暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 数据采集与清洗 | | | | | | |
| 姓 名 | 祁畅 | 专业 | 人工智能 | 班级 | 2 | 学号 | 09118224 |
| 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.对已爬取的各省一分一段表的json，进行同年份同省份合并，文理科分开，并作初步清洗，整合为格式统一的json文件  2.检查一部分数据的正确性  3.补齐了2018河南理科的缺失分数段数据 | | | | | | | |
| **三、个人任务需求分析**  1.对已爬取的各省一分一段表的json，进行同年份同省份合并，文理科分开，并作初步清洗，整合为格式统一的json文件：为了方便后续数据的使用等，需要将短学期爬好的各省一分一段表整理好。  2.检查一部分数据的正确性：后续数据的使用中，发现不少数据有错误或格式不统一等问题，所以为了消除这些错误，安排每人检查一部分大学的数据是否有误，并反馈在在线表格当中。  3.补齐了2018河南理科的缺失分数段数据：在错误查找完成之后，需要对有错或缺失的数据进行弥补。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1.对已爬取的各省一分一段表的json，进行同年份同省份合并，文理科分开，并作初步清洗，整合为格式统一的json文件：  因为很多省都是分作好几个不同同学完成，再加上大部分同学以学号姓名命名文件（如图所示），极不方便查找，所以只能手动将其分类归纳，最终是以某年某省文\理科的json文件为单元。    2.检查一部分数据的正确性：  我负责检查了四所大学的录取分数线信息以及五个一分一段表，并发现了错误且反馈。  3.补齐了2018河南理科的缺失分数段数据：  我负责的是2018河南理科的599分以下的数据补全 | | | | | | | |
| **五、实验结果与分析**  1.对已爬取的各省一分一段表的json，进行同年份同省份合并，文理科分开，并作初步清洗，整合为格式统一的json文件：  分类整理后的结果如下图所示    分类完成之后，条目清晰，方便后续使用  2.检查一部分数据的正确性：        3.补齐了2018河南理科的缺失分数段数据 | | | | | | | |
| **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.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.所学的python的前端和后端开发知识很少用到，希望以后有机会更多运用到所学的知识。  4.希望以后有更多实践课程，提高自己的实践能力。 | | | | | | | |

2020年9月制