Phase 1: Scanning

In this first phase, your job is to implement a scanner for the SnuPL/1 language as specified in the term project overview.

The input to the scanner is a character stream. The output of the scanner is a stream of tokens. The scanner must correctly recognize and tokenize keywords, identifiers, numbers, operators (assignment, binary and relational), comments, and the syntax elements of the language. The scanner should scan the input until the end of the input character stream is reached or an error is detected.

The scanner must scan but ignore single-line comments. Character constants enclosed in single quotes and strings (enclosed in double quotes) must be recognized as a whole and returned to the parser as a single token with the token value set to the single character/character sequence. Make sure to properly scan escape sequences. Numbers must also be scanned as a whole and returned to the parser as a single token. The conversion string → number is done in the parser (i.e., you should simple return the lexeme making up the number).

We provide a skeleton for a scanner/parser framework so that you can focus on the interesting parts. In addition, we also provide a full working example for "SnuPL/-1" that shows you how to use the framework (the EBNF for SnuPL/-1 is provided below).

The scanner skeleton can be found in the files snuplc/src/scanner.[h/cpp]. In its unmodified form, the scanner scans SnuPL/-1.

The header file first defines the token types (EToken); two corresponding data structures (ETokenName and ETokenStr) are located in the C++ file. Here, you will need to add additional tokens to implement SnuPL/0. ETokenName is used to print the token type only; ETokenStr prints the token name along with the lexeme. The elements of ETokenStr are fed to printf, so you can print the lexeme by inserting a '%s' placeholder somewhere in that string.

The token class (CToken) is fully implemented and functional, you do not have to modify it. The scanner (CScanner) is also fully functional, but only recognizes SnuPL/-1. You will need to modify the function CToken* Scanner::Scan() to accept all possible tokens of SnuPL/1.

The sources also contain a simple test program for the scanner. It creates a scanner instance and repeatedly calls CScanner::Get() to retrieve the next token until the end of file is reached. Retrieved tokens are printed to standard out.

Run

snuplc $ make test_scanner

to build it. To invoke it run it with a file name as an argument:

snuplc $ ./test_scanner ../test/scanner/test01.mod

In the directory test/scanner/ you can find a number of test files for the scanner. We advise you to create your own test cases to test special cases; we have our own set of test files to test (and grade) your scanner.

Hints: the first phase is pretty straight-forward to implement. Two points are noteworthy:

- error recovery: unrecognized lexemes should be handled by returning a tUndefined token with the illegal lexeme or a custom error message as its attribute
- handling of comment and whitespace: consume all whitespace and comments in the scanner (i.e., do not return a token for whitespace or comment)

Materials to submit:

- source code of the scanner (use Doxygen-style comments)
- a report describing your implementation of the scanner (PDF)
  (the report is almost as important as your code. Make sure to put sufficient effort into it!)

Submission:

- the deadline for the first phase is **Friday, September 22, 2017 at 14:00**.
- email your submission to the TA (compiler@csap.snu.ac.kr). The arrival time of your email counts as the submission time.

Do not hesitate to ask questions in class/on eTL; implementing a compiler is not an easy task. Also, start as soon as possible; if you wait until a few days before the deadline we cannot help you much and you may not be able to finish in time.

Happy coding!

Appendix: EBNF Syntax Definition of SnuPL/-1

```
module          =  statSequence ".".

digit           =  "0".."9".
letter          =  "a".."z".

factOp          =  "*" | "/"
termOp          =  "+" | "-"
relOp           =  "=" | "#"

factor          =  digit | "(" expression ")".
term            =  factor { factOp factor }.
simpleexpr      =  term { termOp term }.
expression      =  simpleexpr [ relOp simplexpr ].

assignment      =  letter ":=" expression.
statement       =  assignment.
statSequence    =  [ statement { ";" statement } ].

whitespace      =  { " " | \n }+.
```