**23-24(2)《程序设计2》大作业要求**

**2024-04-23**

**1.简述**

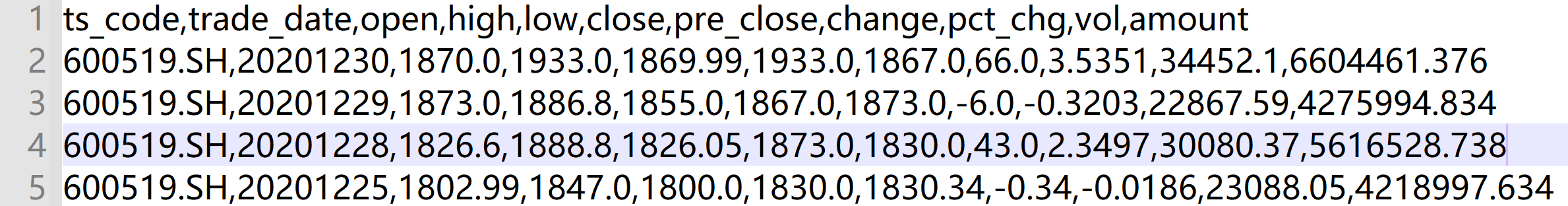
对股票数据进行外排序、创建索引、统计分析、价格预测与可视化等处理。大作业的实现，必须采用面向对象技术，体现面向对象的三大特性。

**2.数据集描述**

数据由老师提供，数据文件名为“China\_Stock.csv”。该文件包含A股股票市场之前的Daily股票数据。数据文件的内容和格式，简述如下：

数据文件由表头（首行）和具体一行行的股票数据构成。

第一行为表头，ts\_code表示股票代码，trade\_date表示交易日期，open表示开盘价，high表示最高价，low表示最低价，close表示收盘价，pre\_close表示临近收盘价，change表示涨跌幅，pct\_chg表示涨跌百分比，vol表示交易量，amount表示交易金额。



其余各行为实际数据，每一行的各个数据项(字段)用“,”隔开。

日期的格式为“YYYYMMDD”，比如20181228表示2018年12月28日。节假日或者周末等非交易日，没有价格数据。

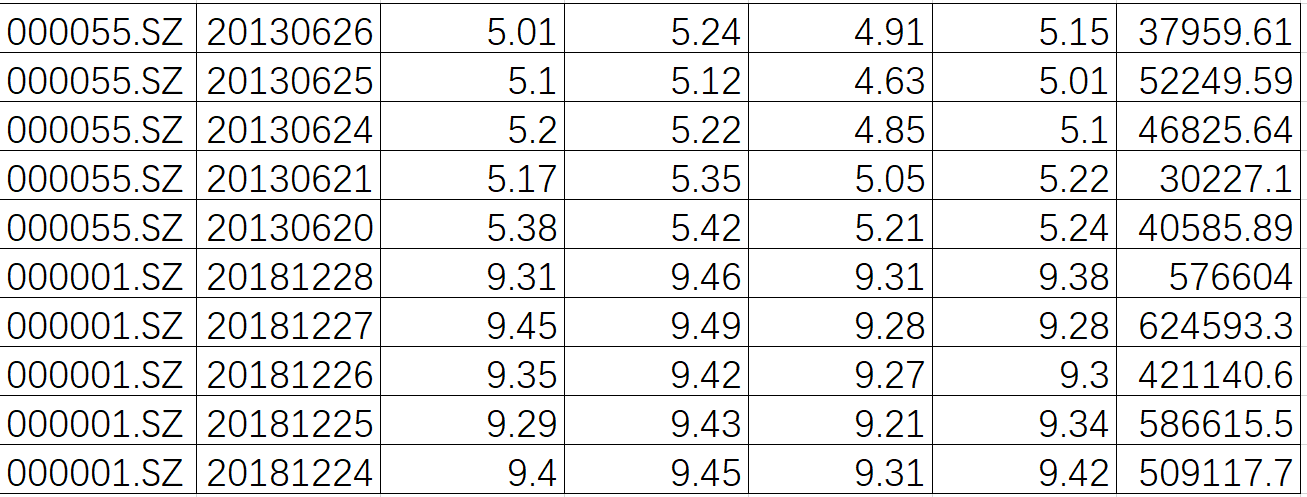
注意：当我们显示某只股票的K线图时，一般只用到股票代码、交易日期、开盘价、最高价、最低价、收盘价、以及交易量。

**3. 任务描述**

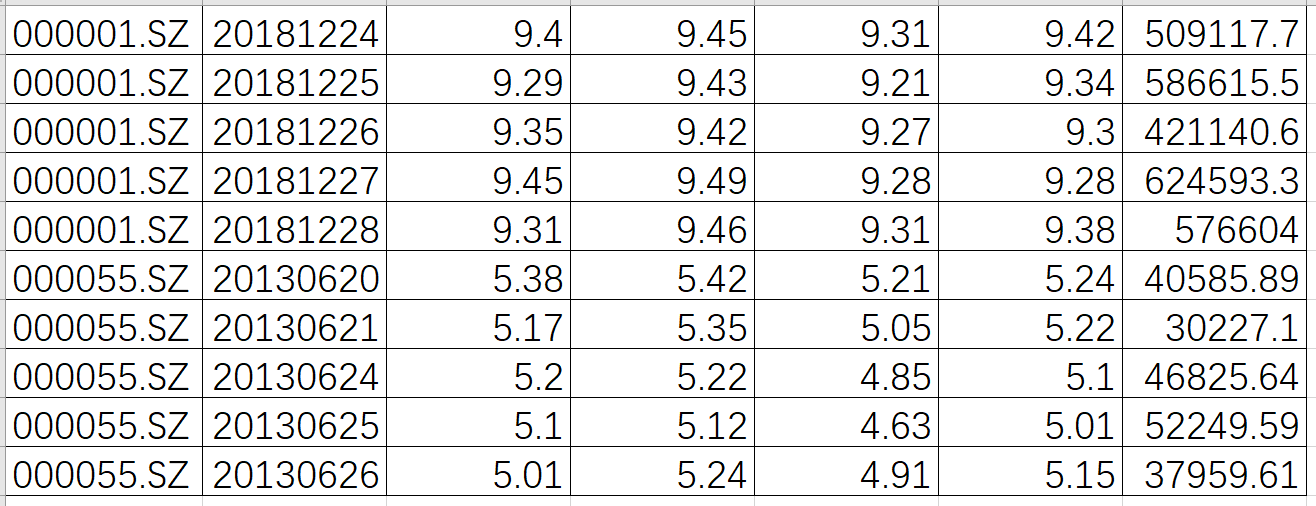
**3.1 对股票数据进行排序 (必须使用外排序，用于排序的内存大小限制为32MB，此限制的检查，需要通过资源管理器展现。)（30分）**

对原文件“input1000.csv” 进行外排序，排序标准是首先按照股票代码的字典序进行排序，如果股票代码相同则按照日期从小到大进行排序，输出文件取名output.txt。

比如源文件input1000.csv如下(注意在文件里，字段间用逗号隔开)。



那么经过排序后，结果output.txt如下(注意在文件里，字段间用逗号隔开)。



即全部价格数据已经按照股票代码的字典序、以及交易日期从小到大排序了。

这里，每行数据只给出了股票代码、交易日期、开盘价、最高价、最低价、收盘价、交易量等；实际排序时，数据格式以“2. 数据集描述”一节的描述为准。

**3.2 创建索引，加快数据访问（20分）**

我们经常要访问某只股票（比如000055.SZ）某个月份（比如201306）的数据，现在1000只股票的若干年的各个月份的数据都放在一个文件output.txt中。如果没有相关技术手段，我们访问某只股票某个月份的数据，只能从头开始扫描该文件。

此处，要求大家创建索引，加快数据访问。针对已排序的output.txt文件，根据每只股票、每个月份的第一条价格数据在output.txt文件里的偏移量，建立索引项，生成文件index.txt。注意，以二进制文件方式打开output.txt，即可正确提取和设置文件的偏移量。

例如，对于上述output.txt(实际数据比这里展示的更多，但是不影响原理解释)，由于其数据里包含000001.SZ和000055.SZ两只股票各1个月的数据，于是index.txt文件的内容如下。

|  |
| --- |
| 000001.SZ,201812,{在output.txt文件里000001.SZ股票201812月第一条数据的文件偏移量}  000055.SZ,201306,{在output.txt文件里000055.SZ股票201306月第一条数据的文件偏移量} |

由于output.txt是经过排序的，index.txt也是有序的。通过索引文件，可以快速定位和读取某只某月股票数据。

**3.3对股票价格进行可视化（20分）**

开发QT程序，由用户输入股票代码（比如000055.SZ）和月份（比如201306），采用3.2所述索引方案的帮助，快速定位某只股票、某个月的数据，在QT界面显示该月的价格数据的K线图。

借助index.txt文件的帮助，我们可以定位某只股票、某个月第一条数据在output.txt文件中的文件偏移量。打开output.txt文件，定位到相应偏移量，开始一行、一行地读取数据，可以从output.txt文件提取该股票该月份的数据。

读完一个月的数据后可视化出来。参考QTCustomPlot程序，编写可视化代码。

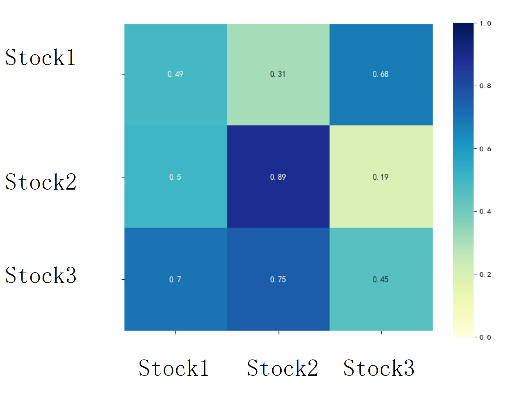
备注：index.txt的内容可以完全载入内存，在内存建立*<股票代码、月份、文件偏移量>*的对照表，以便快速查找。

**3.4 相关系数计算与热力图展示（20分）**

用户输入某年某月（比如201306），然后从列表框里选择、或者手工输入10只股票的股票代码。

从output.txt文件提取这10只股票某年某月的数据，计算每只股票的隔天收益率。隔天收益率的计算公式为：隔天收益率=（今天价格-昨天价格）/昨天价格，第一天没有隔天收益率。这时候，我们得到每只股票的隔天收益率序列。

请计算10只股票的任意两只股票之间的皮尔森相关系数。使用QT实现该功能，支持用户输入必要信息，以及经过计算以后，显示相关系数矩阵，并且显示相关系数热力图（如下图所示）。



3.5 **预测价格变动并展示（10分）**

指定某只股票代码（比如000055.SZ）某个月份（比如201306），请自行选用某种预测方法，预测该股票下一个月的价格（注意这里仅仅使用收盘价）。把预测价格和实际价格作对比，在QT界面上显示实际的价格和预测的价格（这里可以使用多线图），并且计算RMSE均方误差指标。

比如，指定股票600519.SH，指定2021年01月，请预测该股票2021年02月份的价格为，与600519.SH的2021年02月份的实际价格y作对比，计算均方误差指标，其中RMSE=

对于采用的预测办法不做限制，可以是随机预测、根据前几天的涨跌天数和涨跌幅进行判断、进行线性回归预测等。

3.6 **附加功能: 真实股票市场价格预测PK（附加分）**

由教师指定对特定股票，利用预测算法和相关数据，预测提交日后此股票K个交易日的股票价格（收盘价），形成预测输出文件。课程组将计算真实价格与预测价格的均方误差指标，并据此对预测结果进行排序。预测准确的前N名同学，将会酌情加分。

此功能的体现，将会通过网络系统支持，形成排名展示。

**4.提交办法**

(1)试验报告(word格式)

试验报告内容包括前言、各个任务实现的描述、总结等。

针对每个任务，描述其实现策略（包括类的设计、主要的数据结构和算法）、输入界面、输出界面等。界面要给出必要说明。在报告中，给出面向对象技术的体现，详细说明如何采用封装、继承和多态，实现了相关功能。

(2)源代码

请拷贝整个项目文件夹。源代码要有开发过程的体现。所有文件都是同一时间的工程，将视为拷贝抄袭。

(3)视频

对大作业完成情况的介绍和演示，最长5分钟，以便限制视频文件大小。视频展示中，要有外排时内存空间使用情况的体现（可以录制内存管理器运行期的内存占用情况）。 视频中要有开发过程的语音解说。

把(1)、(2)和(3)打包压缩，命名为“学号姓名《程序设计2》大作业.zip”提交。

**注意：** 此大作业软件开发的功能点评分分值中，包含了系统完整性、可扩展性、算法效率、面向对象设计、可视化效果的综合考量，各功能点的分值存在相互关联的情况。实现大作业功能时，设计和开发需要有系统化的考虑。