**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 1 月1日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院** | **计算机科学与网络工程学院** | **年级/专业/班** | **软件183** | **姓名** | 郭阳洲 | **学号** | 1806300011 |
| **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**

郭阳洲,张世华,黄伟峰

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：gyzdata1.xlsx

数据源2：gyzdata2.txt

**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按要求对数据进行清洗，同时调用了xlrd2、math、pandas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，通过导入最后生成的总数据表excel表中的数据，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：**因为给的两个数据源一个是.xlsx，一个是txt文件，所以需要用两种不同的方法对数据进行整合，先将数据源1gyzdata1.xlsx文件读取并整合数据后再生成一个新的文件gyzdata1\_new.xlsx，然后再通过对.txt文件的操作将数据整合生成一个新的文件gyzdata2\_new.xlsx，最后将gyzdata1\_new.xlsx和gyzdata2\_new.xlsx整合成一张数据表AllData.xlsx。

**数据清洗：**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。在合并两个文件之前就先将女性统一命名为“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；对身高列数据进行标准化，统一为以cm为单位。之后就将两个文件的数据导出合并，然后去重，只保留重复的第一个数据，然后以ID进行升序排序后存入AllData.xlsx表中。

1. **难题与解决**

Python相对其他语言来说更容易一些，所以通过python做了这次实验，本来刚开始是打算建立数据库进行实验的，但做着做着搞不下去应该是自己的数据库出现了问题，所以贪图快点做实验就没有用数据库。刚开始的难点是不知道怎么将文档格式进行转换和编程，后来通过百度查资料后解决了，接着不知道所谓的数据集成、清理与统计要实现的是什么，但慢慢的问同学和查资料也就解决了，最后的问题就是相关性的问题，最后通过同学的指导也了解并解决。