利用LEX计算文本文件的字符数等

Task One

姓名： 周炜

学号： 32010103790

一、实验要求

编写一个LEX输入文件，使之生成可计算文本文件的字符、单词和行数且能报告这些数字的程序。单词为不带标点或空格的字母和数字的序列。标点和空白格不计算为单词。

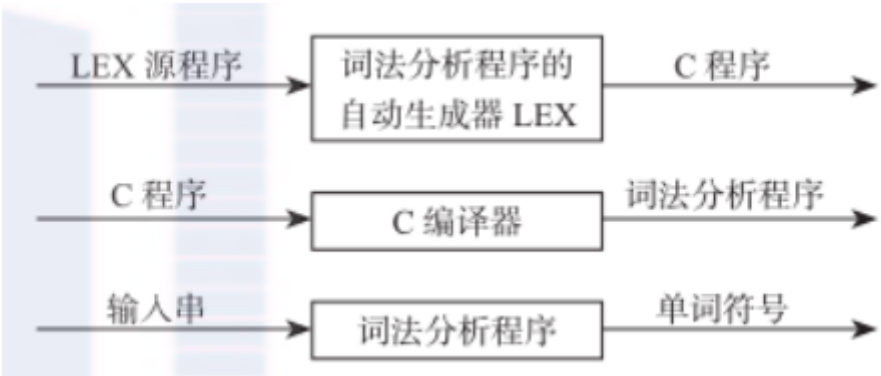
二、实验简介

词法分析器

由于程序设计语言中的单词基本上都可用一组正规式来描述，因此，人们希望构造一个自动生成系统：对于一个给定的高级语言，只要给出用来描述其各类单词词法结构的一组正则表达式，以及识别各类单词时词法分析程序应采取的语义动作，该系统便可自动产生此语言的词法分析程序。1975年美国贝尔实验室的M. Lesk和Schmidt基于正规式与有限自动机的理论研究，用C语言研制了一个词法分析程序的自动生成工具LEX。对任何高级程序语言，用户只需用正规式描述该语言的各个词法类（这一描述称为LEX的源程序），LEX就可以自动生成该语言的词法分析程序。

LEX源文件结构

LEX 的输入是用LEX源语言编写的程序，它是扩展名为.l的文件。LEX源程序经LEX 系统处理后输出一个C程序文件，此文件含有两部分内容：一个是依据正规式所构建的状态转移表； 另一个是用来驱动该状态转移表的总控程序yylex ()。该文件再经过C编译器的编译就产生一个实际可以运行的词法分析程序。



正如课堂上所讲的一个LEX 源程序由“% %”分隔的三个部分组成，其书写格式为：

```
定义部分
% %
识别规则部分
% %
辅助函数部分
```

其中，定义部分和辅助函数部分是任选的，识别规则部分则是必备的。如果辅助函数部分缺省， 则第二个分隔号“% %”可以省略；但由于第一个分隔号% %用来指示识别规则部分的开始，故即使没有定义部分， 也不能将其省略。下面将对这三部分的内容及其书写格式作一概括性介绍。

三、环境配置

我使用的是 Unbuntu 20， 使用下列命令配置了相关环境

```
sudo apt-get install flex bison
sudo apt-get install build-essential
gcc --version
```

我为代码写了shell脚本，只需要运行

```
bash count.sh
```

四、程序编写

首先是定义部分，定义了头文件表、常数定义、全局变量定义、正规表达式定义等（除宏定义外，定义部分的其余代码需用符号“%{”和“%}”括起来）。本次实验的目的是计算文本文件的字符、单词和行数

所以定义3个正则表达式

```
%option noyywrap
%{
    int LineNum = 1, CharNum = 0, WordNum = 0, wordaccept = 1;
%}
wordChar    [0-9a-zA-Z]
otherChar   [^{wordChar}]
```

然后定义了识别规则

```
%%
\n  {
    LineNum++;
    CharNum++;
    if (wordaccept)
        WordNum++;
    wordaccept = 0;
}
{wordChar}  {
            CharNum++;
            if (wordaccept)
                WordNum++;
            wordaccept = 0;
}
```

```
{otherChar} CharNum++; wordaccept = 1;
```

最后是辅助函数部分

表 3-2　LEX 中常用的一些变量和函数

yyin	FILE * 类型，指向 LEX 输入文件，缺省情况下指向标准输入
yyout	FILE * 类型，它指向 LEX 输出文件。缺省情况下指向标准输出
yytext	char * 类型，指向与识别规则中的一个正规式匹配的单词的首字符
yyleng	int 类型，记录与识别规则中正规式匹配的单词的长度
yylex()	从该函数开始分析，由 LEX 自动生成
yywrap()	文件结束处理函数，如果其返回值是 1，就停止解析。
echo	将 yytext 打印到 yyout

这里使用了`yylex()`来标记开始分析

```
int main() {
    yylex();
    printf("There have %d chars\n %d words\n %d lines\n", CharNum, WordNum,
LineNum);
    return 0;
}
```

五、效果展示

1. 写好符合功能要求的.l文件，我将其命名为test.l
2. 运行`lex counter.l`，会生成lex.yy.c文件
3. 运行`gcc lex.yy.c -lfl`，生成a.out文件
4. 运行`./a.out <test.txt`

我把上述过程写为了shell脚本，直接运行如下命令

```
bash count.sh
```

效果为：

```
zhouwei@ubuntu:~/lab1$ bash count.sh
There have 608 chars
 96 words
 2 lines
```

与word中的结果相同

由于word会不计入最后一个\n，所以实际上是607+1=608个char，并且word不会计算数字，在我的样本里面有4个数字，因此有93+3=96个word

The University of California, Berkeley (UC Berkeley, Berkeley, Cal, or California),[11][12] is a public grant research university in Berkeley, California. Founded in 1868 as the University of California and named after Anglo philosopher George Berkeley, it is the state's first grant university and the founding campus of the University of California system.↵

Berkeley is also a founding member of the Association of American Universities and was one of the original eight Public Ivy schools, a group of public universities considered as providing a quality of education comparable to those of the Ivy League.↵