**广州大学学生实验报告**

**开课学院及实验室：**计算机科学与网络工程学院软件实验室 **2020年 11 月20日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院** | **计算机科学与网络工程学院** | **年级/专业/班** | **18级软件工程3班** | **姓名** | 李子轩 | **学号** | 1806300106 |
| **实验课程名称** | **机器学习与数据挖掘实验** | | | | | **成绩** |  |
| **实验项目名称** | 多源数据集成、清洗和统计 | | | | | **指导老师** | 彭伟龙 |

**实验一**

1. **组员信息**

李子轩（小组共1人）

1. **作业环境（文件说明，函数说明，调用的函数库以及涉及哪些技术）**

**文件说明：**

数据源1：一.数据源1.xlsx

数据源2：一.数据源2-逗号间隔.txt

格式转换后的数据源1：一.数据源1.csv

格式转换后的数据源2：一.数据源2-逗号间隔.csv

存放中间结果的临时文件（简单将转换后的两组数据源进行合并）：tempFile.csv

最终的目标文件（存放完成数据集成、清洗后的数据）：objectFile.csv

**函数说明：**

在代码中均有详细标识和说明，在此不重复描述，主要调用了pandas下处理数据的函数和方法。

**调用的数据库以及涉及的技术：**

未调用数据库。

涉及的技术主要包括基于pandas的手段对给定的数据源进行集成，并按要求对数据进行清洗，同时调用了glob、csv等模块协助进行处理，最后由于题目要求计算部分不能使用库函数，通过手动对计算函数进行编程的方式对集成和清洗后的数据进行统计。

1. **作业题目和内容**

**题目**：广州大学某班有同学100人，现要从两个数据源汇总学生数据。第一个数据源在数据库中，第二个数据源在txt文件中，两个数据源课程存在缺失、冗余和不一致性，请用C/C++/Java程序实现对两个数据源的一致性合并以及每个学生样本的数值量化。

* 数据库表：ID (int), 姓名(string), 家乡(string:限定为Beijing / Guangzhou / Shenzhen / Shanghai), 性别（string:boy/girl）、身高（float:单位是cm)）、课程1成绩（float）、课程2成绩（float）、...、课程10成绩(float)、体能测试成绩（string：bad/general/good/excellent）；其中课程1-课程5为百分制，课程6-课程10为十分制。
* txt文件：ID(string：6位学号)，性别（string:male/female）、身高（string:单位是m)）、课程1成绩（string）、课程2成绩（string）、...、课程10成绩(string)、体能测试成绩（string：差/一般/良好/优秀）；其中课程1-课程5为百分制，课程6-课程10为十分制。

#### 参考：

数据库中Stu表数据

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Name | City | Gender | Height | C1 | ... | C10 | Constitution |
| 1 | Sun | Beijing | boy | 160 | 87 |  | 9 | good |
| 2 | Zhu | Shenzhen | girl | 177 | 66 |  | 8 | excellent |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

student.txt中

|  |
| --- |
| ID Name City Gender Height C1 。。。 C10 Constitution  202001 Sun Beijing male 1.80 87 。。。 9 good  202003 Tang Hanghai male 1.56 91 。。。 10 general  ... ... ... .. ... .. ... ... ... |
|  |

两个数据源合并后读入内存，并统计：

#### 1. 学生中家乡在Beijing的所有课程的平均成绩。

#### 2. 学生中家乡在广州，课程1在80分以上，且课程9在9分以上的男同学的数量。(备注：该处做了修正，课程10数据为空，更改为课程9)

#### 3. 比较广州和上海两地女生的平均体能测试成绩，哪个地区的更强些？

#### 4. 学习成绩和体能测试成绩，两者的相关性是多少？（九门课的成绩分别与体能成绩计算相关性）

#### 提示：

参考数据结构：

Student{

int id;

string name;

vector<float> data;

}

可能用到的公式：

|  |  |
| --- | --- |
| 均值公式 |  |
| 协方差公式 |  |
| z-score规范化 |  |
| 数组A和数组B的相关性 | 这里A=[a1, a2,...ak,..., an],  B=[b1, b2,...bk,..., bn],  mean(A)代表A中元素的平均值  std是标准差，即对协方差的开平方。  点乘的定义：IMG_256 |

注意：计算部分不能调用库函数；画图/可视化显示可以用可视化API或工具实现

**具体实验内容：**

**合并文件：**由于给定的两个数据一个为.xlsx，另一个为.txt，首先列出文件夹中的相关数据，找到其中.txt的文件重新组合文件名和后缀名,修改为csv格式，然后利用read\_excel和read\_csv读取文件，在分别利用to\_csv将两个文件均已csv格式保存下来；接着查看同文件夹下的csv文件数，循环读取同文件夹下的csv文件，将两个文件进行合并，将合并结果保存为tempFile.csv并打印出合并的结果。

**数据清洗：**首先要找到文件中缺失的数据项，通过对比两个文件中的数据，将合并后同一项的数据中查询到的已知值赋值给缺失值，具体功能实现通过fillVacancy()完成。由于C10列元素全为空，则将C10列全部填为0，将Constitution列的缺失值填补为unknow，若完成填充后仍存在缺失值，则将数据中所有NAN值转换为0。

由于两个文件中的属性单位等并不统一，需要对各列数据进行标准化。两个文件的性别命名并不统一，在合并后的表中，我们将女性统一命名为“girl”，男性统一命名为“boy”；对ID列数据进行标准化处理，统一为“202\*\*\*”的格式，通过standardId()完成；对身高列数据进行标准化，统一为以cm为单位。还需要注意的是，最后需要将数据转换成float，便于后续去重。紧接着便是对文件进行去重，即去除文件中的重复行，去重完成后还需要判断ID号是否存在相同的，只要检测到ID相同的则判定为数据失效，将所有ID相同的数据项全部删除，最后将清洗后的结果保存在objectFile.csv中并打印出来。

**数据统计：**该部分不详述，定义了avg()用于计算平均值，consGrade()用于将Constitution列的数据转换为数值型，便于后面的计算，std()用于计算标准差，correlation()用于计算相关系数。

1. **难题与解决**

由于对python的运用还不是很熟练，在初期编程过程中遇到了较大的困难，但是在不断的尝试之中找到了方法，通过请教身边对数据处理有相关经验的同学以及自己上网查阅相关的资料，结合自己以前的编程经验使得在实验的后期无论是效率还是质量都得到了快速的提升，在此特别感谢同学对我实验过程的帮助。

其次由于对实验内容的掌握在前期的工作中没有梳理清晰，导致自己对题目没有一个很好的把握，没有注意到很多细节，特别是在对数据的合并过程中出现的错误没有得到及时的解决，以至于在后续处理数据源的缺失、冗余和不一致性上形成了较大的麻烦，但在掌握到了整体的程序布局后马上展开了相关的改善工作，结合自己对题目的认识对代码进行纠错和改进，达到了实验要求的同时还增强了代码正确性与完整性。

1. **总结**

通过本次实验，让我对python对于无论是文件还是数据的处理都有了质的提高，让我对于python强大的功能有了更加细致的了解，其丰富的模块功能能够满足我们的大量的需求，python对于数据的处理的多样性的方法极大的扩展了我对其的理解；其次在实验过程中通过与身边的同学进行交流和探讨并且结合自己的自主学习，独立地去完成该次的实验极大的锻炼了我的自主实践能力，同时也暴露了一些我仍存在的不足，例如对于问题思考的方向还比较局限，难以摆脱固定的思维模式，对于问题解决的思路还不够清晰等等，需要我在后续的实验过程中对这些方面进行有针对性的补强，也感谢老师和同学对我此次实验课所提供的帮助。