

# **CS224-Theory of Automata**

**Project Report:** Lexical Analyzer

#### **Group Members:**

Raja Bilal Khurram (2023591)

Muhammad Ammar Saleem (2023378)

Course: CS224 - Theory of Automata

Submitted to Sajjid Ali

# C++ Tokenizer

#### Introduction

A **Lexical Analyzer** (**LA**) is a fundamental component of a compiler's front-end. It processes the source code to identify and classify sequences of characters into meaningful tokens, such as keywords, identifiers, operators, and literals. This tokenization simplifies the parsing process and aids in syntax and semantic analysis.

In this project, we have developed a Lexical Analyzer for the **C++ language** using **Flex**, a tool for generating scanners. Our analyzer reads C++ source files, identifies various tokens, and outputs their type, value, and line number.

### **Features of Our Lexical Analyzer**

- Recognition of C++ keywords (e.g., int, float, if, else, return)
- Identification of **identifiers**, **constants** (integers, floats, characters, strings)
- Parsing of **operators**: arithmetic, logical, bitwise, assignment, relational
- Detection of **punctuation marks** and **delimiters** (;, ,, (), {}, [])
- Handling of **comments**: single-line (//) and multi-line (/\*...\*/)
- Ignoring of whitespace while maintaining accurate line numbers
- Output of **token type**, **value**, and **line number** to the console

### **Regular Expressions for Token Classes**

digit [0-9] letter [A-Za-z\_]

identifier  ${\{letter\}|\{digit\}\}}^*$ 

data\_type int|float|double|char|bool|string|long|short|void

accessspecifier private|protected|public

keyword

if | else | while | for | return | break | continue | switch | case | default | size of | do | goto | enum | type def | struct | class | const| static | volatile | signed | unsigned | try | catch | throw | new | delete | const| static | volatile | signed | unsigned | try | catch | throw | new | delete | const| static | volatile | signed | unsigned | try | catch | throw | new | delete | const| static | volatile | signed | unsigned | try | catch | throw | new | delete | const| static | volatile | signed | try | catch | throw | t

```
brackets
                   [\{\}\[\]\(\)]
line comment
                     \/\/[^\n]*
                      "\\([^]|\+[^/])\+\/"
block comment
integer_constant
                     {digit}+
float constant
                    {digit}+"."{digit}+
character_constant \langle ([^{\}]] \rangle
                  \"([^\\\"]|\\.)*\"
string literal
library
               \#include[ \t]*<[^>]+>
delimiter
                 [;,]
assignment_operator =
arithmetic_operator [+\-*/%]
relational_operator (==|!=|<=|>=|<|>)
logical_operator
                     (||\&\&|!)
bitwise_operator
                     (\&|\cdot||\cdot|<<|>>)
increment_decrement (\+\+\-)
```

# **Sample Code and Output Explanation**

#### Sample Input (test1.cpp)

```
int main() {
  float x = 3.14;
  // This is a comment
  if (x > 0) {
    x = x + 1;
  }
  return 0;
}
```

#### **Sample Output**

```
Line 1: Token = int → Data Type
Line 1: Token = main → Identifier
Line 1: Token = ( → Bracket/Parenthesis
Line 1: Token = ) \rightarrow Bracket/Parenthesis
Line 1: Token = { → Bracket/Parenthesis
Line 2: Token = float → Data Type
Line 2: Token = x \rightarrow Identifier
Line 2: Token = \Rightarrow Assignment Operator
Line 2: Token = 3.14 \rightarrow \text{Float Constant}
Line 2: Token = ; \rightarrow Delimiter
Line 3: Token = // This is a comment → Line Comment
Line 4: Token = if → Keyword
Line 4: Token = ( → Bracket/Parenthesis
Line 4: Token = x \rightarrow Identifier
Line 4: Token = \rightarrow Relational Operator
Line 4: Token = 0 \rightarrow Integer Constant
Line 4: Token = ) → Bracket/Parenthesis
```

```
Line 4: Token = { → Bracket/Parenthesis

Line 5: Token = x → Identifier

Line 5: Token = = → Assignment Operator

Line 5: Token = x → Identifier

Line 5: Token = + → Arithmetic Operator

Line 5: Token = 1 → Integer Constant

Line 5: Token = ; → Delimiter

Line 6: Token = } → Bracket/Parenthesis

Line 7: Token = return → Keyword

Line 7: Token = 0 → Integer Constant

Line 7: Token = ; → Delimiter

Line 8: Token = } → Bracket/Parenthesis
```

# **Main Code Explanation**

#### int main Function

```
int main(int argc, char **argv) {
    if (argc < 2) {
        fprintf(stderr, "Usage: %s <source-file>\\n", argv[0]);
        return EXIT_FAILURE;
    }
    yyin = fopen(argv[1], "r");
    if (!yyin) {
        perror("fopen");
        return EXIT_FAILURE;
    }
    yylex();
    fclose(yyin);
    return EXIT_SUCCESS;
}
```

This function checks for the input source file passed as a command-line argument. It opens the file and passes it to Flex's yylex() function, which begins tokenizing. After processing, the file is closed.

### **Use of Theory of Automata in Our Project**

In our Lexical Analyzer project, we leveraged several concepts from the theory of automata to effectively tokenize and classify C++ source code. The theory of automata provides the foundation for understanding how regular expressions can be transformed into finite state machines (FSMs), which are essential for recognizing patterns in the input text. Specifically, we utilized **Deterministic Finite Automata (DFA)** to process the regular expressions corresponding to different token classes (such as identifiers, keywords, operators, etc.).

Each token type in the C++ language can be described by a regular expression, and the Lex tool, which we used for the project, automatically constructs an NFA (Non-deterministic Finite Automaton) from these regular expressions. This NFA is then converted into an efficient DFA that can quickly match patterns in the input text. The use of DFA ensures that we process each character of the input source code exactly once, making the lexical analysis both time and space-efficient.

Additionally, the concept of **state transitions** in automata was directly applied to track the progress of the lexical analyzer as it reads characters from the input. For example, when processing identifiers, the automaton starts in an initial state and transitions through various states as it encounters valid characters (letters and digits), ultimately recognizing a valid identifier.

By applying automata theory, we ensured that our lexical analyzer is capable of efficiently recognizing and classifying tokens while maintaining the order and structure of the source code. This implementation also helps handle complex constructs such as nested comments and multicharacter operators, showcasing the power of automata theory in real-world applications like compiler design.

#### **Conclusion**

In conclusion, this project successfully demonstrates the development of a Lexical Analyzer for the C++ programming language using Flex. By defining appropriate regular expressions and token patterns, we were able to accurately identify and classify a wide variety of lexical components such as keywords, identifiers, constants, operators, and delimiters. The analyzer effectively handles real-world C++ syntax, including comments and whitespace, while maintaining accurate line tracking. This foundational work plays a critical role in compiler design and provides practical insight into how source code is broken down during the initial phases of compilation.

#### Our Lexical Analyzer Code:

```
{ printf("Line %d: Line Comment = %s\n", line_no, yytext); }
                                         { printf("Line %d: Block Comment = %s\n", line_no, yytext); }
          {block comment}
                                        { printf("Line %d: Access Specifier = %s\n", line_no, yytext); }
                                         { printf("Line %d: Data Type = %s\n", line_no, yytext); }
          {keyword}
                                       { printf("Line %d: Assignment Operator = %s\n", line_no, yytext); }
          {assignment operator}
                                       { printf("Line %d: Increment/Decrement Operator = %s\n", line_no, yytext); }
                                        { printf("Line %d: Arithmetic Operator = %s\n", line_no, yytext); }
          {arithmetic operator}
          {relational_operator} { printf("Line %d: Relational Operator = %s\n", line_no, yytext); }
                                        { printf("Line %d: Logical Operator = %s\n", line_no, yytext); }
                                        { printf("Line %d: Bitwise Operator = %s\n", line no, yytext); }
          {bitwise operator}
                                        { printf("Line %d: Character Constant = %s\n", line no, yytext); }
          {character constant}
                                         { printf("Line %d: String Literal = %s\n", line_no, yytext); }
          {identifier}
         \n { 1ine no++; }
der Improve Code Share Code Link Open Website
                                                                                                                        Q Ln 36, Col 1 Spaces: 4 UTF-8 LF {} Plain Text 👸 👍 Al Code Chat 🚨
▷ 🖽 ..
       #include <stdio.h>
#include <stdlib.h>
int line_no = 1;
extern FILE *yyin;
%}
                                  [0-9]
[A-Za-z_]
{letter}{{letter}{digit}}*
int|float|double|char|bool|string|long|short|void
private|protected|public
if|else|while|for|return|break|continue|switch|case|default|sizeof|do|goto|enum|typedef|struct|class|const|static|vol
        identifier
data_type
accessspecifier
        keyword |
brackets
line_comment
block_comment
                                  [\{\}\[\]\(\)]
\/\/[^\n]*
"\/*([^*]|\*+[^*/])*\*+\/"
                                  integer_constant
float_constant
        character_constant
string_literal
library
delimiter
       delimiter [;,]
assignment_operator [+\-*/%]
relational_operator (=\!=\=\=\=\<\>)
logical_operator (\\\\\8&\!)
bitwise_operator (&\\\\\\-\<\>)
increment_decrement (\\\\\-\-)
```

#### Sample 1:

#### Sample 1 (output):

```
Line 1: Library Include = #include <iostream>
Line 2: Library Include = #include <cmath>
Line 3: Identifier = using
Line 3: Identifier = namespace
Line 3: Identifier = std
        Line 3: Delimiter = ;
       Line 5: Delimiter - ;
Line 5: Data Type - int
Line 5: Identifier - main
Line 5: Bracket/Parenthesis = (
Line 5: Bracket/Parenthesis = ]
Line 5: Bracket/Parenthesis = [
Line 6: Data Type - int
Line 6: Data Type - int
       Line 6: Data type = Int
Line 6: Assignment Operator = =
Line 6: Integer Constant = 5
Line 6: Delimiter = ;
Line 7: Data Type = float
Line 7: Identifier = b
        Line 7: Assignment Operator = = Line 7: Float Constant = 10.5
        Line 8: Data Type = double
Line 8: Identifier = c
         Line 8: Assignment Operator = =
        Line 8: Float Constant = 3.14159
Line 8: Delimiter = ;
Line 9: Data Type = char
        Line 9: Identifier = d
Line 9: Assignment Operator = =
       Line 9: Character Constant = X
Line 9: Delimiter =;
Line 10: Data Type = string
Line 10: Identifier = message
Line 10: Assignment Operator = Eline 10: String Literal = "Hello, Lexical Analyzer!"
Line 10: Delimiter = ;
Line 12: Identifier = cout
outher to: Assignment operator = =
Line 10: String Literal = "Hello, Lexical Analyzer!"
Line 10: Delimiter = ;
Line 12: Bitwise Operator = <<
Line 12: String Literal = "Value of a: "
Line 12: Bitwise Operator = <<
Line 12: Identifier = endl
Line 12: Delimiter = ;
Line 13: Bitwise Operator = <<
Line 13: String Literal = "Value of b: "
Line 13: Bitwise Operator = <<
Line 13: Identifier = endl
Line 13: Delimiter = ;
Line 14: Bitwise Operator = <<
Line 14: String Literal = "Value of c: "
Line 14: Bitwise Operator = <<
Line 14: Delimiter = ;
Line 15: Bitwise Operator = <<
Line 15: String Literal = "Character d: "
Line 15: Bitwise Operator = <<
 Line 15: Identifier = d
Line 15: Delimiter = ;
Line 17: Data Type = float
```

```
    output.txt

         Line 17: Data Type = float
         Line 17: Assignment Operator = = Line 17: Identifier = a
         Line 17: Identifier = b
         Line 17: Delimiter = ;
Line 18: Data Type = double
         Line 18: Identifier = area
         Line 18: Assignment Operator = =
         Line 18: Identifier = c
         Line 18: Arithmetic Operator = *
         Line 18: Bracket/Parenthesis = (
Line 18: Identifier = a
         Line 18: Delimiter = ,
         Line 18: Integer Constant = 2
         Line 18: Bracket/Parenthesis = )
         Line 18: Delimiter = ;
         Line 20: Bitwise Operator = <<
Line 20: String Literal = "Sum of a and b: "
Line 20: Bitwise Operator = <<
         Line 20: Identifier = sum
Line 20: Bitwise Operator = <<
Line 20: Identifier = endl
         Line 21: Bitwise Operator = <<
Line 21: String Literal = "Area of circle with radius a: "
         Line 21: Bitwise Operator = <<
         Line 21: Bitwise Operator = <<
        Line 23: Keyword = if
Line 23: Rracket/Parenthesis = (
                                                                             Desktop 2
         Line 21: Bitwise Operator = <<
         Line 21: Delimiter = ;
Line 23: Keyword = if
         Line 23: Identifier = a
Line 23: Relational Operator = >
         Line 23: Integer Constant = 0
         Line 23: Bracket/Parenthesis = {
Line 24: Identifier = cout
         Line 24: Bitwise Operator = <<
Line 24: String Literal = "a is positive."
         Line 24: Bitwise Operator = <<
Line 24: Identifier = endl
         Line 25: Keyword = else
Line 25: Bracket/Parenthesis = {
         Line 26: Bitwise Operator = <<
Line 26: String Literal = "a is non-positive."
Line 26: Bitwise Operator = <<
         Line 26: Delimiter = ;
         Line 27: Bracket/Parenthesis = }
         Line 29: Keyword = for
         Line 29: Data Type = int
Line 29: Identifier = i
         Line 29: Assignment Operator = =
         Line 29: Integer Constant = 0
         Line 29: Delimiter = ;
         Line 29: Identifier = i
Line 29: Integer Constant = 5
Line 29: Delimiter = ;
CyberCoder Improve Code Share Code Link Open Website
                                                                                                                           Q Ln 85, Col 30 Spaces: 4 UTF-16 LE CRLF () Plain Text 😝 🔰 Al Code Chat 🚨
```

### Sample 2:

```
Chectoder Improve Code Stare Code Link Open Website

CyberCoder Link CyberCode Stare Code Link Cyber Website

CyberCoder Link CyberCode Link Cyber Website

CyberCoder Link CyberCode Link Cyber CyberCode Link Cyber CyberCode Link Cyber Cyb
```

#### Sample 2 (output):

```
Line 2: Library Include = #include <tostream
Line 4: Identifier = using
Line 4: Identifier = namespace
Line 4: Identifier = std
               Line 4: Delimiter = ;
Line 6: Data Type = int
Line 6: Identifier = main
               Line 6: Bracket/Parenthesis = (
              Line 6: Bracket/Parenthesis = /
Line 7: Identifier = vector
Line 7: Relational Operator = <
Line 7: Data Type = int
               Line 7: Relational Operator = >
Line 7: Identifier = numbers
Line 7: Assignment Operator = =
               Line 7: Bracket/Parenthesis = {
               Line 7: Delimiter = ,
              Line 7: Integer Constant = 2
Line 7: Integer Constant = 3
Line 7: Delimiter = ,
Line 7: Integer Constant = 4
               Line 7: Integer Constant = 5
Line 7: Bracket/Parenthesis = }
               Line 7: Delimiter = ;
Line 8: Data Type = int
               Line 8: Identifier = sum
Line 8: Assignment Operator = =
              Line 8: Integer Constant = 0
Line 8: Delimiter = ;
               Line 10: Keyword = for
Line 10: Bracket/Parenthesis = (
                                                                                                                                          Q Ln 85, Col 30 Spaces: 4 UTF-16 LE CRLF {} Plain Text 🔠 🕴 Al Code Chat 🚨
      LyberCoder Improve Code Share Code Link Open Website
Line 10: Data Type = int
Line 10: Identifier = i
       Line 10: Assignment Operator = =
       Line 10: Delimiter = ;
Line 10: Identifier = i
       Line 10: Identifier = numbers
       Line 10: Unknown Token = .
       Line 10: Bracket/Parenthesis = )
       Line 10: Delimiter = ;
       Line 10: Increment/Decrement Operator = ++
Line 10: Bracket/Parenthesis = )
Line 10: Bracket/Parenthesis = {
        Line 11: Identifier = sum
       Line 11: Arithmetic Operator = +
       Line 11: Assignment Operator =
        Line 11: Identifier = numbers
       Line 11: Identifier = i
Line 11: Bracket/Parenthesis = ]
       Line 14: Data Type = double
Line 14: Identifier = average
        Line 14: Assignment Operator = =
       Line 14: Identifier = sum
       Line 14: Arithmetic Operator = /
       Line 14: Data Type = double
Line 14: Relational Operator = >
```

🔍 Ln 85, Col 30 Spaces: 4 UTF-16 LE CRLF () Plain Text 🤀 🔰 Al Code Chat 🚨

```
Line 14: Data Type = double
Line 14: Relational Operator = >
      Line 14: Bracket/Parenthesis = (
      Line 14: Identifier = numbers
      Line 14: Unknown Token = .
Line 14: Identifier = size
      Line 14: Bracket/Parenthesis = (
      Line 14: Delimiter = ;
      Line 16: Identifier = cout
      Line 16: Bitwise Operator = <<
Line 16: String Literal = "The sum of the numbers is: "
      Line 16: String Efferal - The
Line 16: Bitwise Operator = <<
Line 16: Identifier = sum
      Line 16: Bitwise Operator = <<
      Line 17: Identifier = cout
      Line 17: Bitwise Operator = <<
Line 17: String Literal = "The average of the numbers is: "
      Line 17: Bitwise Operator = <<
Line 17: Identifier = average
      Line 17: Bitwise Operator = <<
      Line 17: Delimiter = ;
Line 19: Keyword = if
Line 19: Bracket/Parenthesis = (
      Line 19: Identifier = average
      Line 19: Relational Operator = >
      Line 19: Bracket/Parenthesis = )
      Line 19: Bracket/Parenthesis = {
     Line 20: Bitwise Operator = <<
Line 20: String Literal = "Average is greater than 3.0"
berCoder Improve Code Share Code Link Open Website
                                                                                                                   Q Ln 85, Col 30 Spaces: 4 UTF-16 LE CRLF {} Plain Text 😝 🦸 Al Code Chat 🗘
      Line 20: Bitwise Operator = <<
Line 20: String Literal = "Average is greater than 3.0"
Line 20: Bitwise Operator = <<
      Line 21: Bracket/Parenthesis = {
      Line 22: Bitwise Operator = <<
Line 22: String Literal = "Average is less than or equal to 3.0"
      Line 22: Bitwise Operator = <<
       Line 22: Delimiter = ;
      Line 25: Keyword = return
Line 25: Integer Constant = 0
      Line 26: Bracket/Parenthesis = ]
                                                                                                                   Q Ln 85, Col 30 Spaces: 4 UTF-16 LE CRLF {} Plain Text 🔠 🔰 Al Code Chat 🗘
```