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 | | | | | | |
| 姓 名 | 鲁瀚洋 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118230 |
| 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**  在软件实践第一次小作业：收集各高校的录取分数数据，第二次小作业：各省市近三年文理科分数排名的基础上，搭建基于Web的高考志愿推荐系统 | | | | | | | |
| **二、小组任务和个人任务**  小组任务   * 进行需求分析 * 爬取软件实践repo数据的数据 * 进行初步的数据清洗工作，提取csv/json文件 * 与第其他小组进行对接，逐步更新，完善数据库 | | | | | | | |
| **三、个人任务需求分析**  个人任务   * 根据小组长分配的任务，修改数据 * 将修改的数据提交到项目文件中，待小组长/负责人审核 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  - 2020/9/2删除数据库中无用的数据，只保留第一二次作业的文件，负责第37-72项，共109项  - 2020/9/13 解决二级学科csv与一级学科csv的冲突问题，修改csv，合并二级学科与一级学科的数据，上传为新的文件，由09118228沈飞鸿和09118230鲁瀚洋共同完成  - 2020/9/14 收集985/211学校校徽，数据由09118230鲁瀚洋提供，09118229张博宇清洗  - 2020/9/22 筛选整合后的csv中缺失的数据，负责第57-84项，一共109项  - 2020/9/19 收集其他小组使用数据中的问题  - 2020/9/20 协助第十组更新数据库  - 2020/9/26 检查2017宁夏文科和理科，2017辽宁理科695-595分，2017辽宁理科594-480分，2017辽宁文科645-532分数据情况  - 2020/9/27 检查湖南省数据情况  - 2020/9/27检查近三年文理科（或者综合） 最高分到最低分（最起码是到本一线） 一分一段数据 | | | | | | | |
| **五、实验结果与分析**  - 2020/9/2 数据库中数据初步筛选  - 2020/9/20数据库更新，更新内容为：修改东南大学为西南交通大学；增填东南大学，南京农大学，吉林大学，苏州大学数据  - 2020/9/24数据库第二次更新，更新内容为：将所有非本一批次的大学数据删除 约10,000条数据，注意：本次更新只更改了数据库db.sqlite3 ，extfiles中csv表格未作更改，删除了下面三所大学的数据：贵州大学；东北农业大学；东北林业大学  - 2002/9/27数据库第三次更新，更新内容针对majors类，内容为：修正清华大学数据，新增南京理工大学、浙江大学等十四所大学数据  - 2020/9/28数据库第四次更新，更新内容针对majors类，内容为：修改复旦大学、华东师范大学、南京财经大学、西南大学数据 | | | | | | | | |
| **六、实验总结与心得体会**   * 初步了解项目中团队协作方式 * 掌握git原理与操作 * 掌握需求分析步骤 * 提升代码能力 | | | | | | | | |

2020年9月制