DATA 255 Deep Learning Technologies – Homework -4

Deadline – 11.59 PM – 11/30/2023

20 Points

Problem 1: Use the IMDB Movie review dataset:

1. (**1+5 pts**) Build the sentiment analysis model using
   a. Text preprocessing steps: Tokenization, Stopwords removing, HTML removing, Convert to lower case, Lemmatization/stemming
   b. Perform combination of different word embeddings (e.g., Word2Vec, Glove, and so on) and sequential models (e.g., RNN, LSTM, GRU, and so on). Provide a table that include results of all the combinations. Minimum expected accuracy is 95% from at least one of the combinations.

2. (**3+2 = 5 pts**) First part of this question, use Keras embedding layer (+ any sequential models) for sentiment analysis. As a byproduct, you will achieve word embeddings for all the words used to train the model. In the second step, use cosine similarity to find the first five most similar words to "movie".

Problem 2: Use the mobydick chapter four dataset for text generation. In the process of text generation, you must develop a multiclass classification sequential model (e.g., GRU/LSTM/RNN, and so on**). Your model should achieve a minimum accuracy of 20%.** You should use minimum 25 tokens as X features and the immediate next token as y feature. The expectation of the generated text is that: it should not be all identical words and number of generated texts should be minimum of 20.

1. (**1 pts**) Perform Text Preprocessing
   a. Tokenization
   b. Convert to lower case
   c. Expand contraction
   d. Remove punctuation
   e. Lemmatization/stemming
2. (**4 pts**) Text generation model should be developed using keras word embeddings.
3. **(4 pts)** Transfer Learning: Text generation model should be developed using word2vec word embeddings

Useful link-

Data Download: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Gensim- https://radimrehurek.com/gensim/models/word2vec.html

You are required to submit:

1. An MS/PDF/Scanned document:
    a. Include all the steps of your calculations.
    b. Attach screenshots of the code output.
    c. Include the summary of the model
    d. Include a Table - Mention all the hyperparameters you selected: activation function in hidden layer and output layer, weight initializer, number of hidden layers, neurons in hidden layers, loss function, optimizer, number of epochs, batch size, learning rate, evaluation metric
2. Source code:
    a. Python (Jupyter Notebook)
    b. Ensure it is well-organized with comments and proper indentation.

- Failure to submit the source code will result in a deduction of 5 points.
- Format your filenames as follows: "your_last_name_HW1.pdf" for the document and "your_last_name_HW1_source_code.ipynb" for the source code.
- Before submitting the source code, please double-check that it runs without any errors.
- Must submit the files separately.
- Do not compress into a zip file.
- HW submitted more than 24 hours late will not be accepted for credit.