编译原理

课程Project1答疑

COMP130014.01

Project 1 完成情况

所有同学都实现了基本的词法分析功能，但实现细节上存在以下问题：

1. 忘记输出统计tokens总数，或将Comment和Error统计为tokens
2. 字符串的正则表达式错误
3. 注释的正则表达式错误
4. 超长字符串、超长id判断错误
5. 遗漏坏字符的报错，遗漏部分tokens（如大于小于号）
6. 误把负数当成一个token或直接当成错误报错（PCAT文档中定义的integer/real正则表达式只能匹配非负数，如果出现负数应该解释为单目运算符“-”加上一个正数）
7. 行号和列号统计错误，行列号以编辑器显示为准，\t长度按1或4均可
8. 用户不友好，输入文件名写死在代码里面，每次运行要重新编译

Project 1 常见错误

Strings begin and end with a double quote (") and contain any sequence of printable ASCII characters, except double quotes. Note in particular that strings may not contain tabs or newlines. String literals are limited to 255 characters in length, not including the delimiting double quotes.

注意：字符串中不能包含引号、tab和换行，判断超长字符串时应该是yytext长度超过255+2=257。

错误的字符串正则表达式：\"[^"]*\"

正确的字符串正则表达式：\"[^"\n\t]*\"

更简介有效的实现：使用flex提供的有限自动机，定义<STRING>状态

Project 1 常见错误

Comments are enclosed in the pair (* and *); they cannot be nested. Any character is legal in a comment. Of course, the first occurrence of the sequence of characters *) will terminate the comment. Comments may appear anywhere a token may appear; they are self-delimiting; i.e. they do not need to be separated from their surroundings by whitespace.

注意：注释以(*开头，以*)结尾，但不能嵌套，也就是说注释需要进行**非贪婪匹配**。例子：(*123*)*)

错误的注释正则表达式："(*".*"*)"

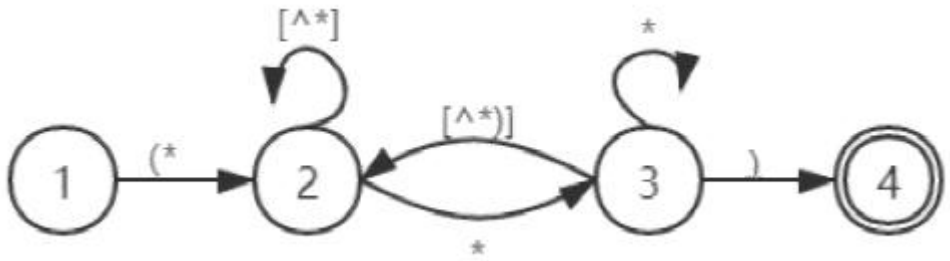
正确的注释正则表达式："(*"([^*]|(*)*[^*\)])*(*)*"*)"

更简介有效的实现：使用flex提供的有限自动机，定义<COMMENT>状态

Project 1 常见错误

实现思路:

画出有限自动机状态图



初始状态只有遇到 (* 进入状态2代表注释开始

不遇到*就一直留在状态2，遇到*进入状态3进一步判定是否结束注释

在状态3下遇到右括号，那当然是注释结束。

在状态3下遇到*，就继续维持判定状态。

在状态3下遇到其他字符，返回状态2。

“(*”　**(　[^*]　|　(*)*[^*\)]　)***　(*)*　"*)"

Project 1 常见错误

也可以使用flex提供的条件式触发功能

- %x COMMENTS　　定义注释标识符
- “(*” {BEGIN(COMMENTS);}　遇到(* 就进入
- <COMMENTS>“*)” {BEGIN(INITIAL);} 遇到*) 就退出

其他特殊处理：

- <COMMENTS>\n　　更新行列号
- <COMMENTS><<EOF>>　遇到结尾，说明未配对

Project 1 报错功能

- 超长整数：atoi越界返回-1，或者先判断位数再逐位比较if(strlen(yytext)>10 || (strlen(yytext)==10 && strcmp(yytext,INTEGER_MAX) > 0))
- 坏字符：都没匹配到的、 [\x00-\x08\x0A-\x1F0x7F]、调用isprintable()都行
- 没有配对的字符串：普通字符串去掉右引号就行，由于长度优先匹配原理，配对字符串是不会先匹配到这条规则的。
- 含\n \t 的字符串：可以把允许\n \t字符串的正则表达式放在不含\n \t的后面，也可以一开始就匹配掉允许\n \t的随后对yytext分析。
- 没有配对的注释：有限自动机去掉最后一个状态即可，或者进入COMMENT状态后遇到EOF还没有退出。因为注释允许换行，没有配对只可能发生在遇到EOF。
- 标识符应当在保留关键词之后识别，因为所有保留字都符合标识符的正则表达式，标识符实际上是：{LETTER}({LETTER}|{DIGIT})*-RESERVED_KEYWORD，可以通过flex优先级来实现差集。（先长度后次序优先）

Project2 建议

- 请大家根据这个PPT里面提到的常见错误修改自己的PJ1代码，因为PJ1直接影响到PJ2的实现，应尽早纠正存在的错误。
- 阅读flex和bison的文档，尽量使用flex和bison提供的接口，API和状态定义，写起来会更加顺手。
- 了解Makefile的使用，不要把输入输出都写死在文件里，而是使用Makefile脚本一键编译和生成，更加方便。

Token数目参考

- Case 1：50 token(s), 0 error(s) found.
- Case 2：88 token(s), 0 error(s) found.
- Case 3：96 token(s), 0 error(s) found.
- Case 4：225 token(s), 0 error(s) found.
- Case 5：97 token(s), 0 error(s) found.
- Case 6：68 token(s), 0 error(s) found.
- Case 7：65 token(s), 0 error(s) found.
- Case 8：108 token(s), 0 error(s) found.
- Case 9：120 token(s), 0 error(s) found.
- Case 10：140 token(s), 0 error(s) found.
- Case 11：22 token(s), 10 error(s) found.