**VIETNAM NATIONAL UNIVERSITY, HANOI**

**UNIVERSITY OF ENGINEERING AND TECHNOLOGY**

****

**MIDTERM SUMMARY REPORT**

**Lớp học phần: INT3402E 20**

**Group 14:**

**Trần Lê Thành Trung – 21020671**

**Vũ Hoàng Hưng – 21020206**

**Phạm Anh Đức – 21020187**

**Trần Hữu Đức – 21020044**

*Hanoi – 4/2024*

1. **LEXER.C**

For implementing Non-deterministic Finite Automata (NFA) and Deterministic Finite Automata (DFA). Here's a breakdown of the code:

**Structures:**

* struct State: Represents a state in the NFA. It has fields for terminality (type of token the state represents), number of transitions, and an array of transitions.
* struct Transition: Represents a transition in the NFA. It has fields for the character that triggers the transition and the successor state.
* struct NFA: Represents a Non-deterministic Finite Automaton. It has fields for the number of states, number of initial states, and an array of states.
* struct DFAState: Represents a state in the DFA. It has fields for terminality and the number of transitions. It also has an array of unique transitions for efficient searching.
* struct PowersetMap: Represents a powerset used during the powerset construction algorithm for DFA generation. It has fields for the bitmap size, number of elements, capacity, an array of keys (represented as bitmaps), and an array of states corresponding to the keys.

**Functions:**

* state\_init, state\_create, state\_copy, state\_mut\_shrink, state\_mut\_sort\_transition, state\_mut\_strip\_transition, state\_mut\_strip\_duplicate: These functions deal with creating, manipulating, and modifying NFA states.
* nfa\_init, nfa\_free, nfa\_define\_token, nfa\_fprint: These functions deal with creating, freeing, defining tokens for, and printing information about NFAs.
* nfa\_verbatim, nfa\_any, nfa\_union, nfa\_concat, nfa\_kleene, nfa\_optional: These functions create NFAs for different regular expression patterns like literal characters, any character, union, concatenation, Kleene star, and optional characters.
* nfa\_mut\_eliminate\_epsilon, nfa\_mut\_strip\_unreachable\_states, nfa\_mut\_strip\_dead\_states, nfa\_mut\_remove\_duplicate\_states, nfa\_mut\_minimize: These functions perform optimizations on the NFA like eliminating epsilon transitions, removing unreachable states, removing dead states, removing duplicate states, and minimizing the NFA.
* dfa\_state\_init, dfa\_state\_search: These functions deal with creating and searching for transitions in DFA states.
* get\_bit, set\_bit, bitmap\_eq, bitmap\_iszero: These are helper functions for manipulating bitmaps used in the powerset construction.
* pm\_create, pm\_release: These functions create and release the powerset data structure.

Overall, this file provides a framework for building and manipulating NFAs and DFAs. It includes functions for creating NFAs from regular expressions, performing optimizations, and potentially converting NFAs to DFAs (although the conversion function is not implemented here).

1. **VC.C**

It scans the source code file and identifies tokens like keywords, identifiers, operators, literals, etc.

**Sections:**

1. **Header Inclusion:**
   * Includes necessary header files like VC.h (presumably containing token definitions), Lexer.h (likely for lexer functionalities), standard C libraries <assert.h>, <stdint.h>, <stdio.h>.
2. **Macros:**
   * Defines constants like MAX\_TOKEN\_LENGTH, BUFFER\_SIZE, and MUNCH\_LIMIT used for token length, input buffer size, and processing limit respectively.
3. **Global Variable:**
   * Declares a global pointer vc\_lexer\_automata to a DFA (Deterministic Finite Automaton) structure, which represents the finite state machine for the lexer.

**Function** vc\_lexer\_automata\_init**:**

* This function initializes the DFA for the lexer.
* It defines Non-deterministic Finite Automata (NFAs) for various token patterns using functions like nfa\_verbatim, nfa\_any, nfa\_union, nfa\_concat, nfa\_kleene, and nfa\_optional. These functions create NFAs for literal characters, any character, unions, concatenations, Kleene star (zero or more repetitions), and optional characters respectively.
* Examples include defining NFAs for comments, non-printable characters, identifiers, keywords, operators, literals (integers, floats, booleans, strings), etc.
* It then minimizes the resulting NFA using nfa\_mut\_minimize.
* Finally, it converts the minimized NFA to a DFA using dfa\_powerset\_construction (not implemented here) and assigns it to the global vc\_lexer\_automata.
* After building the DFA, the function frees the memory used by the intermediate NFAs.

**Function** scanfile**:**

* This function takes the source code file (in) and output token file (out) as arguments.
* It uses a buffer munch of size MUNCH\_LIMIT to read chunks of the source file.
* Variables state, token\_length, tok, and start are used to track the current DFA state, length of the longest recognized token, token type, and starting position of the token in the file respectively.
* It starts by reading a chunk of the file and asserts that the initial state of the DFA accepts empty input.
* The function iterates through the characters in the buffer:
  + For each character:
    - It consumes the character and updates the DFA state using dfa\_consume.
    - If the DFA reaches an error state (SIZE\_MAX), it indicates an unidentified token and exits with an error message.
    - Otherwise, it checks the token type of the current state using dfa\_state\_token.
    - If a token is found:
      * It updates token\_length and tok with the length and type of the discovered token.
      * If the token type is not to be ignored (TOKEN\_IGNORE\_BIT), it writes the token to the output file, followed by its type information from token\_names.
      * The loop breaks as a token is found.
* After processing the entire buffer, it handles the End-of-File (EOF) condition. If no token was found and EOF is reached, it again checks tok and writes the token (if any) to the output file.
* Finally, it asserts that a token was found and rewinds the file to the character after the discovered token. It then recursively calls itself to process the remaining part of the file.

**Function** main**:**

* This function is the program's entry point.
* It checks if the command line arguments are provided correctly (source file and output file).
* It calls vc\_lexer\_automata\_init to initialize the lexer.
* It opens the source and output files.
* It calls scanfile to start the lexical analysis process.
* After processing the entire file, it asserts that the end of file is reached and closes both files.
* Finally, it frees the DFA memory using a presumably implemented vc\_lexer\_automata\_free function and exits successfully.