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 | 邹家馨 | 专业 | 人工智能 | 班级 | 2 | 学号 | 09118221 |
| 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，出于多种原因，数据有冗余，错误，格式不统一和未合并等问题。数据的问题对后面小组做评估，分析，训练模型等均有较大的影响，故本小组的任务为采集，清洗，整理数据，为后面小组提供正确有效的数据。 | | | | | | | |
| **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，并保存为js和csv两种格式  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.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.对高校录取分数线整合.csv进行一次清洗，任务为清洗标签Score,Major,Category  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.整合2020年各省一分一段表.csv，保存为json格式  4.重新爬取东北农业大学数据，作为上学期作业补交 | | | | | | | |
| **三、个人任务需求分析**  1. 高校录取分数线整合.csv的一次清洗: Category一列只能有三种值，文科、理科和不分文理，Major列不能为空，没有专业就写all，score要求是int,现在有浮点数和None，不符合条件的数据都删除。  2. 各省一分一段表.json进行一次清洗:不少数据乱序，对错误数据要进行重新排序，数据为浮点数的要取整，不同json文件的括号缩进要统一，删除数据错误较多和数据个数较少的文件。  3. 采集并整合2020年各省一分一段表:考虑到需要将csv转化为json保存，使用到xlrd和json两个库，将各省一分一段表读取并以json格式保存。  4. 重新爬取东北农业大学：需要通过爬虫爬取东北农业大学官网本科生招生的相关数据，需要使用requests和json库，并进行人工整理。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1.利用excel的筛选，查找和替换功能实现高校录取分数线整合.csv的一次清洗。  2.利用excel的筛选，查找和替换功能以及记事本的查找替换功能实现各省一分一段表.json进行一次清洗。  3.考虑将csv转为json文件，先读取文件    其中的get\_data函数作用为读取文件数据    4. 爬取东北农业大学录取分数线  先设置url并定义字典的键    爬虫爬取所需要的数据 | | | | | | | |
| **五、实验结果与分析**  1.高校录取分数线整合.csv的一次清洗，当时要求未包含删除提前批，艺术批次的相关数据，第一次清洗并不完善，输出表格交给了后面的同学进行进一步的清洗。  2.json文件的清洗，删除了所有引号，将所有的浮点数转为int型，并对肉眼可见的倒序，错乱进行了人工排序，但并不完善，还有不少错误。  3.一分一段表的格式转换部分实验结果截图，实验结果较为理想。    4.爬取东北农业大学录取分数线数据实验结果截图，结果较为理想。 | | | | | | | |
| **六、实验总结与心得体会**  1.初始数据错误较多，乱序，格式错误，数据错误很多，难以写出高准确率的代码进行清洗，进行了人工处理。工作难度低，但工作量很大。  2.一组和十组的同学紧密合作，我的任务分配，任务接收和一些其他团队协作能力在本次实践中得到了提升，  3.利用excel，记事本，python处理数据的能力得到提高，了解并学会一些了xlrd和json库的函数。  4.短学期所学的python的前端和后端开发知识并未用到，有点遗憾，希望以后有机会运用到所学的知识，作出和其他组同学一样优秀的作品。  5.希望以后有更多的软件相关的实践课程，提高自己的实战能力。 | | | | | | | |

2020年9月制