成员：

郑家民、华祖海

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）

环境

总结

通过本次实验，收获了许多，K-Means是聚类算法中的最常用的一种，算法最大的特点是简单，好理解，运算速度快。一开始有点对分类和聚类的概念混淆。在经过两个实验后，终于理清了分类与聚类，它们之间最大的区别就是是否存在预定义；对于分类来说，在对数据集分类时，是知道这个数据集是有多少种类的，而对于聚类来说，在对数据集操作时，我们要做的，是将数据集中相似的数据归纳在一起。