Compiler Construction (CS402)
Phase 1
Lexical Analyzer

Assigned on: January 26, 2016

Instructions:

- Input file is a text file containing the source program in C language.
- Output file would be a text file listing all the tokens.
- Format for the token is <token_class, lexeme, line_no>
- No Symbol Table operations are required in this phase.
- The format for comments is /*comment*/.
- There will be no nested comments.
- All integer constants are unsigned.
- String constants are within " " (i.e. double quotes)
- All string constants are on a single line.
- Char constants are within ' ' (i.e. single quotes)
- Only single char constants are valid.
- After detecting an error the compiler should skip the remaining line (i.e. up to newline)
- The definition for id = **letter (letter | digit | _)***.
- ~, #, etc. are invalid characters.
- Language is case-sensitive.
- Errors to be reported.
    - Non-terminated comments
    - String constants exceed line
    - Char constant too long
    - Undefined symbol

Refer to C grammar for valid and possible token classess (All terminals can be considered for token classes). Any other group of characters that do not match any pattern of these token classes will be considered as an error.) You can use the terminal names given in the C grammar as output tokens.