Pre-lecture brain teaser

Consider the problem of a n-input AND function. The input (x) is a string n-digits long with $\Sigma = \{0, 1\}$ and has an output (y) which is the logical AND of all the elements of x.

Formulate a **language** that describes the above problem.

ECE-374-B: Lecture 2 - Regular Languages

Lecturer: Nickvash Kani

January 19, 2023

University of Illinois at Urbana-Champaign

Pre-lecture brain teaser

Consider the problem of a n-input AND function. The input (x) is a string n-digits long with $\Sigma = \{0, 1\}$ and has an output (y) which is the logical AND of all the elements of x.

Formulate a **language** that describes the above problem.

Pre-lecture brain teaser

Consider the problem of a n-input AND function. The input (x) is a string n-digits long with $\Sigma = \{0, 1\}$ and has an output (y) which is the logical AND of all the elements of x.

Formulate a **language** that describes the above problem.

$$L_{AND_N} = \left\{ \begin{matrix} 0|0, & 1|1, & & \\ 0 \cdot 0|0, & 0 \cdot 1|0, & 1 \cdot 0|0, & 1 \cdot 1|1 \\ \vdots & \vdots & \vdots & \vdots \\ (0\cdot)^n|0, & (0\cdot)^{n-1}1|0, & \dots & (1\cdot)^n|1 \dots \end{matrix} \right\} \tag{1}$$

Pre-lecture brain teaser

Consider the problem of a n-input AND function. The input (x) is a string n-digits long with $\Sigma = \{0, 1\}$ and has an output (y) which is the logical AND of all the elements of x.

Formulate a **language** that describes the above problem.

$$L_{AND_N} = \begin{Bmatrix} 0|0, & 1|1, & & \\ 0 \cdot 0|0, & 0 \cdot 1|0, & 1 \cdot 0|0, & 1 \cdot 1|1 \\ \vdots & \vdots & \vdots & \vdots \\ (0\cdot)^n|0, & (0\cdot)^{n-1}1|0, & \dots & (1\cdot)^n|1 \dots \end{Bmatrix} \tag{1}$$

This is an example of a regular language which we'll be discussing today.

Strings

Alphabet

An alphabet is a **finite** set of symbols.

Examples of alphabets:

- $\Sigma = \{0, 1\}$,
- $\Sigma = \{a, b, c, \ldots, z\}$,
- ASCII.
- UTF8.
- $\Sigma = \{\langle \text{moveforward} \rangle, \langle \text{moveback} \rangle, \langle \text{moveleft} \rangle, \langle \text{moveright} \rangle\}$

String Definition

Definition

1. A string/word over Σ is a **finite sequence** of symbols over Σ. For example, '0101001', '*string*', '$\langle \text{moveback} \rangle \langle \text{rotate90} \rangle$'
2. $x \cdot y \equiv xy$ is the concatenation of two strings
3. The length of a string w (denoted by $|w|$) is the number of symbols in w. For example, $|101| = 3$, $|\epsilon| = 0$
4. For integer $n \geq 0$, Σ^n is set of all strings over Σ of length n. Σ^* is the set of all strings over Σ.
5. Σ^* set of all strings of all lengths including empty string.

Question: $\{'a',' c'\}^* =$

Emptiness

- ϵ is a string containing no symbols. It is not a set
- $\{\epsilon\}$ is a set containing one string: the empty string. It is a set, not a string.
- $\emptyset$ is the empty set. It contains no strings.

Question: What is $\{\emptyset\}$

Concatenation and properties

- If x and y are strings then xy denotes their concatenation.
- Concatenation defined recursively :
  - $xy = y$ if $x = \epsilon$
  - $xy = a(wy)$ if $x = aw$
- xy sometimes written as $x \bullet y$.
- concatenation is associative: $(uv)w = u(vw)$ hence write $uvw \equiv (uv)w = u(vw)$
- **not** commutative: uv not necessarily equal to vu
- The identity element is the empty string ϵ:

$$\epsilon u = u\epsilon = u.$$

Substrings, prefixes, Suffixes

Definition
v is substring of $w \iff$ there exist strings x, y such that $w = xvy$.

- If $x = \epsilon$ then v is a prefix of w
- If $y = \epsilon$ then v is a suffix of w

Subsequence

A subsequence of a string $w[1...n]$ is either a subsequence of $w[2...n]$ or $w[1]$ followed by a subsequence of $w[2...n]$.

Example
kapa is a sub-sequence of *knapsack*

Subsequence

A subsequence of a string $w[1...n]$ is either a subsequence of $w[2...n]$ or $w[1]$ followed by a subsequence of $w[2...n]$.

Example
kapa is a sub-sequence of *knapsack*

Question: How many sub-sequences are there in a string $|w| = 5$?

String exponent

Definition
If w is a string then w^n is defined inductively as follows:

$w^n = \epsilon$ if $n = 0$

$w^n = ww^{n-1}$ if $n > 0$

Question: $(blah)^3 =$.

Rapid-fire questions -strings

Answer the following questions taking $\Sigma = \{0, 1\}$.

1. What is Σ^0?

2. How many elements are there in Σ^n?

3. If $|u| = 2$ and $|v| = 3$ then what is $|u \bullet v|$?

4. Let u be an arbitrary string in Σ^*. What is ϵu? What is $u\epsilon$?

Languages

Languages

Definition
A language L is a set of strings over Σ. In other words $L \subseteq \Sigma^*$.

Languages

Definition
A language L is a set of strings over Σ. In other words $L \subseteq \Sigma^*$.

Standard set operations apply to languages.

- For languages A, B the concatenation of A, B is $AB = \{xy \mid x \in A, y \in B\}$.
- For languages A, B, their union is $A \cup B$, intersection is $A \cap B$, and difference is $A \setminus B$ (also written as $A - B$).
- For language $A \subseteq \Sigma^*$ the complement of A is $\bar{A} = \Sigma^* \setminus A$.

Set Concatenation

Definition
Given two sets X and Y of strings (over some common alphabet Σ) the concatenation of X and Y is

$$XY = \{xy \mid x \in X, y \in Y\} \tag{2}$$

Question: $X = \{\mathit{fido}, \mathit{rover}, \mathit{spot}\}$, $Y = \{\mathit{fluffy}, \mathit{tabby}\} \implies XY = .$

Σ^* and languages

Definition

1. Σ^n is the set of all strings of length n. Defined inductively:
   $\Sigma^n = \{\epsilon\}$ if $n = 0$
   $\Sigma^n = \Sigma\Sigma^{n-1}$ if $n > 0$
2. $\Sigma^* = \cup_{n \geq 0} \Sigma^n$ is the set of all finite length strings
3. $\Sigma^+ = \cup_{n \geq 1} \Sigma^n$ is the set of non-empty strings.

Definition
A language L is a set of strings over Σ. In other words $L \subseteq \Sigma^*$.

Question: Does Σ^* have strings of infinite length?

Rapid-Fire questions - Languages

Problem
Consider languages over $\Sigma = \{0, 1\}$.

1. *What is $\emptyset^0$?*
2. *If $|L| = 2$, then what is $|L^4|$?*
3. *What is $\emptyset^*$, $\{\epsilon\}^*$, ϵ^*?*
4. *For what L is L^* finite?*
5. *What is $\emptyset^+$?*
6. *What is $\{\epsilon\}^+$, ϵ^+?*

Terminology Review

Let's review what we learned.

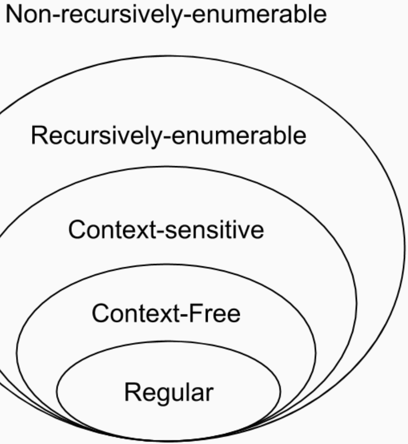
- A character(a, b, c, x) is a unit of information represented by a symbol: (letters, digits, whitespace)
- A alphabet(Σ) is a set of characters
- A string(w) is a sequence of characters
- A language(A, B, C, L) is a set of strings

Terminology Review

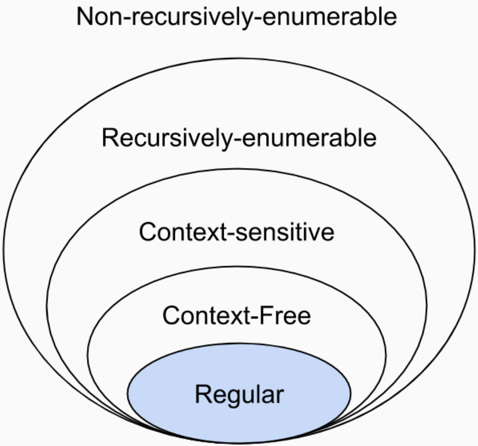
Let's review what we learned.

- A character(a, b, c, x) is a unit of information represented by a symbol: (letters, digits, whitespace)
- A alphabet(Σ) is a set of characters
- A string(w) is a sequence of characters
- A language(A, B, C, L) is a set of strings
- A grammar(G) is a set of rules that defines the strings that belong to a language

Chomsky Hierarchy



Chomsky Hierarchy



Regular Languages

Theorem (Kleene's Theorem)

A language is regular if and only if it can be obtained from finite languages by applying the three operations:

- *Union*
- *Concatenation*
- *Repetition*

a finite number of times.

Regular Languages

A class of simple but useful languages.

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

Base Case

- $\emptyset$ is a regular language.
- $\{\epsilon\}$ is a regular language.
- $\{a\}$ is a regular language for each $a \in \Sigma$. Interpreting a as string of length 1.

Regular Languages

Inductive step:

We can build up languages using a few basic operations:

- If L_1, L_2 are regular then $L_1 \cup L_2$ is regular.
- If L_1, L_2 are regular then $L_1 L_2$ is regular.
- If L is regular, then $L^* = \cup_{n \geq 0} L^n$ is regular.
  The $\cdot^*$ operator name is Kleene star.
- If L is regular, then so is $\overline{L} = \Sigma^* \setminus L$.

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

Some simple regular languages

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

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

Some simple regular languages

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

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

Lemma
Every finite language L is regular.

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

Have basic operations to build regular languages.

Important: Any language generated by a finite sequence of such operations is regular.

Lemma
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.

Regular Languages

Have basic operations to build regular languages.

Important: Any language generated by a finite sequence of such operations is regular.

Lemma
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.

Note:Kleene star (repetition) is a **single** operation!

Regular Languages - Example

Example: The language $L_{01} = 0^i 1^j |$ for all $i, j \geq 0$ is regular:

Rapid-fire questions - regular languages

1. $L_1 = \left\{0^i \mid i = 0, 1, \ldots, \infty\right\}$. The language L_1 is regular. T/F?

Rapid-fire questions - regular languages

1. $L_1 = \left\{0^i \mid i = 0, 1, \ldots, \infty\right\}$. The language L_1 is regular. T/F?
2. $L_2 = \left\{0^{17i} \mid i = 0, 1, \ldots, \infty\right\}$. The language L_2 is regular. T/F?

Rapid-fire questions - regular languages

1. $L_1 = \left\{0^i \mid i = 0, 1, \ldots, \infty\right\}$. The language L_1 is regular. T/F?
2. $L_2 = \left\{0^{17i} \mid i = 0, 1, \ldots, \infty\right\}$. The language L_2 is regular. T/F?
3. $L_3 = \left\{0^i \mid i \text{ is divisible by } 2, 3, \text{or } 5\right\}$. L_3 is regular. T/F?

Rapid-fire questions - regular languages

1. $L_1 = \left\{0^i \mid i = 0, 1, \ldots, \infty\right\}$. The language L_1 is regular. T/F?
2. $L_2 = \left\{0^{17i} \mid i = 0, 1, \ldots, \infty\right\}$. The language L_2 is regular. T/F?
3. $L_3 = \left\{0^i \mid i \text{ is divisible by } 2, 3, \text{or } 5\right\}$. L_3 is regular. T/F?
4. $L_4 = \{w \in \{0, 1\}^* \mid w \text{ has at most } 2 \text{ 1s}\}$. L_4 is regular. T/F?

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 [1].

Inductive Definition

A regular expression $\mathbf{r}$ over an alphabet Σ 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	$\mathbf{r_1} \cdot \mathbf{r_2}$ denotes $R_1 R_2$
R^* is regular if R is	$\mathbf{r}^*$ denote 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 $(1+0)$ denotes same language $\{0,1\}$

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 $(1+0)$ denotes same language $\{0,1\}$
- Two regular expressions $\mathbf{r_1}$ and $\mathbf{r_2}$ are equivalent if $L(\mathbf{r_1}) = L(\mathbf{r_2})$.

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 $(1+0)$ denotes 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: $r^*s + t = ((r^*)s) + t$

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 $(1+0)$ denotes 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: $rst = (rs)t = r(st)$,
  $r + s + t = r + (s + t) = (r + s) + t$.

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 $(1 + 0)$ denotes 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 $\mathbf{r}^+ = \mathbf{rr}^*$. Hence if $L(\mathbf{r}) = R$ then $L(\mathbf{r}^+) = R^+$.

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 $(1+0)$ denotes 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 $\mathbf{r}^+ = \mathbf{rr}^*$. 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 $r \bullet s$.

Some examples of regular expressions

Interpreting regular expressions

1. $(\mathbf{0} + \mathbf{1})^*$:

Interpreting regular expressions

1. $(\mathbf{0}+\mathbf{1})^*$:
2. $(\mathbf{0}+\mathbf{1})^*\mathbf{001}(\mathbf{0}+\mathbf{1})^*$:

Interpreting regular expressions

1. $(\mathbf{0}+\mathbf{1})^*$:
2. $(\mathbf{0}+\mathbf{1})^*\mathbf{001}(\mathbf{0}+\mathbf{1})^*$:
3. $\mathbf{0}^*+(\mathbf{0}^*\mathbf{1}\mathbf{0}^*\mathbf{1}\mathbf{0}^*\mathbf{1}\mathbf{0}^*)^*$: with number of 1's divisible by 3

Interpreting regular expressions

1. $(\mathbf{0}+\mathbf{1})^*$:
2. $(\mathbf{0}+\mathbf{1})^*\mathbf{001}(\mathbf{0}+\mathbf{1})^*$:
3. $\mathbf{0}^*+(\mathbf{0}^*\mathbf{1}\mathbf{0}^*\mathbf{1}\mathbf{0}^*\mathbf{1}\mathbf{0}^*)^*$:   with number of 1's divisible by 3
4. $(\epsilon+\mathbf{1})(\mathbf{01})^*(\epsilon+\mathbf{0})$:

Creating regular expressions

1. All strings that end in 1011?

Creating regular expressions

1. All strings that end in 1011?
2. All strings except 11?

Creating regular expressions

1. All strings that end in 1011?
2. All strings except 11?
3. All strings that do not contain 000 as a subsequence?

Creating regular expressions

1. All strings that end in 1011?
2. All strings except 11?
3. All strings that do not contain 000 as a subsequence?
4. All strings that do not contain the substring 10?

Tying everything together

Consider the problem of a n-input AND function. The input (x) is a string n-digits long with an input alphabet $\Sigma_i = \{0, 1\}$ and has an output (y) which is the logical AND of all the elements of x. We knwo the language used to describe it is:

$$L_{AND_N} = \left\{ \begin{matrix} 0|0, & 1|1, & & \\ 0 \cdot 0|0, & 0 \cdot 1|0, & 1 \cdot 0|0, & 1 \cdot 1|1 \\ \vdots & \vdots & \vdots & \vdots \\ (0\cdot)^n|0, & (0\cdot)^{n-1}1|0, & \dots & (1\cdot)^n|1 \dots \end{matrix} \right\} \tag{3}$$

Formulate the regular expression which describes the above language:

Tying everything together

Consider the problem of a n-input AND function. The input (x) is a string n-digits long with an input alphabet $\Sigma_i = \{0, 1\}$ and has an output (y) which is the logical AND of all the elements of x. We knwo the language used to describe it is:

$$L_{AND_N} = \left\{ \begin{matrix} 0|0, & 1|1, & & \\ 0 \cdot 0|0, & 0 \cdot 1|0, & 1 \cdot 0|0, & 1 \cdot 1|1 \\ \vdots & \vdots & \vdots & \vdots \\ (0\cdot)^n|0, & (0\cdot)^{n-1}1|0, & \ldots & (1\cdot)^n|1 \ldots \end{matrix} \right\} \tag{3}$$

Formulate the regular expression which describes the above language: $\Sigma = \{0, 1, `\cdot', `|'\}$

$$r_{AND_N} = \underbrace{(\text{“}0\cdot\text{”} + \text{“}1\cdot\text{”})^* 0 (\text{“}0\cdot\text{”} + \text{“}1\cdot\text{”})^* \text{“}|0\text{”}} + \overbrace{(\text{“}1\cdot\text{”})^* \text{“}|1\text{”}}^{\text{all output 1 instances}}$$

Regular expressions in programming

One last expression....

Bit strings with odd number of 0s and 1s

Bit strings with odd number of 0s and 1s

The regular expression is

$$\big(00 + 11\big)^* (01 + 10) \Big(00 + 11 + (01 + 10)(00 + 11)^* (01 + 10)\Big)^*$$

Bit strings with odd number of 0s and 1s

The regular expression is

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

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