关于亚马逊平台上关于“中国研究”主题的书籍的用户评论情感分析

数据说明

从亚马逊网站上选择一些关于研究中国的书籍的用户评论
数据说明：在搜索框内输入“china”， 爬取排序较为靠前的10本书的全部评论；存入'./raw/data/'路径下。
数据处理：每一个评论文件表格的‘a-icon-alt’， ‘a-size-base’， ‘a-size-base 3’分别是评分星数（1星为最差，5星为最好）， 评论标题， 评论文本。这是我们主要关注的数据，将把它们提取出来。

数据具体是哪一本书对应的样本量在num_of_review.txt文件中可以看到。

负类样本总量217，正类样本总量1377。（存在数据不平衡的问题）

我们定义评分（‘a-icon-alt’）小于4星的为负类，大于等于4星的为正类。

在get_data.ipynb中把数据处理成./train/pos/、./train/neg/、./train/pos/、./train/neg/四个文件夹分别表示训练集和测试集中的正类数据和负类数据。（把正负类样本的10%划分为测试集）

1d-CNN模型

在sentiment_analysis_1dCNN.ipynb中运用一维CNN模型训练数据、最后建立端到端模型。

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, None)]            0

 embedding (Embedding)       (None, None, 128)         2560000

 dropout (Dropout)           (None, None, 128)         0

 conv1d (Conv1D)             (None, None, 128)         114816

 conv1d_1 (Conv1D)           (None, None, 128)         114816

 global_max_pooling1d (Globa  (None, 128)              0
 lMaxPooling1D)

 dense (Dense)               (None, 128)               16512

 dropout_1 (Dropout)         (None, 128)               0

 predictions (Dense)         (None, 1)                 129

=================================================================
Total params: 2,806,273
Trainable params: 2,806,273
Non-trainable params: 0
```

```
_________________________________________________________________
```

模型效果

```
# 训练反馈
Epoch 1/10
72/72 [==============================] - 5s 58ms/step - loss: 0.4226 - accuracy:
0.8576 - val_loss: 0.3667 - val_accuracy: 0.8889
Epoch 2/10
72/72 [==============================] - 3s 46ms/step - loss: 0.4227 - accuracy:
0.8602 - val_loss: 0.3543 - val_accuracy: 0.8889
...
Epoch 8/10
72/72 [==============================] - 3s 45ms/step - loss: 0.0029 - accuracy:
0.9991 - val_loss: 0.6062 - val_accuracy: 0.9097
Epoch 9/10
72/72 [==============================] - 3s 46ms/step - loss: 0.0024 - accuracy:
0.9991 - val_loss: 0.6165 - val_accuracy: 0.9167
Epoch 10/10
72/72 [==============================] - 3s 44ms/step - loss: 5.9897e-04 -
accuracy: 1.0000 - val_loss: 0.7079 - val_accuracy: 0.9167
```

```
测试集效果：
10/10 [==============================] - 0s 5ms/step - loss: 0.8392 - accuracy:
0.9139
```

由于数据原本两类的数据平衡性就不高，所以91.39%的测试集准确率，效果并不算很好。