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 | 沈飞鸿 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118228 |
| 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. 从之前班级作业中获取各省分数线和各校分数线(序号为73到109)。 2.整理二级学科文件前半部分，使其格式与一级学科一致。 3.将各省分数线的json文件统一格式，并将所有数据类型转化为统一数据类型 4.检查部分省份分数线数据。 5.修改云南文科理科的文件，使其从五分一段转为一分一段表。 | | | | | | | |
| **三、个人任务需求分析**  1将上学期爬取的各校分数线数据下载，删除无用数据，只保留一二次作业的文件（最终格式为1个名字文件夹内只含有csv文件和json文件，序号为73到109）。  2整理前半部分二级学科文件，使其格式与一级学科一致。  3遍历所有2017-2019各省文理科分数的json文件，并统一格式，将所有名次变为统一数据类型。  4检查2017云南文科和理科，2017山西文科和理科，2017青海文科和理科，2017内蒙古文科和理科的数据  5修改云南文科理科的json，使其从五分一段表转为一分一段表 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1将上学期同学们上传到码云的作业下载，删除文件夹中的无用文件，只保留一二次作业的结果，即包含各省市的一分一段表的csv文件，和包含各校的录取分数线的csv文件。将缺失或有问题的同学序号记录下来。(负责序号为73到109的同学的文件夹)  2整理整个合并的二级学科文件，为各级学科添加小标题(一级学科)。使其格式与一级学科的csv一致。由于需添加的量不多，此任务为手动添加。  3遍历所有2017-2019各省文理科分数的json文件，并统一格式，将所有名次变为统一数据类型Int，对齐并统一格式。输出结果到文件夹json\_updated.  代码：  #处理json文件，使所有年份,分数都变成字符,串位次数据都变成Int, import json import os  if not os.path.exists('json\_updated'):  os.mkdir('json\_updated') path=os.getcwd()+'\\json' save\_path=os.getcwd()+'\\json\_updated'  my\_encode='utf8'#编码格式 jsonlist=os.listdir(path)#包含json文件名的列表  #print(jsonlist) #print(jsonlist.\_\_len\_\_())  for i in range(jsonlist.\_\_len\_\_()):  trans\_dict0=dict()  trans\_dict1=dict()  trans\_dict2=dict()  trans\_dict3=dict()    if i==7 or i==8 or i==18 or i==23 or i==24 or i==63 or i==89 or i==100 or i==106:  my\_encode='gbk'  continue#这里8个文件编码格式有问题，手动完成  else:  my\_encode='utf8'    with open(path+'\\'+jsonlist[i],encoding=my\_encode) as json\_temp:  dict\_temp=json.load(json\_temp,encoding=my\_encode)  for k1,v1 in dict\_temp.items():  #print(v1)  k1=str(k1)  for k2,v2 in v1.items():  #print(v2)  for k3,v3 in v2.items():  #print(v3)  for k4,v4 in v3.items():  #print(type(k4))  k4=str(k4)  v4=int(v4)  trans\_dict3.update({k4:v4})  trans\_dict2.update({k3:trans\_dict3})  trans\_dict1.update({k2:trans\_dict2})  trans\_dict0.update({k1:trans\_dict1})   #print(trans\_dict0)    result=json.dumps(trans\_dict0,indent=4,ensure\_ascii=False)  #print(result)  with open(save\_path+'\\'+jsonlist[i],'w',encoding=my\_encode) as json\_file:  json\_file.truncate()#清空原json  json\_file.write(result)  4检查2017云南文科和理科，2017山西文科和理科，2017青海文科和理科，2017内蒙古文科和理科的数据，检查文件是否缺失，是否数据量过少。  5修改云南文科理科的json，使其从五分一段表转为一分一段表。  代码：  import json import os  my\_encode='utf8'#编码格式  trans\_dict0=dict() trans\_dict1=dict() trans\_dict2=dict() trans\_dict3=dict()  path=os.getcwd()  with open(path+"\\2017云南理科.json",encoding=my\_encode) as json\_temp:  dict\_temp=json.load(json\_temp,encoding=my\_encode)     for k1,v1 in dict\_temp.items():  #print(v1)  k1=str(k1)  for k2,v2 in v1.items():  #print(v2)  for k3,v3 in v2.items():  #print(v3)  for k4,v4 in v3.items():  #print(type(k4))  if(int(k4)>=680):   max\_add=4  if(int(k4)<160):  max\_add=2  elif(int(k4)>=160 and int(k4)<680):  max\_add=5  for i in range(5):  if int(k4)<=100:  trans\_dict3.update({int(k4):v4})  break  v4\_new=min(int(v4)+i,v4+max\_add)  print(int(v4))  trans\_dict3.update({int(k4)-i:v4\_new})  trans\_dict2.update({k3:trans\_dict3})  trans\_dict1.update({k2:trans\_dict2})  trans\_dict0.update({k1:trans\_dict1})     result=json.dumps(trans\_dict0,indent=4,ensure\_ascii=False) f=open("2017云南理科\_new.json",'w') f.write(result) f.close() | | | | | | | |
| **五、实验结果与分析**  第一个任务得到73-109号同学的文件夹，内含各省分数线的json和各校分数线的csv文件，另附一个txt文件记录有误或者缺失的名单  第二个任务得到修改过后的学科代码为1-9的二级学科文件。  第三个任务得到统一格式后的Json\_updated文件夹  第四个任务为检查文件  第五个任务得到云南文科理科一分一段的json文件 | | | | | | | | |
| **六、实验总结与心得体会**  本次实验任务分配到个人身上较为简单，并且一半任务是手动完成的，另一半的代码难度也不大。主体还是解决后面各组同学提出的各种问题。  通过这次实验，我深刻体会到了数据在整个项目中的重要性，数据处理环节的问题会严重影响后面的环节，我们也需要不断地解决其他小组提出的问题。本次实验最深刻的印象是，虽然最初几天我们已经初步“完成”任务，但后面还是会不断涌入问题，需要不断与他组协商，了解问题，解决问题。总而言之，此次实验不仅让我明白了数据处理和清洗的任务，也让我大致了解了一个项目的产生过程，令我受益颇多。 | | | | | | | | |

2020年9月制