西安电子科技大学

Java 程序设计 课程实验报告

实验名称 文本文件字符统计程序

计算机科学与技术 学院 2103051 班

姓名 张平 学号 21030540006

同作者_____

实验日期 2022 年 05 月 22 日

成 绩

指导教师评语： 指导教师： _____年____月____日
实验报告内容基本要求及参考格式 一、实验目的 二、实验内容 三、实验过程 四、实验结果分析 五、实验小结（实验过程感受和建议）

一、实验目的

1. 熟悉 File 对象的操作，文件信息的获取与测试；

2. 掌握典型的流式输入输出（文件流、缓存流、数据流、标准输入输出流），典型的流接口的使用。

3. 了解 java.util.Scanner 类以及输入输出的重定向方法。

二、实验内容

1．编写一个程序，程序实现对用户指定的文本文件中的英文字符和字符串的个数进行统计的功能，并将结果根据用户选择输出至结果文件或屏幕。

1) 构建统计类，该类实现对 I/O 的操纵；实现对文本文件中英文字符、字符串的统计；实现对统计结果的输出。

2) 构建测试类，该类实现与用户的交互，向用户提示操作信息，并接收用户的操作请求。

程序应具有良好的人机交互性能，即：程序应向用户提示功能说明，并可根据用户的功能选择，执行对应的功能，并给出带详细描述信息的最终执行结果。

三、实验过程

1. 实验环境

操作系统：Windows 11

集成开发环境：Eclipse IDE for Enterprise Java and Web Developers (includes Incubating components) 2022-03 (4.23.0)

2. 题目分析

本实验编写一个文本文件字符统计程序，还要求输出统计结果。为此，这里特别限定：**统计英文字母**（不包含英文标点符号、空格、数字等）**和中文汉字**（不包含中文标点符号、空格、数字等）的个数。至于统计字符串什么的，由于字符串的判断标准太不清晰（空格还是.或。隔开？），这里我不决定实现。

3. 代码说明

这里简要说明所写的代码。测试类 CounterTest 如下所示，在 main 方法中提示用户输入文本文件的完整路径，调用计数器类时还使用 try-catch 语句处理可能的异常。使用计数器类时，通过调用 setFile 方法来传入目标文件，调用 englishNum 方法返回英文字符个数，调用 chineseNum 方法返回中文字符个数：

```java
import java.util.Scanner;
import java.io.*;

public class CounterTest {
    public static void main(String[] args) {
        System.out.print("本程序统计指定文本文件中的中英文字符个数，");
        System.out.println("请输入指定文本文件的完整路径：");
        Scanner sc = new Scanner(System.in);
        String filePath = sc.nextLine();

        try {
            // 静态调用文本文件字符计数器类,统计字符个数
            TextFileCharacterCounter.setFile(new File(filePath)); // 设置字符计数的目标文件
            System.out.println("英文字符个数为(不含标点符号): " + TextFileCharacterCounter.englishNum());
            System.out.println("中文字符个数为(不含标点符号): " + TextFileCharacterCounter.chineseNum());
        } catch (Exception e) {
            System.out.println("请输入可读取文本文件的正确路径！");
            e.printStackTrace();
        }
    }
}
```

由于输入是文本文件，文本文件字符计数器类中使用字符输入流 FileReader，为了使读取字符效率更高，代码中还将字符输入流包装为带缓冲的字符输入流 BufferedReader。为了自动关闭输入流，代码中还使用了 try with resource 语法。

```java
private static File targetFile; // 对目标文件的字符进行统计
private static int enNum; // 英文字符
private static int zhNum; // 中文字符

public static void setFile(File f) {
    targetFile = f;
    enNum = zhNum = 0;
    try ( // 使用try with resource语法,在结束时自动关闭流
        FileReader rd = new FileReader(targetFile); // 使用字符输入流来统计字符
        BufferedReader bufRd = new BufferedReader(rd); // 使用缓冲,减少大文件读写的耗时
    ) {
        String curLine; // 每次从流中读取一行
        while ((curLine = bufRd.readLine()) != null) {
            enNum += countEnglishCharacterInLine(curLine);
            zhNum += countChineseCharacterInLine(curLine);
        }
    } catch (Exception e) {
        e.printStackTrace();
    }
}
```

将字符串按行从输入流中读出后，对每行的字符串进行字符统计。此处采用 Character 类的 isUpperCase 方法和 isLowerCase 方法，统计英文字母个数；采用 Character.UnicodeScript.of 方法，判断是否为汉字。此处不使用正则表达式”[\u4e00-\u9fa5]”，是因为现在这一表达式无法匹配到所有汉字；不使用正则表达式”[\u4e00-\u9fff]”，是因为这实际上是 Unicode 的中日韩认同表意文字区，且硬编码不符合个人习惯。

```java
private static int countEnglishCharacterInLine(String line) {
    int ec = 0;
    for (int i = 0; i < line.length(); ++i) {
        char c = line.charAt(i);
        if (Character.isUpperCase(c) || Character.isLowerCase(c)) { // 大写字母或小写字母
            ++ec;
        }
    }
    return ec;
}

private static int countChineseCharacterInLine(String line) {
    int cc = 0;
    for (int i = 0; i < line.length(); ++i) {
        char c = line.charAt(i); // 检查分配给字符的Unicode脚本的枚举常量是否为HAN(即汉字)
        if (Character.UnicodeScript.of(c) == Character.UnicodeScript.HAN) {
            ++cc;
        }
    }
    return cc;
}
```

四、实验结果分析

完成程序编写后，在 Eclipse 中点击 Run，执行结果如下。傲慢与偏见这一文本文件也放在代码目录中：

```
本程序统计指定文本文件中的中英文字符个数，请输入指定文本文件的完整路径：
C:\Users\dell\Desktop\Java实验\实验7_1\src\傲慢与偏见(中英文对照).txt
英文字符个数为(不含标点符号)：537131
中文字符个数为(不含标点符号)：192277
```

为了验证结果是否正确，在网上搜索了一个统计方式较为接近的程序，全文复制进去后，统计结果如下。程序网址见 https://www.a-site.cn/tool/zi/。可见汉字个数是正确的。

后来换了一个网站 https://www.hao353.com/zishutongji，发现字母和汉字个数都是正确的：

虽然达西再三不肯让韦翰到彭伯里来，但是看在伊丽莎白面上，他依旧帮助他找职业。丽迪雅每当丈夫到伦敦去或是到巴思去寻欢作乐的时候，也不时到他们那儿去作客；到于彬格莱家里，他们夫妇老是一住下来就不想走，弄得连彬格莱那样性格温和的人，也觉得不高兴，甚至说，要暗示他们走。
达西结婚的时候，彬格莱小姐万分伤心，可是她又要在彭伯里保持作客的权利，因此便把多少怨气都打消了；她比从前更喜爱乔治安娜，对达西好象依旧一往情深，又把以前对伊丽莎白失礼的地方加以弥补。
乔治安娜现在长住在彭伯里了；姑嫂之间正如达西先生所料到的那么情投意合，互尊互爱，甚至融洽得完全合乎她们自己的理想。乔治安娜非常推崇伊丽莎白，不过，开头看到嫂嫂跟哥哥谈起话来，那么活泼调皮，她不禁大为惊讶，几乎有些担心，因为她一向尊敬哥哥，几乎尊敬得超过了手足的情份，想不到现在他竟成为公开打趣的对象。她以前无论如何也弄不懂的事，现在才恍然大悟了。经过伊丽莎白的陶治，她开始懂得，妻子可以对丈夫放纵，做哥哥的却不能允许一个比自己小十岁的妹妹调皮。
咖苔琳夫人对她姨侄这门婚姻极其气愤。姨侄写信给她报喜，她竟毫不留情，直言无讳，写了封回信把他大骂一顿，对伊丽莎白尤其骂得厉害，于是双方有一个短时期断绝过往来。后来伊丽莎白说服了达西，达西才不再计较这次无礼的事，上门去求和；姨母稍许拒绝了一下便不计旧怨了，这可能是因为疼爱姨侄，也可能是因为她有好奇心，要看看侄媳妇怎样做人。尽管彭伯里因为添了这样一位主妇，而且主妇在城里的那两位舅父母都到这儿来过，因此便门户受到了玷污，但她老人家还是屈尊到彭伯里来拜访。
新夫妇跟嘉丁纳夫妇一直保持着极其深厚的交情。达西和伊丽莎白都衷心喜爱他们，又一直感激他们，原来多亏他们把伊丽莎白带到德比郡来，才成全了新夫妇这一段姻缘。

总字数	899131(个)
总行数	3928(行)
中文字数	192277(个)
中文标点	22134(个)
字母个数	537131(个)
单词个数	123257(个)
英文标点	24342(个)
数字个数	122(个)
数字组	64(个)

小规模的测试也不可少：

程序运行结果如下：

```
本程序统计指定文本文件中的中英文字符个数，请输入指定文本文件的完整路径：
C:\Users\dell\Desktop\Java实验\实验7_1\src\testfile.txt
英文字符个数为(不含标点符号)：6
中文字符个数为(不含标点符号)：18
```

测试网站的结果如下：

总字数	36(个)
总行数	4(行)
中文字数	18(个)
中文标点	3(个)
字母个数	6(个)
单词个数	2(个)
英文标点	3(个)
数字个数	6(个)
数字组	1(个)

由上可见，实验结果正确。

五、实验小结

本次实验实现了对文本文件的读写、对字符信息的统计，使我在一定程度上熟悉了 Java 的 IO 体系和相关使用规范。

在代码中的一个关键点是使用了 try with resource 语法来自动关闭输入输出流，虽然在本程序中没什么用处（因为 GC 会自动回收垃圾），但是对某些使用缓冲区的输出流来说，如果不正确关闭流，可能导致文件内容无法及时写入。当然，这一语法适用的资源对象不止输入输出流，可以说是极大方便了代码的编写。

最后在测试中，偶然发现了一个问题：如果我在 Eclipse 中创建文本文件、输入文本内容，字符统计时结果是正确的；但如果使用记事本来创建文本文件、输入文本内容，字符统计时结果却是错误的。Debug 发现是编码问题，记事本创建的文本文件在 Eclipse 中打开是乱码（它使用的是 UFT-8 编码），而我的 Eclipse 使用的是 GBK 编码。虽然可以在代码中使用语句 new InputStreamReader(new FileInputStream(targetFile), “gbk”)来指定数据流的编码方式，但这不能一劳永逸解决所有问题。暂时我还没有找到解决办法。