**中国大学生计算机设计大赛**

作品信息概要表 (2023版)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 作品编号 | 2023035163 | | 作品名称 | | SIU-基于爬虫和BERT的网络舆情分析 | | | | | | |
| 作品大类 | 大数据应用 | | | | 作品小类 | | | 实践赛 | | | |
| 作品简介(100字以内)：  我们使用预训练的BERT模型对中文文本进行情感分类，分析文本是正向的还是负向的。接着，使用爬虫从B站上爬取数据，使用此模型对文本进行分类。最后，对数据进行分析，例如对均值进行估计以判断舆情对某一方面的总体态度，对方差进行分析以判断舆情对某一个领域的态度的分化程度，以及对各个时间段上的舆情情况的变化进行分析等。 | | | | | | | | | | | |
| 创新描述（100字以内）：  我们的创意主要体现在以下几个方面：  1.在统计分析方面，我们使用了二项分布的期望和方差进行区间估计，然后使用中心极限定理近似二项分布，得到了正态分布的估计值，从而进行了置信区间的计算。同时，我们还使用了假设检验来验证假设的真实性，使用t分布作为枢轴变量进行计算。  2.在系统实现方面，我们的分析系统中，评论分类模型是最重要的组成部分之一。为了提高模型的预测准确率，我们在微调预训练模型的基础上使用Additive Margin Softmax重新设计了损失函数，用于防治过拟合，并使用了一些微调手段来提升模型的效果。同时，我们还设计了多层次微调方法，使得模型在重点任务和其他分类问题上都能表现出色。  3.在作品总结方面，我们强调了我们的作品是基于大数据分析的情感分类系统，可以为政府、企业和个人提供科学的决策支持和舆情监测服务。我们的创意主要体现在统计分析和系统实现两个方面，这也是我们未来进一步提升和应用拓展的方向。 | | | | | | | | | | | |
| 特别说明（1.作品中如有地图，请说明来源，并标注地图审图号；  2.作品如有前期基础请具体说明，并注明本次参赛的主要工作。） | | | | | | | | | | | |
| 作者及其分工比例(“姓名#”请替换为作者姓名，并按实际作者人数增减，不需要的列可清空；表中填写每位作者各项工作量的百分比，项目名称可以调整或增减，可另加行) | | | | | | | | | | | |
| 项目 | 李沅昕 | | | 蔡嘉骏 | | | 刘梓航 | |  | |  |
| 组织协调 | 33% | | | 33% | | | 33% | |  | |  |
| 作品创意 | 33% | | | 33% | | | 33% | |  | |  |
| 竞品分析 | 33% | | | 33% | | | 33% | |  | |  |
| 方案设计 | 33% | | | 33% | | | 33% | |  | |  |
| 技术实现 | 33% | | | 33% | | | 33% | |  | |  |
| 文献阅读 | 33% | | | 33% | | | 33% | |  | |  |
| 测试分析 | 33% | | | 33% | | | 33% | |  | |  |
|  |  | | |  | | |  | |  | |  |
| 指导教师作用 | | □项目创意 ■理论指导 □技术方案 □实验场地 □硬件资源  □数据提供 ■后勤支持 ■宣讲通知 □组织协调 □经费支持  □其他： | | | | | | | | | |
| 开发制作平台 | | ■Windows □Linux □macOS □其他： | | | | | | | | | |
| 运行展示平台 | | ■Windows □Linux □macOS □iOS □Android □其他： | | | | | | | | | |
| 开发制作工具 | | Python 3.9.pycharm, Vscode | | | | | | | | | |
| 参考文献、项目或作品(前3项) | | 1、 陈希孺 概率论与数理统计[M].安徽: 中国科学技术大学出版社, 2009:145.  2、  3、 | | | | | | | | | |
| 提交内容 | | □素材压缩包 ■报告文档 ■演示视频 ■PPT ■源代码 □部署文件  ■数据集 ■模型 ■成品文件 □其他 | | | | | | | | | |
| 相关文件  (包括必须提交的文件，和其他与本作品开发制作相关的文件；可另加行；可能包括的内容有：信息表、设计报告、源代码、素材包、数据集、训练模型、安装配置说明、用户手册等) | | | | | | | | | | | |
| 序号 | 文件名与描述 | | | | | 文件状态 | | | | 版权状态 | |
| 1 | 文件名：b站爬虫  描述：用于爬取舆论数据 | | | | | □已上传到网盘  ■未上传，下载地址：  <https://github.com/Forstant/ComputerDesign/b>站爬虫 | | | | ■自制□未知版权  ■开源□获得授权 | |
| 2 | 文件名：code  描述：文本分类模型 | | | | | □已上传到网盘  ■未上传，下载地址：  https://github.com/Forstant/ComputerDesign/code | | | | ■自制□未知版权  ■开源□获得授权 | |
| 3 | 文件名：data\_analyze  描述：统计分析 | | | | | □已上传到网盘  ■未上传，下载地址：  https://github.com/Forstant/ComputerDesign/data\_analyze | | | | ■自制□未知版权  ■开源□获得授权 | |
| 4 | 文件名：data  描述：测试数据 | | | | | □已上传到网盘  ■未上传，下载地址：  https://github.com/Forstant/ComputerDesign/data | | | | ■自制□未知版权  ■开源□获得授权 | |
| 5 | 文件名：  描述： | | | | | □已上传到网盘  □未上传，下载地址： | | | | □自制□未知版权  □开源□获得授权 | |
| 6 | 文件名：  描述： | | | | | □已上传到网盘  □未上传，下载地址： | | | | □自制□未知版权  □开源□获得授权 | |
| 7 | 文件名：  描述： | | | | | □已上传到网盘  □未上传，下载地址： | | | | □自制□未知版权  □开源□获得授权 | |
| 8 | 文件名：  描述： | | | | | □已上传到网盘  □未上传，下载地址： | | | | □自制□未知版权  □开源□获得授权 | |
| **特别申明：**  本表所列内容是正式参赛内容组成部分，务必真实填写。如不属实，将导致奖项等级降低甚至终止本作品参加比赛。  请仔细阅读参赛作品类别提交要求，并根据要求上传相应的文档、数据等。 | | | | | | | | | | | |

填写说明：

1. 所有□可根据需要变化为■（软键盘输入）；
2. “作者及其分工比例”以及“相关文件”可根据需要增加或减少项目或行数；
3. “作者及其分工比例”中的“姓名1”等，请修改为作者具体姓名；
4. “相关文件”是指提交上传的，或不需要提交上传，但本作品涉及的所有文件，建议分类别填写；
5. 请将**本表以PDF格式**上传到大赛指定的位置；
6. 版权状态一栏，如有来自支持企业授权参赛师生用的数据、模型、文档等，在“授权方：\_\_\_\_\_\_\_”一栏，并填写来源地址。