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 | 尹行健 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118244 |
| 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**  在长学期所做的高校数据爬虫基础上，分组进行数据处理、知识图谱构建、知识推理、算  法设计、UI设计、数据库构建，实现基于Web的高考志愿填报系统。 | | | | | | | |
| **二、小组任务和个人任务**  小组任务：   1. 补充采集数据；2、分析现有数据列举统计指标；3、提供查询接口。   个人任务：  1、将json格式的各省份历年一分一段表转化为csv格式，为后续任务提供更方便的数据；2、数据清洗与纠错。 | | | | | | | |
| **三、个人任务需求分析**  1、json\_to\_csv  读取已有的province文件编码和已有的json文件，根据文件名判断文件中数据的省份、  年份、科类，提取json文件中的分数排名信息，加入省份、年份、科类标签输出为csv文件。  2、数据清洗与纠错  该任务与第十组分工完成，人工检查数据的完整性、合理性，并进行修改。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1、json\_to\_csv.py  import pandas as pd  df = pd.read\_csv("Province.csv",encoding="gbk")  province = df['Unnamed: 1'].tolist()  code = [i for i in range(1,len(province)+1)]  code = dict(zip(province,code))  print(code)  #存放json的文件夹名称  jsonpath = './json'  #输出csv的文件夹名称  csvpath = './json\_csv/'  import os  def scan\_files(directory,prefix=None,postfix=None):  files\_list=[]  for root, sub\_dirs, files in os.walk(directory):  for special\_file in files:  if postfix:  if special\_file.endswith(postfix):  files\_list.append(os.path.join(root,special\_file))  elif prefix:  if special\_file.startswith(prefix):  files\_list.append(os.path.join(root,special\_file))  else:  files\_list.append(os.path.join(root,special\_file))  return files\_list  #读取当前目录的上级目录  temp = os.path.abspath(jsonpath)  all\_files = scan\_files(temp)  import json  with open(all\_files[0],'r',encoding='utf8')as fp:  json\_data = json.load(fp)  print(json\_data)  for a in json\_data.items():  for b in a[1].items():  for c in b[1].items():  print(c[1])  values = [i for i in c[1].values()]  keys = [i for i in c[1].keys()]  print(values)  print(keys)  pro\_num = 0  for file in all\_files:  file\_name = file.split('\\')[-1].split('.')[0]  year = file\_name[0:4]  for pro in province:  if pro in file:  pro\_num = code[pro]  break  if "理科" in file:  subject = "理科"  elif "文科" in file:  subject = "文科"  else:  subject = "综合"  try:  try:  with open(file,'r',encoding='gbk')as fp:  json\_data = json.load(fp)  except:  with open(file,'r',encoding='utf-8-sig')as fp:  json\_data = json.load(fp)  except:  print(file\_name,"failed!")  continue  for a in json\_data.items():  for b in a[1].items():  for c in b[1].items():  print(file," open successfully!"," ",subject)  values = [i for i in c[1].values()]  keys = [i for i in c[1].keys()]  df = pd.DataFrame([keys,values,[pro\_num] \* len(keys),[year] \* len(keys),[subject] \* len(keys)])  df = df.T  df.set\_axis(['分数', '累计人数', '省份','年份','科类'], axis='columns')    new\_path = csvpath + file\_name + '.csv'  df.to\_csv(new\_path,encoding='utf-8-sig')  由于至少存在两种编码格式（utf-8&gbk）并且部分文件存在json格式错误（直接运行会报错），我增加了对每个文件成功读入的判断，当文件存在错误时会有输出提示（如下图）      2、数据清洗与纠错  分别进行了对高校招生分数线数据、一分一段表数据的正确性的检查。 | | | | | | | |
| **五、实验结果与分析**  1、json\_to\_csv  在输出错误之后找到错误的json文件手动进行改正，重新运行后得到csv文件    ……  2、数据清洗与纠错  人工检查了如下数据并对部分数据进行了手动修改与筛选，对北京体育大学的数据在原代码基础上重新爬取，最终得到了较为完整的正确数据。 | | | | | | | | |
| **六、实验总结与心得体会**  1、实验总结  实现了json到csv文件的格式转化，人工清洗了数据，为后续小组的开发提供了较为完整的正确数据。  2、心得体会  通过这次实践课程，我对大型软件项目有了更深入的了解：从开发流程、需求分析、系统设计到任务如何分配、各个小组如何协作……  我觉得需要反思的是每个环节的容错率和需求的明确性。以我做的json\_to\_csv来说，由于需求的改动和数据的错误，前前后后改了五六次，也做了一些无用功，最后才得到了比较正确的结果。而在数据清洗的过程中也由于我的一些检查失误导致了二次改动。所以我觉得在人数较多的项目中，应该从顶层设计到底层实现每个环节每个成员都要明确需求目标、每次提交任务都要经过慎重的检查，这样一个团队才能实现1+1>2的效率提升。 | | | | | | | | |

2020年9月制