**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22日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院** | **计算机科学与网络工程学院** | **年级/专业/班** | **软件181** | **姓名** | 黄晓波 | **学号** | 1806300174 |
| **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** | 实验一《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**

**黄晓波（单干）**

**二、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。（此处不规定语言，可随意选择）

* 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 180 87 。。。 9 good  202003 Tang Hanghai male 156 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 id;

vector<float> data;

}

可能用到的公式：

|  |  |
| --- | --- |
| 均值公式 | https://docimg9.docs.qq.com/image/MZxvju9kaLGQm3PWf8OsgA?w=184&h=89 |
| 协方差公式 | https://docimg4.docs.qq.com/image/WzKGNUBrrSKwDAJURyVRQA?w=448&h=74 |
| z-score规范化 | https://docimg8.docs.qq.com/image/DREX3MFPw-d3hZhtAkyp4g?w=200&h=94 |
| 数组A和数组B的相关性 | https://docimg2.docs.qq.com/image/XwYEh8-AUcIxFt2zVQ-sEg?w=561&h=263  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：https://docimg4.docs.qq.com/image/odNr0YyX8NwtmlJUDvD2lQ?w=248&h=72 |

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。

**三、作业环境**

（1）文件说明：

* 数据源1.xlsx (老师给出)
* 数据源2.txt （老师给出）
* 结果整理的数据.xls （整理获得）
* 数据挖掘实验1.py (代码文件)
* 数据挖掘-实验1-实验报告.docx (实验报告）
* 运行结果截图.docx (代码运行结果）

1. 函数说明：

由于本实验不存在需要可移植的函数，所以没有使用自定义函数解答。实现直接在主程序上完成

调用的函数只有读取xlsx数据和txt数据时才有。使用的是xlrd.open\_workbook（）函数和open（）函数。

（3）调用的函数库：

实验要求，所以在计算的过程中没有调用任何现有库，只有在读取数据的时候才调用了读取数据的函数。

但是使用的一些基本的数学公式，包括求根公式sqrt（），以及求次方公式（\*\*2）python写法。

（4）涉及哪些技术：

* 包括将xlsx文件读入内存中，使用了包括 xlrd.open\_workbook(file1\_addr)和xlrd\_file.sheets()[0]的函数。
* 将txt文件读入内容的技术，包括open函数，readlines函数以及write函数等。
* 程序中使用到的主要数据结构包括列表以及字典。列表包含了插入的函数append以及删除的函数pop，字典主要用于查找是否读过此数据，在python中使用了 （if search\_key in 源列表）的语句，也可以使用count函数查找这个待查值是否读取过。
* 将列表中的数据导入一个xls文件中，这里使用到了创建文件的函数，open（），
* 还有用到了python中一些特有的功能，包括在大部分情况下可以任意转换变量的类型，列表可以有负数下标，列表最后一个的下标可以是-1，

**四、难题与解决**

开始有一个最大的难题也是最难以解决的问题，就是我不知道该用什么语言来编写这次的实验程序，我自己在c++方面比较了解，在python上可以说是还没入门。但是c++相关的可视化方面我也还没学习，而python在这方面相对比较简单，并且有相当多的可使用的库（相对于c++），为了下面的实验，我觉得还是使用python来解决相对好一些。

在最开始的文件读取上我就遇到了麻烦，由于之前都是在txt文件上读入数据，没有尝试过读取Excel文件的数据，使用c++读取非常的麻烦，我哦也不想将文件转化为.csv格式，还好python有专门解决的库来帮助我解决这个问题，这里需要感谢软件183的肖小恩同学，我正是查看了她写的源代码

（https://github.com/tiger-xiao31/Machine\_Learning--experiment/tree/main）以及自己在网上查找资料才将读取的预备工作完成。

c++中的map在python总被字典所替代，我也在这里选择了字典做为一种数据结构，还有就是python的列表非常的好用，可以存放任意的数据类型，并且相当于一个数组般（有下标），比其他很多的数据结构清晰简单。

强制类型转换在这个实验中也是被我运用的很广泛，在清理整合数据的时候不断的使用str-->float 类型的转换。

我认为在这个实验中比较难的一部分就是如何清理数据，由于老师没有说明清理的规则，所以我使用了跟肖小恩同学一样的方法，先对名字去重，在对学号进行修改，但是有不同的方面，在她的做法中，她是将可能有重复的名字放在一起，但其实有可能不是在相邻的两行，我使用了字典存放出现过的名字，只要读取到出现过的名字，就将其删除，由于查找到相同立即删除而不是存储起来，还减少了少量的内存消耗。还有一个比较大的不同之处在于合并两个文件的内容时，在使用名字查找另一个表，并将第一个表格中的空缺的数据（可能是一个成绩，或者是身高等）补充完完整后，应该再将只出现在第二个表格中的信息再次填入第一个表格list1中，在这里我还是使用了字典记录出现过的名字，如果在list2表中找到未被记录的信息，就将其添加入list1中。由于清理整合方法的不同，所以得出的答案也跟别人的有所不同，但是大致上没有差别。（毕竟数量不算少）

在清理整合完数据后，就没有太大的问题了，第一个求解平均成绩的可以直接遍历一遍求和在除以总人数既可。

第二问可以使用一个判断语句就可以将符合条件的人数累加，当然，在这里我还有记录下符合条件的人的姓名，另外用了一个列表存放姓名。在最后的结果中有所显示。

第三个问题中就是将一，二问结合起来，在符合条件的基础上计算平均值。

第四个问题就是使用指导书上给出的公式，创建结果列表，存放使用攻击计算出来的结果，进行判断相关性。

**五、实验总结**

这次的实验，使用了python来编写，相当于重新学习了Python的基本语法以及使用，对于列表的增删改查方面有了比较深刻的理解，虽然在完成实验的时候遇到了很多的困难，但是在自己的摸索下还是完成了这个实验，总的来说收获还是很大（毕竟之前一直都是使用c++来编写程序），这次实验的最大收获应该就是初步的掌握了Excel文件的读写以及储存工作，在之前并没有完后过相对应的实验，希望在接下来的实验中也可以有比较大的收获。