## A7-新闻文本分类算法

### 时间：2021-03-26        来 源：中国软件杯

**第十届中国软件杯大赛--A组赛题**

**赛题名称:新闻文本分类算法**

**组类： A组（ 本科、研究生、高职）**

**本赛题已更新，详情见文中红字（3.31）**

|  |  |
| --- | --- |
| 赛题简介：介绍整个赛题的实现目标、实用价值、涉及技术和整体要求 | 新闻发展越来越快，每天各种各样的新闻令人目不暇接，  对新闻进行科学的分类既能够方便不同的阅读群体根据需求快速选取自身感兴趣的新闻，也能够有效满足对海量的新闻素材提供科学的检索需求。 |
| 赛题业务场景：描述赛题相关的真实企业业务背景。从真实场景中，适当简化或者提炼出适合比赛的赛题场景 | 赛题以新闻数据为赛题数据，整合划分出如下候选分类类别：财经、房产、教育、科技、军事、汽车、体育、游戏、娱乐和其他共十类的新闻文本数据。选手根据新闻标题和内容，进行分类。  输入为新闻的标题和正文内容，输出为新闻的分类。 |
| 基本功能要求 | 1、输出分类的准确率不低于80%  2、提供简单的可视化界面。能够输入单条新闻，输出新闻的分类，或者支持本地上传csv/xlsx文件，批量输入新闻，并输出新闻分类。 |
| 非功能性要求 | 执行效率：单条新闻，程序从输入到输出的执行时间不超过5s |
| 实现条件：开发环境、实验平台、开发语言、数据库、编译器、涉及硬件等实现条件 | 开发软件：不限制  开发语言：Java或Python  服务器操作系统：windows  运行环境：如使用python,需使用python3.6以上版本 |
| 测试数据或平台：提供给参赛者的测试环境和测试数据 | 提供了十类新闻的测试数据，供选手分析使用。鼓励选手可自行通过互联网收集数据集进行训练，并可通过项目、文档、演示视频等形式来呈现收集过程  **数据下载：**[训练数据样本.xlsx](http://www.cnsoftbei.com/ueditor/php/upload/file/20210401/1617241934831197.xlsx" \o "训练数据样本.xlsx) |
| 开发所需设备及设备指标需求说明 | 开发设备：市场上常规可见的PC机即可 |
| 文档及其他要求 | 代码规范，可读性强、文档说明清晰  不能使用各种在线api接口服务 |
| 各评分项及大致占比 | 1、代码的规范性以及技术文档的完整性  2、预测结果的f1\_score均值  3、代码的执行效率  其中，预测结果的准确率为主要的评审要点，占主要评分占比。 |
| 初赛作品提交要求 | 1、提供源文件、说明使用文档以及算法的实现原理说明，如果使用开源算法，请注明。  2、结果验证：根据后期给定的测试集数据，通过参赛选手的程序进行分类，得出f1\_score均值，作为主要参考。  3、演示视频（7分钟内） |

**出题企业：浪潮云信息技术股份公司**

答疑老师：邢老师  答疑QQ群：490719434