README

1. 任务描述

本实践来源于天池大赛，同学们可以切身体会一场NLP大赛。利用课程所学的深度学习知识，对新闻文本进行分类。一共有14个分类类别：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。

最终将测试集的预测结果上传至大赛官网，可查看排名。详细提交步骤请查看大赛官网

零基础入门NLP - 新闻文本分类-天池大赛-阿里云天池

零基础入门NLP - 新闻文本分类本次新人赛是Datawhale与天池联合发起的零基础入门系列赛事第三场 -- 零基础入门NLP赛事之新闻文本分类，本赛题以自然语言处理为背景，要求选手根据新闻文本字符对新闻的类别进行分类。这是一个经典文本分类问题。

https://tianchi.aliyun.com/competition/entrance/531810/introduction

2. 数据集

- 训练集：20w
- 测试集：5w

数据集中标签的对应关系如下：

```
{'科技': 0, '股票': 1, '体育': 2, '娱乐': 3, '时政': 4, '社会': 5, '教育': 6, '财经':
 '7','家居': 8, '游戏': 9. '房产': 10, '时尚': 11, '彩票': 12, '星座': 13}
```

注：本次数据集已经事先将文本字符转换成数字，不需再进行预处理

例子：

label	text
2	2967 6758 339 2021 1854 3731 4109 3792 4149 1519 2058 3912 2465 2410 1219 6654 7539 264 2456 4811 1292 2109 6905 5520 7058 6045 3634 6591 3530 6508

3. 评价指标

评价标准为类别f1_score的均值，提交结果与实际测试集的类别进行对比
计算公式：

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

报告+代码提交注意事项

本次提交的材料不仅包括代码和报告，还要将在大赛官网上提交的结果截图一并加入
压缩包内（不要求结果领先，但求真才实学）
材料提交截止时间：2022-11-27 23.59
材料上传地址：
https://bhpan.buaa.edu.cn:443/link/0F46AC5863BCA4D83A1477490301DB7A
有效期限：2022-11-27 23:59
附件打包示例

```
├──20XXXXXX-张三.zip/.rar
│  ├──20XXXXXX-张三-中文文本分类.docx/.pdf
│  ├──代码
│  │  ├──XXX.py
│  │  └──XXX.py
│  └──结果截图.png
```