2017 年暑期夏令营上机测试题目

题目描述：

请你开发一个简单的英文全文搜索系统，可以满足以下功能：

1. 针对用户指定的文本文件进行全文搜索。可以在该文件中查找单个单词，列出包含搜索项的总行数，并逐条列出所有包含搜索项的行。

2. 该系统必须支持某种形式的布尔查询语言，在该系统中将支持：

   1) && 在一行中这两个单词不仅存在而且相邻

   2) || 在一行中这两个单词至少有一个存在

   3) ！在一行中该单词不存在

   4) ( ) 把子查询组合起来的方式

   例如，下面的表达式：

   This && (evening || night)表示搜索所有包含 This evening 或 This night 的行。

3. 在进行搜索时，应该做如下处理：

   1) 不区分大小写, 即大写字母和小写字母当作相同的字母处理, 例如, Cat、CAT 和 cat 都应该被当作同一个词处理。

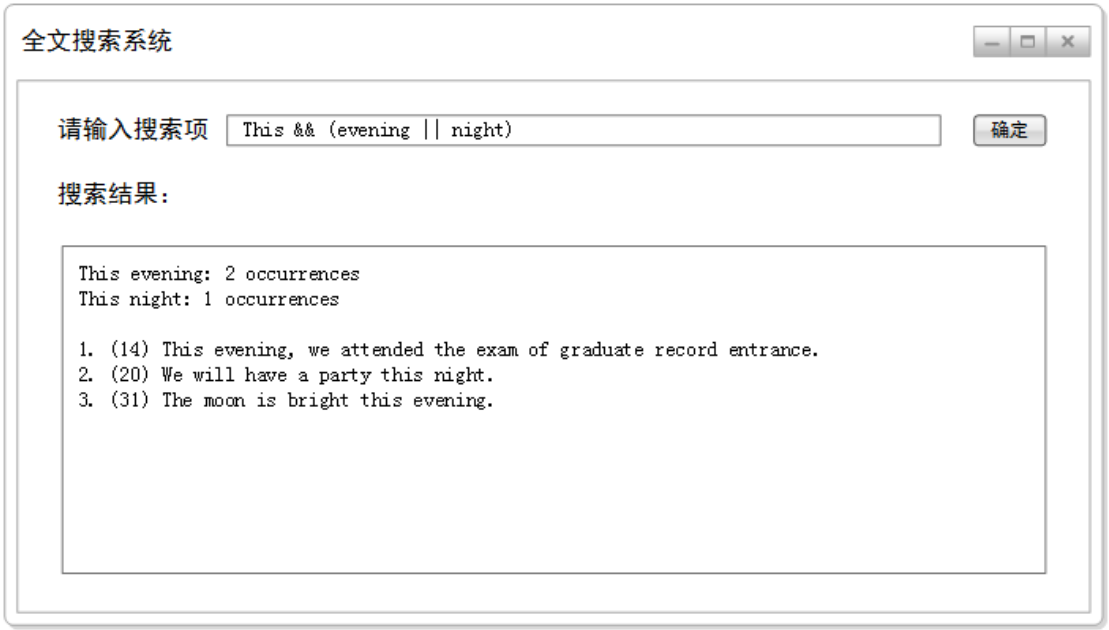
   2) 名词的单复数应该当作同一个词处理，例如，cat 和 cats 应该是同一个词, tomato 和 tomatoes 应该当作同一个词处理。不规则变化可以不处理，例如 datum 和 data。

   3) 动词的第三人称单数、现在分词和规则的过去式与过去分词都应该当作同一个词处理，例如，display、displays、displaying、displayed 应该当作同一个词处理。

4. 应该整个搜索系统提供输入输出图形化界面。如果无法按照要求完成图形化界面的功能，可以用命令行界面替代，但无法获得图形化界面的分数。详见“考核要求”。具体要求如下：

   选择文件操作的 UI 类似下图所示：

   文件搜索输入/输出操作的 UI 类似下图所示：



   在搜索结果中，前面两行分别显示了 This evening 出现了 2 次，This night 出

现了 1 次。在后面显示的每一行中，第一个数字表示找到的结果序号，第二个在圆括号中的数字表示文档中的行号。

考核要求：

1. （36 分）无论使用图形化界面还是命令行界面，能够实现题目第 1 点描述中所列举的基本功能，使全文搜索系统能够正确运行和使用，具体要求为：
   a) 能够读入指定的文本文件（6 分）：用户可以选择任意的包含英文的文本文件。
   b) 能够将词频统计的结果输出为文件（10 分）：将英文文本的词频统计结果输出为文本文件。
   c) 能够对单个单词进行全文搜索，显示出现的次数（10 分）。
   d) 能够对单个单词进行全文搜索，显示所有包含所搜索单词的英文行号和内容（10 分）。
2. （34 分）无论使用图形化界面还是命令行界面，能够实现题目第 2 点描述中所列举的基本功能，使全文搜索系统能够正确运行和使用，具体要求为:
   a) 支持单个&&操作（12 分）
   b) 支持单个 || 和 ！操作各 6 分（12 分）
   c) 支持这三种布尔操作符组合操作（10 分）。
3. （15 分）无论使用图形化界面还是命令行界面，能够实现题目第 3 点描述中所列举的基本功能，使全文搜索系统能够正确运行和使用，具体要求为:
   a) 不区分大小写（5 分）。
   b) 名词的单复数当作同一个词处理（5 分）。
   c) 动词的第三人称单数、现在分词和规则的过去式与过去分词都当作同一

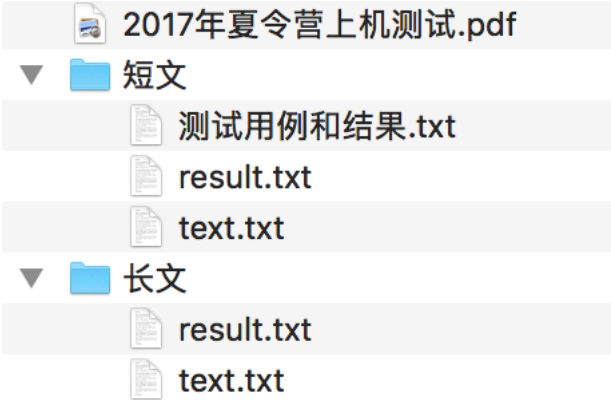
个词处理（5 分）。

4. （15 分）提供图形化界面完成第 1、2、3 点中的各项输入输出功能要求。

如果提供了图形化界面，则无需再提供命令行界面。

题目资料：

在你拿到的 U 盘中，根目录下有一个 pdf 文件“2017 年夏令营上机测试.pdf”，就是你目前正在阅读的文件。此外，还有两个目录，如下图所示：



1. 短文：这其中包含了三个文件

   a) text.txt，这是测试用例文件，包含了一篇英文短文

   b) result.txt，这是对 text.txt 文件进行词频统计后产生的结果

   c) 测试用例与结果.txt，包含了针对 text.txt 的一些布尔查询测试用例与结果

   上述这些文件可以供你作为功能调试时的依据，也是进行验收时评判你的程序功能的依据。

2. 长文：这其中包含了两个文件

   a) text.txt，这是测试用例文件，包含了一篇英文长文

   b) result.txt，这是对 text.txt 文件进行词频统计后产生的结果

上述这些文件可以供你作为性能调试时的依据，也是进行验收时评判你的程序性能的依据。我们以你的程序做词频统计并产生输出结果文件的时间作为程序性能测试的依据。

提交要求：

请将你的程序，包括工程中的源码、编译出来的可执行文件和产生的英文长文与短文的词频统计结果文件压缩成 .rar 或 .zip 文件。文件名为“所在大学名_姓名”，例如“上海交通大学_张三.rar”，将该文件存储到 U 盘的根目录中。

如果你的程序需要特别说明使用方式，请编写一个使用文档，存储在 U 盘的根目录中。

验收完成后，请将你的 U 盘交给助教。