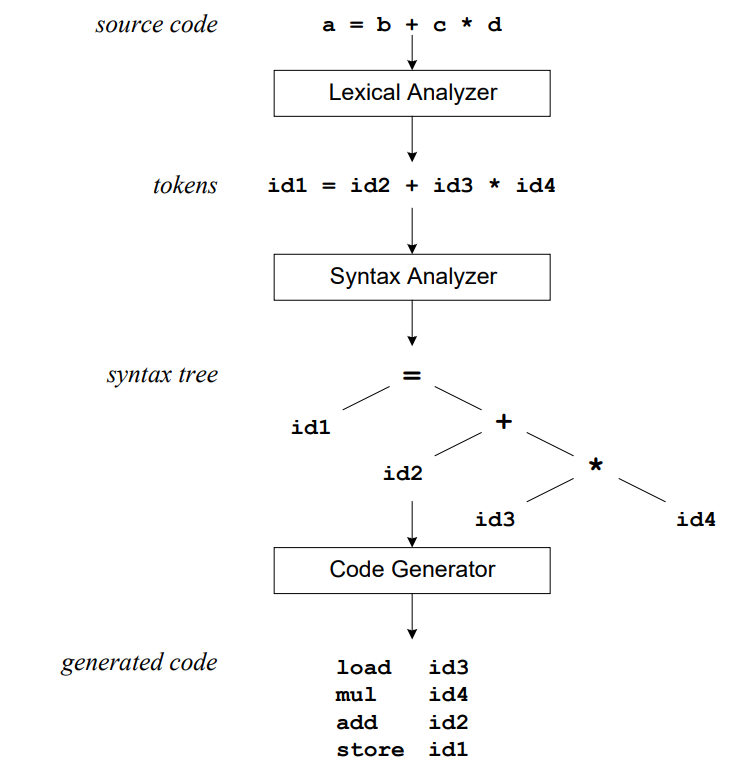
**LAB SESSION 1: INTRODUCTION TO LEX**

**AIM**: To implement a basic program using the Lex Tool.

**PROBLEM DEFINITION:** Implement the following programs:

1. A LEX program to find if the input is integer, real number or word
2. A LEX program to convert decimal numbers to hexadecimal numbers.
3. A Lex program to include line numbers in a given source program
4. A LEX program to compute the average of a given set of numbers.

**THEORY:** The unix utility lex parses a file of characters. It uses regular expression matching; typically it is used to ‘tokenize’ the contents of the file. In that context, it is often used together with the yacc utility. However, there are many other applications possible.

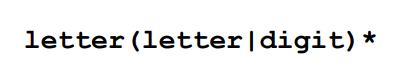


Lex generates C code for a lexical analyzer, or scanner. It uses patterns that match strings in the input and converts the strings to tokens. Tokens are numerical representations of strings, and simplify processing. This is illustrated in Figure.

As lex finds identifiers in the input stream, it enters them in a symbol table. The symbol table may also contain other information such as data type (integer or real) and location of the variable in memory. All subsequent references to identifiers refer to the appropriate symbol table index.

The first phase in a compiler reads the input source and converts strings in the source to tokens. Using regular expressions, we can specify patterns to lex that allow it to scan and match strings in the input. Each pattern in lex has an associated action. Typically an action returns a token, representing the matched string, for subsequent use by the parser.

To begin with, however, we will simply print the matched string rather than return a token value. We may scan for identifiers using the regular expression.



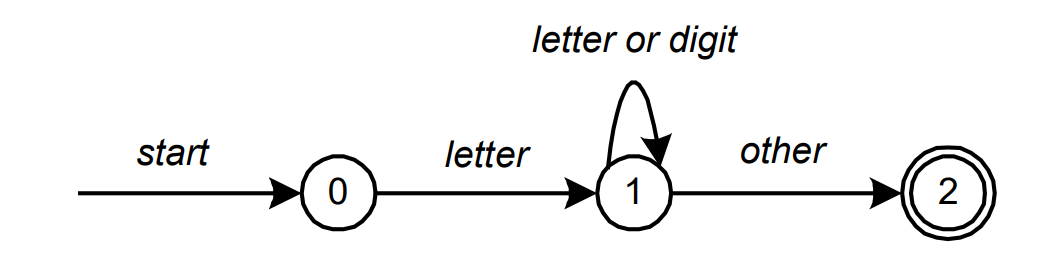
This pattern matches a string of characters that begins with a single letter, and is followed by zero or more letters or digits. This example nicely illustrates operations allowed in regular expressions:

• repetition, expressed by the “\*” operator

• alternation, expressed by the “|” operator

• concatenation

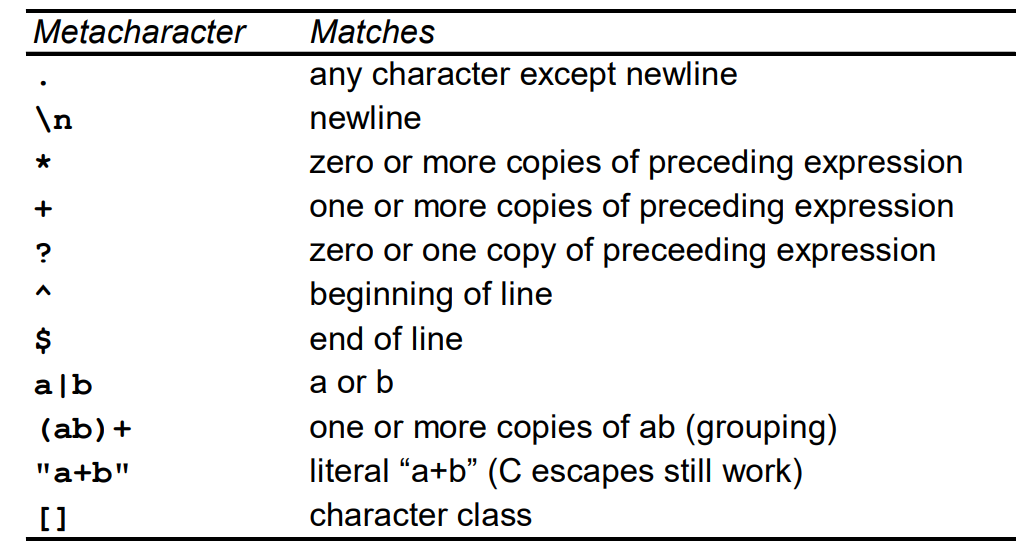
Any regular expressions may be expressed as a finite state automaton (FSA). We can represent an FSA using states, and transitions between states. There is one start state, and one or more final or accepting states.



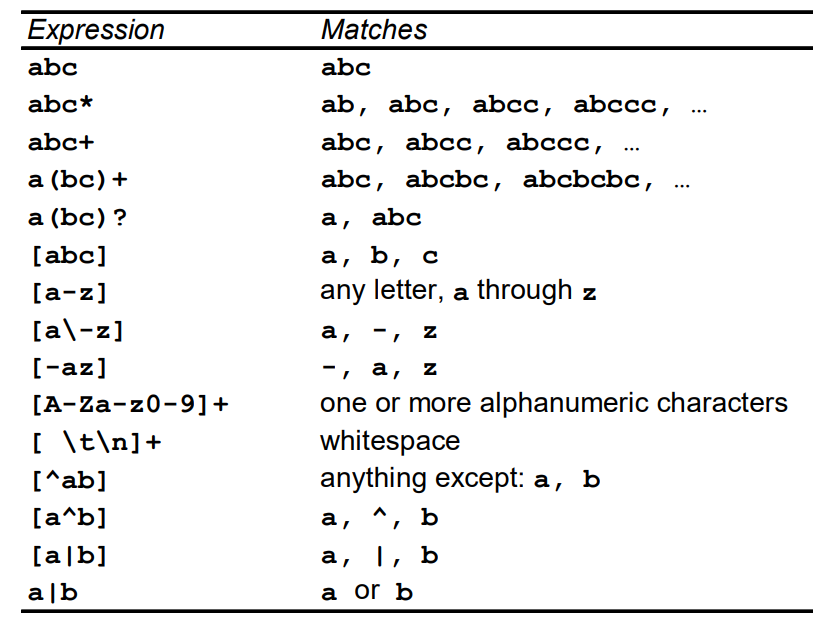
In Figure, state 0 is the start state, and state 2 is the accepting state. As characters are read, we make a transition from one state to another. When the first letter is read, we transition to state 1. We remain in state 1 as more letters or digits are read. When we read a character other than a letter or digit, we transition to state 2, the accepting state. Any FSA may be expressed as a computer program.

This is the technique used by lex. Regular expressions are translated by lex to a computer program that mimics an FSA. Using the next input character, and current state, the next state is easily determined by indexing into a computer-generated state table.

Now we can easily understand some of Lex's limitations. For example, lex cannot be used to recognize nested structures such as parentheses. Nested structures are handled by incorporating a stack. Whenever we encounter a “(”, we push it on the stack. When a “)” is encountered, we match it with the top of the stack, and pop the stack. Lex, however, only has states and transitions between states. Since it has no stack, it is not well suited for parsing nested structures. Yacc augments an FSA with a stack, and can process constructs such as parentheses with ease. The important thing is to use the right tool for the job. Lex is good at pattern matching. Yacc is appropriate for more challenging tasks.



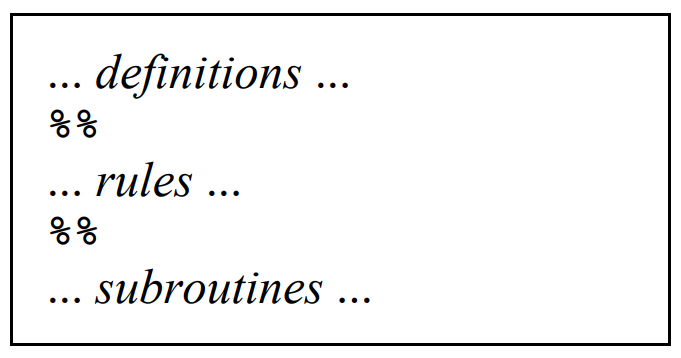
Metacharacters in Lex



Sample Regular Expressions in Lex

Regular expressions in lex are composed of metacharacters. Pattern matching examples are shown in Table. Within a character class, normal operators lose their meaning. Two operators allowed in a character class are the hyphen (“-”) and circumflex (“^”). When used between two characters, the hyphen represents a range of characters.

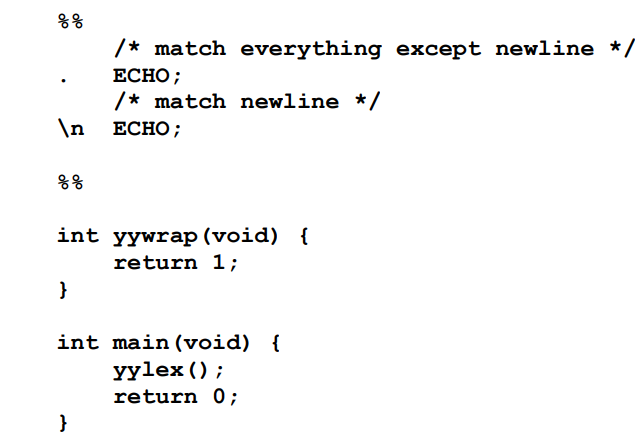
The circumflex, when used as the first character, negates the expression. If two patterns match the same string, the longest match wins. In case both matches are the same length, then the first pattern listed is used.



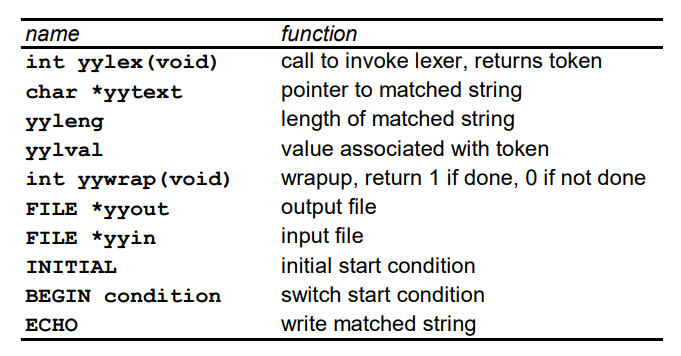
Input to Lex is divided into three sections, with %% dividing the sections. This is best illustrated by example. The first example is the shortest possible lex file: %%

Input is copied to output, one character at a time. The first %% is always required, as there must always be a rules section. However, if we don’t specify any rules, then the default action is to match everything and copy it to output. Defaults for input and output are stdin and stdout, respectively.

Here is the same example, with defaults explicitly coded:



Two patterns have been specified in the rules section. Each pattern must begin in column one. This is followed by whitespace (space, tab or newline), and an optional action associated with the pattern. The action may be a single C statement, or multiple C statements enclosed in braces. Anything not starting in column one is copied verbatim to the generated C file. We may take advantage of this behavior to specify comments in our lex file. In this example there are two patterns, “.” and “\n”, with an ECHO action associated for each pattern. Several macros and variables are predefined by lex. ECHO is a macro that writes code matched by the pattern. This is the default action for any unmatched strings.



Pre defined variables in Lex

**How to execute**

1. Type lex lexfile.l
2. Type gcc lex.yy.c.
3. Type ./a.exe

**PROGRAM**

**OUTPUT**

**CONCLUSION**