****

**JAVA进阶编程第二次实验报告**



**学 院 智能与计算学部**

**专 业 软件工程**

**年 级 2017**

**姓 名 李名扬**

**2019年 3 月 23 日**

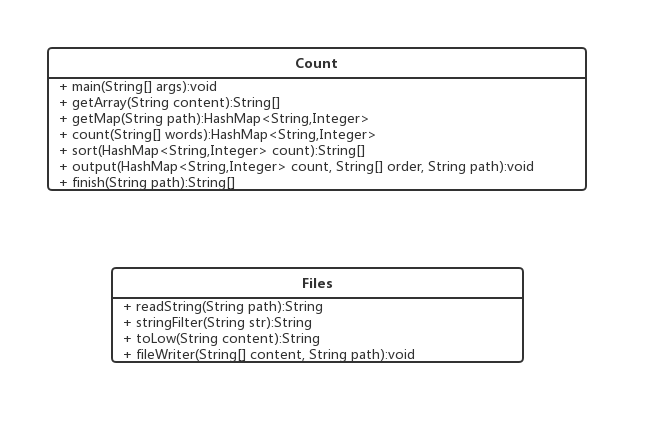
# JAVA进阶编程第二次实验报告

1. 需求分析（描述具体需求）

输入一个英文小说的txt文本，根据每个单词出现的次数进行词频统计，输出一个叫output.txt的文件，用于记录词频，

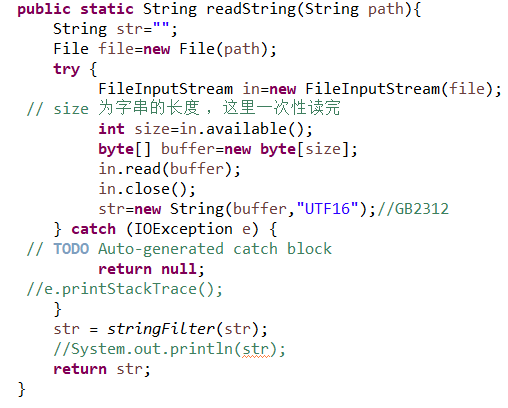
1. 概要设计（简单描述设计思路，配合UML图）

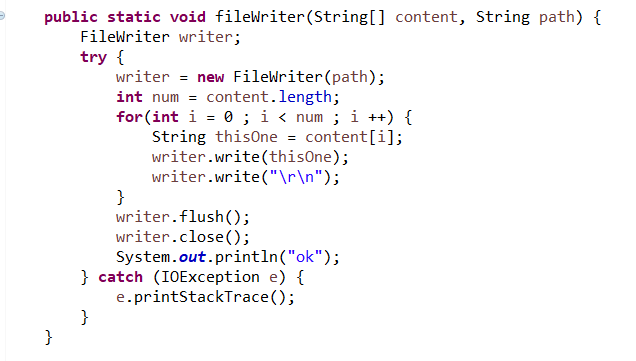
统计和文件读写分成两个类，类之下再做详细的方法划分功能。



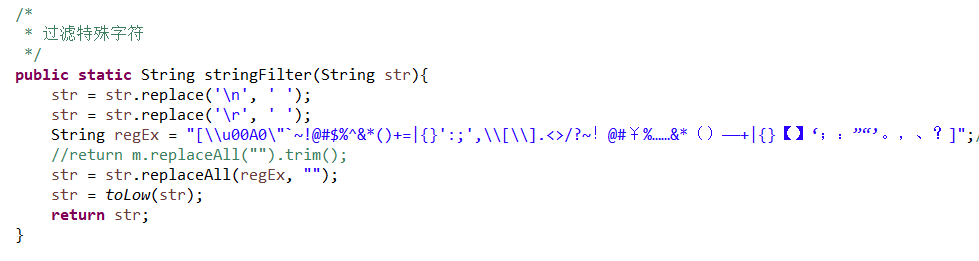
1. 详细设计（详细描述具体如何实现，附代码及说明）

读写方法分别为read String和 file Writer

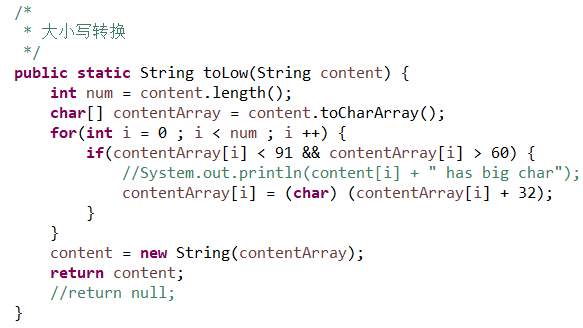




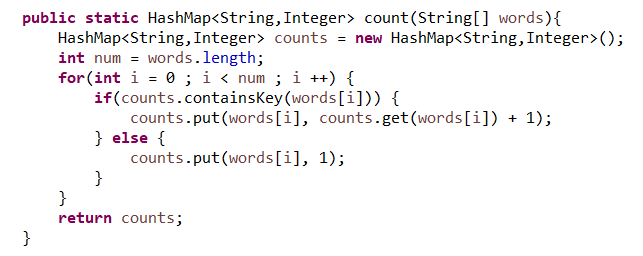
由于给出的文章内容有不少瑕疵，故应对其内容实现过滤。将多余符号清除。故实现Files中的string Filter方法



再其次，应当视只含大小写区别的单词为同一单词，故在Files的to Low方法中全部转换为小写。



对于统计，使用HashMap，在Count类中实现了count方法



排序是自己写的。

1. 调试分析（在实验过程中遇到的问题以及如何解决）

一开始编码问题使人十分困扰，因为英文版采用的编码可能是UTF16，比特长度是某些常见编码的二倍长，但又由于全是英文字母，左面的一半字节大概切开是代表空格。故写入文件后每1个字符后就有一个空格。

开始的解决方法是将原有的小说文本转换成了ANSI编码，提交前重新使用原版的文本，读取文件方法更改为UTF16编码。

另外，对文本的清洗上出现了诸多不同的方法，大家意见不同，导致的结果也不同，但考虑到本次实验的重点不在精确的统计而是文件的读写和容器的使用，因此我保留了自己的清洗方法。尽管这样，连续空格导致的空白分段，开始我以为是换行符等特殊字符，但是清洗后仍未解决，后来才想到原因，故在统计时直接跳过空字符串(也有人可能清洗时把多个空格的子串转为单空格，也可以解决)。

1. 测试结果（描述输入和输出）

有点长，输入另附文件input.txt，输出我只截取前面几个

OUTPUT：

the 1970

and 1281

a 1244

i 1023

of 1015

to 1001

in 702

he 688

was 665

it 472

that 471

at 448

his 417

you 410

with 357

her 333

she 331

had 329

on 309

me 275

for 261

as 241

him 230

said 212

1. 总结

本次实验主要熟悉了文件的读写，Map的使用。

在实验过程中解决了一些额外的问题，例如文本的清洗，编码的差异等。

通过本次实验，提升了我对JAVA相关知识的了解和实践能力。