作业1：个人项目（初始开发）

目 录

2021-09-09 20:56:37 来源：本站编辑 阅读量：132

> 我们要构建一些基本的测试用例来保证程序的基本功能不会在不断的扩展中出问题输出某个英文文本文件中 26 字母出现的频率字母频率 = 这个字母出现的次数 / （所有A-Z一个用于统计文本文件中的英语单词出

课程目标

1. 课程目标3：使用Git 等协作管理工具进行团队开发的能力

作业目标

1. 熟悉 IDE 开发环境
2. 磨炼个人编程技能

作业要求

1. 自己独立完成，杜绝抄袭；
2. 提交到 gitee 地址
3. 按照实验报告，编写作业报告；
4. 以学号-姓名-作业二.docx 命名，提交到雨课堂 **软件工程作业一**

作业内容：

1. 基本源代码控制的用法， 逐步扩展的程序设计，对字符，字符串的处理，英语分词，排序，程序的测试，回归测试，C/C++/C#，Java 等基本语言的运用和 debug。 考虑到同学的基础参差不齐，这个作业提供了多种要求，请按先易后难的次序实现。
2. 每一步都至少要签入源代码控制 （github 或gitee）一次，同时把回归测试的测试用例也写好签入到适当的目录中。

用户需求：

英语的26 个字母的频率在一本小说中是如何分布的？某类型文章中常出现的单词是什么？某作家最常用的词汇是什么？《哈利波特》 中最常用的短语是什么，等等。 我们就写一些程序来解决这个问题，满足一下我们的好奇心。

假设我们的命令行程序叫 WF.exe (WF: Word Frequence)

第0步：

输出某个英文文本文件中 26 字母出现的频率，由高到低排列，并显示字母出现的百分比，精确到小数点后面两位。

wf.exe -c

字母频率 = 这个字母出现的次数 / （所有A-Z，a-z字母出现的总数）

如果两个字母出现的频率一样，那么就按照字典序排列。 如果 S 和 T 出现频率都是 10.21%， 那么，S 要排在T 的前面。

这个程序容易写吧？ 如果要处理一本大部头小说 （例如 Gone With The Wind), 你的程序效率如何？有没有什么可以优化的地方？

第一步：

输出单个文件中的前 N 个最常出现的英语单词。

作用：一个用于统计文本文件中的英语单词出现频率的控制台程序

单词：以英文字母开头，由英文字母和字母数字符号组成的字符串视为一个单词。单词以分隔符分割且不区分大小写。在输出时，所有单词都用小写字符表示。

英文字母：A-Z，a-z

字母数字符号：A-Z，a-z，0-9

分割符：空格,非字母数字符号 例：good123是一个单词，123good不是一个单词。good，Good和GOOD是同一个单词

功能列表：

功能1： wf.exe -f

输出文件中所有不重复的单词，按照出现次数由多到少排列，出现次数同样多的，以字典序排列。

功能2： wf.exe -d 指定文件目录，对目录下每一个文件执行 wf.exe -f 的操作。

wf.exe -d -s 同上， 但是会递归遍历目录下的所有子目录。

功能3： 支持 -n 参数，输出出现次数最多的前 n 个单词， 例如， -n 10 就是输出最常出现单词的前 10 名。 当没有指明数量的时候，我们默认列出所有单词的频率。

现在我们这个程序已经有一点复杂度了， 我们要构建一些基本的测试用例来保证程序的基本功能不会在不断的扩展中出问题。 请看 《构建之法》【回归测试】的内容，构建一些测试来保证基本功能的正确性。

第二步:

支持 stop words

我们从第一步的结果看出，在一本小说里， 频率出现最高的单词一般都是 “a”， “it”， “the”， “and”， “this”, 这些词， 我们并不感兴趣. 我们可以做一个 stop word 文件 （停词表）， 在统计词汇的时候，跳过这些词。 我们把这个文件叫 “stopwords.txt” file.

功能 4： 支持新的命令行参数, 例如： wf.exe -x -f

在这一步我们要增加什么回归测试呢？

第三步:

我们想看看常用的短语是什么， 怎么办呢？

先定义短语：” 两个或多个英语单词， 它们之间只有空格分隔”. 请看下面的例子：

hello, world 才这不是一个短语

功能 5： 支持新的命令行参数 -p

参数 说明要输出多少个词的短语，并按照出现频率排列。同一频率的词组， 按照字典序来排列。

在这一步我们要增加什么回归测试呢?

第四步：

把动词形态都统一之后再计数。

我们想找到常用的单词和短语，但是发现英语动词经常有时态和语态的变化，导致同一个词，同一个短语却被认为是不同的。 怎么解决这个问题呢?

假设我们有这样一个文本文件，这个文件的每一行都是这样构成：

动词原型 动词变形1 动词变形2...

词之间用空格分开。

e.g. 动词 TAKE 有下面的各种变形

take takes took taken taking

我们希望在实现上面的各种功能的时候，有一个选项， 就是把动词的各种变形都归为它的原型来统计。

功能 6： 支持动词形态的归一化，参数为 -v

wf.exe -v 其中 是纪录动词形态的文本文件。

实现这些功能， 分析程序的效能 。

上一篇：作业4：面向对象的... 下一篇：作业5：个人项目（...

评论

已有0条评论

0/150

提交

热门评论