**词法分析程序说明文档**

**141250149 吴秦月**

目录

[1. 实验目的 3](#_Toc403096214)

[2. 内容描述 3](#_Toc403096215)

[3. 思路方法 3](#_Toc403096216)

[4. 假设 3](#_Toc403096217)

[5. 相关FA描述 4](#_Toc403096218)

[6. 重要数据结构 4](#_Toc403096219)

[7. 核心算法 6](#_Toc403096220)

[8. 运行截图 12](#_Toc403096221)

[9. 问题与解决 13](#_Toc403096222)

[10. 感受与总结 14](#_Toc403096223)

# 实验目的

编写、调试一个词法分析程序，并对语句进行词法分析，掌握记号、模式与单词，掌握正规式与正规集，掌握有限自动机，掌握如何从正规式到词法分析器的各种算法，从而更好理解词法分析原理。

# 内容描述

此程序用Java编写。

1. input：程序读取一个文本文件，并对其中内容进行词法分析，此程序实现了对java程序的词法识别，可识别关键字、标识符（关键字优先于标识符）

、操作符、分隔符 、整数、注释符、浮点数

1. output：输出格式为：种类，识别的单词符号的TOKEN序列
2. error报告：对未识别字符、整型过大（超过java最大的4字节）、文件过大进行异常报错。

# 思路方法

1. 对要识别的单词符号写出正则表达式RE
2. 根据正则表达式RE写出对应的NFA
3. 合并所有NFA并化简为DFA
4. 最小化DFA
5. 基于最小化DFA编程。

# 假设

假设输入的文件内容是正常的java程序，即包含合法的关键字和操作符。

Token分为标识符、关键字、操作符、分隔符、整数和浮点数。

标识符：identifier = letter (letter | digit)\*

关键字：public|class|static|void|main|while|if|else|for|switch|case|package|String|int|float

|double|char|boolean|true|false；

操作符：[,],~,(,),\*,/,.,’,”,+,-,<,>,=,!,&,|,+=,-=,<=,>=,!=,&&,|| ,++,--；

分隔符：;|{|}｜，；

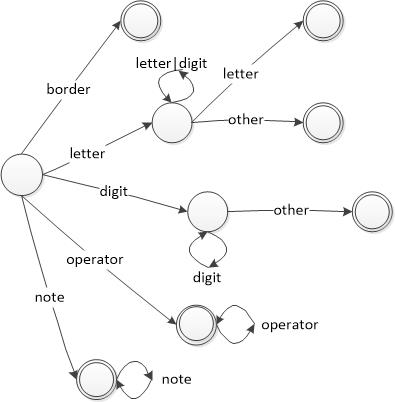
注释符：//|/\*|\*/；

整数：digit (digit)\*

浮点数：digit (digit)\*. digit (digit)\*

另外，空格、\n、\t、\r在词法分析阶段忽略。

# 相关FA描述



ID->letter (letter | digit)\*

Int->digit digit\*

Double->digit digit\* . digit digit\*

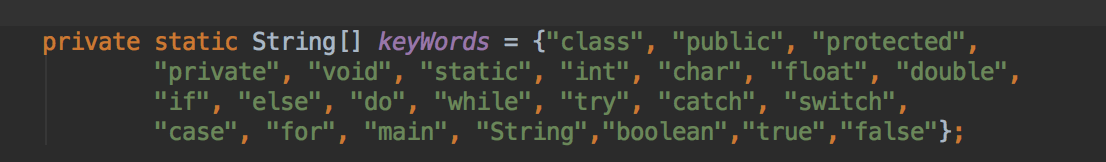
digit->0|1|2|3|4|5|6|7|8|9

letter->a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x

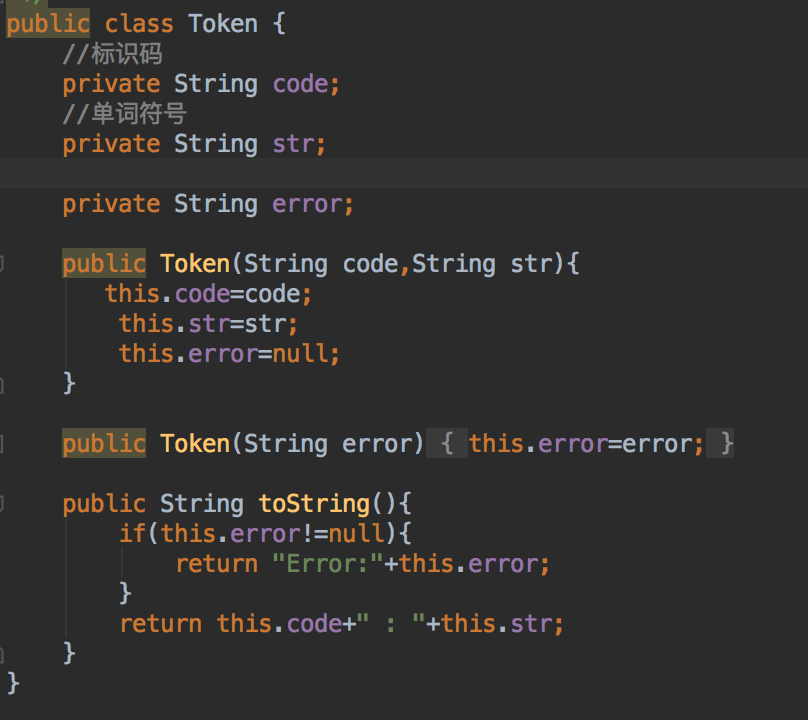
|y|z

# 重要数据结构





//对于每个输出都实例化一个token对象，并加入输出链表



# 核心算法

程序主要的方法有：

readFile()——读取文件得到输入

scanner()——扫描输入进行分析，核心算法

output()——将结果输出到控制台和输出文件中

下面介绍scanner()方法：

Scanner()是该程序分析词法的核心方法，此方法每次只能识别一个单词符号，所以程序通过反复调用scanner()方法分析整个输入。

通过读取的第一个字符的类型，预测接下来的单词符号可能的类型。读到英文字符，可能为保留字或变量名（类型一）；读到数字，就是常数（正数）（类型二）；读到其他字符（类型三），则可能是操作符或边界符或注释符，如果是‘-’符，后面是数字就组成了负数，当然也可能是换行符或是未定义的字符。

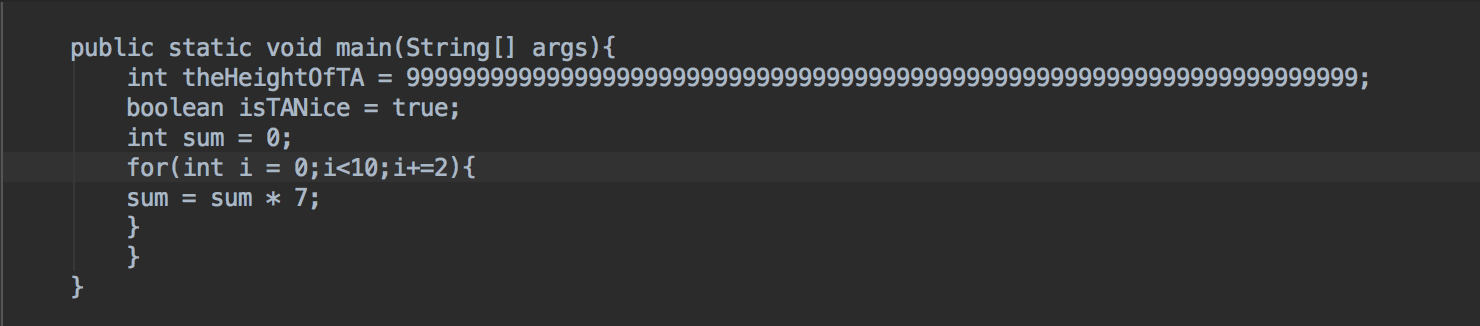
类型一，则继续读取，每读一位，都判断是否属于保留字，若是就直接输出（因为保留字优先于变量名），否则一直读到不是英文字符为止，此时需要退回一位，并输出变量名。

类型二，一直读到不是数字为止，同样需要退回一位，并输出正常数。

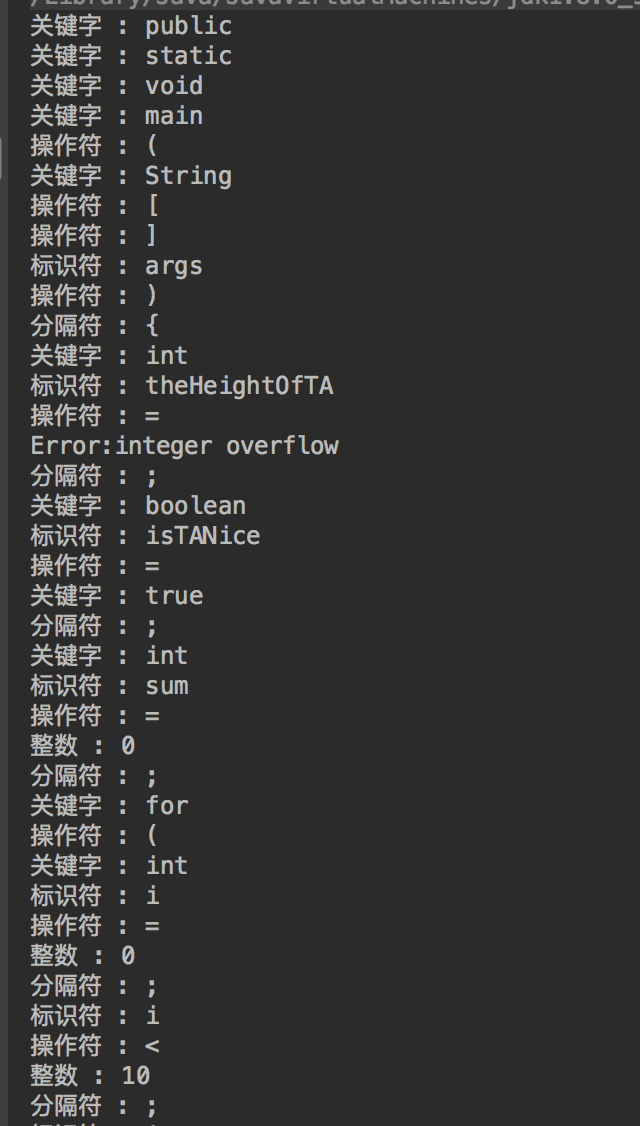
类型三相对复杂一点，如果单字符就可以确定种别，则直接输出；否则继续读取下一位，直到可以确定种别为止，若负号（减号）后面是数字，按类型二读取，最后需要输出负数。同样多读的需要退回。

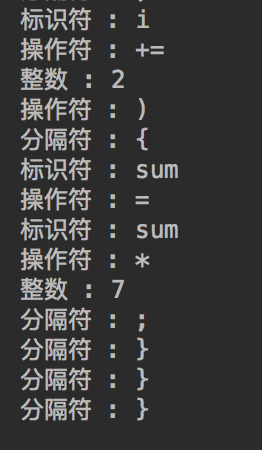
# 运行截图

Input.txt：



Output：





并输出了output .txt文件，保存结果

# 问题与解决

1. 控制台输入较麻烦，且不易有多行输入。所以进行读取文本来读取输入，这样对于读取源码文件也较为方便。
2. 对于int类型常数，如果超过了max\_value则会变为负数，所以根据其结果是否为负，判断是否溢出。

# 感受与总结

经过自己动手查找资料、编写简单的词法分析程序，有助于对词法分析过程和方法有更深入的理解，受益匪浅！