Compiler Labl

Pascal Scanner ReadMe

B063040061
陳少洋

1. Lex 版本：flex 2.6.4
2. 作業平台：Ubuntu 18.04.4
3. 執行方式：
   i. make (make all)
   ii. ./a.out < testfile.pas
   iii. make clean
4. 如何處理這份規格書上的問題：
   i. “reserved word”：
      由於保留字不分大小寫，每個字都給予一個大小寫皆可選擇的 regular expression，例如 and：[Aa][Nn][Dd]，最後再將所有保留字集合起來，並給予最高的優先權。
   ii. “ID”：
      Identifier 為底線或英文字母開頭且不超過 15 字元的字串，給予 regular expression：[_a-zA-Z][_a-zA-Z0-9]*，以底線與英文字母開頭後，數字也能加入 ID 中，並在其中判斷是否超過 15 字元。錯誤的 ID 為非底線、英文字母與特殊符號開頭且其中包含英文字母或底線的字串，判斷其為錯誤的 ID。
   iii. “symbols”：
      Symbol 分為兩種規則類別撰寫，一個是單一字元的 symbol：[;:()><=\[\]+\-*/,\.]，另一個是有兩個字元的 symbol：[:<>=]=，其中特別抓出在{int}[+-]{int}中的”+-“是 symbol。
   iv. “integer”：
      Integer 的正負符號可有可無，後面可接 1 個 0 或是以非 0 開頭接一串數字則為整數，其錯誤的整數為以 0 開頭的一串數字。
   v. “real”：
      Real 的優先權次於 Integer，其中整數部分與 Integer 一樣，而小數點與 exp 有其中一項則判斷為 Real，其中 exp 後面必定帶有[+-]{integer}，小數點部分則可由任意數字組成，但避免尾數多於 1 個 0 的情況，其中錯誤的 Real 分為五

個部分，第一種錯為整數部分的錯，與 error integer 一樣，第二種為小數點後結尾多餘 1 個以上的 0，第三與第四種錯分別為小數點前後沒有任何數字，第五種錯為 exp 後的整數部分錯誤，將其錯誤結合。

vi. “string”：

Quoted String 為兩個單引號內的所有字元形成的字串，其中不包刮換行字元，所有出現單引號必定成雙成對出現，錯誤型式為只有單邊左或右引號，擷取至空白、換行或特殊符號為止。

vii. “comment”：

Comment 為”(*”與”*)”所括出來的任意字元所組成，其中包含換行字元，我的處理方式為以這兩個括號為開頭與結束，其中不能同時出現”*)”，否則視為 comment 結束，而錯誤的註解則為僅有齊頭與結尾，其中以換行或空白為斷點。

5. 寫這個作業所遇到的問題：

第一次接觸 Lex 語法，難免生澀，需要重新學習一個新的表達模式，其中也在 regular expression 進行了很多次的 try and error，去嘗試並熟悉規則的撰寫。

我認為自己想了很多有可能出現的測試資料，讓我在撰寫的過程中想了很多使用者可能會寫出來的結果，也盡量將所有可能出現的結果避免掉，所以在除了一般的要求外，多做了很多不同的測試，也將我能想到的所有錯誤避免掉，花了較多的時間。

6. 測試檔執行出來的結果：

測資 1.pas

```
yang@ubuntu-virtual-machine:~/Compiler/testfile_lab1_2020/example$ ./a.out < 1.pas
Line: 1, 1st char: 1, "program" is a "reserved word".
Line: 1, 1st char: 9, "test" is a "ID".
Line: 1, 1st char: 13, ";" is a "symbol".
Line: 2, 1st char: 1, "var" is a "reserved word".
Line: 3, 1st char: 3, "i" is a "ID".
Line: 3, 1st char: 5, ":" is a "symbol".
Line: 3, 1st char: 7, "integer" is a "reserved word".
Line: 3, 1st char: 14, ";" is a "symbol".
Line: 4, 1st char: 1, "begin" is a "reserved word".
Line: 5, 1st char: 3, "read" is a "reserved word".
Line: 5, 1st char: 7, "(" is a "symbol".
Line: 5, 1st char: 8, "i" is a "ID".
Line: 5, 1st char: 9, ")" is a "symbol".
Line: 5, 1st char: 10, ";" is a "symbol".
Line: 6, 1st char: 1, "end" is a "reserved word".
Line: 6, 1st char: 4, ";" is a "symbol".
```

測資 2. pas

```
yang@ubuntu-virtual-machine:~/Compiler/testfile_lab1_2020/example$ ./a.out < 2.pas
Line: 1, 1st char: 1, "program" is a "reserved word".
Line: 1, 1st char: 9, "test" is a "ID".
Line: 1, 1st char: 13, ";" is a "symbol".
Line: 2, 1st char: 1, "var" is a "reserved word".
Line: 3, 1st char: 3, "3i" is an invalid "ID".
Line: 3, 1st char: 6, ":" is a "symbol".
Line: 3, 1st char: 8, "string" is a "reserved word".
Line: 3, 1st char: 14, ";" is a "symbol".
Line: 4, 1st char: 1, "begin" is a "reserved word".
Line: 5, 1st char: 3, "3i" is an invalid "ID".
Line: 5, 1st char: 6, ":=" is a "symbol".
Line: 5, 1st char: 9, "'ab" is an invalid "string".
Line: 5, 1st char: 12, ";" is a "symbol".
Line: 6, 1st char: 1, "end" is a "reserved word".
Line: 6, 1st char: 4, ";" is a "symbol".
```

測資 3. pas

```
yang@ubuntu-virtual-machine:~/Compiler/testfile_lab1_2020/example$ ./a.out < 3.pas
Line: 1, 1st char: 1, "(* comment 1
   comment 2 *)" is a "comment".
Line: 3, 1st char: 1, "program" is a "reserved word".
Line: 3, 1st char: 9, "test" is a "ID".
Line: 3, 1st char: 13, ";" is a "symbol".
Line: 4, 1st char: 1, "var" is a "reserved word".
Line: 5, 1st char: 3, "i" is a "ID".
Line: 5, 1st char: 5, ":" is a "symbol".
Line: 5, 1st char: 7, "integer" is a "reserved word".
Line: 5, 1st char: 14, ";" is a "symbol".
Line: 6, 1st char: 1, "begin" is a "reserved word".
Line: 7, 1st char: 3, "read" is a "reserved word".
Line: 7, 1st char: 7, "(" is a "symbol".
Line: 7, 1st char: 8, "i" is a "ID".
Line: 7, 1st char: 9, ")" is a "symbol".
Line: 7, 1st char: 10, ";" is a "symbol".
Line: 8, 1st char: 1, "end" is a "reserved word".
Line: 8, 1st char: 4, ";" is a "symbol".
```

測資 4. pas

```
yang@ubuntu-virtual-machine:~/Compiler/testfile_lab1_2020/example$ ./a.out < 4.pas
Line: 1, 1st char: 1, "program" is a "reserved word".
Line: 1, 1st char: 9, "test" is a "ID".
Line: 1, 1st char: 13, ";" is a "symbol".
Line: 2, 1st char: 1, "var" is a "reserved word".
Line: 3, 1st char: 3, "f" is a "ID".
Line: 3, 1st char: 5, ":" is a "symbol".
Line: 3, 1st char: 7, "float" is a "reserved word".
Line: 3, 1st char: 12, ";" is a "symbol".
Line: 4, 1st char: 1, "begin" is a "reserved word".
Line: 5, 1st char: 3, "f" is a "ID".
Line: 5, 1st char: 5, ":=" is a "symbol".
Line: 5, 1st char: 8, "12.25e+6" is a "real".
Line: 5, 1st char: 16, ";" is a "symbol".
Line: 6, 1st char: 1, "end" is a "reserved word".
Line: 6, 1st char: 4, ";" is a "symbol".
```

測資 5.pas

```
yang@ubuntu-virtual-machine:~/Compiler/testfile_lab1_2020/example$ ./a.out < 5.pas
Line: 1, 1st char: 1, "(* a**b) *)" is a "comment".
Line: 2, 1st char: 1, "program" is a "reserved word".
Line: 2, 1st char: 9, "test" is a "ID".
Line: 2, 1st char: 13, ";" is a "symbol".
Line: 3, 1st char: 1, "var" is a "reserved word".
Line: 4, 1st char: 3, "i" is a "ID".
Line: 4, 1st char: 5, ":" is a "symbol".
Line: 4, 1st char: 7, "integer" is a "reserved word".
Line: 4, 1st char: 14, ";" is a "symbol".
Line: 5, 1st char: 3, "_s" is a "ID".
Line: 5, 1st char: 5, "," is a "symbol".
Line: 5, 1st char: 7, "_s2" is a "ID".
Line: 5, 1st char: 10, "," is a "symbol".
Line: 5, 1st char: 12, "_s3" is a "ID".
Line: 5, 1st char: 15, "," is a "symbol".
Line: 5, 1st char: 17, "_s4" is a "ID".
Line: 5, 1st char: 20, "," is a "symbol".
Line: 5, 1st char: 22, "_s5" is a "ID".
Line: 5, 1st char: 26, ":" is a "symbol".
Line: 5, 1st char: 28, "string" is a "reserved word".
Line: 5, 1st char: 34, ";" is a "symbol".
Line: 6, 1st char: 1, "begin" is a "reserved word".
Line: 7, 1st char: 3, "i" is a "ID".
Line: 7, 1st char: 5, ":=" is a "symbol".
Line: 7, 1st char: 8, "-100" is a "integer".
Line: 7, 1st char: 12, ";" is a "symbol".
Line: 8, 1st char: 3, "_s" is a "ID".
Line: 8, 1st char: 6, ":=" is a "symbol".
Line: 8, 1st char: 9, "'db lab'" is a "string".
Line: 8, 1st char: 17, ";" is a "symbol".
Line: 9, 1st char: 3, "_s2" is a "ID".
Line: 9, 1st char: 7, ":=" is a "symbol".
Line: 9, 1st char: 10, "'You''ll see'" is a "string".
Line: 9, 1st char: 23, ";" is a "symbol".
Line: 10, 1st char: 3, "_s3" is a "ID".
Line: 10, 1st char: 7, ":=" is a "symbol".
Line: 10, 1st char: 10, "''" is a "string".
Line: 10, 1st char: 12, ";" is a "symbol".
Line: 11, 1st char: 3, "_s4" is a "ID".
Line: 11, 1st char: 7, ":=" is a "symbol".
Line: 11, 1st char: 10, "'''''" is a "string".
Line: 11, 1st char: 14, ";" is a "symbol".
Line: 12, 1st char: 3, "_s5" is a "ID".
Line: 12, 1st char: 7, ":=" is a "symbol".
Line: 12, 1st char: 10, "' '" is a "string".
Line: 12, 1st char: 13, ";" is a "symbol".
Line: 13, 1st char: 1, "end" is a "reserved word".
Line: 13, 1st char: 4, ":" is a "symbol".
```

測資 6.pas

```
yang@ubuntu-virtual-machine:~/Compiler/testfile_lab1_2020/example$ ./a.out < 6.pas
Line: 1, 1st char: 1, "ProGram" is a "reserved word".
Line: 1, 1st char: 9, "test" is a "ID".
Line: 1, 1st char: 13, ";" is a "symbol".
Line: 2, 1st char: 1, "var" is a "reserved word".
Line: 3, 1st char: 3, "#db" is an invalid "ID".
Line: 3, 1st char: 7, ":" is a "symbol".
Line: 3, 1st char: 9, "float" is a "reserved word".
Line: 3, 1st char: 14, ";" is a "symbol".
Line: 4, 1st char: 3, "_f2" is a "ID".
Line: 4, 1st char: 7, ":" is a "symbol".
Line: 4, 1st char: 9, "float" is a "reserved word".
Line: 4, 1st char: 14, ";" is a "symbol".
Line: 5, 1st char: 1, "begin" is a "reserved word".
Line: 6, 1st char: 3, "#db" is an invalid "ID".
Line: 6, 1st char: 7, ":=" is a "symbol".
Line: 6, 1st char: 10, ".1" is an invalid "real".
Line: 6, 1st char: 12, ";" is a "symbol".
Line: 7, 1st char: 3, "_f2" is a "ID".
Line: 7, 1st char: 7, ":=" is a "symbol".
Line: 7, 1st char: 10, "12.100" is an invalid "real".
Line: 7, 1st char: 16, ";" is a "symbol".
Line: 8, 1st char: 1, "end" is a "reserved word".
Line: 8, 1st char: 4, ";" is a "symbol".
```

測資 7.pas

```
yang@ubuntu-virtual-machine:~/Compiler/testfile_lab1_2020/example$ ./a.out < 7.pas
Line: 1, 1st char: 1, "(* This line is a comment. *)" is a "comment".
Line: 2, 1st char: 1, "program" is a "reserved word".
Line: 2, 1st char: 9, "test" is a "ID".
Line: 2, 1st char: 13, ";" is a "symbol".
Line: 3, 1st char: 1, "var" is a "reserved word".
Line: 4, 1st char: 3, "i" is a "ID".
Line: 4, 1st char: 5, ":" is a "symbol".
Line: 4, 1st char: 7, "integer" is a "reserved word".
Line: 4, 1st char: 14, ";" is a "symbol".
Line: 5, 1st char: 1, "begin" is a "reserved word".
Line: 6, 1st char: 3, "i" is a "ID".
Line: 6, 1st char: 5, ":=" is a "symbol".
Line: 6, 1st char: 8, "1" is a "integer".
Line: 6, 1st char: 9, "+" is a "symbol".
Line: 6, 1st char: 10, "2" is a "integer".
Line: 6, 1st char: 8, ";" is a "symbol".
Line: 7, 1st char: 1, "end" is a "reserved word".
Line: 7, 1st char: 4, ";" is a "symbol".
```

7. 加分部分(額外用自己的測資　如下圖)

自己多想了很多不同的錯誤處理，例如在 21-23 行中，不只能完成 1+1 為 int、 symbol、int，也能完成 1.0e+1+1.0e+1 的 real、symbol、real 的型式，即便在 1 +1 中有一個空白，仍能表示+為 symbol 而不會因為有空白變成兩個 integer。另外，我認為在我自己的測資中第 43-45 行的部分是較多人容易忽略的，因為在 comment 中可能會有換行，而我們就必須在接下來的 Line number 中多+1，且若有換行後的 comment 結束，後面還有其他字串，也應該有對應正確的 1st char number。

lab1TEST.txt

```
(* (* *) *)
001
(* + (* + *) + *) - *) - *) (* +++ *)
' qfqfqf (* *) '
(* '''' qfqfqf '' *)
123.456E+6.7
12+34
12
+34
1a
^db
#db
abcdefghijklmnopqrstuvwxyz12345
peter
_db
a1
1+2
1e+2
2E2
1.1
1 +1
1.0+1.0
1.0e+1+1.0e+1
1.0
3.14
7E-2
12.25e+6
-7.5E+3
1.00
03.0
12.100
.1
1.
''
'''
asdf
' '
'You''ll see'
'ab;
ab'
********)
(* comment *)(*asdf*)
(* comment
dfeghhhhhh
second line *) abcdefg 12
(*****)
(* a**b) *)
(*ab*)**)
123
asdfsadf
'''''
'absadlkaskdfhasdfasdfsdfjksdjfkjsdkfsdf'
(*asdf(*asdf*)
(*******asdf
absolute and begin break case const continue do else end
12++60
0123.111000E+123
0123.11E+789
123.12300e+789.123
12.100
0123.123E+0123
```

自己的測資測試結果：

```
yang@ubuntu-virtual-machine:~/Compiler/testfile_lab1_2020/example$ ./a.out < lab1TEST.txt
Line: 1, 1st char: 1, "(* (* *)" is a "comment".
Line: 1, 1st char: 1, "*)" is an invalid "comment".
Line: 2, 1st char: 1, "001" is an invalid "integer".
Line: 3, 1st char: 1, "(* + (* + *)" is a "comment".
Line: 3, 1st char: 1, "+" is a "symbol".
Line: 3, 1st char: 3, "*)" is an invalid "comment".
Line: 3, 1st char: 6, "-" is a "symbol".
Line: 3, 1st char: 8, "*)" is an invalid "comment".
Line: 3, 1st char: 11, "-" is a "symbol".
Line: 3, 1st char: 13, "*)" is an invalid "comment".
Line: 3, 1st char: 16, "(* +++ *)" is a "comment".
Line: 4, 1st char: 1, "' qfqfqf (* *) '" is a "string".
Line: 5, 1st char: 1, "(* '''' qfqfqf '' *)" is a "comment".
Line: 6, 1st char: 1, "123.456E+6" is a "real".
Line: 6, 1st char: 11, ".7" is an invalid "real".
Line: 7, 1st char: 1, "12" is a "integer".
Line: 7, 1st char: 3, "+" is a "symbol".
Line: 7, 1st char: 4, "34" is a "integer".
Line: 8, 1st char: 1, "12" is a "integer".
Line: 9, 1st char: 1, "+34" is a "integer".
Line: 10, 1st char: 1, "1a" is an invalid "ID".
Line: 11, 1st char: 1, "^db" is an invalid "ID".
Line: 12, 1st char: 1, "#db" is an invalid "ID".
Line: 13, 1st char: 1, "abcdefghijklmnopqrstuvwxyz12345" is an invalid "ID".
Line: 14, 1st char: 1, "peter" is a "ID".
Line: 15, 1st char: 1, "_db" is a "ID".
Line: 16, 1st char: 1, "a1" is a "ID".
Line: 17, 1st char: 1, "1" is a "integer".
Line: 17, 1st char: 2, "+" is a "symbol".
Line: 17, 1st char: 3, "2" is a "integer".
Line: 18, 1st char: 1, "1e+2" is a "real".
Line: 19, 1st char: 1, "2E2" is an invalid "ID".
Line: 20, 1st char: 1, "1.1" is a "real".
Line: 21, 1st char: 1, "1" is a "integer".
Line: 21, 1st char: 3, "+" is a "symbol".
Line: 21, 1st char: 4, "1" is a "integer".
Line: 22, 1st char: 1, "1.0" is a "real".
Line: 22, 1st char: 4, "+" is a "symbol".
Line: 22, 1st char: 5, "1.0" is a "real".
Line: 23, 1st char: 1, "1.0e+1" is a "real".
Line: 23, 1st char: 7, "+" is a "symbol".
Line: 23, 1st char: 8, "1.0e+1" is a "real".
Line: 24, 1st char: 1, "1.0" is a "real".
Line: 25, 1st char: 1, "3.14" is a "real".
Line: 26, 1st char: 1, "7E-2" is a "real".
Line: 27, 1st char: 1, "12.25e+6" is a "real".
Line: 28, 1st char: 1, "-7.5E+3" is a "real".
```

```
Line: 29, 1st char: 1, "1.00" is an invalid "real".
Line: 30, 1st char: 1, "03.0" is an invalid "real".
Line: 31, 1st char: 1, "12.100" is an invalid "real".
Line: 32, 1st char: 1, ".1" is an invalid "real".
Line: 33, 1st char: 1, "1." is an invalid "real".
Line: 34, 1st char: 1, "''" is a "string".
Line: 35, 1st char: 1, "''" is a "string".
Line: 35, 1st char: 3, "'" is an invalid "string".
Line: 36, 1st char: 1, "asdf" is a "ID".
Line: 37, 1st char: 1, "' '" is a "string".
Line: 38, 1st char: 1, "'You''ll see'" is a "string".
Line: 39, 1st char: 1, "'ab" is an invalid "string".
Line: 39, 1st char: 4, ";" is a "symbol".
Line: 40, 1st char: 1, "ab'" is an invalid "string".
Line: 41, 1st char: 1, "********)" is an invalid "comment".
Line: 42, 1st char: 1, "(* comment *)" is a "comment".
Line: 42, 1st char: 0, "(*asdf*)" is a "comment".
Line: 43, 1st char: 1, "(* comment
dfeghhhhhh
second line *)" is a "comment".
Line: 45, 1st char: 16, "abcdefg" is a "ID".
Line: 45, 1st char: 24, "12" is a "integer".
Line: 46, 1st char: 1, "(*****)" is a "comment".
Line: 47, 1st char: 1, "(* a**b) *)" is a "comment".
Line: 48, 1st char: 1, "(*ab*)" is a "comment".
Line: 48, 1st char: 15, "**)" is an invalid "comment".
Line: 49, 1st char: 1, "123" is a "integer".
Line: 50, 1st char: 1, "asdfsadf" is a "ID".
Line: 51, 1st char: 1, "''''" is a "string".
Line: 51, 1st char: 5, "'" is an invalid "string".
Line: 52, 1st char: 1, "'absadlkaskdfhasdfasdfsdfjksdjfkjsdkfsdf'" is an invalid "string".
Line: 53, 1st char: 1, "(*asdf(*asdf*)" is a "comment".
Line: 54, 1st char: 1, "(*******asdf" is an invalid "comment".
Line: 55, 1st char: 1, "absolute" is a "reserved word".
Line: 55, 1st char: 10, "and" is a "reserved word".
Line: 55, 1st char: 14, "begin" is a "reserved word".
Line: 55, 1st char: 20, "break" is a "reserved word".
Line: 55, 1st char: 26, "case" is a "reserved word".
Line: 55, 1st char: 31, "const" is a "reserved word".
Line: 55, 1st char: 37, "continue" is a "reserved word".
Line: 55, 1st char: 46, "do" is a "reserved word".
Line: 55, 1st char: 49, "else" is a "reserved word".
Line: 55, 1st char: 54, "end" is a "reserved word".
Line: 56, 1st char: 1, "12" is a "integer".
Line: 56, 1st char: 3, "+" is a "symbol".
Line: 56, 1st char: 4, "+60" is a "integer".
Line: 57, 1st char: 1, "0123.111000E+123" is an invalid "real".
Line: 58, 1st char: 1, "0123.11E+789" is an invalid "real".
```

```
Line: 59, 1st char: 1, "123.12300e+789" is an invalid "real".
Line: 59, 1st char: 15, ".123" is an invalid "real".
Line: 60, 1st char: 1, "12.100" is an invalid "real".
Line: 61, 1st char: 1, "0123.123E+0123" is an invalid "real".
```