CS657A: Information Retrieval

Assignment 2 (110 marks)

Due on: 4th April, 2022, 11:00pm

For this assignment, use the language you indicated your preference as. (See email.)

1. (20 marks) Download pre-trained word vectors for GloVe, Word2Vec (both CBOW and Skip-Gram) and FastText from
   https://www.cfilt.iitb.ac.in/~diptesh/embeddings/monolingual/non-contextual/

   Use them for the *word similarity* task. Do not us any library for implementing the similarity function. (In short, try to code up your own cosine_similarity function.) Report the accuracy.

   The word similarity datasets are available from
   https://drive.google.com/drive/folders/1VovzSE1-zXH0bKCar2M8peL4-62BSlZJ?usp=sharing

   The task of word similarity is defined as "given two words, are they similar?". For example, are *bhupati* and *nripati* similar?

   Use different thresholds of 0.4, 0.5, 0.6, 0.7, and 0.8, and report the accuracies for each threshold.

2. (40 marks) Use the pre-trained IndicBERT model available from AI4Bharat https://indicnlp.ai4bharat.org/indic-bert/ and fine-tune it using the NER task.

   The NER datasets are available from
   https://drive.google.com/file/d/1S5TOqIC37dxWCeQbA9VpplXOGAB7cIMV/view?usp=sharing

   Define the `forward` method for the NER task and the training loop yourself. Usage of libraries in this case will only be given partial marks.

   Report training, validation and testing F-scores.

3. Download the language corpora from AI4Bharat. https://indicnlp.ai4bharat.org/corpora/

   (a) (10 marks) Find the top-100 most frequent unigrams, bigrams, trigrams and quadrigrams for characters. Distinguish between vowels when they are used stand-alone versus when they are used as consonant endings. Correct the Unicode mistake of not encoding the *halanta* character.

   (b) (10 marks) Find the top-100 most frequent unigrams, bigrams and trigrams for words.

   (c) (10 marks) Find the top-100 most frequent unigrams, bigrams and trigrams for syllables. A *syllable* is a sequence of characters till the next vowel or end of word.
   For example, if the word is *kshatriya*, the syllables are '*ksha*', '*tri*' and '*ya*'.

   (d) (10 marks) Test if frequency of characters, syllables and words follow Zipfian distribution.

4. (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-assignment2.zip` in the course portal (hello.iitk.ac.in) within the deadline.
If your code fails to compile or run properly, you will get ZERO for that question.
The programs MUST run in the Linux operating system.