CSE 304-Compiler Lab

Assignment 2 (Tokenization)

Department of CSE, MIST

---

Tokenization is a way of separating a piece of text into smaller units called tokens. In this lab, you will have to tokenize a sample C source code.

**Token:** A token is a pair consisting of a token name and an optional attribute value. <TOKEN, ATTRIBUTE>
**Lexeme:** A Lexeme is a sequence of characters (actual character set)
**Pattern:** A pattern is a description of the form that the lexemes of a token may take.
**Symbol Table:** A symbol-table is a data structure maintained by compilers in order to store information about the occurrence of various identifier, functions, objects etc.
**Lexical error:** if any lexeme do not match with any pattern described.

**Tasks:**

1. Scan the input program and identify Tokens
2. Insert tokens into Symbol Table, print the whole symbol table in console for each insertion
3. Generate different files for different Tokens mentioning the lexeme and its line number
4. Generates lexical errors with line number and print it in the console

| Serial | Token | Tokens to be handled |
|---|---|---|
| 1 | KEYWORD | Identify the following keywords<br>`if, else, for, while, break, int, char, float, double` |
| 2 | FUNCTION | Identify functions: For all types of function calling and declarations. |
| 3 | IDENTIFIER | Identify identifiers |
| 4 | OPERATOR | Identify arithmetic, logical, assignment operators :<br>Arithmetic operators : `+, -, *, /, %`<br>Logical operators: `&&, ||, ==, !=, !`<br>Assignment operators: `=, +=, -=, *=, /=, %=` |

**Note:** First you have to check Keywords in your code. Do not check Function name before it. If you check function name before keyword, then "**if ( )**" will be detected as a function name.

****HANDLE THE ABOVE MENTIONED OPERATORS ONLY ****