Project 3 文本分析工具

【问题描述】

文本编辑器中，通常会提供针对文本内容的分析工具。例如统计文档的长度、字数、词频等，又或者代码编辑器中统计函数、注释、空行等信息。而实现这些分析的功能，往往需要灵活实现字符串的操作。

【实现要求】

假设文本信息存储于一个文本文件中，待统计的词汇集合要求一次输入完毕，即统计工作必须在程序运行一次之后，就全部完成。程序的基本输出结果，要求是每个词的出现次数，以及出现位置所在的行号。

具体输出格式可以自行设计，可采用多种数据展示方式。

【测试数据】

测试数据可以尝试多种文本格式的文档，例如：

(1) 以 txt 格式存储的英文短文，统计词频值 Top-K 的单词。

(2) C/C++语言编写的源代码，以保留字符集作为待统计的词汇集。

注：测试数据仅是简单举例，更多的文本类型可以参考 VS Code，UltraEdit 等编辑器。

【实现提示】

(1) 建议从简单实例开始进行实现，然后逐步增加对特殊字符、特殊格式等的功能支持。文本的内容组织方式也可以根据实际情况，自行定义。

(2) 约定文本中的词汇一律不跨行，每读入一行，就统计每个词在该行中的出现次数。而出现位置所在行的行号，则可以用线性表进行存储。若某行中出现了不止一次，则不必存多个相同的行号。

(3) 模式匹配算法不能仅实现最简单的朴素匹配，要求至少实现 KMP 算法，或其他模式匹配算法。

(4) 在统计信息的输出方式上，既可以探索字符为基础的输出方式，也可以图形化的数据可视化输出方式。鼓励大家根据实际的时间精力，进行分析功能上的扩展。

【检查时间和要求】

2022 年秋季学期第 8 周实验课（10 月 20 日）。

评分要求：功能实现(50%)，程序输入界面(30%)，代码规范(20%)。

将代码、可执行文件和实验报告，打包为 zip 文件，发送到邮箱：homework-szh@qq.com

ZIP 和邮件命名格式：学号+姓名+Project3 (示例：21332001+张三+Project3)

附实验报告内容模板参考：

Project 3 实验报告

学号　姓名

1、程序功能简要说明。

2、程序运行截图，包括计算功能演示、部分实际运行结果展示、命令行或交互式界面效果等。

3、部分关键代码及其说明。

4、程序运行方式简要说明。