实验目的

通过自己编写、调试一个词法分析程序，并对语句进行词法分析，进行更好理解的词法分析原理。

内容描述

本程序目的是进行一个对 java 语言程序的词法识别，可识别保留字、变量名、操作符、数字等等，并输出格式为 （id ， content） 的 TOKEN 序列，同时对未定义字符、整型过大、变量名格式错误等异常进行异常报错。

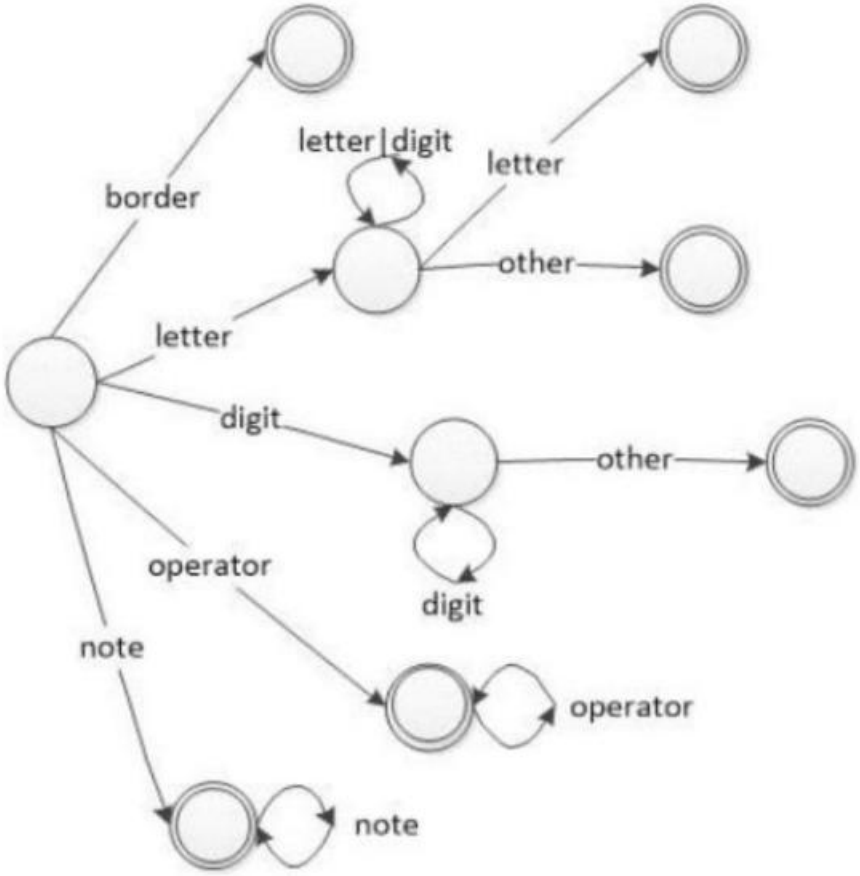
思路方法

1. 针对要识别的单词符号写出正则表达式
2. 构造出正则表达式对应的 NFA
3. 合并所有 NFA 并化简为 DFA
4. 基于 DFA 编写代码
5. 代码中具体的实现：先读取一个输入字符，判断其可能的类别，再读取下一个继续判断，若已经识别出则添加到输出链，指针指向下一个字符位置，否则继续依次读取下一个字符。

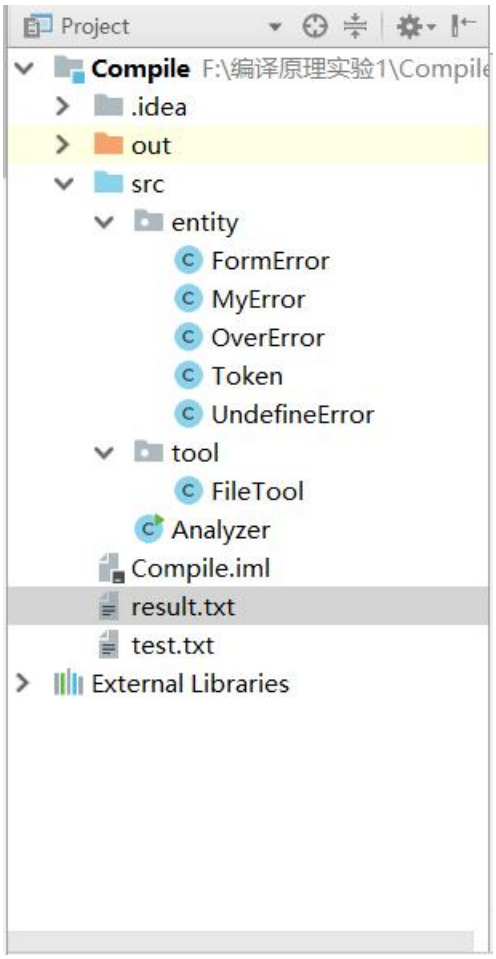
假设

假设输出的文件内容是正常的 java 程序，即包含合法的保留字和运算符。

相关 FA


letter|digit
letter
border
other
letter
digit
other
operator
digit
note
operator
note

重要数据结构描述



总共有 7 个类。其中 MyError 是 FormError,OverError 和 UndefineError 的父类。

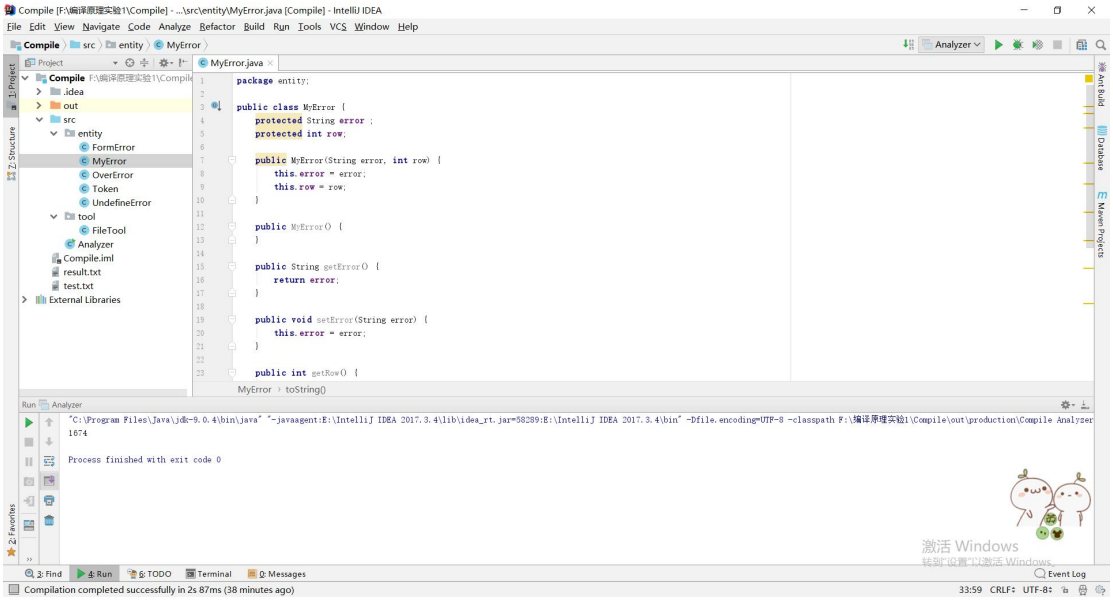
FormError：变量名称格式错误；

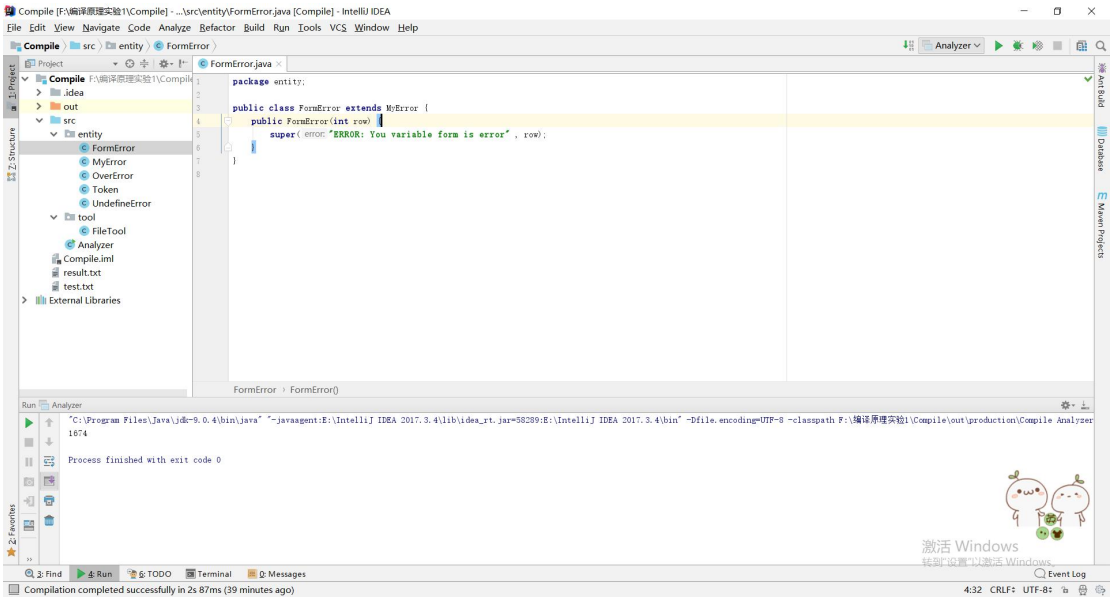
OverError：数字过大错误；

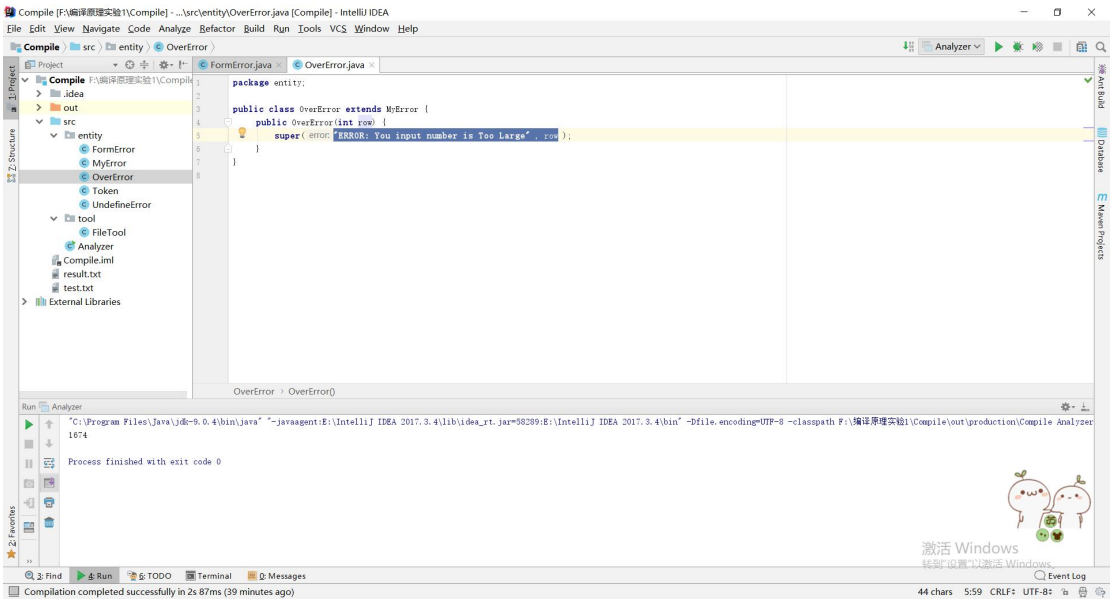
UndefineError：未定义字符错误

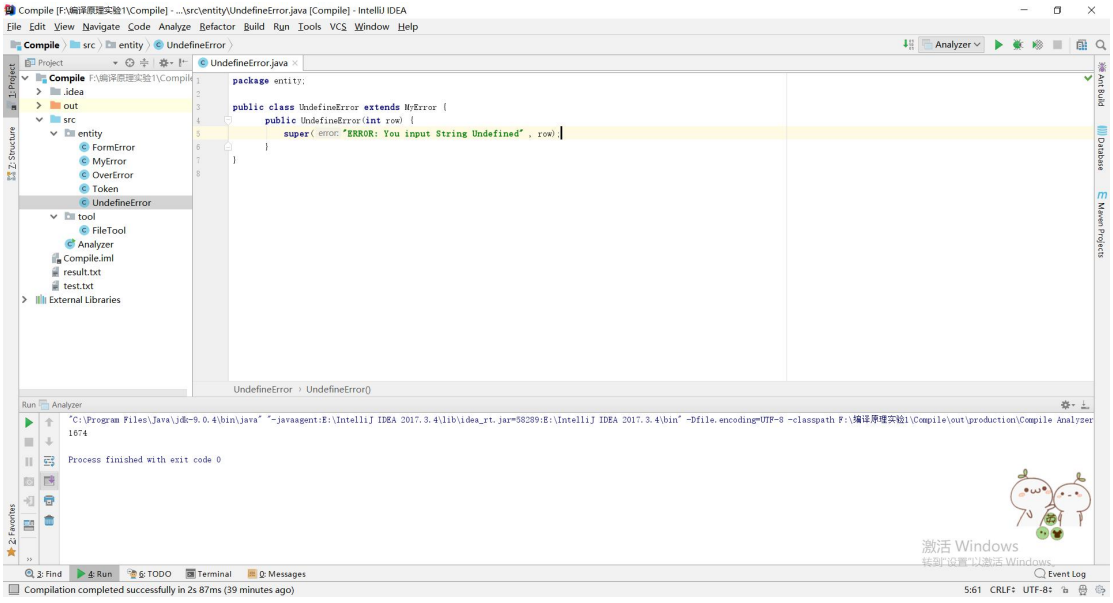
Error 属性：错误类型

Row 属性：错误行数





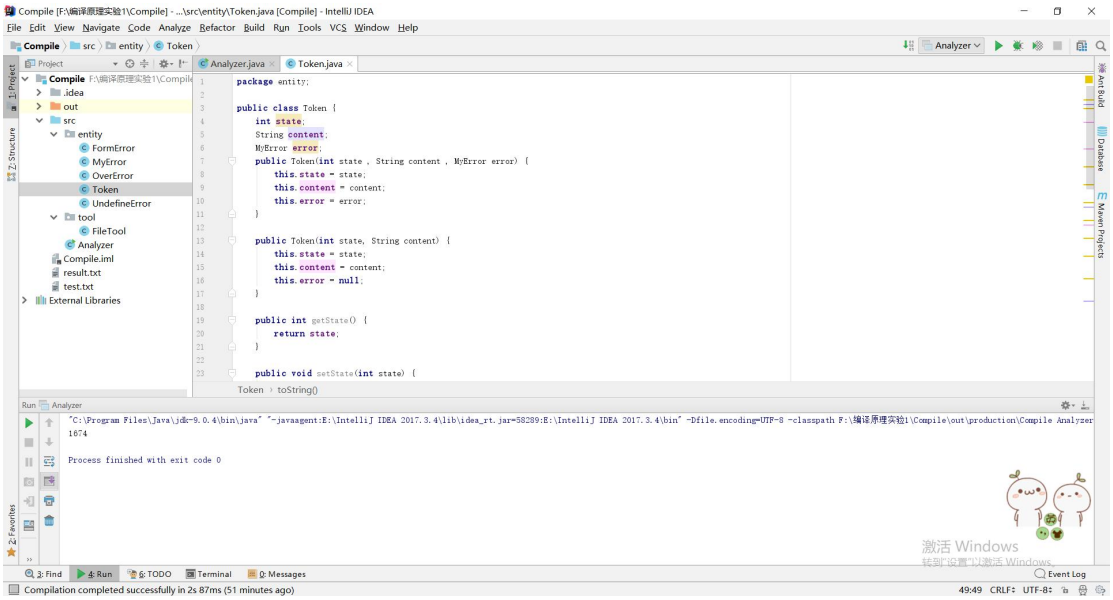

Compile [F:\编译原理实验1\Compile] - ...\src\entity\OverError.java [Compile] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Compile src entity OverError
Analyzer
Project
FormError.java
OverError.java
Compile F:\编译原理实验1\Compile
.idea
out
src
entity
FormError
MyError
OverError
Token
UndefineError
tool
FileTool
Analyzer
Compile.iml
result.txt
test.txt
External Libraries
1 package entity;
2
3 public class OverError extends MyError {
4 public OverError(int row) {
5 super(error: "ERROR: You input number is Too Large" , row);
6 }
7 }
8
OverError › OverError()
Run Analyzer
"C:\Program Files\Java\jdk-9.0.4\bin\java" "-javaagent:E:\IntelliJ IDEA 2017.3.4\lib\idea_rt.jar=58289:E:\IntelliJ IDEA 2017.3.4\bin" -Dfile.encoding=UTF-8 -classpath F:\编译原理实验1\Compile\out\production\Compile Analyzer
1674
Process finished with exit code 0
3: Find 4: Run 6: TODO Terminal 0: Messages Event Log
Compilation completed successfully in 2s 87ms (39 minutes ago)
44 chars 5:59 CRLF: UTF-8:
激活 Windows
转到"设置"以激活 Windows。


Compile [F:\编译原理实验1\Compile] - ...\src\entity\UndefineError.java [Compile] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Compile src entity UndefineError
Analyzer
Project
UndefineError.java
Compile F:\编译原理实验1\Compile
.idea
out
src
entity
FormError
MyError
OverError
Token
UndefineError
tool
FileTool
Analyzer
Compile.iml
result.txt
test.txt
External Libraries
1 package entity;
2
3 public class UndefineError extends MyError {
4 public UndefineError(int row) {
5 super(error: "ERROR: You input String Undefined" , row);
6 }
7 }
8
UndefineError › UndefineError()
Run Analyzer
"C:\Program Files\Java\jdk-9.0.4\bin\java" "-javaagent:E:\IntelliJ IDEA 2017.3.4\lib\idea_rt.jar=58289:E:\IntelliJ IDEA 2017.3.4\bin" -Dfile.encoding=UTF-8 -classpath F:\编译原理实验1\Compile\out\production\Compile Analyzer
1674
Process finished with exit code 0
3: Find 4: Run 6: TODO Terminal 0: Messages Event Log
Compilation completed successfully in 2s 87ms (39 minutes ago)
5:61 CRLF: UTF-8:
激活 Windows
转到"设置"以激活 Windows。

Token 类：输出的 token 序列

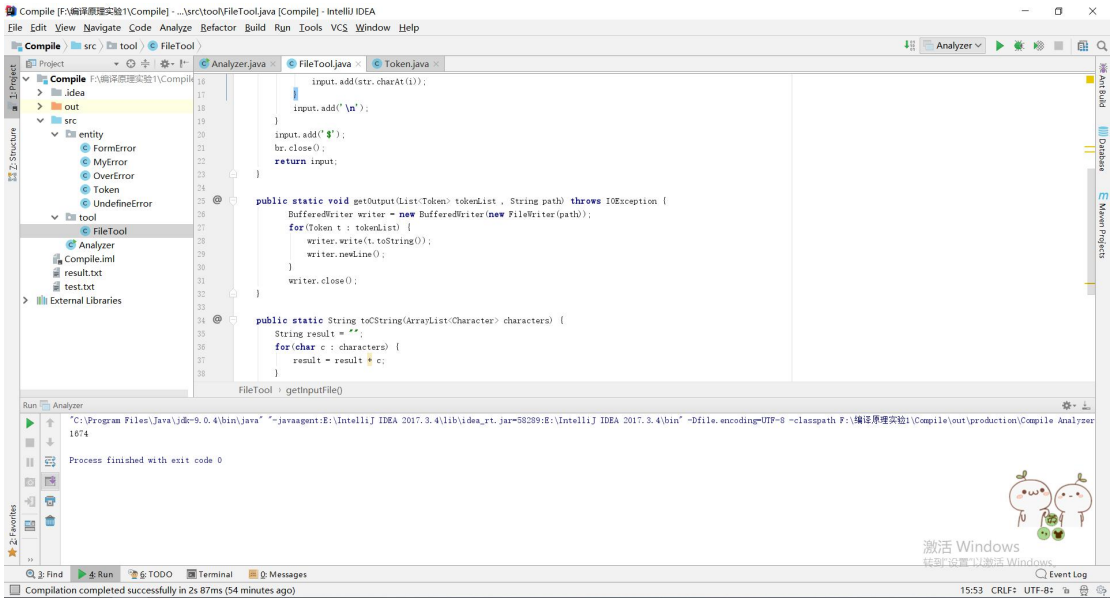
State 和 content 对应输出的合法 token 序列内容。

Error 类属性是用来打印错误信息。

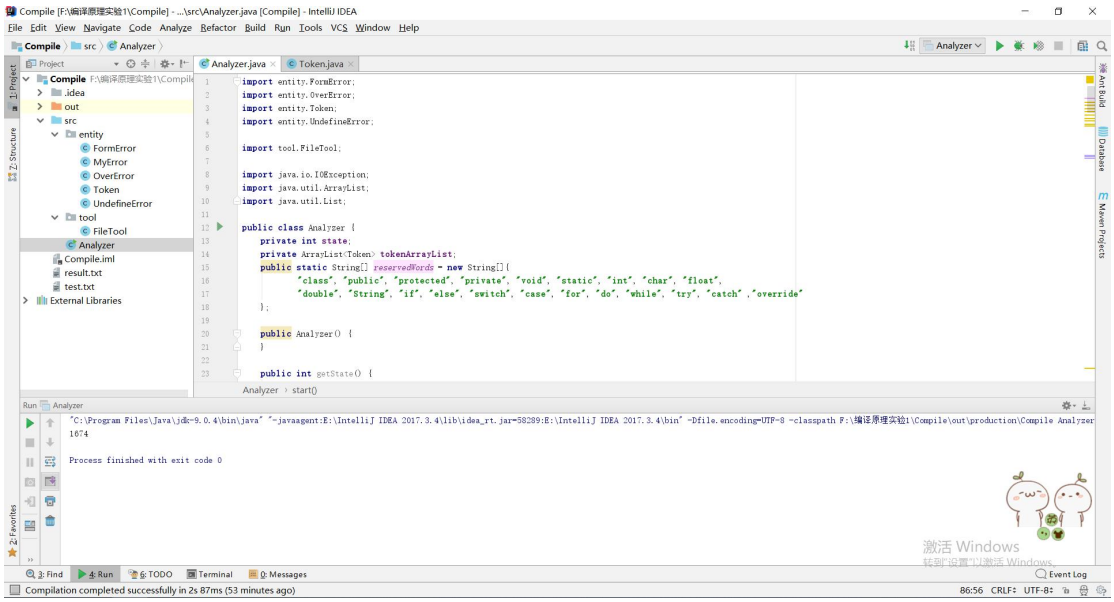


FileTool 类：功能:读取文件，输出文件和字符 array 转字符串

内含三个静态方法。



Analyzer 类：实现类

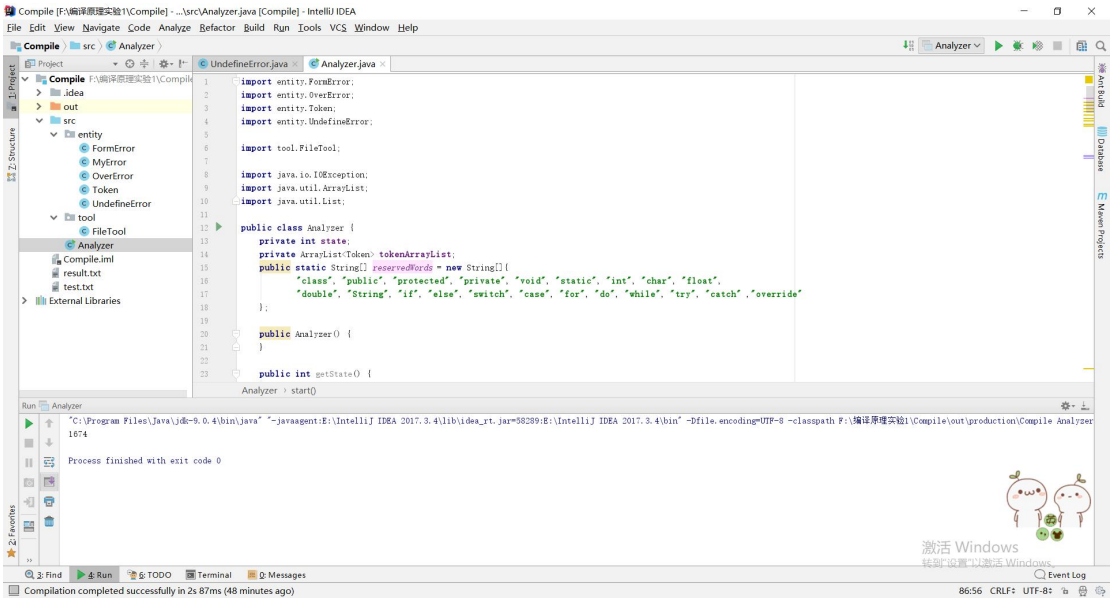


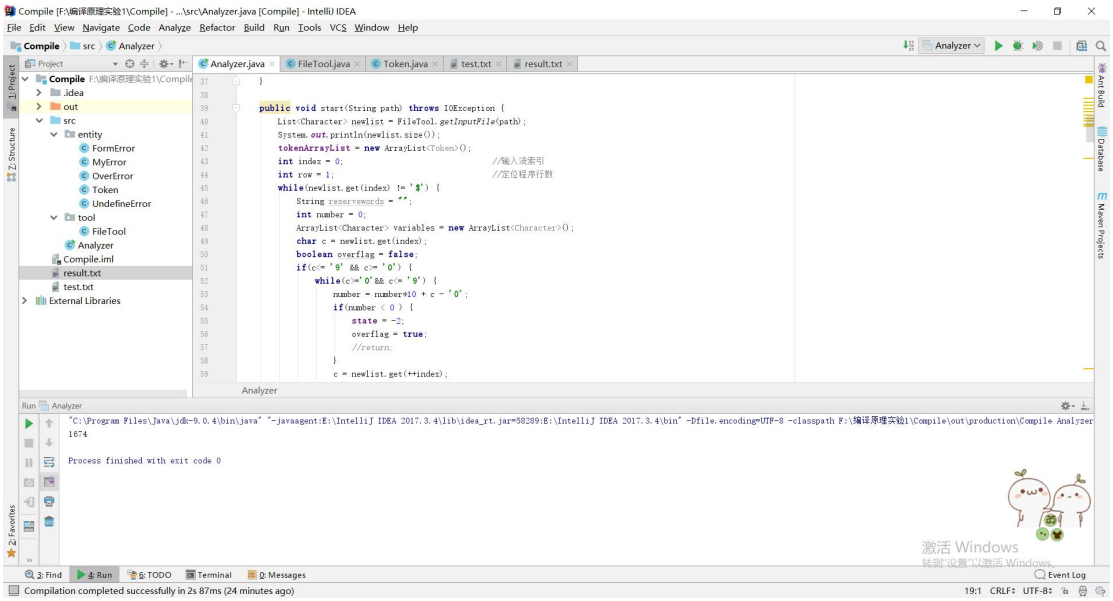
State ： 当前分析词法单元的 id；

tokeArraylist: 输出的 token 序列；

ReservedWords：设定好的静态保留字数组。

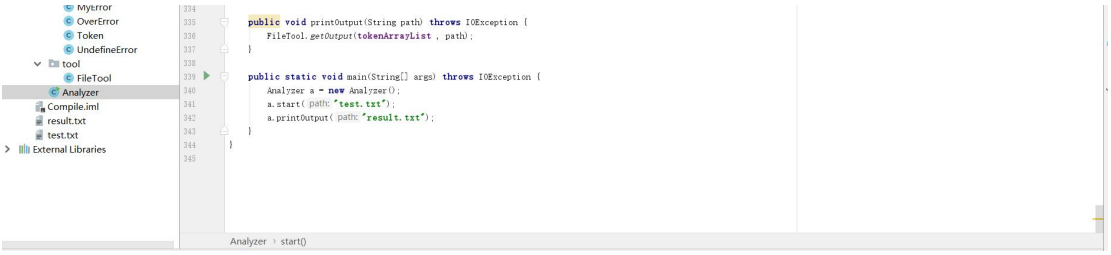
核心算法





整个实现类是 Analyzer 类，核心逻辑放在 start 函数中。输入测试文件的路径 path，通过 start 方法开始分析整个文件，因此该方法逻辑十分复杂。使用 while 循环，不停的读取字符。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，可能是常数（正数）也可能是非法的变量名定义（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，指针指到下一位，并输出变量名。 类型二，一直读到不是数字为止，并输出正常数，指针指到下一位；若不是正常数字，指针指到下一位，并**报错 OverError；**若多读的一位是字符的话，则继续读下去，直到指针不是字符位置，**并报错 FormError；**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；若不是正常数字，指针指到下一位，并**报错 OverError；**如果碰上未定义的字符，指针指到下一位，并**报错 UndefinedError。**



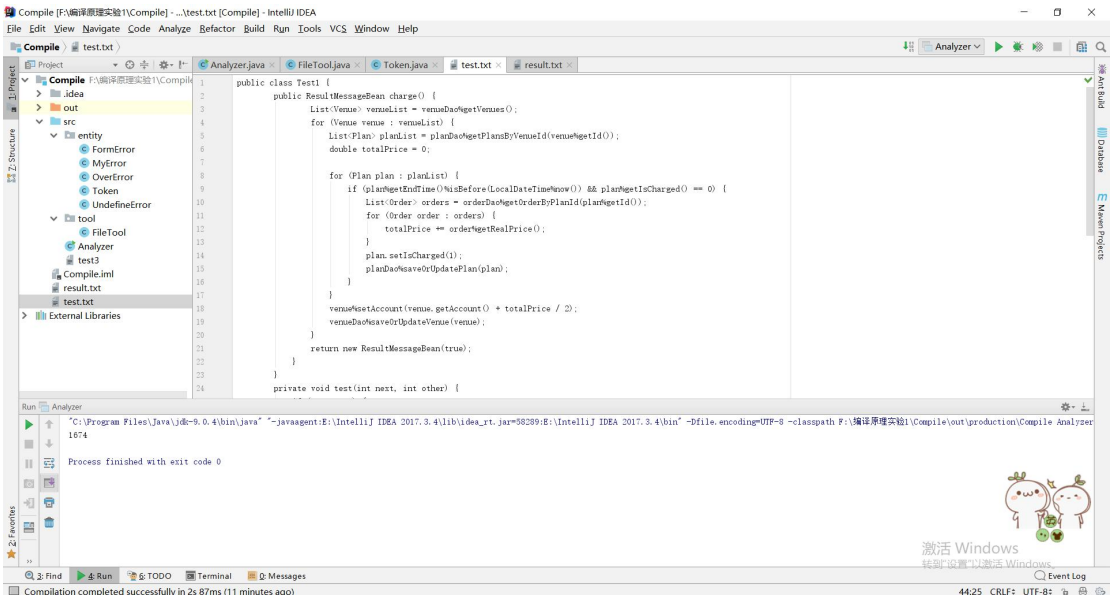
结果通过调用 printOutput 输出。

附录：id 和 content 的对应关系

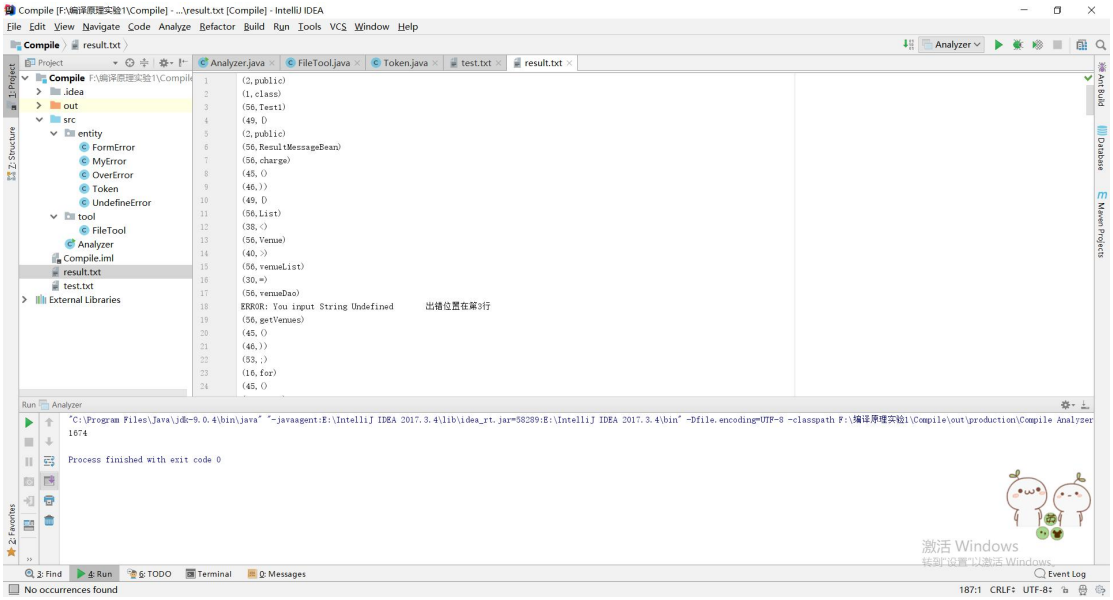
Class	1	/=	29
Public	2	=	30
Protected	3	==	31
Private	4	&	32
Void	5	&&	33
Static	6	|	34
Int	7	||	35
Char	8	!	36
Float	9	!=	37
Double	10	<	38
String	11	<=	39
If	12	>	40
Else	13	>=	41
Do	14	//	42
While	15	/*	43
Try	16	*/	44
Catch	17	(	45
Switch	18	)	46
Case	19	[	47
for	20	]	48
	21	{	49
+	22	}	50
+=	23	,	51
-	24	:	52
-=	25	;	53
*	26	'	54
*=	27	"	55
/	28	Letter(letter|digit)*	56
Digit|digit*	57		

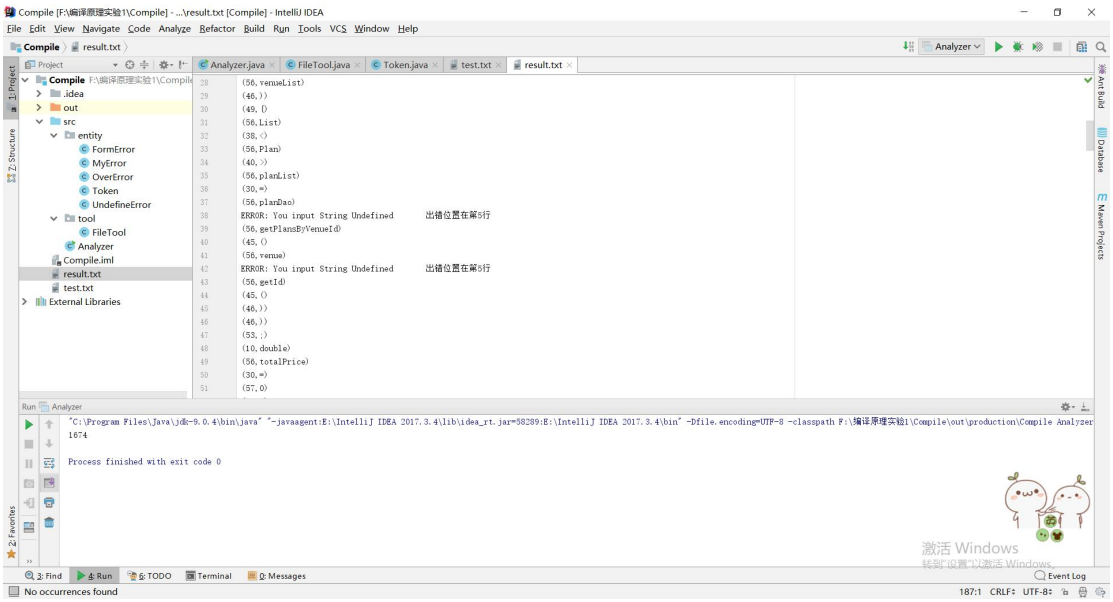
运行截图

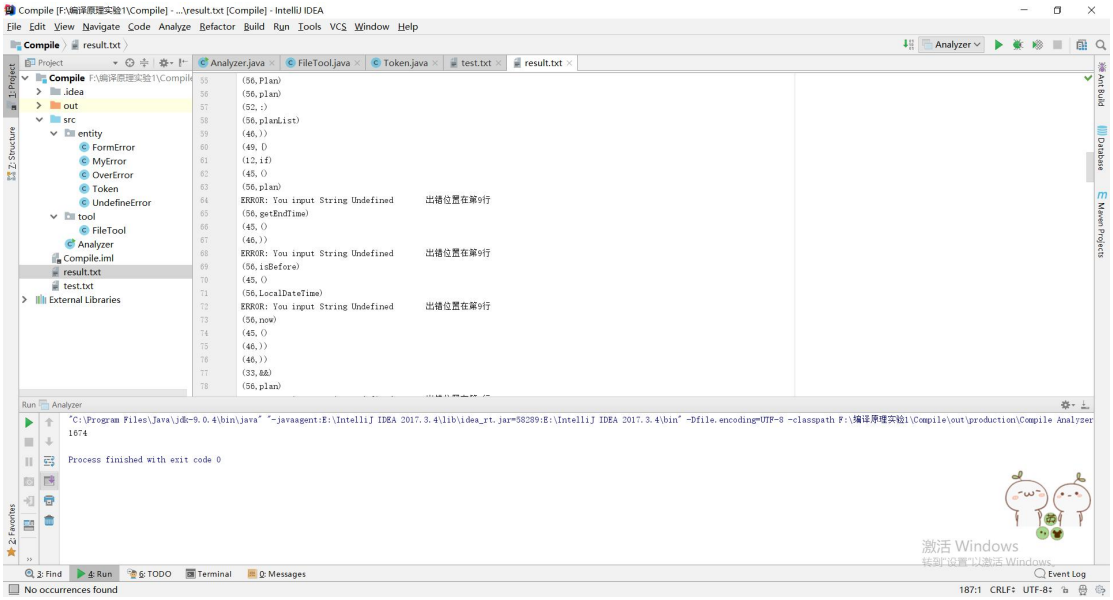
输入 test.txt：

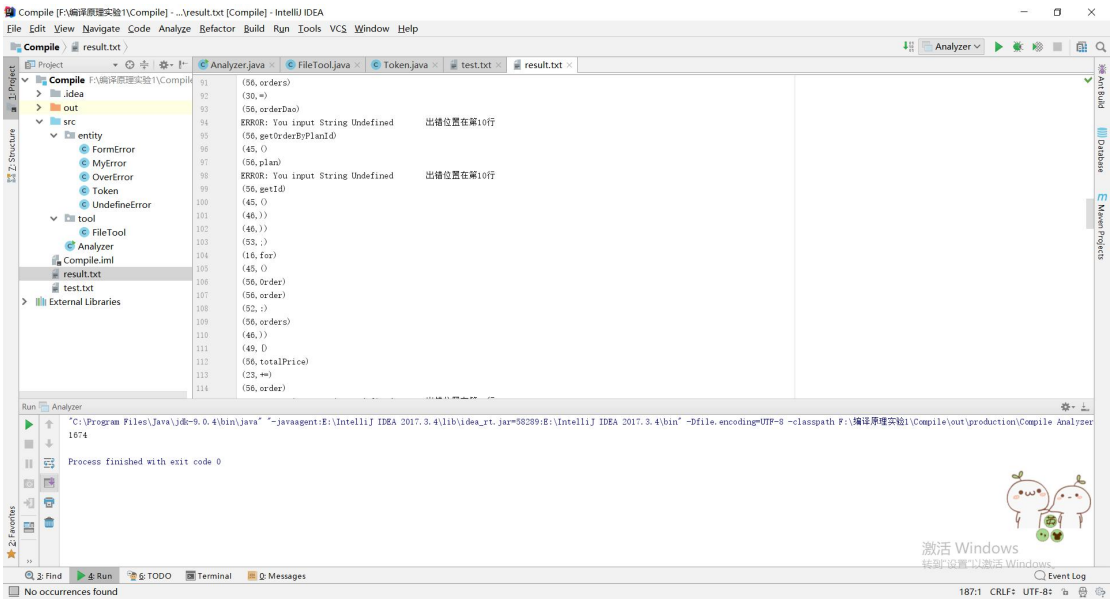


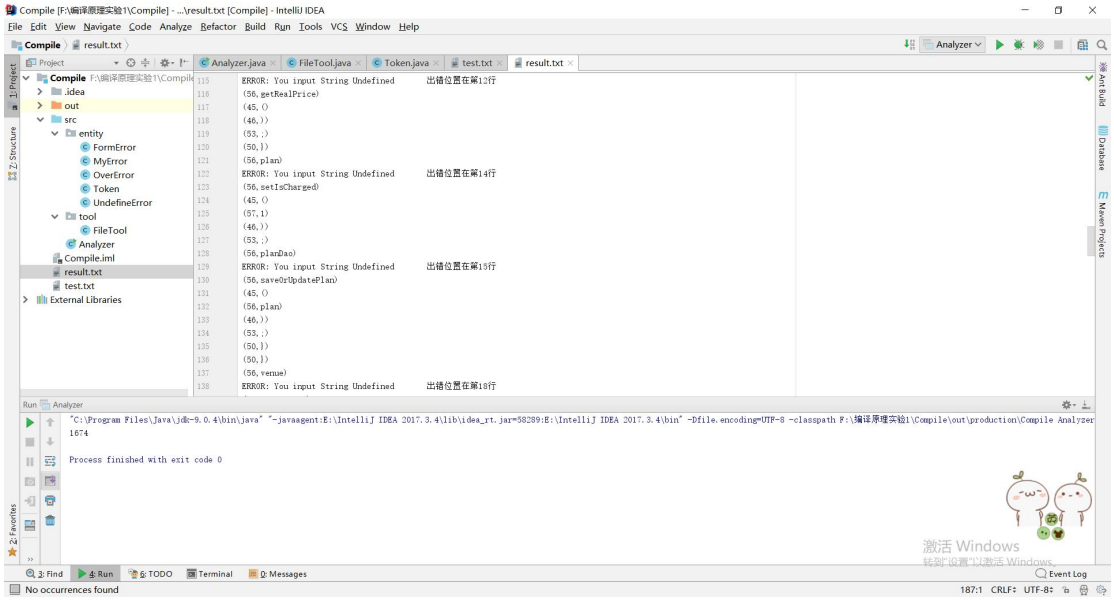
将结果打印到 result.txt 文件中：

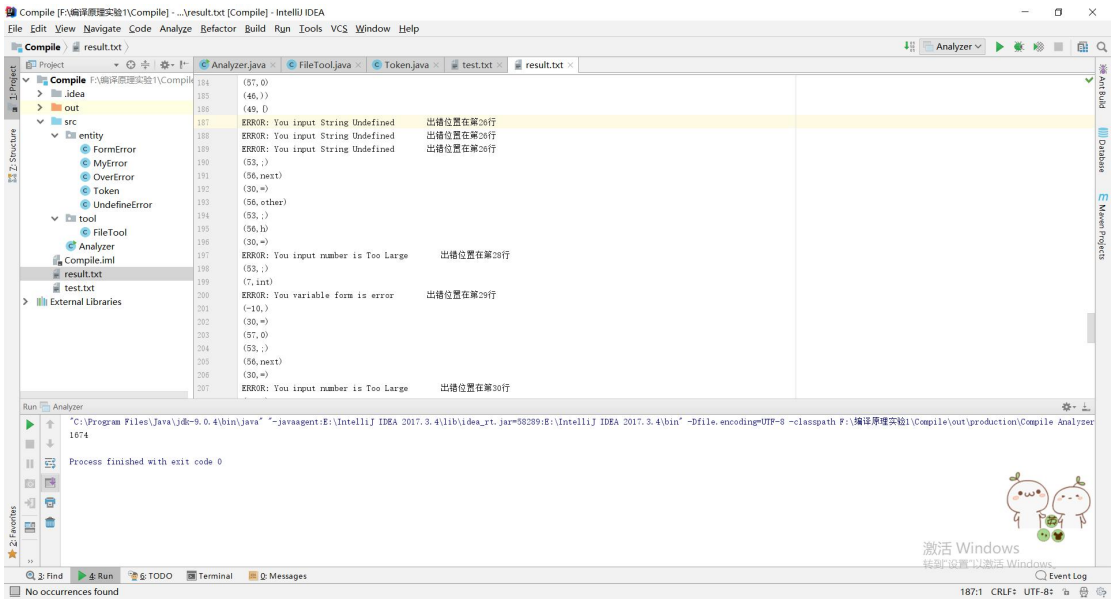

Compile [F:\编译原理实验1\Compile] - ...\result.txt [Compile] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Compile result.txt
Analyzer
Project
Analyzer.java FileTool.java Token.java test.txt result.txt
Compile F:\编译原理实验1\Compile
.idea
out
src
entity
FormError
MyError
OverError
Token
UndefineError
tool
FileTool
Analyzer
Compile.iml
result.txt
test.txt
External Libraries
(2,public)
(1,class)
(56,Test1)
(49,{)
(2,public)
(56,ResultMessageBean)
(56,charge)
(45,()
(46,))
(49,{)
(56,List)
(38,<)
(56,Venue)
(40,>)
(56,venueList)
(30,=)
(56,venueDao)
ERROR: You input String Undefined 出错位置在第3行
(56,getVenues)
(45,()
(46,))
(53,;)
(16,for)
(45,()
Run Analyzer
"C:\Program Files\Java\jdk-9.0.4\bin\java" "-javaagent:E:\IntelliJ IDEA 2017.3.4\lib\idea_rt.jar=58289:E:\IntelliJ IDEA 2017.3.4\bin" -Dfile.encoding=UTF-8 -classpath F:\编译原理实验1\Compile\out\production\Compile Analyzer
1674
Process finished with exit code 0
3: Find 4: Run 6: TODO Terminal 0: Messages Event Log
No occurrences found
187:1 CRLF: UTF-8:


Compile [F:\编译原理实验1\Compile] - ...\result.txt [Compile] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Compile result.txt
Analyzer
Project
Analyzer.java FileTool.java Token.java test.txt result.txt
Compile F:\编译原理实验1\Compile
.idea
out
src
entity
FormError
MyError
OverError
Token
UndefineError
tool
FileTool
Analyzer
Compile.iml
result.txt
test.txt
External Libraries
(56,venueList)
(46,))
(49,{)
(56,List)
(38,<)
(56,Plan)
(40,>)
(56,planList)
(30,=)
(56,planDao)
ERROR: You input String Undefined 出错位置在第5行
(56,getPlansByVenueId)
(45,()
(56,venue)
ERROR: You input String Undefined 出错位置在第5行
(56,getId)
(45,()
(46,))
(46,))
(53,;)
(10,double)
(56,totalPrice)
(30,=)
(57,0)
Run Analyzer
"C:\Program Files\Java\jdk-9.0.4\bin\java" "-javaagent:E:\IntelliJ IDEA 2017.3.4\lib\idea_rt.jar=58289:E:\IntelliJ IDEA 2017.3.4\bin" -Dfile.encoding=UTF-8 -classpath F:\编译原理实验1\Compile\out\production\Compile Analyzer
1674
Process finished with exit code 0
3: Find 4: Run 6: TODO Terminal 0: Messages Event Log
No occurrences found
187:1 CRLF: UTF-8:


Compile [F:\编译原理实验1\Compile] - ...\result.txt [Compile] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Compile result.txt
Analyzer
Project
Analyzer.java FileTool.java Token.java test.txt result.txt
Compile F:\编译原理实验1\Compile
.idea
out
src
entity
FormError
MyError
OverError
Token
UndefineError
tool
FileTool
Analyzer
Compile.iml
result.txt
test.txt
External Libraries
(56,Plan)
(56,plan)
(52,:)
(56,planList)
(46,))
(49,{)
(12,if)
(45,()
(56,plan)
ERROR: You input String Undefined 出错位置在第9行
(56,getEndTime)
(45,()
(46,))
ERROR: You input String Undefined 出错位置在第9行
(56,isBefore)
(45,()
(56,LocalDateTime)
ERROR: You input String Undefined 出错位置在第9行
(56,now)
(45,()
(46,))
(46,))
(33,&&)
(56,plan)
Run Analyzer
"C:\Program Files\Java\jdk-9.0.4\bin\java" "-javaagent:E:\IntelliJ IDEA 2017.3.4\lib\idea_rt.jar=58289:E:\IntelliJ IDEA 2017.3.4\bin" -Dfile.encoding=UTF-8 -classpath F:\编译原理实验1\Compile\out\production\Compile Analyzer
1674
Process finished with exit code 0
3: Find 4: Run 6: TODO Terminal 0: Messages Event Log
No occurrences found
187:1 CRLF: UTF-8:
激活 Windows
转到"设置"以激活 Windows。


Compile [F:\编译原理实验1\Compile] - ...\result.txt [Compile] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Compile result.txt
Analyzer
Project
Analyzer.java FileTool.java Token.java test.txt result.txt
Compile F:\编译原理实验1\Compile
.idea
out
src
entity
FormError
MyError
OverError
Token
UndefineError
tool
FileTool
Analyzer
Compile.iml
result.txt
test.txt
External Libraries
(56,orders)
(30,=)
(56,orderDao)
ERROR: You input String Undefined 出错位置在第10行
(56,getOrderByPlanId)
(45,()
(56,plan)
ERROR: You input String Undefined 出错位置在第10行
(56,getId)
(45,()
(46,))
(46,))
(53,:)
(16,for)
(45,()
(56,Order)
(56,order)
(52,:)
(56,orders)
(46,))
(49,{)
(56,totalPrice)
(23,+=)
(56,order)
Run Analyzer
"C:\Program Files\Java\jdk-9.0.4\bin\java" "-javaagent:E:\IntelliJ IDEA 2017.3.4\lib\idea_rt.jar=58289:E:\IntelliJ IDEA 2017.3.4\bin" -Dfile.encoding=UTF-8 -classpath F:\编译原理实验1\Compile\out\production\Compile Analyzer
1674
Process finished with exit code 0
3: Find 4: Run 6: TODO Terminal 0: Messages Event Log
No occurrences found
187:1 CRLF: UTF-8:
激活 Windows
转到"设置"以激活 Windows。





总共分析了 1647 个词法单元：

问题与解决

暂无

感受与总结

实现这样一个简单的小小编译器，就有着那么复杂的逻辑。让我深刻地感受到要想实现一个识别完备的编译器，是多么的不容易。与此同时，编写词法分析程序，有助于对词法分析过程和方法有更深入的理解。