisper API.  
  
  
  
**Tasks:**  
  
  
**1. Data  
Pre-processing:**  
  
  
a)      
Download  
the Urdu Speech dataset from the given GitHub link. (<https://github.com/siddiquelatif/URDU-Dataset/tree/master>  
)  
  
  
b)      
Merge the  
speeches of the same speaker into one file.  
  
  
c)     Convert  
the merged audio files into text using OpenAI Whisper API.  
  
  
**2. Exploratory  
Data Analysis:**  
  
  
a)      
Identify  
the top 10 most frequent words in the transcribed text for each emotion and  
plot them in a bar graph.  
  
  
b)      
Perform  
Word Cloud analysis to visualize the most frequent words in the transcribed  
text for each emotion.  
  
  
c)     Calculate  
the average length of the transcribed text for each emotion and plot them in a  
bar graph.  
  
  
d)      
Analyze  
the sentiment polarity of the transcribed text for each emotion using Vader  
Sentiment Analysis and plot the results in a line graph.  
  
  
e)      
Identify  
the top 10 most frequent bigrams and trigrams in the transcribed text for each  
emotion and plot them in a bar graph.  
  
  
**3. Downstream  
NLP Tasks (you may choose any one, need not be from the list below). You may use any library or an API for this part**  
  
  
a)      
Perform  
Sentiment Analysis on the generated transcription using TextBlob library, or  
  
  
b)      
Perform  
Named Entity Recognition on the generated transcription using spaCy library, or  
  
  
c)     Perform  
Text Classification on the generated transcription using scikit-learn library, or  
d)    Perform any of the NLP tasks such as Part of Speech Tagging, Topic Modelling, Language Modelling, or any other of your choice.