实验二

基于词向量，使用CNN或RNN进行文本分类。

准备

可以使用在实验一中训练得到的词向量文件。

若效果不好，则可以根据你的需要下载以下不同维度的词向量。

英文词向量建议Glove (https://nlp.stanford.edu/projects/glove/)，中文词向量建议Chinese Word Vectors (https://github.com/Embedding/Chinese-Word-Vectors)。

数据集

英文数据集

英文数据集，使用IMDB电影评论数据集 (http://ai.stanford.edu/~amaas/data/sentiment/)(百度网盘地址 (https://pan.baidu.com/s/1QuMUH81PA3doaiG_JeCv6Q) 提取码:l3k3)

数据集提供了25000条训练数据和25000条测试数据。训练数据和测试数据都包含两级分化明显的正面评价和负面评价，各50%。

中文数据集

中文数据集，使用THUCNews (http://thuctc.thunlp.org/#%E4%B8%AD%E6%96%87%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB
(百度网盘地址 (https://pan.baidu.com/s/1Z6E9Rahk-NnpnfoohUwwSQ) 提取码:tkvz)

完整数据集 (http://thuctc.thunlp.org/message)包含74万篇新闻文档，均为UTF-8纯文本格式。我们在原始新浪新闻分类体系的基础上，重新整合划分出14个候选分类类别：财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐。

模型

构建你的模型，完成基于上述数据集的一个文本分类任务。

你可以参考这些项目：

1. text-classification-cnn-rnn (https://github.com/gaussic/text-classification-cnn-rnn)
2. text-classification (https://github.com/brightmart/text_classification)

训练得到分类模型后，请编写一个评估函数**evaluate_model()**，使用模型对测试集进行分类并输出混淆矩阵，在报告中附上测试截图和相关说明。

报告

提供了报告模板