Model details	Test accuracy
Model in the reference book with 600 review words	0.886
Question_1 cutoff reviews after 150 words	0.846
Question_2 restricts training sample to 160	0.534
Question_3 10,000 validation data	0.864
Question_4 only top 10,000 tokens	0.877
Question_5_1 20,000 training with embedding layer	0.874
Question_5_2 20,000 training with pretrained word embedding	0.876
Question_5_3 15,000 training with embedding layer	0.867
Question_5_4 22,500 training with embedding layer	0.88
Question_5_5 160 training with pretrained word embedding	0.573
Question_5_6 160 training with embedding layer	0.655
Question_5_7 2,2500 training with pretrained word embedding	0.865

Table summary for Assignment_4.

Summary:

From Question_1 to Question_4 with one-hot encoding, more reviews, more tokens, more training samples could reach higher test accuracy for the model.

The review words and tokens used in Question_5 is 600, 20,000, respectively.

When the training data is 20,000, using the embedding layer and pretrained word embedding has a similar result, around 0.87.

As for small training dataset, like 160 training samples, although embedding test accuracy is higher than pretrained word embedding, embedding layer shows overfit soon. The training accuracy almost reached 100% while validation accuracy was only around 50%.

When the training data size is larger, for example, 2,2500, the test accuracy for embedding layer is higher than pretrained word embedding. Since the dataset has enough samples to learn, leveraging pretrained embeddings is not very helpful in this case. But for small training samples in the above illustration, pretrained embeddings worked.