暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 组1——数据采集和清洗 | | | | | | |
| 姓 名 | 陈嘉源 | 专业 | 人工智能 | 班级 | 2 | 学号 | 09118241 |
| 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**  爬取了各高校近三年的招生分数线，利用这些数据预测2020年的分数线，建立知识图谱，用AI算法为学生提供志愿填报推荐。  对项目进行分工与合作，让同学们感受大厂分工的安排，学习并适应如何在组内和组外互相沟通，最终完成项目，为将来的工作与实习打下基础 | | | | | | | |
| 1. **小组任务和个人任务**   小组任务：  数据处理，统计现有数据，按照需求补充新数据，数据清洗。  个人任务：  对仓库中两次小作业的文件（部分）清洗，删除无用数据。  将各个省的json文件合并  爬取数据库中所包含的高校（部分）的校徽  检查第一次作业高校分配表与数据库中的文件是否对应  部分省、学校数据检查 | | | | | | | |
| **三、个人任务需求分析**  前两次作业爬取了某些大学的录取分数线以及各省文科的一分一段表，而在同学们提交的  作业中包括作业的源代码以及一些无用的数据文件，因此需要将文件（部分）下载到本地后进行清洗，保留csv文件和json文件后以建立数据库。  由于为保证每个人的作业量相当，各个省的文科或理科或某些分段的一分一段表被分开到一些json文件中，因此将这些文件按照省份进行合并。  由于UI组需要高校的校徽，需要爬取分配到的部分高校的校徽。  部分同学提交的作业中的内容错误，需要进行更正，同时有些同学所爬取的学校与其分配的学校不对应，需要汇总这些信息并更改。  得到AI组的反馈，部分省以及某些学校爬取的分数线存在错误，包括一本线分数不完整，高校录取分数线包含提前批等问题，需要进行检查更正 | | | | | | | |
| 1. **实验过程（需附上关键代码及相关说明）**   （部分任务为人工处理，无需编程实现，不在此赘述）  1、  截屏2020-10-07 下午4.41.52  2、  **截屏2020-10-07 下午5.27.46**  **截屏2020-10-07 下午5.21.38**  ***Reference：<https://blog.csdn.net/qq_37954086/article/details/88722852>*** | | | | | | | |
| 1. **实验结果与分析**   部分数据汇总结果：  截屏2020-10-07 下午5.43.09 截屏2020-10-07 下午9.49.09  将大学录取分数线汇总后pr到仓库到****extFiles**文件夹中**  在检查高校数据中发现南师大的录取分数线中包含艺术类；河北省2018年一本线数据不完整；部分csv文件未按要求格式给出。 | | | | | | | |
| **六、实验总结与心得体会**  这次16+48学时的软件实践经历，让我收获颇丰。  在两年的大学课程学习过程中，学习了编程语言、数据结构等编写代码方面的知识，但缺乏一个完整的软件项目开发经历，让我感觉自己并非一个真正的计软智学院的学生。通过这次暑期学校的学习，与同专业同学一同学习、合作完成的高考填报助手项目，经历了完整的项目开发流程。同时，作为负责数据采集和清洗的小组成员之一，不断接收别的小组的反馈从而完善、改正错误或满足新的需要的过程，不仅为项目的开发贡献自己的力量，也是对自己码代码的基本素养、合作能力的极大提升。  我们组负责数据采集和清洗，需要为整个项目提供足够且准确的数据。从整合高校、各省的分数线，到数据的检查更正，接受别的组的反馈不断完善数据量，每一步都需要认真完成，为自己提供的数据负责到底，才能给整个软件开发项目打下牢固基石。  在检查同学所交作业的时候，如果发现有些错误，需要查看他们提交的作业源代码进行调整，这时我深刻意识到，编写代码时的规范性以及注释的重要作用。按照规范的格式编写代码，命名变量，添加合理的注释，能在回过头检查的时候带来很大的便利。  在后面进行数据清洗的时候，看似简单的一部分，实则需要检查和修改的任务也很繁重，而且保证数据的准确和完整对整个项目来说也是至关重要的。随着软件开发的进行，在与小组成员合作的过程中，我认识到，团队开发成员间如何更好地交流、协同工作以提升开发的效率在整个开发过程中尤为关键。  、 | | | | | | | |

2020年9月制