Regular Languages (Section 11.1)

Regular Languages

- *Problem:* Suppose the input strings to a program must be strings over the alphabet $\{a, b\}$ that contain exactly one substring bb. In other words, the strings must be of the form $xbby$, where x and y are strings over {a,b} that do not contain bb, x does not end in b, and y does not begin with b. Later we will see how to describe such strings formally.
- A regular language ove the alphabet A is a language constructed by the following rules:
  - $\varnothing$ and $\{\Lambda\}$ are regular languages.
  - $\{a\}$ is a regular language for all $a \in A$.
  - If M and N are regular languages, then so are $M \cup N$, MN, and M^*.

An example

Example: Let $A = \{a, b\}$. Then the following languages are a sampling of the regular languages over A:

- $\varnothing, \{\Lambda\}, \{a\}, \{b\}, \{a, b\}, \{ab\}, \{a\}^* = \{\Lambda, a, aa, aaa, \ldots, a^n, \ldots\}$

Regular expressions

- A regular expression over alphabet A is an expression constructed by the following rules:
  - $\varnothing$ and Λ are regular expressions.
  - a is a regular expression for all $a \in A$.
  - If R and S are regular expressions, then so are $(R), R + S, R \cdot S$, and R^*.

  The hierarchy in the absence of parentheses is: $*$ (do it first), $\cdot$, $+$ (do it last). Juxtaposition will be used instead of $\cdot$.
- *Example:* Let $A = \{a, b\}$. Then the following expressions are a sampling of the regular expressions over A:
  - $\varnothing, \Lambda, a, b, ab, a + ab, (a + b)^*$

Regular expressions represent regular languages

Regular expressions represent regular languages by the following correspondence, where $L(R)$ denotes the regular language of the expression R:

- $L(\varnothing) = \varnothing$
- $L(\Lambda) = \{\Lambda\}$
- $L(a) = \{a\}$ for all $a \in A$
- $L(R + S) = L(R) \cup L(S)$
- $L(RS) = L(R)L(S)$
- $L(R^*) = L(R)^*$

Regular expression example

Example: The regular expression $ab + a^*$ represents the following regular language:

$$\begin{aligned} L(ab + a^*) &= L(ab) \cup L(a^*) \\ &= L(a)L(b) \cup L(a)^* \\ &= \{a\}\{b\} \cup \{a\}^* \\ &= \{ab\} \cup \{\Lambda, a, aa, aaa, \ldots, a^n, \ldots\} \\ &= \{ab, \Lambda, a, aa, aaa, \ldots, a^n, \ldots\} \end{aligned}$$

Another example

Example: The regular expression $(a+b)^*$ represents the following regular language:

- $L((a+b)^*) = (L(a+b))^* = \{a, b\}^*$, the set of all possible strings over $\{a, b\}$.

Original Problem

- *Back to the problem:* Suppose the input strings to a program must be strings over the alphabet $\{a, b\}$ that contain exactly one substring bb. In other words, the strings must be of the form $xbby$, where x and y are strings over $\{a, b\}$ that do not contain bb, x does not end in b, and y does not begin with b. How can we describe the set of inputs formally?

Original Problem

- *Back to the problem:* Suppose the input strings to a program must be strings over the alphabet $\{a, b\}$ that contain exactly one substring bb. In other words, the strings must be of the form $xbby$, where x and y are strings over $\{a, b\}$ that do not contain bb, x does not end in b, and y does not begin with b. How can we describe the set of inputs formally?
- *Solution:* Let $x = (a + ba)^*$ and $y = (a + ab)^*$.

Original Problem

- *Back to the problem:* Suppose the input strings to a program must be strings over the alphabet $\{a, b\}$ that contain exactly one substring bb. In other words, the strings must be of the form $xbby$, where x and y are strings over $\{a, b\}$ that do not contain bb, x does not end in b, and y does not begin with b. How can we describe the set of inputs formally?
- *Solution:* Let $x = (a + ba)^*$ and $y = (a + ab)^*$.
- So the entire answer is: $(a + ba)^* bb(a + ab)^*$

The Algebra of Regular Expressions

- *Equality:* Regular expressions R and S are *equal*, written $R = S$, when $L(R) = L(S)$.
- *Examples:* $a + b = b + a, a + a = a, aa^* = a^*a, ab \neq ba$
- *Properties of* $+, \cdot,$ *and closure:*
  - $+$ is commutative, associative, $\varnothing$ is identity for $+$, and $R + R = R$.
  - $\cdot$ is associative, Λ is identity for $\cdot$, and $\varnothing$ is zero for $\cdot$.
  - $\cdot$ distributes over $+$

Simplification

Simplify the regular expression $aa(b^* + a) + a(ab^* + aa)$

$$
\begin{aligned}
& aa(b^* + a) + a(ab^* + aa) & \\
&= aa(b^* + a) + aa(b^* + a) & \cdot \text{ distributes over } + \\
&= aa(b^* + a) & R + R = R.
\end{aligned}
$$

Another Simplification

Show that $(a + aa)(a + b)^* = a(a + b)^*$

$$
\begin{aligned}
& (a + aa)(a + b)^* & \\
&= (a + aa)a^*(ba^*)^* & (R + S)^* = R^*(SR^*)^* \\
&= a(\Lambda + a)a^*(ba^*)^* & R = R\Lambda \text{ and } \cdot \text{ dist over } + \\
&= aa^*(ba^*)^* & (\Lambda + R)R^* = R^* \\
&= a(a + b)^* & (R + S)^* = R^*(SR^*)^*
\end{aligned}
$$

QED