Programming Language Syntax

ORGANIZATION OF PROGRAMMING LANGUAGES

JUCHEOL MOON

Decimal Numeral Specification in Java

- 1234_5678
- A decimal numeral is either the single ASCII digit 0, representing the integer zero, or consists of an ASCII digit from 1 to 9 optionally followed by one or more ASCII digits from 0 to 9 interspersed with underscores, representing a positive integer.
- https://docs.oracle.com/javase/specs/jls/se8/html/jls-3.html#jls-3.10.1

Language Syntax

- Programming Language must have a clearly specified syntax.
  - Programmers can learn the syntax.
  - Programmers know what is allowed and what is not allowed.
- Compiler writers can understand programs and enforce the syntax.

Lexical Analysis

- Input is a stream of characters.
  - How to analyze a stream of characters?

```
if(x==y)
    z=0;
else
    z=1;
```

➔ if(x==y)\n\tz=0;\nelse\n\tz=1;

- We must first analyze the stream of characters into something that a program can understand.

Token Class

- In English
  - Noun, verb, adjective, …
- In a programming language
  - Keywords
    - new, switch, for, while, if, try, exception, go to …
  - Identifiers
    - String of letters or digits
  - Integers
    - string of digits

Basic Definitions

- Alphabet (Σ)
  - Any finite set of symbols.
    - Σ = {a, b}
  - Σ^* is the set of all strings over Σ.
    - $\Sigma^* = \{a, b, ab, ba, aa, bb, aba, aabb, \cdots, \emptyset\}$
- String over an alphabet
  - A finite sequence of symbols drawn from that alphabet.
  - ϵ is the empty string
    - Concatenating ϵ with a string s gives s
      - $s\epsilon = \epsilon s = s$ ex. $ab\epsilon = ab = \epsilon ab$

What is Languages?

- Language (L)
  - A language over Σ is a set of strings of characters drawn from Σ.
  - Is Σ infinite? - No
  - Is Σ* infinite? - Yes
  - Is L infinite? - Yes

Regular Expressions

- A regular expression is one of the following:
  - $\emptyset$ = { }
  - ϵ
  - c where $c \in \Sigma$
  - A B: Two regular expressions concatenated
  - A | B: Two regular expressions separated by | (i.e., or)
  - A* : A regular expression followed by the Kleene star * (concatenation of zero or more strings)

Regular Language

- A regular expression defines a language (the set of all strings that the regular expression describes)
- The language $L(R)$ of regular expression R is given by:
  - $L(\emptyset) = \emptyset$
  - $L(\epsilon) = \{\epsilon\}$
  - $L(c) = \{c\}$
  - $L(R_1 | R_2) = L(R_1) \cup L(R_2)$
  - $L(R_1 R_2) = L(R_1)\, L(R_2)$

Regular Language

- $L(R_1 \mid R_2) = L(R_1) \cup L(R_2)$
  - $L(a \mid b) = \{a, b\} = L(a) \cup L(b) = \{a\} \cup \{b\}$
  - $L(a \mid b \mid c) = \{a, b, c\}$
  - $L(a \mid \boldsymbol{\varepsilon}) = \{a, \varepsilon\}$
  - $L(\boldsymbol{\varepsilon} \mid \boldsymbol{\varepsilon}) = \{\varepsilon\}$

Regular Language

- $L(R_1\, R_2) = L(R_1)\, L(R_2)$
  - $A = \{aa, b\}, B = \{a, b\}$
  - $A\, B = \{aaa, aab, ba, bb\}$
  - $A = \{aa, b, \boldsymbol{\varepsilon}\}, B = \{a, b\}$
  - $A\, B = \{aaa, aab, ba, bb, a, b\}$