新闻文本分类赛题宣讲

浪潮集团 – 数据分析师 - 邢生阳

2021年5月8日

inspur 浪潮

一、题目的简介

基于自然语言处理的分类算法

侧重于实践的新闻文本分类算法

从**数据的收集、数据的清洗、算法的设计、验证**等完整的流程

二、题目的区分度

市面上存在大量的文本分类相关的题目以及开源算法

本赛题的区分度

1、本题目结果**分为十个类别，包括九个常见的类别以及“其他”（贴近实际）**
2、题目提供的**测试数据存在数据噪声，需要进行数据清洗和分析**
3、部分**测试数据的样本较小，需要参赛选手自行收集更广泛的数据**

三、评价方式

1、**初赛作品提交前两周 公布验证集数据**

2、通过运行选手提交的验证集数据，计算**评价指标**

3、通过选手提交的**程序代码，运行并验证**

四、题目的评分标准

1、**预测结果的f1_score值（核心评分）**
2、代码的规范性以及技术文档的完整性
3、代码的执行效率（5s）

五、常见问题

1、**不能使用在线api**

-----离线环境演示与验证

2、简单的可视化界面

-----便于决赛演示算法的效果（单条输入与批量）

3、测试集“其他”类别没有数据

-----选手自行处理

三、常见问题

4、**数据清洗以及数据收集的过程是否需要阐述**

-----需要阐述如何通过上述过程达到更好的效果

5、**能否使用开源的数据集或者算法**

-----可以使用。开源的数据集往往时效性不够

6、能否使用除python外的其他语言

-----除python外，可以使用java。**最终的演示和验证环境是windows。**

三、常见问题

7、python版本

-----3.6及以上版本，**程序中的功能对版本有要求的需要注明**

8、验证集的格式

-----与提供的测试集字段一致，分类结果列是空的

9、测试集以及验证集的数据来源

-----各大互联网新闻，如百度、腾讯、新浪、网易等

-----官方媒体，如人民网、新华网、央广网

inspur 浪潮