**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183** | **姓名** | 肖小恩 | **学号** | 1806300128 |
| **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**

**一、组员信息**

* 肖小恩（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：xxe1.xlsx，

数据源2：xxe2.txt

集成、清洗后的数据：xxe\_final\_data.xls

（2）函数说明： 无函数使用，故没有说明。

（3）调用的函数库：

* sqrt(x)函数：返回数字x的平方根。
* 其他都是求平均数、方差、相关系数等都是一步步纯计算机计算，没有调用库函数

（4）涉及哪些技术：

* 将excel文件、txt文件数据按指定格式导入都python中的一个列表中，每一行数据本身是一个小列表，同时是列表的一个元素，然后便可通过下标索引来访问元素
* 对于两个数据源转成了对于两个列表操作，通过下标索引，循环比较，补充等，完成实验基本要求。进而进行题目1-4的计算。
* 同时将处理后的数据导出到一个excel文件中，方便查阅，以验证集成、清洗的正确性。

**四、难题与解决**

1. 首先是之前没有接触到这方面的知识或者实验，有些无从下手…
2. 后来自己不断摸索，看题，不懂的点就百度，然后老师也给了两份数据源，然后自己就不断尝试，最终得到了本次实验的大概思路：选择Python语言，先把两份数据源导入Python中，构造两个列表，然后通过列表下表索引操作里面的数据，最终得出处理后的数据。题目1-4就是按公式循环操作列表计算结果。
3. 得到思路后，先是去学习了Python语言中，两种文件如何导入列表，这个也尝试了不少时间，毕竟是初次接触这些，而且也是这学期开始才Python入门。后来发现其实导入挺简单的。
4. 导入后就是数据的去冗余、去不一致，以及数据缺失要处理，我是通过下标索引访问的列表元素，不断摸索，不断改进，不断debug，最终终于得到了处理成功的数据（按自己的标准，这个在github上有讲到。）
5. 然后就是进行题目1-4的计算，虽然说是按公式，前面三问问题不大，第四问花费时间稍微多一点，因为一开始没有很理解题目意思，理解了之后就去按公式通过循环一步步计算出来，虽然看起来比较简单，但也比较繁琐哈哈哈，因为不能调用库函数。

**五、实验总结**

通过这次实验，我学习到了Python中是如何导入文件数据，如何把数据导出到文件，以及对于列表操作更加熟悉了，也更加注意到了使用下标索引进行访问元素需要很谨慎，否则可能会失之毫厘，差之千里。对于那些平均值、方差、标准差、相关系数、协方差的计算公式进一步掌握与记忆。这次实验过程总体来说还是坎坷的，经历了从不想做🡪虽想做但无从下手🡪开始有思路，不断尝试🡪思路初步实现，不断摸索、改进🡪最终成功处理了两个数据源存在缺失、冗余和不一致性的问题🡪进行题目1-4的计算🡪现在在写实验报告了✌

题目-14中题目4花费时间较多，因为刚开始不是很理解题目，然后跟老师、同学交流后懂了，开始按公式写代码一步步计算，因为不能直接调用库函数，所以计算量比较大，那些下标索引刚开始也是有问题的，后来不断debug，最终得到了题目4的结果，即九门课程学习成绩与体能测试成绩的相关系数，一共九个。

至此，本次实验一圆满完成！感谢老师的指导、同学们的热心交流。