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 | 朱浩鹏 | 专业 | 人工智能 | 班级 | 091181 | 学号 | 09118138 |
| 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。在全班一起完成任务的基础下，数据的处理与融合是必不可少的一部分。 | | | | | | | |
| **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. 修改并清洗了之前同学爬取得大学数据，同时对一些错误的数据进行了再爬取  2. 修改并清洗了同学所制作的一分一段表数据  3. 制作了每个大学所对应的是否为985,211，以及对应省市的表格 | | | | | | | |
| **三、个人任务需求分析**  任务需求：   1. 检查一分一段表中数据是否有数字的错误，以及逗号是否为中文逗号 2. 对于一份一段表中数据存在错误的情况，使用Excel对其进行重新计算。   2. 对于爬取完全错误，或者数据缺失的省份，进行数据的在爬取，所爬取的两个学校均为动态url类型的网站，需要在代码中运用时间戳对网站进行爬取，最后将爬取的数据按照格式进行排列  3. 制作各个学校是否为985,211的表格，并显示学校的学校编号，这部分内容因为网上没有合适的内容爬取，所以采用手动查询的方式，并手动录入数据 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  除了爬虫之外的工作基本通过手动修改，爬虫的代码如下：  import requests import pandas as pd import csv   *# 用于遍历的列表* province\_list = [**'北京'**, **'天津'**, **'河北'**, **'山西'**, **'内蒙古'**, **'辽宁'**, **'吉林'**, **'黑龙江'**, **'上海'**, **'江苏'**, **'浙江'**, **'安徽'**, **'福建'**,  **'江西'**, **'山东'**, **'河南'**, **'湖北'**, **'湖南'**, **'广东'**, **'广西'**, **'海南'**, **'重庆'**, **'四川'**, **'贵州'**, **'云南'**, **'西藏'**,  **'陕西'**, **'甘肃'**, **'青海'**, **'宁夏'**, **'新疆'**, **'新疆内地班'**, **'西藏内地班'**, **'港澳台侨联招'**] year\_list = [**'2019'**, **'2018'**, **'2017'**]  *# 爬虫的header* header1 = {  **'Accept'**: **'text/html,application/xhtml+xml,application/xml;q=0.9,**\  **image/webp,image/apng,\*/\*;q=0.8,application/signed-exchange;v=b3;q=0.9'**,  **'Accept - Encoding'**: **'gzip, deflate, br'**,  **'Accept-Language'**: **'zh-CN,zh;q=0.9'**,  **'Connection'**: **'Keep-Alive'**,  **'Host'**: **'wsbm.ecust.edu.cn'**,  **'User-Agent'**: **'Mozilla/5.0 (Windows NT 10.0; Win64; x64)** \  **AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.87 Safari/537.36'** }   def request\_url(url, data1):  try:  response = requests.post(url, data=data1, headers=header1)  if response.status\_code == 200:  return response.text  except requests.RequestException:  return None   def main(province, year):  data1 = {  **'province'**: province,  **'nf'**: year,  } *# 对页面的动态申请* url = **'https://wsbm.ecust.edu.cn/lqfs.htm'** html = request\_url(url, data1)  df = pd.read\_html(html) *# 使用pandas的html解析功能对表格进行解析* table = df[6] *# 选择我们要处理的表* print(table)  for ii in range(2, table.shape[0]):  csv\_writer.writerow([**"华东理工大学"**, year, province, table[4][ii], table[6][ii], table[9][ii], **"09118138朱浩鹏"**])   if \_\_name\_\_ == **'\_\_main\_\_'**:  f = open(**'09118138朱浩鹏-华东理工大学.csv'**, **'w'**, newline=**''**, encoding=**'ANSI'**)  csv\_writer = csv.writer(f)  csv\_writer.writerow([**"College"**, **"Year"**, **"Province"**, **"Category"**, **"Major"**, **"Score"**, **"Contributor"**])  for i in province\_list:  for j in year\_list:  main(i, j) | | | | | | | |
| **五、实验结果与分析**  1. 对错误的数据进行修正后得到的正确数据，对之后的组来说，可以进行后续的操作 | | | | | | | |
| **六、实验总结与心得体会**  在这次的软件实践中，我参与了许多数据清洗的工作，对数据进行了检查与修改，同时对缺失的数据进行了查询，在整个过程当中，体会到了很多甲方乙方对接的感觉，不断满足所对接的小组所给出的需求，从中体会到了一些联合开发的感觉。  在进行数据补充时，使用爬虫对缺失的部分数据进行了爬取，所爬取的网站与之前所爬取的类型不同，其url是随着时间动态变化的，我也从中了解了时间戳的概念，对爬虫有了更深入的理解。  这次软件实践中，我学到了很多东西，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10月制