CS689A: Computational Linguistics for Indian Languages

Assignment 1 (100 marks)

Due on: 14th September, 2022, 11:00pm

Choose the corpus file according to your mother tongue (Gujarati, Telugu, Marathi, Malayalam, Bangla, Hindi).

The corpus files are available from https://indicnlp.ai4bharat.org/corpora/#downloads.

The lemma and tagged files are available from https://drive.google.com/drive/folders/1cyOlGodBR86EftZPwKkMwUdxmkoqwBAS?usp=sharing.

1. (5 marks) Perform the Unicode correction as discussed in the class. You may transliterate to ITRANS format after performing the correction.

2. (10 marks) Consider a token as a white-space separated sequence of characters. For each token, 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.

3. (a) (15 marks) Run BPE on the corpus with different vocabulary sizes ($V = 1k$, $10k$, $30k$, $50k$, $100k$).

   (b) (5 marks) For each token this 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.

4. (10 marks) Assume that the set of tokens from Question 2 is the ground truth set. For each vocabulary size of BPE, find the precision, recall and F-score of the BPE-output token set as found in Question 3.

5. (5 marks) Extract a list of lemmas found from the UD-tagged files.

6. (15 marks) Check if the frequency of whitespace-separated words, BPE tokens, syllables, characters, lemmas (found in Questions 2, 3 and 5) follow Zipfian distributions.

7. (a) (15 marks) Given a lemma and the corresponding surface form, derive the suffix. Do a suffix stripping from the surface form till the lemma or a subset of the lemma is reached (choose the longer one). Call the stripped part as the *suffix*. List the 50 most common suffixes ordered in this manner.

   (b) (10 marks) From your knowledge of language, mark the ones that are correct.

8. (10 marks) The submission MUST contain a README file and a Makefile. The code must have documentation with appropriate comments.

Instructions

Submit the assignment as one zip file `rollno-assignment1.zip` in the course portal (canvas.cse.iitk.ac.in) within the deadline.