WordCount 需求说明

1 WordCount 功能需求说明

WordCount 程序的需求可以概括为：根据输入参数对记事本（txt）文件进行内容读取，并执行相应的功能，并要求能够快速地完成指定功能。

可执行程序命名为：wc.exe，该程序处理用户需求的模式为：

```
wc.exe  -w [input_file_name] // 统计输入文件的单词总数，结果输出到 result.txt 中，与 wc.exe 同目录
wc.exe  -s [input_file_name] // 统计输入文件的单词词频，并从大到小排序，结果输出到 result.txt 中，与
                             //wc.exe 同目录
wc.exe  -o [output_file_name] // 将结果输出到指定输出文件 output_file_name.txt 中，与 wc.exe 同目录
```

具体细则说明如下。

2 被处理的文件

第一，仅处理 txt 文件，不处理其他类型的文件。

第二，一次仅处理一个文件，不同时处理多个文件。

3 文件内容

文件中仅包含单词（a-z）、常见字符、数字(0-9)。不包含其他内容。

4 对单词的规定

满足如下两个条件中的任意一个条件，则视为单词，

第一，由连续的若干个英文字母组成的字符串，例如，software，

第二，以英文字母开头和结尾，并用 1 个或多个连续的连字符（即短横线）所连接的若干个英文单词，也视为 1 个单词，例如，content-based，视为 1 个单词，Gravity-center-based，视为 1 个单词，而-based，或 content-，则将分别视 based 和 content 为单词，短横线不包含在内。

注意，单词不区分大小写，不考虑英文以外的其他语言，且仅考虑半角。

不做单词有效性校验，例如，thes 将视为一个单词。

5 对常见字符的规定

文件中包含的常见字符如下表所示，该表以外的特殊字符不考虑。

字符	~	`	!	#	%	^	&	*	_	…
含义	波浪号	重音符号	感叹号	井号	百分号	指数符号	与符号	星号	下划线	省略号
字符	( )	[ ]	+	=	-	:	;	“	‘	|
含义	左右小括号	左右方括号	加号	等于号	短横线	冒号	分号	双引号	单引号	竖线
字符	<	>	,	.	/	?				
含义	小于号	大于号	逗号	点	反斜杠	问号	空格	换行符	水平制表符	

6 对词频的定义

词频即某单词在文档中出现的次数。

7 对输入命令行参数的规定

输入命令中仅包含文件名，不包含扩展名。
-s，-w 参数不能同时出现。
-s 或-w 必须与文件名同时使用，且输入文件必须紧跟在-s 或-w 参数后面，不允许单独使用-s 或-w 参数。
-o 必须与文件名同时使用，且输出文件必须紧跟在-o 参数后面，不允许单独使用-o 参数。
-o 参数必须与-s 或-w 参数同时使用，形如：wc.exe –w inputFile –o outputFile
-s,-w, -o 命令参数不区分大小写。

8 对输出的规定

8.1 单词总数统计

对于-w 参数，仅输出文件中的单词总数，形式如下：

```
Number of words: 600
```

注意：输出文件末尾多余的换行符应去除。

8.2 词频统计

对于-s 参数，仅输出单词词频从高到低排序的前 100 个（从 1 到 100），每行分别给出一个单词及其词频，单词全部按小写形式给出，单词和词频之间空一格。对于单词词频相同的情况，按照单词所包含的每个字母从 a 到 z 的次序依次排列。输出文件末尾多余的换行符应去除。
形式如下：

```
this 200
i 180
ij 180
the 180
```