Assignment 0 - C Tokenizer

Karun Kanda (kk951), Junxian Cai (jc2411)

October 4, 2020

1 Name of Program

tokenizer - tokenizes a string input to individual tokens built in the C programming language.

2 Usage

./tokenizer <string input>

3 Description of Algorithm

This program takes a single string of an arbitrary length and breaks them up into types of tokens, including word, decimal integer, octal integer, hexadecimal integer, floating point, C operator. Additionally, the program can recognize the reserved keywords in C as distinct tokens, skip // and /* */ comments, and recognize strings in quotes as single tokens. The program reads the input string char by char from the start to the end, and prints out every single token in the form: <token type >: "<token name >".

4 Unique Features of our Implementation

This program runs fast designed to run in linear time and reads each char only once in most cases. It's implemented with the idea of state machine using switch cases, hence the decision process of the type of a token is straight forward and won't run into irrelevant types.

5 Efficiency Analysis

With the algorithm we implemented to tokenize the string has a worst-case running time of O(n) because the tokens are printed until the null terminator is found and the string is a unknown size which we can consider as a n input size.