Alphabets, Strings, and Languages

Chapter 2

Alphabets and Strings

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

A ***string*** or ***word*** is a finite sequence, possibly empty, of symbols drawn from some alphabet Σ.

- ε is the empty string.
- Σ* is the set of all possible strings over the alphabet Σ.

Alphabet name	***Alphabet symbols***	***Example strings***
English alphabet	{a, b, c, …, z}	ε, aabbcg, aaaaa
Binary alphabet	{0, 1}	ε, 0, 001100
A star alphabet	{★ , ✪ , ✬ , ✯, ✡, ✰}	ε, ✪✪, ✪✯✯✰★✰

String Operations

Length: $|s|$ is the number of symbols in string s.

$|\varepsilon| = 0$ $\quad$ $|\texttt{1001101}| = 7$

Concatenation: xy is the ***concatenation*** of x and y.

If x = `good` and y = `bye`, then xy = `goodbye`

ε is the identity for concatenation. So, $x\,\varepsilon = \varepsilon\, x = x$

Replication: For each string w and each natural number i, the string w^i is defined as follows:

$w^0 = \varepsilon$

And if i > 0:

w^i = i copies of w concatenated together

So: $\texttt{a}^3 = \texttt{aaa}$

$(\texttt{bye})^2 = \texttt{byebye}$

$\texttt{a}^0\texttt{b}^3 = \texttt{bbb}$

More String Operations

Reverse: For each string w, w^R is the reverse of w.

Concatenation and Reverse: If w and x are strings, then $(wx)^R = x^R w^R$.

Example:

$(\texttt{nametag})^R = (\texttt{tag})^R (\texttt{name})^R = \texttt{gateman}$

Substrings

The string *x* is a **substring** of string *w* if it is a contiguous sequence of characters in *w*. If $|x| < |w|$, it is a ***proper substring***.

Note 1: Every string is a substring of itself.

Note 2: The empty string ε is a substring of every string.

Prefixes

p is a ***prefix*** of *t* if *t* = *px*

p is a ***proper prefix*** of *t* iff: *p* is a prefix of *t* and *p* ≠ *t*.

Note 1: Every string is a prefix of itself.

Note 2: The empty string ε is a prefix of every string.

Examples:

The ***prefixes*** of `abbb` are: ε, `a`, `ab`, `abb`, `abbb`.
The ***proper prefixes*** of `abbb` are: ε, `a`, `ab`, `abb`.

Suffixes

s is a ***suffix*** of *t* if *t* = *xs*

s is a ***proper suffix*** of *t* iff: *s* is a suffix of *t* and $s \neq t$.

Note 1: Every string is a suffix of itself.

Note 2: The empty string ε is a suffix of every string.

Examples:

The ***suffixes*** of `abbb` are: ε, `b`, `bb`, `bbb`, `abbb`.
The ***proper suffixes*** of `abbb` are: ε, `b`, `bb`, `bbb`.

Sets

- The set with no members is the ***empty set***, denoted {} or $\varnothing$
- If every element of set *A* is a member of set *B*, we say that *A* is a ***subset*** of *B*, denoted $A \subseteq B$
- If set *A* is a subset of set *B* and $A \neq B$, then *A* is a ***proper subset*** of *B.*
- The empty set is a subset of any set.
- Set membership: $x \in S$ denotes that *x* is in the set ***S***

More on Sets

- The ***union*** of two sets, $A \cup B$, is the set that contains everything that is in *A*, in *B*, or in both.

- The ***intersection*** of two sets, $A \cap B$, is the set that contains only those elements that are in both *A* and *B*.

- Two sets *A* and *B* are ***disjoint*** if they have no elements in common, i.e. $A \cap B = \varnothing$

- The ***set difference*** of *A* and *B*, *A* – *B*, is the set that contains everything that is in *A* but not in *B*.

Even More on Sets

- The ***complement*** of *A*, written $\overline{A}$, is the set containing everything that is in some universal set *U* that is not in *A*.

- The ***cardinality*** of a set *A*, written $|A|$, is the number of elements in *A*.

- The ***power set*** of *A, denoted P(A),* is the set of all subsets of *A*. Note that $|P(A)| = 2^{|A|}$.

Defining a Language

A ***language*** is a (finite or infinite) set of strings over an alphabet Σ.

Examples:

Let Σ = {a, b}

Some languages over Σ:

- ∅, (the language with no strings in it)
- {ε}, (the language containing just the empty string)
- {a, b},
- {ε, a, aa, aaa, aaaa, aaaaa}

How Large is a Language?

The smallest language over any Σ is $\emptyset$, i.e. the language with no strings in it, so size 0.

The language Σ^* contains a countably infinite number of strings: the empty string ε and all strings of any length containing only the characters `a` and `b`.

Example Language Definition

L = {*x* ∈ {a, b}* : all a's precede all b's}

Then*:*

abbb, aabb, and aaaaaaabbbbb are in *L*.

aba, ba, and abc are *not* in *L*.

What about: ε, a, aa, and bb?

Exercises

What are the following languages?

$L = \{x : \exists y \in \{\mathtt{a}, \mathtt{b}\}^* : x = y\mathtt{a}\}$

$L = \{w \in \{\mathtt{a}, \mathtt{b}\}^*$: no prefix of w contains $\mathtt{b}\}$

$L = \{w \in \{\mathtt{a}, \mathtt{b}\}^*$: no prefix of w starts with $\mathtt{a}\}$

$L = \{w \in \{\mathtt{a}, \mathtt{b}\}^*$: every prefix of w starts with $\mathtt{a}\}$

Functions on Languages

- Set operations
  - Union
  - Intersection
  - Complement
- Language operations
  - Concatenation
  - Kleene star

Concatenation of Languages

If L_1 and L_2 are languages over Σ:

$L_1L_2 = \{w \in \Sigma^* : \exists s \in L_1 \text{ and } \exists t \in L_2 \text{ such that } w = st\}$

Example:
L_1 = {`cat`, `dog`}
L_2 = {`apple`, `pear`}
L_1 L_2 ={`catapple`, `catpear`, `dogapple`, `dogpear`}

$\{\varepsilon\}$ is the identity for concatenation:
$L\{\varepsilon\} = \{\varepsilon\}L = L$

$\varnothing$ is the “zero” for concatenation:
$L\,\varnothing = \varnothing\, L = \varnothing$

Be Careful Concatenating Languages Defined Using Variables

$L_1 = \{\mathtt{a}^n : n \geq 0\}$
$L_2 = \{\mathtt{b}^n : n \geq 0\}$

$L_1L_2 = \{\mathtt{a}^n\mathtt{b}^m : n, m \geq 0\}$

$L_1L_2 \neq \{\mathtt{a}^n\mathtt{b}^n : n \geq 0\}$

Kleene Star

$L^* = \{\varepsilon\} \cup$

$\{w \in \Sigma^* : w = w_1\, w_2 \ldots w_k$ and

$w_1, w_2, \ldots w_k \in L\}$

Example:

$L = \{$dog, cat, fish$\}$

$L^* = \{\varepsilon$, dog, cat, fish, dogdog,

dogcat, fishcatfish,

fishdogdogfishcat, …$\}$

$L^+ = L\, L^*$

$L^+ = L^* - \{\varepsilon\}$ iff $\varepsilon \notin L$

Lexicographic Ordering

Lexicographic Order:

- Shortest first
- Within a given length, “dictionary” order

What is the lexicographic enumeration of:

$\{w \in \{\mathtt{a}, \mathtt{b}\}^* : |w| \text{ is even}\}$