暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 第一小组，数据处理，数据融合 | | | | | | |
| 姓 名 | 李辰浩 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118226 |
| 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**  实验背景：本次课的课程设计是全体同学协作实现一个基于Python的高考志愿推荐网站。为提高同学们的动手实践能力。同时高考志愿一直以来都是家长们十分关心的问题，如何报一个好的大学，让自己的分数不浪费是我们构建高考填报志愿助手的背景。 | | | | | | | |
| **二、小组任务和个人任务**  **小组任务：**1. 对同学们收集到的大学数据进行整理，删除其中定向，港澳台等特殊考生的成绩。2. 整理一分一段表，并加入2019缺失大学的一分一段表以及2020部分大学的一分一段表。3. 满足别的小组所提需求，获得校徽，一类学科等数据。  **个人任务：**1. 和老师以及别的组交流对接，得到对于数据的需求以及反馈，对数据任务进行简单的分析改进 2. 整理一二级学科等相关数据 | | | | | | | |
| **三、个人任务需求分析**  任务需求：1. 和老师以及别的组的组长对接，分析数据任务的具体要求以及重点难点，并将任务合理分配给组员。  2. 数据的预处理，对各大学在各省市的分数进行筛选，选择出符合数据要求的数据。  3. 解决数据中存在的问题，比如一些数据同学们收集数据时没有符合老师所给的格式，我们把数据调整，将文史，理工等科目换成文科，理科，同时删除定向、港澳台、艺体生等数据。对于混在正常本科批次的数据，我们尽可能的删除，不能删除的，我们和第十组一起，重新爬取了相关的数据。同时感谢我的组员，他们把本该大家爬取的学校的数据有些进行了重新爬取。  4. 和王士一同学一起和UI组对接，通过在界面上用数据库的查询，可以直接显示录取分数线或者一分一档表的结果，在我们和UI组对接时发现查询结果不能用DataFrame形式展示，我们想办法将查询的结果用图片形式展示出来。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  手动居多。  import os  import pandas as pd  import numpy as np  import json  rank\_file=os.getcwd()+r'\2018安徽理科.xlsx'  df=pd.read\_excel(rank\_file,sheet\_name=0)  for col in df.columns:      df[col]=df[col].apply(lambda x: x if str(x).isdigit() else np.nan)  df\_science=df.iloc[:,0:3:2].dropna(axis=0)  #分数和排名转换给字典  dict\_science=df\_science.set\_index(df\_science.columns[0]).T.to\_dict('list')  for k in dict\_science:      dict\_science[k]=dict\_science[k][0]  year=2018  province='安徽'  category='理科'  data\_dict={year:{province:{category:dict\_science}}}  json\_str=json.dumps(data\_dict,indent=4,ensure\_ascii=False)  print(json\_str)  with open('2018安徽理科.json','w')as json\_file:      json\_file.write(json\_str) | | | | | | | |
| **五、实验结果与分析**  处理好的数据 | | | | | | | |
| **六、实验总结与心得体会**  本次实验本组工作总的来说做的不算好，数据清洗好多不是很干净，但小组内的成员都十分努力，参与了数据处理的大部分工作，好多时候都是自己手动来做，手动清洗，有些数据处理了一遍又一遍，同时还经常受到别的组的抱怨。  当然，我们组自身有不小的问题，一些艺体生的分数缺失会严重影响分析，给后面的组造成不便，但正如组员所讲，我们只是装修队，现在却让装修队去造房子，确实有些难办。同时，有些同学大学的数据缺失，一份一段表比较残缺，我们只能去返工，这些工作量都比较大。  在这次软件实践，我还是学了很多东西，掌握了一些团队协作的技巧，然后也熟悉了企业编写一个大程序的流程，同时也掌握了许多数据处理的方法。 | | | | | | | |

2020年9月制