DISCRETE MATHEMATICS

LECTURE 9

Plan: Introduction to Regular expressions & finite automata

Reference:

Pages 13 - 14 and Section 1.3 of book:

Introduction to the Theory of Computation

(third edition)

by Michael Sipser

Regular expressions:

- A syntax for describing string patterns
- Used for text search in text editors
- Also present in some programming languages

Example: $a\, b^*\, c$ - Pattern consisting of an 'a' followed by some no. of 'b's, followed by a 'c'

eg. abbc, abbbc, etc.

In this course:

- Formal understanding of regular expressions
- An algorithm for searching patterns in text.

LANGUAGES :

Σ : an alphabet

→ is a finite set of symbols

$$\Sigma_1 = \{a, b\}$$

$$\Sigma_2 = \{0, 1\}$$

$$\Sigma_e = \{a, b, c, \ldots\ x, y, z\}$$

Word: a word is a finite sequence of symbols.

Assume $\Sigma = \{a, b\}$

Eg: aab, abb baba, bbbaaa, ...

Language : a set of words.

L_1 : { set of all words that start with 'a' }

{ a, aa, ab, aaa, - - - - - - .

infinitely many words in L_1

L_2 = words that contain 'ab' as a substring

{ ab, aab, aba, ... }

L_1, L_2 are examples of languages.

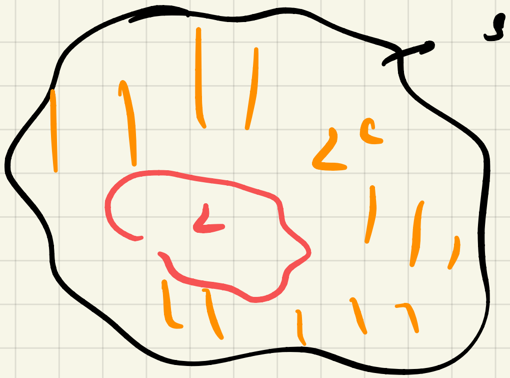
OPERATIONS ON LANGUAGES:

Union: $L_1 \cup L_2 = \{ w \mid w \in L_1 \text{ or } w \in L_2 \}$

(set theoretic union)

Intersection: $L_1 \cap L_2$

Complementation: L^c



set of all words over $\{a, b\}$

Star: L^*

Concatenation: $L_1 \cdot L_2 = \{ w_1 w_2 \mid w_1 \in L_1 , w_2 \in L_2 \}$

$L_1 = \{ ab, aa \}$

$L_2 = \{ b, ba \}$

$L_1 \cdot L_2 = \{ abb, abba, aab, aaba \}$

L_1 = words that start with 'a' L_2: words that end with 'b'

$L_1 \cdot L_2 =$ words that start with 'a' and end with 'b'

Is $L_1 \cdot L_2 = L_2 \cdot L_1$ in general? No.

Question: Give two languages L_1, L_2 s.t.

$$L_1 \cdot L_2 = L_2 \cdot L_1$$

Wrong:

L_1 = starts with 'a'

L_2 = end with 'a'

$L_1 \cdot L_2$ = start with 'a' and end with 'a'

$L_2 \cdot L_1$: $\underbrace{ba}_{\in L_2}\underbrace{ab}_{\in L_1} \in L_2 \cdot L_1$

One correct answer: L_1 = words of length 2, L_2 = words of length 3.

$L_1 \cdot L_2 = L_2 \cdot L_1$ = words of length 5.

L^* :

L .

$L \cdot L = L^2$

$L^3 = L \cdot L^2$

$\vdots$

$L^n = \{ w_1 w_2 \ldots w_n \mid w_i \in L \}$

$\downarrow$

We pick 'n' words from L and concatenate them.

L^0 : ?

Empty string: ϵ

length (w) = no. of symbols present

$|abba| = 4 \qquad |ab| = 2$ ..

ϵ : is defined to be a word of length 0

a string of length 0

→ called Empty string

$\underline{L^*}$:

$$L = L^1$$

$$L \cdot L = L^2$$

$$L^3 = L \cdot L^2$$

$$\vdots$$

$$L^n = \{ w_1 w_2 \ldots w_n \mid w_i \in L \}$$

$\downarrow$

We pick 'n' words from L and concatenate them.

L^0 : $\{ \epsilon \}$

$$L^* = L^0 \cup L^1 \cup L^2 \cup L^3 \ldots\ldots$$

$$\bigcup_{k \geq 0} L^k$$

$$L^* = L^0 \cup L^1 \cup L^2 \cup L^3 \cdots\cdots$$

$$\bigcup_{k \geq 0} L^k$$

$L = \{a\}$ $\quad L^* : \{\epsilon, a, aa, aaa, \cdots\}$

$L = \{a, b\}$ $\quad L^* : \epsilon \cup$ set of all words over $\{a, b\}$

$L^0 = \epsilon$

$L = \{a, b\}$

$L^2 = \{aa, ab, ba, bb\}$ (set of words over $\{a, b\}$ of length 2)

$\vdots$

$L^n =$ words of length 'n' over $\{a, b\}$

So far:

- Alphabet $\Sigma = \{a, b\}$
- Languages
- Union, intersection, complementation,
  Concatenation, Star

Coming next: An informal introduction to regular expressions

Example 1:

Describe the language described by the following regular expressions. $\Sigma = \{a, b\}$

1.1. $a^* b$

1.2. $ab + ba$

1.3 $b^* a + a^* b + a^*$

1.4 $(a + b)(a + b)$

1.5 $(a+b)^*$

1.6 $\Sigma^* a \Sigma^*$

1.7. $a + aa + b^*$

1.1. $a^* b$

$$a^* \cdot b$$

$$a^* = \{ \epsilon, a, aa, aaa, \ldots \}$$

$$a^* b = \{ \epsilon, a, aa, \ldots, a^n, \ldots \} \cdot \{b\}$$

$$= b, ab, a^2 b, \ldots a^n b, \ldots$$

1.2. $ab + ba = \{ab, ba\}$

1.3. $b^* a + a^* b + a^*$

$= \{a, ba, b^2a \ldots b^i a \ldots\} \cup \{b, ab, \ldots a^i b \ldots\}$

$\cup \{\epsilon, a, a^2, a^3, \ldots\}$

1.4 $(a+b)(a+b) = \{aa, ab, ba, bb\}$

1.5 $(a+b)^* = \{a, b\}^* = \cancel{\epsilon} \cup$ set of all words over $\{a, b\}$

$\{\epsilon\}$ ϵ

1.6 $\Sigma^* a \Sigma^*$

1.7. $a + aa + b^*$

$\Sigma^* a \Sigma^*$: set of all words that contain an 'a'

Words in $\Sigma^* a \Sigma^* = w_1 \, a \, w_2$

→ they will contain an 'a'

If v is a word that contains an 'a', we can break it as

$$\underbrace{\qquad\qquad}_{w_1} \boxed{a} \underbrace{\qquad\qquad}_{w_2} \in \Sigma^* a \Sigma^*$$

1.6. $a + aa + b^*$

$= \{a, \ aa, \ \epsilon, b, bb, b^3, \cdots \}$

Example 2: Write regular expressions for the following languages:

$\Sigma = \{a, b\}$

2.1. Set of words that contain 'ab' as a substring

2.2. words that contain both 'a' and 'b'

2.3. $\{aa, ab, ba, bb\}$

2.4. words that contain no 'a's

2.5 all words ending with an 'a'.

2.1. Set of words that contain 'ab' as a substring

2.2. words that contain both 'a' and 'b'

2.3. { aa, ab, ba, bb }

2.4. words that contain no 'a's

2.5 all words ending with an 'a'

$$\Sigma = a + b$$

2.1. $(a+b)^* ab (a+b)^*$

2.2.
- $(a+b)^* a (a+b)^* b (a+b)^*$
  $+ (a+b)^* b (a+b)^* a (a+b)^*$
- $\Sigma^* a \Sigma^* b \Sigma^* + \Sigma^* b \Sigma^* a \Sigma^*$
- $\Sigma^* ab + \Sigma^* ba \cancel{+}$
- $\Sigma^* ab \Sigma^* + \Sigma^* ba \Sigma^*$
- $\Sigma^* (ab + ba) \Sigma^*$

ab

ba

2.1. Set of words that contain 'ab' as a substring

2.2. words that contain both 'a' and 'b'

2.3. $\{aa, ab, ba, bb\}$

2.4. words that contain no 'a's

2.5 all words ending with an 'a'

2.3. $aa + ab + ba + bb$ $\quad (a+b).(a+b)$ $\quad \Sigma \cdot \Sigma$

2.4. b^*

2.5. $(a+b)^* a$

$\Sigma^* a$

Regular expressions : SYNTAX

R is a regular expression if R is :

- 1. a for some $a \in \Sigma$
- 2. ε
- 3. Φ
- 4. $R_1 + R_2$,
- 5. $R_1 \circ R_2$ (also written as $R_1 R_2$)
- 6. R_1^*

where R_1 and R_2 are regular expressions

Regular expressions : SEMANTICS

$\mathcal{L}(R)$: language associated with regular expr. R

- 1. $\mathcal{L}(a) = \{a\}$
- 2. $\mathcal{L}(\varepsilon) = \{\epsilon\}$
- 3. $\mathcal{L}(\Phi) = \Phi$
- 4. $\mathcal{L}(R_1 + R_2) = \mathcal{L}(R_1) \cup \mathcal{L}(R_2)$
- 5. $\mathcal{L}(R_1 R_2) = \{x_1 x_2 \mid x_1 \in R_1, x_2 \in R_2\} = \mathcal{L}(R_1).\mathcal{L}(R_2)$
- 6. $\mathcal{L}(R_1^*) = \{x_1 x_2 \ldots x_k \mid k \geq 0 \text{ and } x_i \in R_1 \text{ for } 1 \leq i \leq k\}$

$$= [\mathcal{L}(R_1)]^*$$

- 1. $L(a) = \{a\}$
- 2. $L(\varepsilon) = \{\epsilon\}$
- 3. $L(\Phi) = \Phi$
- 4. $L(R_1 + R_2) = L(R_1) \cup L(R_2)$
- 5. $L(R_1 R_2) = \{x_1 x_2 \mid x_1 \in R_1, x_2 \in R_2\}$
- 6. $L(R_1^*) = \{x_1 x_2 \dots x_k \mid k \geq 0 \text{ and } x_i \in R_1 \text{ for } 1 \leq i \leq k\}$

$$
\begin{aligned}
L(ab + ba) &= L(ab) \cup L(ba) \\
&= L(a) \cdot L(b) \cup L(b) \cdot L(a) \\
&= \{a\}\{b\} \cup \{b\} \cdot \{a\} \\
&= \{ab, ba\}
\end{aligned}
$$

- 1. $L(a) = \{a\}$
- 2. $L(\epsilon) = \{\epsilon\}$
- 3. $L(\Phi) = \Phi$
- 4. $L(R_1 + R_2) = L(R_1) \cup L(R_2)$
- 5. $L(R_1 R_2) = \{x_1 x_2 \mid x_1 \in R_1, x_2 \in R_2\}$
- 6. $L(R_1^*) = \{x_1 x_2 \dots x_k \mid k \geq 0 \text{ and } x_i \in R_1 \text{ for } 1 \leq i \leq k\}$

$$L(a^*) = [L(a)]^*$$

$$= \underline{\underline{\{a\}}}^*$$

$$= \{\epsilon, a, aa, \dots\}$$

- 1. $L(a) = \{a\}$
- 2. $L(\varepsilon) = \{\epsilon\}$
- 3. $L(\Phi) = \Phi$
- 4. $L(R_1 + R_2) = L(R_1) \cup L(R_2)$
- 5. $L(R_1 R_2) = \{ x_1 x_2 \mid x_1 \in R_1 ,\ x_2 \in R_2 \}$
- 6. $L(R_1^*) = \{ x_1 x_2 \ldots x_k \mid k \geq 0 \text{ and } x_i \in R_1 \text{ for } 1 \leq i \leq k \}$

$$L(ab^* + b)$$

Example 3: Write regular expressions for the following:

$\Sigma = \{a, b\}$

3.1. Words that do not contain 'ab'

3.2. Words with even length

3.2. Words that contain even no. of 'a's.

Example 4: Write the language corresponding to the foll. expr.

4.1. $ab + \epsilon$

4.2. $ab + \Phi$

4.3. $a \cdot \epsilon$

4.4. $a \cdot \Phi$

4.5. $(a + \epsilon) \cdot b$

4.6. Φ^*

Example 4: Write the language corresponding to the foll. expr.

4.1. $ab + \epsilon$

4.2. $ab + \Phi$

4.3. $a \cdot \epsilon$

4.4. $a \cdot \Phi$

4.5. $(a + \epsilon) \cdot b$

4.6. Φ^*

Solutions:

3.1. Words that do not contain 'ab'

3.2. Words with even length

3.1: $b^* a^*$

3.2. $[(a+b).(a+b)]^*$

$$(\Sigma . \Sigma)^*$$

$(aa)^* + (bb)^* + (ab)^* + (bb)^*$ Not correct

$\neq$

aaaaaaaa - - - $= (aa + bb + ab + bb)^*$

aa bb

Summary:

- Languages, operations on languages
- Regular expressions: a notation for describing languages
  - Syntax, semantics, examples.