CS689A: Computational Linguistics for Indian Languages

Assignment 1 (125 marks)

Due on: 15th 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. You may sample *sentences* (not lines) from the corpus to bring down the size to at least 1 GB. You must mention your sampling method in the report.

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

1. (a) (10 marks) Perform the Unicode correction as discussed in the class.
   (b) (5 marks) Clean the corpus, i.e., get rid of non-language characters.
2. (a) (10 marks) Consider a *word* as a white-space separated sequence of characters. For each word, find the characters and the syllables. Store a list of them and the words in descending order of their frequencies.
   (b) (5 marks) Find the bi-gram frequencies of syllables and characters.
3. (a) (15 marks) Run BPE on the corpus with different vocabulary sizes ($V = 1k$, $2k$, $5k$).
   (b) (10 marks) Bonus: Run BPE for vocabulary sizes of $10k$, $30k$ and $50k$.
   (c) (5 marks) For each *token* thus found, find the characters and the syllables. Find the unigram and 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 and the corresponding surface forms found from the UD-tagged files.
6. (a) (10 marks) Draw if the frequency of whitespace-separated words, BPE tokens, syllables, characters, lemmas (found in Questions 2, 3 and 5).
   (b) (5 marks) Perform statistical tests to check if they follow Zipfian distributions.
7. (a) (15 marks) Given a lemma and the corresponding surface form, derive the suffix. Do an end 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.