Regular Languages and Expressions

Lecture 2

Thursday, August 29, 2024

LaTeXed: September 2, 2024 22:30

2.1
Regular Languages

Regular Languages

A class of simple but useful languages.

The set of regular languages over some alphabet Σ is defined inductively as:

1. $\emptyset$ is a regular language.
2. $\{\epsilon\}$ is a regular language.
3. $\{a\}$ is a regular language for each $a \in \Sigma$. Interpreting a as string of length 1.
4. If L_1, L_2 are regular then $L_1 \cup L_2$ is regular.
5. If L_1, L_2 are regular then $L_1 L_2$ is regular.
6. If L is regular, then $L^* = \cup_{n \geq 0} L^n$ is regular.
   The $\cdot^*$ operator name is **Kleene star**.
7. If L is regular, then so is $\overline{L} = \Sigma^* \setminus L$.

Regular languages are closed under operations of union, concatenation and Kleene star.

Regular Languages

Have basic operations to build regular languages.
Important: Any language generated by a finite sequence of such operations is regular.

Lemma 2.1.

Let $L_1, L_2, \ldots,$ be regular languages over alphabet Σ. Then the language $\cup_{i=1}^{\infty} L_i$ is not necessarily regular.

Some simple regular languages

Lemma 2.2.

If w is a string then $L = \{w\}$ is regular.

Example: $\{aba\}$ or $\{abbabbab\}$. Why?

Lemma 2.3.

Every finite language L is regular.

Examples: $L = \{a, abaab, aba\}$. $L = \{w \mid |w| \leq 100\}$. Why?

More Examples

- $\{w \mid w$ is a keyword in Python program$\}$
- $\{w \mid w$ is a valid date of the form mm/dd/yy$\}$
- $\{w \mid w$ describes a valid Roman numeral$\}$
  $\{I, II, III, IV, V, VI, VII, VIII, IX, X, XI, \ldots\}$.
- $\{w \mid w$ contains "CS374" as a substring$\}$.

Review questions

1. $L_1 \subseteq \{0,1\}^*$ be a finite language. L_1 is a set with finite number of strings. T/F?
2. $L_2 = \left\{0^i \mid i = 0, 1, \ldots, \infty\right\}$. The language L_2 is regular. T/F?
3. $L_3 = \left\{0^{2i} \mid i = 0, 1, \ldots, \infty\right\}$. The language L_3 is regular. T/F?
4. $L_4 = \left\{0^{17i} \mid i = 0, 1, \ldots, \infty\right\}$. The language L_4 is regular. T/F?
5. $L_5 = \left\{0^i \mid i \text{ is } \textbf{not} \text{ divisible by } 17\right\}$. L_5 is regular. T/F?
6. $L_6 = \left\{0^i \mid i \text{ is divisible by } 2, 3, \text{or } 5\right\}$. L_6 is regular. T/F?
7. $L_7 = \left\{0^i \mid i \text{ is divisible by } 2, 3, \text{and } 5\right\}$. L_7 is regular. T/F?
8. $L_8 = \left\{0^i \mid i \text{ is divisible by } 2, 3, \text{but not } 5\right\}$. L_8 is regular. T/F?
9. $L_9 = \left\{0^i1^i \mid i \text{ is divisible by } 2, 3, \text{but not } 5\right\}$. L_9 is regular. T/F?
10. $L_{10} = \left\{w \in \{0,1\}^* \mid w \text{ has at most 374 } 1\text{s}\right\}$. L_{10} is regular. T/F?

2.1.1
Regular Languages: Review questions

Review questions

1. $L_1 \subseteq \{0,1\}^*$ be a finite language. L_1 is a set with finite number of strings. T/F?
2. $L_2 = \{0^i \mid i = 0, 1, \ldots, \infty\}$. The language L_2 is regular. T/F?
3. $L_3 = \{0^{2i} \mid i = 0, 1, \ldots, \infty\}$. The language L_3 is regular. T/F?
4. $L_4 = \{0^{17i} \mid i = 0, 1, \ldots, \infty\}$. The language L_4 is regular. T/F?
5. $L_5 = \{0^i \mid i$ is **not** divisible by $17\}$. L_5 is regular. T/F?
6. $L_6 = \{0^i \mid i$ is divisible by $2, 3,$ or $5\}$. L_6 is regular. T/F?
7. $L_7 = \{0^i \mid i$ is divisible by $2, 3,$ and $5\}$. L_7 is regular. T/F?
8. $L_8 = \{0^i \mid i$ is divisible by $2, 3,$ but not $5\}$. L_8 is regular. T/F?
9. $L_9 = \{0^i 1^i \mid i$ is divisible by $2, 3,$ but not $5\}$. L_9 is regular. T/F?
10. $L_{10} = \{w \in \{0,1\}^* \mid w$ has at most 374 1s$\}$. L_{10} is regular. T/F?

2.2
Regular Expressions

Regular Expressions

A way to denote regular languages

- simple patterns to describe related strings
- useful in
  - text search (editors, Unix/grep, emacs)
  - compilers: lexical analysis
  - compact way to represent interesting/useful languages
  - dates back to 50's: Stephen Kleene
    who has a star names after him.

Inductive Definition

A regular expression $\mathbf{r}$ over an alphabet $\boldsymbol{\Sigma}$ is one of the following:

Base cases:

- $\emptyset$ denotes the language $\emptyset$
- ϵ denotes the language $\{\epsilon\}$.
- a denote the language $\{a\}$.

Inductive cases: If $\mathbf{r}_1$ and $\mathbf{r}_2$ are regular expressions denoting languages R_1 and R_2 respectively then,

- $(\mathbf{r}_1 + \mathbf{r}_2)$ denotes the language $R_1 \cup R_2$
- $(\mathbf{r}_1 \bullet \mathbf{r}_2) = r_1 \bullet r_2 = (\mathbf{r}_1\mathbf{r}_2)$ denotes the language $R_1 R_2$
- $(\mathbf{r}_1)^*$ denotes the language R_1^*

Regular Languages vs Regular Expressions

Regular Languages	**Regular Expressions**
$\emptyset$ regular	$\emptyset$ denotes $\emptyset$
$\{\epsilon\}$ regular	ϵ denotes $\{\epsilon\}$
$\{a\}$ regular for $a \in \Sigma$	$\mathbf{a}$ denote $\{a\}$
$R_1 \cup R_2$ regular if both are	$\mathbf{r_1} + \mathbf{r_2}$ denotes $R_1 \cup R_2$
$R_1 R_2$ regular if both are	$\mathbf{r_1} \bullet \mathbf{r_2}$ denotes $R_1 R_2$
R^* is regular if R is	$\mathbf{r}^*$ denote R^*
$\overline{R} = \Sigma^* \setminus R$	—-

Regular expressions denote regular languages — they explicitly show the operations that were used to form the language

Notation and Parenthesis

- For a regular expression $\mathbf{r}$, $L(\mathbf{r})$ is the language denoted by $\mathbf{r}$. Multiple regular expressions can denote the same language!
  Example: $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathbf{r}_1$ and $\mathbf{r}_2$ are equivalent if $L(\mathbf{r}_1) = L(\mathbf{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  Example: $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  Example: $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathbf{r}^+ = \mathbf{r}\mathbf{r}^*$. Hence if $L(\mathbf{r}) = R$ then $L(\mathbf{r}^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. rs is sometimes written as $r \bullet s$.

Notation and Parenthesis

- For a regular expression $\mathbf{r}$, $L(\mathbf{r})$ is the language denoted by $\mathbf{r}$. Multiple regular expressions can denote the same language!
  Example: $(\mathbf{0}+\mathbf{1})$ and $(\mathbf{1}+\mathbf{0})$ denote same language $\{\mathbf{0},\mathbf{1}\}$
- Two regular expressions $\mathbf{r}_1$ and $\mathbf{r}_2$ are equivalent if $L(\mathbf{r}_1) = L(\mathbf{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  Example: $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  Example: $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathbf{r}^+ = \mathbf{r}\mathbf{r}^*$. Hence if $L(\mathbf{r}) = R$ then $L(\mathbf{r}^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. rs is sometimes written as $r \bullet s$.

Notation and Parenthesis

- For a regular expression $\mathbf{r}$, $L(\mathbf{r})$ is the language denoted by $\mathbf{r}$. Multiple regular expressions can denote the same language!
  Example: $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathbf{r}_1$ and $\mathbf{r}_2$ are equivalent if $L(\mathbf{r}_1) = L(\mathbf{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  Example: $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  Example: $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathbf{r}^+ = \mathbf{r}\mathbf{r}^*$. Hence if $L(\mathbf{r}) = R$ then $L(\mathbf{r}^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. rs is sometimes written as $r \bullet s$.

Notation and Parenthesis

- For a regular expression $\mathbf{r}$, $L(\mathbf{r})$ is the language denoted by $\mathbf{r}$. Multiple regular expressions can denote the same language!
  Example: $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathbf{r}_1$ and $\mathbf{r}_2$ are equivalent if $L(\mathbf{r}_1) = L(\mathbf{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  Example: $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  Example: $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathbf{r}^+ = \mathbf{r}\mathbf{r}^*$. Hence if $L(\mathbf{r}) = R$ then $L(\mathbf{r}^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. rs is sometimes written as $r \bullet s$.

Notation and Parenthesis

- For a regular expression $\mathbf{r}$, $L(\mathbf{r})$ is the language denoted by $\mathbf{r}$. Multiple regular expressions can denote the same language!
  Example: $(0 + 1)$ and $(1 + 0)$ denote same language $\{0, 1\}$
- Two regular expressions $\mathbf{r}_1$ and $\mathbf{r}_2$ are equivalent if $L(\mathbf{r}_1) = L(\mathbf{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  Example: $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  Example: $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathbf{r}^+ = \mathbf{r}\mathbf{r}^*$. Hence if $L(\mathbf{r}) = R$ then $L(\mathbf{r}^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. rs is sometimes written as $r \bullet s$.

Notation and Parenthesis

- For a regular expression $\mathbf{r}$, $L(\mathbf{r})$ is the language denoted by $\mathbf{r}$. Multiple regular expressions can denote the same language!
  Example: $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathbf{r}_1$ and $\mathbf{r}_2$ are equivalent if $L(\mathbf{r}_1) = L(\mathbf{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  Example: $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  Example: $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathbf{r}^+ = \mathbf{r}\mathbf{r}^*$. Hence if $L(\mathbf{r}) = R$ then $L(\mathbf{r}^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. rs is sometimes written as $r \bullet s$.

Skills

- Given a language L "in mind" (say an English description) we would like to write a regular expression for L (if possible)
- Given a regular expression $\mathbf{r}$ we would like to "understand" $L(\mathbf{r})$ (say by giving an English description)

Skills

- Given a language L "in mind" (say an English description) we would like to write a regular expression for L (if possible)
- Given a regular expression $\mathbf{r}$ we would like to "understand" $L(\mathbf{r})$ (say by giving an English description)

2.2.1
Some examples of regular expressions

Understanding regular expressions

- $(0+1)^*$: set of all strings over $\{0,1\}$
- $(0+1)^*001(0+1)^*$: strings with 001 as substring
- $0^* + (0^*10^*10^*10^*)^*$: strings with number of 1's divisible by 3
- $\emptyset 0$: $\{\}$
- $(\epsilon+1)(01)^*(\epsilon+0)$: alternating 0s and 1s. Alternatively, no two consecutive 0s and no two consecutive 1s
- $(\epsilon+0)(1+10)^*$: strings without two consecutive 0s.

Creating regular expressions

- bitstrings with the pattern $\mathbf{001}$ or the pattern $\mathbf{100}$ occurring as a substring
  one answer: $\mathbf{(0+1)^*001(0+1)^* + (0+1)^*100(0+1)^*}$
- bitstrings with an even number of $\mathbf{1}$'s
  one answer: $\mathbf{0^* + (0^*10^*10^*)^*}$
- bitstrings with an odd number of $\mathbf{1}$'s
  one answer: $\mathbf{0^*1}\boldsymbol{r}$ where $\boldsymbol{r}$ is solution to previous part
- bitstrings that do <u>not</u> contain $\mathbf{011}$ as a substring
- Hard: bitstrings with an odd number of 1s <u>and</u> an odd number of 0s.

Bit strings with odd number of **0**s and **1**s

The regular expression is

$$(\mathbf{00}+\mathbf{11})^*(\mathbf{01}+\mathbf{10})\Big(\mathbf{00}+\mathbf{11}+(\mathbf{01}+\mathbf{10})(\mathbf{00}+\mathbf{11})^*(\mathbf{01}+\mathbf{10})\Big)^*$$

(Solved using techniques to be presented in the following lectures...)

Regular expression identities

- $r^*r^* = r^*$ meaning for any regular expression r, $L(r^*r^*) = L(r^*)$
- $(r^*)^* = r^*$
- $rr^* = r^*r$
- $(rs)^*r = r(sr)^*$
- $(r+s)^* = (r^*s^*)^* = (r^* + s^*)^* = (r + s^*)^* = \dots$

Question: How does on prove an identity?
By induction. On what? Length of r since r is a string obtained from specific inductive rules.

2.2.2
An example of a non-regular language

A non-regular language and other closure properties

Consider $L = \{0^n 1^n \mid n \geq 0\} = \{\epsilon, 01, 0011, 000111, \ldots\}$.

Theorem 2.1.

$L = \{0^n 1^n \mid n \geq 0\} = \{\epsilon, 01, 0011, 000111, \ldots\}$.
The language L is **not** *a regular language.*

How do we prove it?

Other questions:

- Suppose R_1 is regular and R_2 is regular. Is $R_1 \cap R_2$ regular?
- Suppose R_1 is regular is $\overline{R_1}$ (complement of R_1) regular?

A sketchy proof

Theorem 2.2.

$L = \{0^n 1^n \mid n \geq 0\} = \{\epsilon, 01, 0011, 000111, \ldots\}$.

The language L *is* **not** *a regular language.*