CSI 5340 Homework Exercise 3

IMDB Movie Review Dataset is a standard dataset for text classification or sentiment analysis, where each document (a movie review) is labeled either by a positive label or by a negative label (indicating the positivity of the review). The dataset and its description can be found at:

`https://ai.stanford.edu/~amaas/data/sentiment/`

In this exercise, you are to develop two text classification models using Vanilla RNN and LSTM for this dataset.

For each of the models, you only need to consider the architecture in which the recurrent units are connected as a chain. You can consider either taking the final state/output of the chain as the extracted text feature, or taking the a mean pooling of all outputs of the recurrent unit as the extracted feature. The extracted feature is then passed to a logistic regression classifier. Inevitably, each input word in a document needs to enter the RNN models as a word embedding vector. The word embedding vector can be pre-trained, which you can download (e.g., from `https://nlp.stanford.edu/projects/glove` for the embedding vectors trained via Glove), or a randomly assigned vector.

You can use either TensorFlow or Pytorch deep learning library in this homework.

A key hyper-parameter in the setup of your models is the state dimension, for which you should investigate the following options: 20, 50, 100, 200, 500. For each setting of state dimension, tune the hyper-parameter of each model to obtain the best classification result (on the testing set), and report these results in a table. In your report, also describe the setup and hyper-parameter settings of each model. Submit your report and together with your code in a single zip file.