Q1: I have done the Unicode Correction. It is written in the Q1_Q2.ipynb file.

Q2: I have printed the top 20 Uni-gram and Bi-gram frequent Characters and Syllable in the Q1_Q2.ipynb file.

Q3: The annotated file is downloaded as “cs689_assignment.txt” in this folder. The first line is the sentence and the below line itscorresponding split for words or group of words seperated by comma.

Q4: I have found unigram frequencies of tokens and bi-gram frequencies of tokens, syllables, and characters for each of the tokenizers in each of the separate .ipynb file. I have find those metrics for each tokenizer in its corresponding .ipynb file: Q4_IndicBERT.ipynb, Q4_WhitespaceTokenizer.ipynb, Q4_Unigram_Tokenizer.ipynb, Q4_mBERT_tokenizer.ipynb. All the files are almost similar. Q4_IndicBERT is commented properly. Others also commented.

Q5: I have found the tokenization of the cutom given corpus by each of the tokenizer towards the end of its corresponding .ipynb file. I calculated the metric i.e. Precission, Recall and F1, at the end of the each corresponding .ipynb File.
The