期中作业：IMDB电影评论文本分类

IMDB文本分类

❑ 数据集简介

1. IMDB数据集包含了50000条电影评论和它们对应的情感极性标签（positive or negative），可以建模为一个文本二分类问题
2. 数据下载地址：https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
3. 数据划分：为了方便验收，统一数据集的划分标准。具体为0-30000条为训练集、30001-40000条为验证集和40001-50000条为测试集。模型的性能以测试集的结果为最终标准

IMDB文本分类

数据清洗和文本预处理

1. 原始数据是爬虫得到的初步结果，里面包含了HTML标签和URL等与文本情感无关的噪音，需要进行数据清洗
2. 自行决定是否要统一单词大小写、去除停用词和低频词以及标点符号等

文本清洗可以调用现有工具实现

IMDB文本分类

分类模型

1. 三类特征：分别尝试词频TF、TF-IDF、 word2vec特征
2. 一种分类方法：前馈神经网络、卷积神经网路或循环神经网络
3. 不允许使用现成的线上情感分析平台（API）

IMDB文本分类

- 通过作业能收获

1. 熟悉和掌握基本的文本清洗步骤和实践
2. 熟悉文本的表示方法
3. 熟悉将常用的深度学习模型应用于文本分类问题的流程和范式

IMDB文本分类

- 提交
  - 源代码（至少包含文本清洗、模型训练、模型测试过程以及训练好的checkpoint）
  - 文档（pdf）（至少包含方法、实验结果分析以及心得体会）
  - 压缩文件并命名：“2021NLP-mid-term-project-学号-姓名.zip/rar”
  - 邮件主题：2021NLP-mid-term-project-学号-姓名
  - 提交邮箱：sysucsers@163.com
  - **Deadline**: 2021-11-30, 24:00pm