Regular Expressions

CSCI 2670: Introduction to Theory of Computing

Farhan Jiva

Outline

- Introduction
- Vocabulary Recap
- Definitions
  - Constants
  - Operators
- Operator Precedence
- Examples
- Group quiz

Introduction

- A regular expression provides for a concise and flexible means for “matching” strings of text
  - i.e. certain characters, words, patterns of characters
- The concept of regular expressions was first popularized by Unix tools
  - ed, grep
- Invented by Stephen Cole Kleene in the 1950s

Vocabulary Recap

- Recall from our last class:
  - A character is some type of symbol (letter, digit)
  - An alphabet (denoted by Σ) is any set of characters
  - A word (or string) is any finite sequence of characters over an alphabet Σ

Definition - Constants

- Given a finite alphabet Σ
  - Empty set (∅) is a set containing zero elements
  - Empty string (λ) is a string containing zero characters
  - Literal character (a) is a set containing only the character a

Definition - Concatenation Operator

- Simplest type of regular expression
- Formed by concatenating many characters together
- Formal definition
  - RS denoting the set { αβ | α in R and β in S }
- Examples
  - Alphabet Σ = {a,b}
  - Language = ΣΣ
  - String in the language: aa, ab, ba, bb

Definition - Alteration Operator

- Lets us choose from one of many possibilities
  - Similar to the logical OR operator
- Formal definition
  - R|S denoting the set union of R and S
- Examples
  - Alphabet Σ = {a,b}
  - Language = aa | bb
  - Strings in the language: exactly two strings, aa and bb

Definition – Kleene Star Operator

- Also known as the “replication” operator
- Formal definition
  - R* denoted the set of all strings that can be made by concatenating any finite number (including zero) of strings from R
- Examples
  - Alphabet Σ = {a,b}
  - Language = a*
  - Strings in the language: λ, a, aa, aaa, aaaa, ...

Definition – Plus Operator

- Not a fundamental operator, however used extensively in the real world.
- Similar to the Kleene Star operator, however the preceding character must match at least once
- Examples
  - Alphabet Σ = {a,b}
  - Language = a+b+
  - Strings in the language: ab, aabb, abb, aab, …
  - Note that the empty string is NOT in the language described above
- Plus operator can also be written using the Kleene Star
  - a+ <=> aa*

Definition - Grouping Operator

- Allows us to specify precedence (with parentheses) to the various operators
- Examples
  - Alphabet Σ = {a,b}
  - Language = (ab)*(a|b)
  - Strings in the language: a, aba, ababb, b, ...

Operator Precedence

- Order of operations for regular expression operators
  - Parentheses
  - Kleene star/plus operator
  - Concatenation
  - Alteration
- If there are no ambiguities, parentheses may be omitted
  - (ab)c = abc
  - a|(b(c*)) = a|bc*

Examples

- Alphabet Σ = {a,b}
- a = a
- a|b = a,b
- (a|b)(b|a) = aa, ab, bb, ba
- a*b* = λ, a,aa,aaa,b,bb,bbb,ab,aabb, …
- (bb)* = λ, bb, bbbb, bbbbbb, …
- (ab)+|a+ = ab, abababab, a, aa, aaa, …

Group Quiz

- Alphabet Σ = {a,b}
- Group 1
  - (a+b+)(ab)*
- Group 2
  - ba*b+|ab*a+
- Group 3
  - a*b*|ba+
- Group 4
  - (b+(a+)*(ab)+)|(a|b)*