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 | 卢甲浩 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118243 |
| 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**  这次项目实验的核心是对上学期同学们收集到的数据进行清洗整理，保证数据的正确性和合理性，调整为合适的格式， 为其他组后续开发工作做铺垫，满足后面小组提出来的要求 | | | | | | | |
| **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.手动整理一二级专业  2.同小组其他成员共同清洗检查高校三年的录取分数线和三年一分一段表json文件内的数据是否合理，格式是否正确，若不正确则修改  3.更改高校三年录取情况中大家爬取数据里省份不统一，科目不统一，数据矛盾等问题  4.二次检查清洗后的一分一段表json文件，修改错误，统一编码格式，并将其转为csv文件 | | | | | | | |
| **三、个人任务需求分析**  1. 手动整理一二级专业  关于录取过程中一二级专业，在此前并没有可直接读取，可代码出来数据的文件，需要手动整理至表格中，在整理至excel文件后由于编码问题，使用python读取后重新写入，在写入时选定编码  2. 同小组其他成员共同清洗检查高校三年的录取分数线和三年一分一段表json文件内的数据是否合理，格式是否正确，若不正确则修改  由于初次爬取数据的同学存在不统一、数据错误、格式错误等问题，我们小组分工，对照各大学本科招生网站数据检查高校三年录取分数线，对照原始excel文件检查三年一分一段表，来着手解决数据缺失、错误等问题。  3. 更改高校三年录取情况中大家爬取数据里省份不统一，科目不统一，数据矛盾等问题  大家在爬取数据中由于出现例如{新疆、新疆省、新疆区、新疆维区、新疆内地、新疆维吾尔自治区}等不统一省份名称的问题，导致显示的省份达一百以上，需要对20万条数据的 所有省份名称来进行统一，科目名称也同理  4. 二次检查清洗后的一分一段表json文件，修改错误，统一编码格式，并将其转为csv文件  第一次检查清洗后的一分一段表json文件仍不能保证完全正确，因此由我来二次检查清洗，修改其中的错误，统一编码格式，最后转为统一的csv文件 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1. 手动整理一二级专业  手动整理一二级相关数据至表格中，在整理至excel文件后由于编码问题，使用python读取后重新写入，在写入时选定编码  2. 同小组其他成员共同清洗检查高校三年的录取分数线和三年一分一段表json文件内的数据是否合理，格式是否正确，若不正确则修改  我们小组分工，对照各大学本科招生网站数据检查高校三年录取分数线，对照原始excel文件检查三年一分一段表，来着手解决数据缺失、错误等问题。  3. 更改高校三年录取情况中大家爬取数据里省份不统一，科目不统一，数据矛盾等问题  大家在爬取数据中由于出现不统一省份名称的问题，导致显示的省份达一百以上，需要对20万条数据的所有省份名称来进行统一，我使用excel的筛选功能，手动删选对应错误的省份名称数据，然后对其进行更改，也用同样的方式来处理科目名称的问题  4. 二次检查清洗后的一分一段表json文件，修改错误，统一编码格式，并将其转为csv文件    **首先手动找出错误数据，排查出22个编码不统一的文件，排查出4个json文件中仍存在数据错误的问题**  **将json文件读取为字典格式，使用多重循环读取到最内层的分数和累计人数，用pandas保存数据，最后保存至csv文件中，编码格式统一为uft-8-sig** | | | | | | | |
| **五、实验结果与分析**  1. 手动整理一二级专业  找出记录一二级学科专业的docx文件，人工读取数据并写入excel文件，用代码读取并重新写入来控制编码格式  2. 同小组其他成员共同清洗检查高校三年的录取分数线和三年一分一段表json文件内的数据是否合理，格式是否正确，若不正确则修改  检查3年江苏文理科的一分一段表  检查2017年上海综合604-504分一分一段表  检查2017年上海综合503-402分一分一段表  检查2017年江苏文科理科选测BC一分一段表  检查2017年浙江综合一分一段表  检查黄一凡同学爬取的北京外国语大学数据  检查徐浩卿同学爬取的中国地质大学（北京）数据  检查闻浩同学爬取的中国矿业大学（北京）数据  检查邵一展同学爬取的新疆大学数据  3. 更改高校三年录取情况中大家爬取数据里省份不统一，科目不统一，数据矛盾等问题  将一百多省份缩减回34个，将科目控制为{文科、理科、综合}三门  4. 二次检查清洗后的一分一段表json文件，修改错误，统一编码格式，并将其转为csv文件  初次清洗后的json文件仍存在22个编码不统一，4个文件存在格式问题    更改后的一分一段表json文件已处理为csv文件，文件编码为’utf-8-sig’，已上传且发送给薛翔天同学 | | | | | | | | |
| **六、实验总结与心得体会**  1. 数据清洗工作对代码水平要求不高，需要的是面对大量数据保持冷静与耐心，认真逐行校对数据，一旦失误，则大概率需要重新清洗，因此很多时候往往在重复一遍又一遍的文字工作。  2. 数据清洗工作相比于写代码更需要队友的合作与信任，良好的分工可以减少大家的疲惫，提高工作效率  3. 这次试验部分磨练了我的耐心和认真，作为项目的基础，既要保证后续开发工作数据的正确性，又要及时根据开发小组要求来提供对应数据  4. 数据清洗工作是必要的，但是我仍能感觉出所用的时间比学到的知识效率要低些，希望下次可以尝试其他类型的工作 | | | | | | | | |

2020年9月制