暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 第一小组（数据搜集与清洗） | | | | | | | |
| 姓 名 | 王士一 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118234 | |
| 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**  为了能够建立一个给高三毕业生高考出分后使用的大学推荐系统网站，并且锻炼我们的软件实践能力以及相关AI算法的应用能力，我们使用Django的模型框架搭建了一个大学推荐系统的网站。 | | | | | | | | |
| **二、小组任务和个人任务**  整合并清洗同学们在第一二次作业中获得到的大学录取信息的csv文件以及各个省市的一分一段表的json文件。在此过程中将不合理的数据剔除；  建立一级学科二级学科的对应表单，给每个专业增添了对应的学科类型；  存取并录入了985，211，双一流等名牌学校的信息。并按照省份进行划分；  为爬取到的大学都找到了对应校徽的图片；并获得了大学的排名；  收集了部分2020的一分一段表以及在修改同学们不正确的内容上又自己爬取了一些缺失的数据；  编写查询函数和前端小组进行对接函数的查询功能，并对查询的结果进行可视化；  以及在最后报告之后配合第十组进行了多轮的数据清洗与核查； | | | | | | | | |
| **三、个人任务需求分析**  我们是第一小组一个小分队的组长，带领我们组员一起完成了查找985，211以及双一流名校并记录的任务。在最开始我们还被安排建立从各个大学自己的学科名到对应一级学科的映射，但由于沟通问题，效果并不理想，故未使用。然后我们组又为985，211大学增添了省份信息，并且从网上查找了多种大学的排名信息。并且在汇报结束后配合第十组进行数据的清洗与核查；  我个人还接手了对接第九组（前端小组）的工作，负责给前端小组编写具有不同查询功能的响应函数，从文件和数据库中均获得应用。并在开始阶段为查询函数提供了可视化的工具函数。 | | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  *for* cate,num *in* zip(csv['Major'],csv['Number']):  *for* category *in* my\_category:  *if* category[:-1] *in* str(cate):  num=my\_category[category]  *break*  这段代码用于添加不同学科名对应的一级学科名称，但实际准确率只有百分之70，故舍弃。    这个函数用以从给定的大学排名的网站来进行爬取大学的排名数据    这个函数用以查询给定省份，年份，科目类别后某大学近3年的录取情况；    这个函数用以查询某分是在某年上下浮动wave分的录取情况    这个函数用以查询某年某省的一分一段表    并且在main函数中实现了接收用户不同输入的自动调用功能，通过接收用户不同的的输入来自主判断所要调用的查询函数。  同样的查询功能刚开始在文件里查询也有，但没有使用。还有一个将查询结果可视化的函数也没使用（已pr） | | | | | | | | |
| **五、实验结果与分析**  1.数据库查询的效率比从文件中查询的效率快的多。只需要不到1s的时间就可以完成查询，但连接数据库耗时较长。  2.自动添加一级二级学科的内容逻辑不够明确，添加的结果仅有70%的准确率；  3.录入的结果问题不大。  4.可视化绘制表格可以赏玩；如下 | | | | | | | |
| **六、实验总结与心得体会**  1.组与组之间的交流沟通十分重要，对接的要求适合在一开始确定下来，以免出现最后关头临时更改任务的事情。组长在一个团队作业中的任务十分重要，良好且明确的分工可以给全组同学带来便利；  2.数据库的查询和维护的成本都比原始数据好得多，更容易也方便。推荐这种有数据量的工作都统一使用数据库，而不是一直使用原始文件；  3.前端的接口设计包括各个算法组的接口都应该在最开始列出一个明确的需求来反馈给数据库组，来让数据的同学可以对症下药；  4.数据可以少，但是不可以错，数据少只会结果不好；但数据错误的损失太大了； | | | | | | | |

2020年9月制