

SoftwareTeacher

现代软件工程 教学博客 邹欣

首页

新随笔

联系

订阅

管理

现代软件工程 作业 文本文件中英语单词的频率

这是《构建之法》现代软件工程课的个人项目作业

1. 个人项目 Individual Project: 一个人独立完成.
2. 时间: 可以考虑在第一周就给同学们这个项目; 也可以考虑分为两部分, 个人做第一部分, 然后两人结对, 选两个人中较好的程序, 再继续开发其他功能。
3. 考核内容

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, 效能测试C/C++/C# 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。

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。

命令行参数是:

wf.exe -c <file name>

一些同学要复习一下程序如何处理命令行参数, [请看别人的经验](#)。

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? 如果要处理一大本大部头小说 (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

分隔符: 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:

公告

Flag Counter

昵称: SoftwareTeacher

园龄: 10年3个月

荣誉: 推荐博客

粉丝: 6507

关注: 176

+加关注

2011年11月						
日	一	二	三	四	五	六
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3
4	5	6	7	8	9	10

搜索

找找看

谷歌搜索

常用链接

我的随笔

我的评论

我的参与

最新评论

我的标签

我的标签

软件工程(74)

作业(20)

构建之法(15)

软件工程专业(5)

团队作业(3)

软件测试(3)

教学计划(3)

结对编程(3)

software engineering(3)

创新(3)

更多

积分与排名

积分 - 417927

排名 - 1076

- 功能1：wf.exe -f <file>
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 <directory> 指定文件目录，对目录下每一个文件执行 wf.exe -f <file> 的操作。
 - wf.exe -d -s <directory> 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 <stopwordfile> -f <file>

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 //这是一个短语

hello, world //这不是一个短语

功能 5：支持新的命令行参数 -p <number>

参数 <number> 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 <verb file> 其中 <verb file> 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 (link: [现代软件工程讲义 2 开发技术 - 效能分析](#))，

每一步都至少要签入源代码控制（github 或其他工具）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。

标签: [软件工程](#), [作业](#), [词频分析](#)

[好文要顶](#)[关注我](#)[收藏该文](#)



SoftwareTeacher
关注 - 176
粉丝 - 6507

0



0



随笔档案 (227)

- 2019年1月(1)
- 2018年11月(2)
- 2018年10月(2)
- 2018年9月(4)
- 2018年7月(1)
- 2018年4月(1)
- 2017年9月(1)
- 2017年8月(3)
- 2017年7月(1)
- 2017年5月(1)
- 2017年1月(1)
- 2016年12月(1)
- 2016年11月(2)
- 2016年10月(2)
- 2016年8月(1)
- 更多

BUAA 2012

Shine
代码厨房
FightingSnail
100Years
TeamSH*T
76er
www-Buaa
CSE
superbro
MagicCode

BUAA 2014

Echo
SixSix
P#
C705
罗杰
NewBe
Sevens
DX
hots

MS创新学院

码连锁
码共和
枪嘸玫
皇家码衣
司马扣
大坏狼
2015 - 团队Azure
2015 - 团队Dae-De-Lus
2015 - 团队CodeHunters
2015 - 团队Altas

Tsinghua 2011

- 1 - Seven
- 2 - 霸王移山
- 3 - 锄艳
- 4 - Take it and go
- 5 - Banana

USTC

学生团队1: MicroTeam
学生团队2: USTC_MSRA_Ase
学生团队3: SE_Team
学生团队4: CodingCrazy
进修的蒋老师

 [登录后才能发表评论，立即 登录 或 注册， 访问 网站首页](#)

【推荐】阿里云Java训练营第3期-实战Spring Cloud，结营抢小米耳机

【推荐】更好的世界，更好的你-阿里巴巴2021实习生招聘专场来啦！

【推荐】阿里云Java训练营第2期-实战Spring Boot 2.5，抢智能音箱

【推荐】大型组态、工控、仿真、CAD\GIS 50万行VC++源码免费下载!

【推荐】实战应用实时计算 Flink 开发技能，4天突破，抢天猫精灵！

【推荐】注册 Amazon Web Services(AWS) 账号，成为博客园赞助者

【推荐】HarmonyOS开发者创新大赛，一起创造无限可能



AWS免费产品：

- 如何在AWS上免费构建网站
- AWS免费云存储解决方案
- 在AWS上免费构建数据库
- AWS上的免费机器学习



最新新闻：

- 李丰独家授课：华为、戴森、三顿半崛起背后的新消费底层逻辑
 - 完美日记舍命狂奔，上坡路明显更难走了
 - Keep试错狂奔，难解流量焦虑
 - 谷歌因Chrome无痕模式问题面临诉讼 或罚50亿美元
 - 京东关联公司公开“货运无人机”相关专利
- » 更多新闻...

USTC 2011

第四组 OMG!

第三组 southseven(南七)

第二组 meng-meng(萌萌)

第一组 Rosting(螺丝钉)

我的其他角色

新浪微博

豆瓣

移山之道

最新评论

1. Re:最新软件工程总结，项目模板，软工作
业下载

知乎上面的总结。

--SoftwareTeacher

2. Re:现代软件工程讲义 1 软件工程概论

我在：2021年 1月 15日 18:40:33 看过本篇博
客！

--努力变胖-HWP

3. Re:现代软件工程 结对/团队作业 - 汉字的 2
048 + 俄罗斯方块

来转转~~~老师的博客

--君君的BigHeadDaddy

4. Re:现代软件工程讲义 源代码管理

孟宁老师提供的好文章：

--SoftwareTeacher

5. Re:现代软件工程讲义 目录 《构建之法》

「学习一门新学科就好比在我的头脑中建造一
座新房子；理解它就好比熟悉我头脑中的新房
子的内部状况；而研究它的问题就好比是在布
置家具。思考它的问题就好比在房子里里外外
生活。作为学习者，我在心中创造了一个世
界...

--SoftwareTeacher

6. Re:现代软件工程讲义 目录

邹老师，你好。我已经毕业了，买了《构建之
法》的多看版本。对于我这种离校的人，如何
实践这本书呢？有相关的微信群吗？

--ttkltl

7. Re:现代软件工程系列 学生的精彩文章 (6)

我们其实还不懂互联网

链接失效了。

--yeka

8. Re:现代软件工程 第八章 【需求分析】练习
与讨论

A：是的，然后平台就可以坐着收钱。一个例
子就是淘宝。平台也有很多坑。一个反例是共
享单车。淘宝服务100个商家和10000个商家，
需要增加的成本是非常少的，就加一些服务器
就好了（暂且不讨论服务1亿个...

--SoftwareTeacher

9. Re:现代软件工程 第八章 【需求分析】练习
与讨论

讨论：如果解决用户的需求，就会有成功的
软件&服务，那么，用户在使用搜索引擎的时
候，很烦那些广告，为何没有一个只有高质量
搜索结果，而没有广告的产品呢？而且用户并
没有要求：我搜索的时候能看到一些广...

--SoftwareTeacher

10. Re:现代软件工程讲义 源代码管理

试试手气：

--fll

阅读排行榜

1. 现代软件工程讲义 目录 《构建之法》(89324)
2. 现代软件工程 课件 软件工程师能力自我评价表(23891)
3. 《构建之法》参考书和链接汇总(23523)
4. 技能的反面 - 魔方和模仿(23122)
5. 现代软件工程讲义 源代码管理(22669)
6. 现代软件工程讲义 0 教学方法(20863)
7. 软件工程讲义 0 微博上的软件工程(19666)
8. 顶级程序员的心得—Coders at Work(18436)
9. 现代软件工程课件 需求分析 如何提出靠谱的项目建议 NABCD(13902)
10. 现代软件工程讲义 2 工程师的能力评估和发展(13858)

评论排行榜

1. 2012 夏季高校微软俱乐部活动 - 开门创新(50)
2. 软件工程 敏捷的酒后问答(34)
3. 现代软件工程讲义 目录 《构建之法》(33)
4. 创新的时机 - 黄金点游戏(31)
5. 对大学 IT 专业教育的反馈(30)
6. 创新 - 王屋村的魔方们(29)
7. 现代软件工程讲义 0 教学方法(28)
8. 现代软件工程 10 绩效管理(25)
9. 感恩回馈——你评博客，我送好书(24)
10. 微软学术搜索项目 10个版本的历程(24)

推荐排行榜

1. 现代软件工程 课件 软件工程师能力自我评价表(46)
2. 现代软件工程讲义 目录 《构建之法》(39)
3. 现代软件工程讲义 0 教学方法(35)
4. 技能的反面 - 魔方和模仿(20)
5. 创新 - 王屋村的魔方们(20)