Assignment

ECSE344L (NLP)

Text Classification Problem

In the assignment 1, we have gone through Tokenization, steaming, lemmatization, BOW and TF_IDF on the given Times of India news headline dataset.

In this lab, we need to perform following task:

1. Create one hot encoding for the sentence and news category
   a. Perform tokenization
   b. Find number of unique tokens
   c. Find maximum length sentence to determine number of neurons in the input node
   d. Find number of news category to determine number of neurons in the output node
   e. Create One hot encoding for sentence using SUM of One hot encoding on each word
2. Split your data into train, test and validation (80:10:10)
3. Create a neural network with different hyperparameters (your choice) using keras/tensorflow/pytorch
4. Train and test your network
5. Once you satisfied with the performance of the model then evaluate model with validation data
6. Now you take headline of your mother tong newspaper convert with google API feed to your model and check the accuracy
   a. Include UNK token in your corpus to deal with unknown words
   b. Preprocess your translated sentence before feeding to network