CS689A: Computational Linguistics for Indian Languages

Assignment 1 (75 marks)

Due on: 17th February, 2024, 11:00pm

Choose the corpus file according to your mother tongue.

The corpus files are available from https://bangla.iitk.ac.in/assignment_cs689/cs689_100mbs/. You may use `wget` to download these 100 MB files.

1. (5 marks) Perform the Unicode correction as discussed in class. Essentially, consonants with a halant character should be counted as 1, while those without that should be counted as 2. You may transliterate the corpus to ISO15919 format or ITRANS before and/or after performing the correction.

2. (5 marks) Find all characters and syllables. Store a list of them in descending order of their frequencies.

   Find the top-20 frequent uni-gram and bi-gram frequencies of characters and syllables.

3. (25 marks) For a random set of 25 sentences, find all the *word groups* in them. Remember that a word group is a semantic unit that includes inflections, verb auxiliaries, and compounds, as discussed in class.

4. (25 marks) Run the Unigram, BPE (vocabulary sizes, $V = 1k$, $2k$), mBERT (max_length = $1k$, $2k$), IndicBERT (max_length = $1k$, $2k$), and White-space tokenizers on the entire corpus. You may use Sentence Piece or similar libraries for this purpose.

   For each token found, find the characters and the syllables. Store a list of them and the tokens in descending order of their frequencies.

   Find the bi-gram frequencies of tokens, syllables, and characters comparing both the above models.

5. (5 marks) Assume that the set of tokens from Question 3 is the ground truth set. For each tokenizer in Question 4, find the precision, recall and F-score for the 25 sentences.

6. (10 marks) What do you learn from this comparison?

Instructions

Submit the assignment as one zip file `rollno-assignment1.zip` in the course portal (https://canvas.cse.iitk.ac.in/) within the deadline. The submission MUST contain a README file and a Makefile. The code must have documentation with appropriate comments.