暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 数据采集和清洗 | | | | | | | |
| 姓 名 | 张文杰 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118227 | |
| 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**  云端协作开发的应用日渐成熟，高考结束后，对有关志愿填报的辅助的需求日益增多，在此背景下，暑期学校实验项目目标定为构建基于Web的高考志愿推荐系统 | | | | | | | | |
| **二、小组任务和个人任务**  小组任务：数据采集和清洗  个人任务：大学所在省份采集  大学排名数据采集  各省份985，211高校数据采集与检查  部分查询函数  部分缺失大学数据采集与部分省份一分一档表检查 | | | | | | | | |
| **三、个人任务需求分析**  i)大学所在省份采集通过编程实现很麻烦，但简单重复的人工操作却更容易完成。  ii)大学排名数据各网站并不统一且无官方数据，这里选择一个合理即可  iii)部分缺失大学数据采集的爬取工作由于官网存在的部分问题导致数据获取困难，可通  过其他网站获取  iv)985，211高校数据按省份排列 | | | | | | | | |
| 1. **实验过程（需附上关键代码及相关说明）**   def get\_rank(Year,Category,Province,score):##输入省份年份与分数类别查询对应的排名     year=str(Year)     name=year+Province+Category     path="E:\作业1\一分一档表"+"\\"+name+".csv"     df=pd.read\_csv(path)     temp=df[df['分数'] == score]['累计人数']     rank=temp.values[0]     return rank  def score\_iqury(Year,Category,Province):##输入省份年份与类别查询对应的分数和排名     year=str(Year)     name=year+Province+Category     path="E:\作业1\一分一档表"+"\\"+name+".csv"     df1=pd.read\_csv(path)     df2=df1.drop(['Unnamed: 0','省份','年份','科类'],axis=1)     return df2  ######################爬取大学排名数据#############  soup = BeautifulSoup(html, "html.parser") # soup是解析后的树形结构对象  for item in soup.find\_all('tr'): # 查找符合要求的字符串形成列表  #print(item) #测试查看item全部  data = [] # 保存一个学校的所有信息  item = str(item)  #排名  paiming1 = re.findall(paiming, item) #re正则表达式  #print(paiming1)  def askURL(url):  # 我的初始访问user agent  head = { "User-Agent": "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.116 Safari/537.36"  }  request = urllib.request.Request(url, headers=head) # 携带头部信息访问url  # 用request对象访问  response = urllib.request.urlopen(request) # 用urlopen传递封装好的request对象  html = response.read().decode("utf-8") | | | | | | | | |
| **五、实验结果与分析**  i)实现了大学所在省份与各大学排名的数据采集工作，导出了对应的csv文件  ii)完成了985，211高校数据采集与检查，导出了对应的txt文件  iii)检查并修改了部分省份一份一段表与部分大学录取分数存在的数据丢失、不规范问题  iv)在网上查找到的信息需要进行二次筛选，以剔除垃圾信息，部分网站数据甚至存在错  误，需要进行多方比对。  v)整理数据时需统一格式、类型，避免因类型、格式不当出现数据无法录入的情况 | | | | | | | |
| 1. **实验总结与心得体会**   i)网上的数据存在各种各样的问题，需要进行检查与核对。  ii)可尽量爬取官方数据以保证准确性。  iii)本次实验中我负责的部分大多需要人工手动操作，工作时要细致，不然会给后面操作的同学带来很多不必要的麻烦。 | | | | | | | |