暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 第一小组：数据清洗与融合小组 | | | | | | |
| 姓 名 | 张博宇 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118229 |
| 实验时间 | 2020.8.31-2020.9.23 | | 指导教师 | 孔祥龙 | | 成绩 |  |
| 1. **实验背景和目的**   背景：高考热度逐年升高，如何根据自身排名与分数选择一个合适的高校是考生当前急需解决的问题。在这样的背景下，我们构建基于Web的高考志愿推荐系统，以帮助考生们更科学、更高效地做出志愿选择。  目的：加深对软件实践课上所学开发模式的理解，运用软件实践课程所学知识进行实践，提高实践能力、Python编程能力、团队协作能力等。 | | | | | | | |
| **二、小组任务和个人任务**  小组任务：①清洗同学们两次作业中的数据，分别整理各高校近三年录取分数线与各省市近三年高考成绩一分一段表。  ②根据各组提供的需求，将数据重新增添改查，修改成其他组需要的呈现模式。  ③负责收集仓库外的数据，如：未交作业同学本应负责的数据，校徽图片等  个人任务：①清洗同学们两次作业中的数据，整理各高校近三年录取分数线，对数据进行初步统计。  ②检查各高校近三年录取分数线与各省市近三年高考成绩一分一段表数据是否合理，改正不合理/爬取错误的数据。查找未交作业同学的数据。  ③收集校徽。 | | | | | | | |
| **三、个人任务需求分析**  任务①：这需要整合所有同学的作业，在作业仓库中先去除csv、json格式外的所有文件，然后将json文件与csv文件分别整合在一个文件夹中。由于json文件根据各组需求的不同会产生很多变化，可能有一省三年分数的要求也可能有一省一年的要求，所以暂时先不合并json，根据后续其他组的要求再做处理。对于csv文件，可以直接进行整合，将所有csv合并成一个文件。还需要注意的是，很多同学的做作业时，并没有按照老师的要求去做，会出现属性名因为错别字而增多、编码转换、整型数据写成字符数据等等问题，所以在合并时，需要先写工具函数将这些问题解决。  任务②：这是根据各组反馈回来的问题再对数据进行检查，这部分需要认真仔细地检验，可以利用Python/Excel函数进行判断，必要时，还需要手动检查。  任务③：收集校徽是一个有趣的工作。实现思路是先下载全国所有高校的校徽（全部高校的校徽很容易找到），然后在Python中根据所需高校的名字列表对图片进行筛选，将需要的高校图片输出为结果。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**     1. 数据清洗与整合   定义工具函数进行数据的预处理与统计工作    根据数据输出整合好的高校分数文件csv，并且展示数据初步统计的结果     1. 爬取校徽图片   根据所需高校的名字列表对图片进行筛选，将需要的高校图片输出为结果 | | | | | | | |
| **五、实验结果与分析**  ①得到整合在一个csv文件里全部高校近三年录取分数线，此文件包含所有按照老师所要求的正确格式完成作业的同学们的数据。还有一些高校数据因为同学们没交作业/未按照要求写作业而无法使用，后续我们会根据各组的需求来完善与修改这部分有问题的数据。  ②得到所有高校的校徽图片。  ③检查浙江2017-2019一分一段表，发现2018一分一段表不完善，将其分数段补全到一本线。检查北京大学、清华大学、北京交通大学、中国人民大学的各专业录取分数线，发现清华大学数据中本一批与提前批混杂了了一起，故删去提前批只保留本一批。 | | | | | | | |
| **六、实验总结与心得体会**  刚开始的时候，我们并没有想象到第一组的任务会如此的艰难Orz。很多困难其实都是来源于同学们没有按照要求完成作业，但是最重要的还是我们没有经验导致的。我们在整理完数据后收到了各组的更改反馈，这些反馈大多都是发现几条数据数值不合理，这些问题我们是没有想到的，所以只能在各组反馈后再去更改。实际上，我们在一开始的时候就应该把好关，将数据从头到尾细致检查一遍，因为当时觉得这样做太耗费人力与时间，并且抱着同学们的作业都很完善的侥幸心理，所以就放弃了这样的检查。但是最后，我们还是对数据做了全面的检查。这次艰难体验让我明白了后端工作者的不容易，同时也让我初步了解了团队协作开发项目的流程。  从个人角度来说，这是对我python编程能力的一次锻炼，我进一步熟悉了Pandas、Shutil等库的使用方法。同时，我也明白了团队之间保持良好且及时的沟通是极为重要的，无论是组内还是组外。因为项目开发的过程中几乎总会出现各种各样的问题，而当突发情况发生时，有效的沟通可以使问题快速被对症下药。  回顾整个软件实践课程，我收获了许多经验和教益。无论当初做得是否出色，最后都成为了宝贵的历练。最后，感谢一直带领我们的孔老师，您耐心的指导和无数付出成就了我们在这门课上的收获。回首这四周软件实践，我们受益无穷。 | | | | | | | |

2020年9月制