**一、组员：蔡焕斑**

**二、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 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 |  |
| 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： |

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：

压缩包中包含5个文件

dataSource1.xlsx和dataSource2.txt为老师给的数据源

code.docx为作业的代码

final\_data.xlsx为处理后最终得到的数据

README.docx为说明文件

②函数说明：

//处理xlsx数据源

public void parseXlsx(String fileName) {}

//处理txt数据源

public void parseTxt(String fileName){}

//根据传递的参数来对 txt/xlsx 根据 ID/Name 去重

public void delByNameOrId(String dataSource,String choice)

//以txt数据源为基准，合并两个数据源

public void mergeData()

//对数据中的空值进行处理，置为nullpublic void nullHandle()

//格式化Height，单位为m

public void formatHeight()

//格式化ID为202\*\*\*的格式

public void formatID()

//格式化Gender，男性和女性分别为male和female

public void formatGender()

//对数据进行格式化处理，包括 置null、格式化Height、ID和Gender

public void formatData()

//对数据进行去重处理

public void del()

1.统计学生中家乡在Beijing的所有课程的平均成绩

public void count1()

// 2.统计学生中家乡在广州，课程1在80分以上，且课程9在9分以上的男同学的数量

public void count2()

// 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bad,general,good,excellent的值分别定为：25,50,75,100

public void count3()

// *4.求学习成绩和体能测试成绩，两者的相关性*

public void count4()

//计算所有科目的平均值，打包成集合返回public ArrayList<Double> Average()

//将相同科目的成绩归纳到一个ArrayList，再将所有科目归纳到一个ArrayListpublic ArrayList<ArrayList<Double>> subjectsToArray()

//求协方差public ArrayList<Double> Covariance()

//求标准差public ArrayList<Double> Std()

//求Z-score规范化public ArrayList<ArrayList<Double>> Zscore()

//主函数

public static void main(String[] args)

③调用的函数库

import org.apache.poi.ss.usermodel.\*;  
import org.apache.poi.xssf.usermodel.XSSFRow;  
import org.apache.poi.xssf.usermodel.XSSFSheet;  
import org.apache.poi.xssf.usermodel.XSSFWorkbook;  
import java.io.\*;  
import java.text.NumberFormat;  
import java.util.\*;

**四、难题与解决**

对我来说，这个作业的难题主要在对数据的处理上，观察两个数据源后，我发现两个数据源**存在几个问题**：ID重复、Name重复、部分数据为空、两个数据源的ID、Height和Gender的格式不统一。思考之后，我采用了以下方法来处理数据：

1. 对于重复的ID和Name，去除掉第一个数据

②以dataSource2.txt数据源来基准，用dataSource1.xlsx数据源来补充dataSource2.txt数据源中的空白部分，补充之后依然为空的，置为null

③统一ID的格式为202\*\*\*\*，统一Height的单位为米(m)，统一Gender的值male/female

**五、总结**

因为最近刚好在学习java，所以就想着使用java来完成作业，当作对所学内容的一次巩固。

整个作业加上注释，接近600行代码。完成这个作业花费了我很多时间，其中大部分时间花在了处理bug上。Java不像python，一个变量可以是多种类型的，java是强类型的语言，在完成作业的整个过程，需要频繁地对数据进行各种类型的转化，经常出现类型错误的情况，这个调试的过程就花了我很长时间。

这个作业并没有遇到很大的难题，但是要处理的数据量稍微有点大，处理数据是这个作业主要的内容，比较繁杂。

在完成实验的过程中我也是学到了很多新的知识，除了对java知识的巩固之外，我还第一次接触到了以前从未使用过的断点调试，并且学会了使用断点调试，断点调试给我的调试工作带来了很大的便利。