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 | 沈毅 | 专业 | 人工智能 | 班级 | 2 | 学号 | 09118222 |
| 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. 爬取2020所有高校的一分一段表，并整理成符合格式要求的csv文件  2. 检索并排错先前的一分一段数据  3. csv与json格式按要求转换、输出 | | | | | | | |
| **三、个人任务需求分析**  1. 多种方法爬取收集2020年各高校一分一段信息，将原格式转换成文本信息，再按照其他组给定的格式要求编纂成表。  2. 校对，清洗，按格式输出一分一段表。特别注意高分学生隐私保护、少数民族军人家属优待、特招等等对分数的影响。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1. 手动收集各高校一分一段表，按其他组要求的格式   |  |  |  |  |  | | --- | --- | --- | --- | --- | | 分数 | 累计人数 | 省份 | 年份 | 科类 |   保存为csv文件  2.将爬取的csv转为json文件    其中的get\_data函数作用为读取文件数据 | | | | | | | |
| **五、实验结果与分析** | | | | | | | |
| **六、实验总结与心得体会**  1. 数据来源纷杂，各个高校给出的数据格式（图片表格、文本、电子表格、pdf文件等）不一，无法进行批量自动化操作，但后期处理格式以及字节处理可以写脚本快速完成任务。  2. **下游任务组应提高自身代码鲁棒性，不要事事苛求数据收集组，特别是第十组。事多只能凸显自身的无能**  3. 刷commit是很不礼貌、无用、无聊、不符合规范的行为  4. 对git命令的熟练度提升了  5. 初步体验了多人协同开发 | | | | | | | |

2020年9月制