Assignment 4: Text and Sequence Data **Qiao Wang**

By using various models, the test results in terms of Accuracy are shown in the table below. As we can see, the models with embedding layer are doing a little bit better than the models with pretrained embeddings. This could be because the data has enough samples for the model to learn from scratch, and pretrained models could be useful when dataset is very small.

We can also see that a basic sequence model that uses one-hot encoded vector sequences have the best result when training sample is 10002, when the training sample increased to 15000, the accuracy slightly decreased. This might be because training sample of 10002 is enough for the model to learn and perform well, increasing the sample size may not help much.

Model	Batch Size	Number of Training Samples	Number of Validation Samples	Number of Test Samples	Epochs	Test Accuracy
one-hot encoded vector sequences-basic sequence model	32	102	10000	25000	10	0.607
one-hot encoded vector sequences-basic sequence model	32	1002	10000	25000	10	0.759
one-hot encoded vector sequences-basic sequence model	32	10002	10000	25000	10	0.833
one-hot encoded vector sequences basic sequence model	32	15000	10000	25000	10	0.827
Embedding layer trained from scratch	32	102	10000	25000	10	0.600
Embedding layer trained from scratch	32	1002	10000	25000	10	0.706
Embedding layer trained from scratch	32	10002	10000	25000	10	0.820
Embedding layer trained from scratch	32	15000	10000	25000	10	0.833
pretrained word embeddings	32	102	10000	25000	10	0.546
pretrained word embeddings	32	1002	10000	25000	10	0.676
pretrained word embeddings	32	10002	10000	25000	10	0.814
pretrained word embeddings	32	15000	10000	25000	10	0.831

Please refer to the models and results in Github.