CS 316 Fall 2017
Observe **course policies** in undertaking this project.
All programs must be written in Oracle Standard Edition compliant Java or ANSI/ISO standard compliant C++.

PROJECT 1: Lexical Analyzer
Due: 10/13/17, Friday, 11 PM
Late projects will not be accepted.

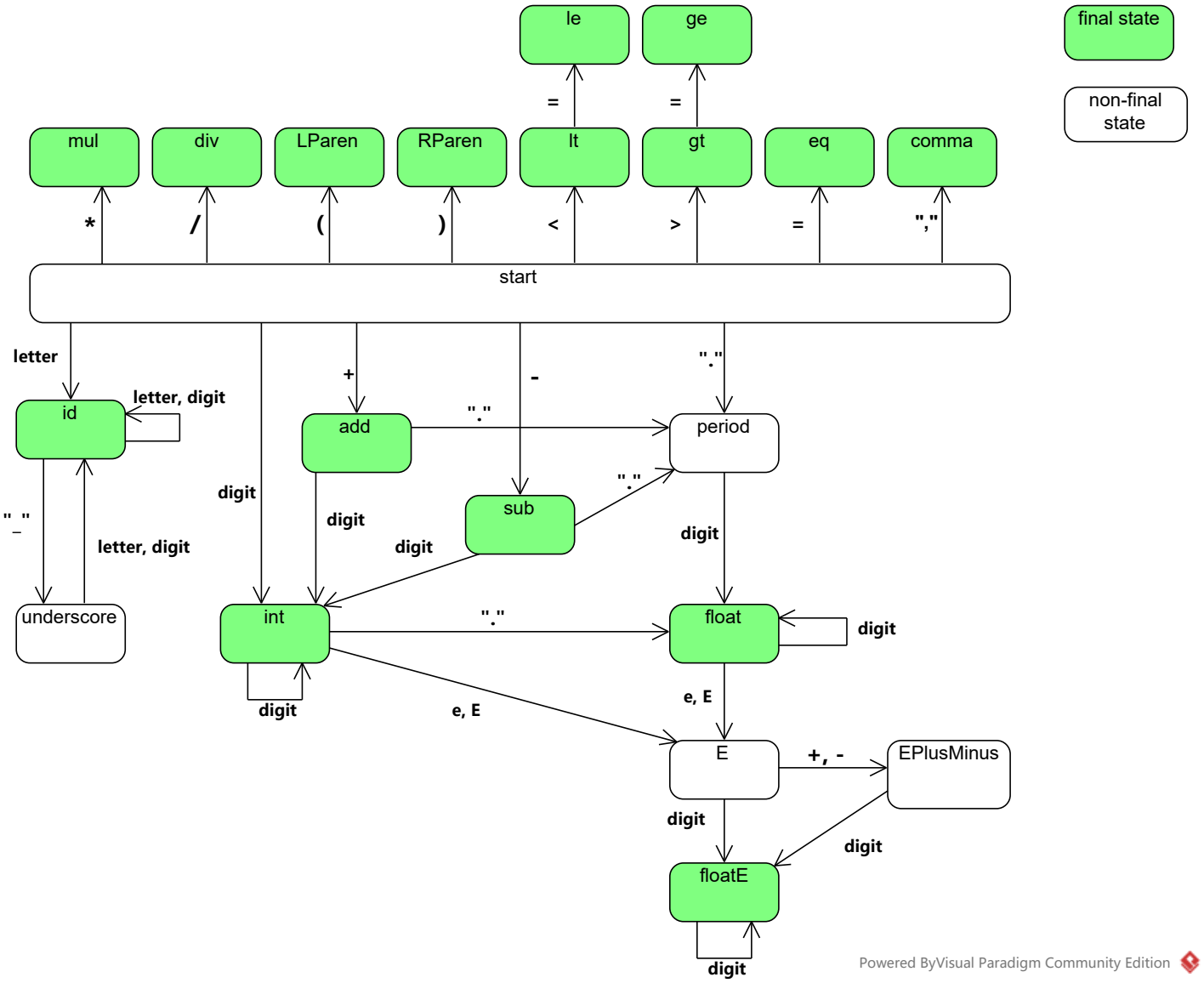
Consider the following EBNF defining 16 token categories ⟨id⟩ through ⟨comma⟩:

⟨letter⟩ → a | b | ... | z | A | B | ... | Z
⟨digit⟩ → 0 | 1 | ... | 9
⟨basic id⟩ → ⟨letter⟩ {⟨letter⟩ | ⟨digit⟩}
⟨letters and digits⟩ → {⟨letter⟩ | ⟨digit⟩}$^+$
⟨id⟩ → ⟨basic id⟩ { "_" ⟨letters and digits⟩ } // Note: "_" is the underscore char
⟨int⟩ → [+|−] {⟨digit⟩}$^+$
⟨float⟩ → [+|−] ({⟨digit⟩}$^+$ "." {⟨digit⟩} | "." {⟨digit⟩}$^+$)
⟨floatE⟩ → (⟨float⟩ | ⟨int⟩) (e|E) [+|−] {⟨digit⟩}$^+$
⟨add⟩ → +
⟨sub⟩ → −
⟨mul⟩ → *
⟨div⟩ → /
⟨lt⟩ → "<"
⟨le⟩ → "<="
⟨gt⟩ → ">"
⟨ge⟩ → ">="
⟨eq⟩ → "="
⟨LParen⟩ → "("
⟨RParen⟩ → ")"
⟨comma⟩ → ","

⟨letter⟩, ⟨digit⟩, ⟨basic id⟩, ⟨letters and digits⟩ are not token categories by themselves; rather, they are auxiliary categories to assist the definitions of the tokens ⟨id⟩, ⟨int⟩, ⟨float⟩, ⟨floatE⟩.

According to the above definitions, the integers and floating-point numbers may be signed with "+" or "−". Moreover, the integer or fractional part, but not both, of a string in ⟨float⟩ may be empty.

The following is a DFA to accept the 16 token categories.



The objective of this project is to implement a lexical analyzer that accepts the 16 token categories **plus the following keywords, all in lowercase letters only**:

int, float, boolean, if, then, else, and, or, not, false, true

These keywords cannot be used as identifiers, but can be parts of identifiers, like "iff" and "delse". In this and the next three projects, we assume that **the identifiers and keywords are case-sensitive**. The implementation should be based on the above DFA. Your lexical analyzer program should clearly separate the driver and the state-transition function so that the driver will remain invariant and only state-transition functions will change from DFA to DFA. The enumerated or integer type is suggested for representation of states.

The following keyword recognition method is adequate for this project.

1. Create 11 additional DFA states for the keywords.
2. The DFA initially accepts the keywords as identifiers.
3. Each time the DFA accepts an identifier, check if it is one of the keywords, and if so, move the DFA to the corresponding state.

The lexical analyzer program is to read an input text file, extract the tokens in it, and write them out one by one on separate lines. Each token should be flagged with its category. The output should be sent to an output text file. Whenever invalid tokens are found, error messages should be printed, and the reading process should continue. To make grading efficient and uniform, the program is to read the input/output file names as external arguments to the main function. How to set external arguments to Java main function in Eclipse.

You may modify one of these sample Java programs into your solution; if you do so, modify the comments suitably as well.

Here's a sample set of test input/output files:

in1 | out1
in2 | out2
in3 | out3
in4 | out4
in5 | out5
in6 | out6
in7 | out7
in8 | out8

You should make your own additional input files to test the program.

Since the purpose of this project is to reinforce, firsthand, the understanding of the internal mechanism of lexical analyzers built from finite automata as opposed to viewing them as black boxes, you are **not** allowed to use any library functions/tools for lexical analysis (like the Java StringTokenizer).

The above token set is used for a small typed functional language designed for our projects. Our project plan for the semester is to implement an interpreter of this language: a top-down parser in Project 2, a type checker in Project 3, and an interpreter in Project 4.

Submission

Your source program must be emailed to keitaro.yukawa@gmail.com with the subject header:

CS 316, Project 1, your full name

Include concise instructions for how to compile and run your program. You may email the entire materials in a .zip or .rar compressed file.

The due date is 10/13/17, Friday, 11 PM. No late projects will be accepted. If you haven't been able to complete the project, you may send an incomplete program for partial credit. In this case, include a description of what is and is not working in your program along with what you believe to be the sources of the problems.