CPSC 323 Compilers & Languages (Spring 2024)

PROJECT 1

Deadline: March 6th, 2024

The first project assignment is to write a lexical analyzer (Lexer).

Lexical analysis is the process of reading the stream of characters making up the source code of a program and dividing the input into tokens.

The Lexer:

Writing a procedure (Function) - lexer () that returns a token when it's required will be a key part of your assignment. Your lexer() function should return a record with two fields: one for the token and one for the token's actual "value" (lexeme), or the instance of a token.

Write a lexer from scratch that returns the tokens from the simple source code in C/C++/Java/Python in the given file “input_scode.txt”.

1. You need to design and implement FSA for tokens

   (a) identifier
   (b) integer

   Otherwise, there will be a deduction of 3 points! You must also write regular expressions (REs) and draw the corresponding FSA for the required tokens (identifier and integer) and save this information in a document named "FSA_mydesign.doc".

   2. Your executable program, which is supposed to be developed in C/C++/Java/Python, should represent the implementation of your idea. To read and return the subsequent token, use the lexer () method.

   3. Print out the result into two columns, one for tokens and another for lexemes, and save it into a document named as “output” (see an example I/O as below).

   4. You must write a “readme” file to briefly specify documentation & how to setup/run your program if needed.

   5. Your submission must have Five (5) files: the given “input_scode.txt”, your design file, your program file, output file, and a readme file.

Your program should read a file containing the source code of input_scode.txt to generate tokens and write out the results to an output file. Make sure that you print both the tokens, and lexemes.

NOTE: Please keep in mind that you have the freedom to define your token classes, names and associate them with their respective lexemes.

Example: The lexer's job is to convert the original text into a series of tokens by removing newlines, blank spaces, and comments. Please review the input and output samples that are given below.

Input Source code (input_scode.txt)

while (k < lower) s = 33.00;

Output:

token	Lexeme
keyword	while
Separator	(
Identifier	k
Operator	<
Identifier	lower
Separator	)
Identifier	s
Operator	=
Real	33.00
Separator	;