暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 数据采集和清洗小组 | | | | | | |
| 姓 名 | 孔译轮 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118233 |
| 实验时间 | 2020.8.31-2020.9.23 | | 指导教师 | 孔祥龙 | | 成绩 |  |
| **一、实验背景和目的**  本次软件实践的数据基于上学期同学们所爬取的近三年的大学专业录取分数和各省一分一段表。  由于部分同学的完成质量不是太高，所以我们需要对数据进行清洗和补充，并根据后面的小组对  数据格式的要求对所有数据进行整改。 | | | | | | | |
| **二、小组任务和个人任务**  小组任务: 1.整合并清洗仓库内的近三年大学专业录取分数和各省一分一段表  2.补充缺失的大学专业录取分数和2020年各省一分一段表  3.根据各组的要求将数据转为其所需要的形式  3.采集各省的985,211学校和各学校的一级、二级专业，以及各学校校徽  个人任务:1.为各个年份的每个学校添加各专业的分数线  2.对各省一分一段表.json进行一次清洗，任务为去错误数据，纠正缩进，浮点数取整  3.检查各高校三年内录取分数线与各省份一分一段表是否合理，对不合理之处进行修改  4.重新整理2018年甘肃文科json数据 | | | | | | | |
| **三、个人任务需求分析**  1. 为各个年份的每个学校添加各专业的分数线  因后面的小组需要根据分数线进行学校与专业的预测，而从网上爬取的专业分数线又不全，所以需要补全各学校各专业的分数线。在与对应小组沟通后，采用在合理分数区间内取随机数的方法补全数据。这样做可能会导致分数线与实际情况不符，从而预测结果与实际有所偏差，但因为不是所有学校官网都有各专业的分数线，所以由人力对3年内所有学校的所有专业的专业线进行统计是不可取的。最终选择了随机数这个方法。  2. 对各省一分一段表.json进行一次清洗，任务为去错误数据，纠正缩进，浮点数取整  由于部分同学上学期作业完成情况不太良好，所以我们需要对json文件进行手动的清洗  3. 检查各高校三年内录取分数线与各省份一分一段表是否合理，对不合理之处进行修改  根据其他小组的反馈，数据中存在不合理的情况。所以我们需要根据excel中的函数对数据进行检查并修改，必要时需要人工检查修改  4.重新整理2018年甘肃文科json数据  由于2018年甘肃文科的json数据缺失，故需要重新整理 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1. 为各个年份的每个学校添加各专业的分数线    检查没有具体专业的学校，添加各专业和随机数生成的专业线  专业线根据不同的省份设置，如江苏省的专业分数线在最低分数线往上10分以内随机选取  2. 对各省一分一段表.json进行一次清洗，任务为去错误数据，纠正缩进，浮点数取整  3.检查各高校三年内录取分数线与各省份一分一段表是否合理，对不合理之处进行修改  4.重新整理2018年甘肃文科json数据  C:\Users\Lenovo\Documents\Tencent Files\1332368139\Image\C2C\%6@N$VT0@N`QDAF0FRVS5NA.png  与上学期类似 | | | | | | | |
| **五、实验结果与分析**  1. 为各个年份的每个学校添加各专业的分数线    在文理科的基础上，添加了具体的专业及分数线。可以看到，理科的分数线略大于文科，且均在合理的范围内高于学校分数线。所以结果较为合理  但不可避免的是与真实情况的偏差。因为不是所有学校官网都有各专业的分数线，所以由人力对3年内所有学校的所有专业的专业线进行统计是不现实的，故只能采取随机数的方法。可能会导致最后的预测结果有误差。  2.对各省一分一段表.json进行一次清洗，任务为去错误数据，纠正缩进，浮点数取整  部分json文件中存在分数线缺失的情况，需进行补全。完成情况良好  3.检查各高校三年内录取分数线与各省份一分一段表是否合理，对不合理之处进行修改  对有些明显偏低的数据进行清除。  对接的小组后来又要求去除各学校提前批专业，故需要上官网进行查询，通过excel筛选的功能清除提前批数据。  4.重新整理2018年甘肃文科json数据  将2018甘肃文科的csv文件转为json文件 | | | | | | | | |
| **六、实验总结与心得体会**  1. 数据清洗工作听起来简单，但初始数据错误较多，乱序，格式错误，数据错误很多，很难写出代码进行清洗，所以只能进行人工处理。工作难度低，但工作量很大。  2. 在与其他小组进行对接合作时，一定要搞清具体的任务及要求是什么，避免出现因数据不符合预期而造成需要二次加工的情况  3. 由于只负责数据清洗，所以这次没有接触到前端的开发和AI部分设计，有点遗憾。但一个团队中一定需要对数据进行处理的成员。  4. 这次实验使我对pandas和excel更为熟悉，也学到了很多新的库函数  5. 希望以后有更多小组合作开发软件的机会 | | | | | | | | |

2020年9月制