Lecture 2

Strings and Sets

Decision Problems Versus Functions

A *decision problem* is a function with a one-bit output: "yes" or "no." To specify a decision problem, one must specify

- the set A of possible inputs, and
- the subset $B \subseteq A$ of "yes" instances.

For example, to decide if a given graph is connected, the set of possible inputs is the set of all (encodings of) graphs, and the "yes" instances are the connected graphs. To decide if a given number is a prime, the set of possible inputs is the set of all (binary encodings of) integers, and the "yes" instances are the primes.

In this course we will mostly consider decision problems as opposed to functions with more general outputs. We do this for mathematical simplicity and because the behavior we want to study is already present at this level.

Strings

Now to our first abstraction: we will always take the set of possible inputs to a decision problem to be the set of finite-length strings over some fixed finite

D. C. Kozen, *Automata and Computability*

alphabet (formal definitions below). We do this for uniformity and simplicity. Other types of data—graphs, the natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$, trees, even programs—can be encoded naturally as strings. By making this abstraction, we have to deal with only one data type and a few basic operations.

Definition 2.1

- An *alphabet* is any finite set. For example, we might use the alphabet $\{0, 1, 2, \ldots, 9\}$ if we are talking about decimal numbers; the set of all ASCII characters if talking about text; $\{0, 1\}$ if talking about bit strings. The only restriction is that the alphabet be finite. When speaking about an arbitrary finite alphabet abstractly, we usually denote it by the Greek letter Σ. We call elements of Σ *letters* or *symbols* and denote them by $a, b, c, \ldots$. We usually do not care at all about the nature of the elements of Σ, only that there are finitely many of them.

- A *string* over Σ is any finite-length sequence of elements of Σ. Example: if $\Sigma = \{a, b\}$, then $aabab$ is a string over Σ of length five. We use $x, y, z, \ldots$ to refer to strings.

- The *length* of a string x is the number of symbols in x. The length of x is denoted $|x|$. For example, $|aabab| = 5$.

- There is a unique string of length 0 over Σ called the *null string* or *empty string* and denoted by ϵ (Greek epsilon, not to be confused with the symbol for set containment $\in$). Thus $|\epsilon| = 0$.

- We write a^n for a string of a's of length n. For example, $a^5 = aaaaa$, $a^1 = a$, and $a^0 = \epsilon$. Formally, a^n is defined inductively:

$$\begin{aligned} a^0 &\overset{\text{def}}{=} \epsilon, \\ a^{n+1} &\overset{\text{def}}{=} a^n a. \end{aligned}$$

- The set of all strings over alphabet Σ is denoted Σ^*. For example,

$$\begin{aligned} \{a, b\}^* &= \{\epsilon, a, b, aa, ab, ba, bb, aaa, aab, \ldots\}, \\ \{a\}^* &= \{\epsilon, a, aa, aaa, aaaa, \ldots\} \\ &= \{a^n \mid n \geq 0\}. \end{aligned}$$ □

By convention, we take

$$\varnothing^* \overset{\text{def}}{=} \{\epsilon\},$$

where $\varnothing$ denotes the empty set. This may seem a bit strange, but there is good mathematical justification for it, which will become apparent shortly.

If Σ is nonempty, then Σ^* is an infinite set of finite-length strings. Be careful not to confuse strings and sets. We won't see any infinite strings until much later in the course. Here are some differences between strings and sets:

- $\{a, b\} = \{b, a\}$, but $ab \neq ba$;
- $\{a, a, b\} = \{a, b\}$, but $aab \neq ab$.

Note also that $\varnothing$, $\{\epsilon\}$, and ϵ are three different things. The first is a set with no elements; the second is a set with one element, namely ϵ; and the last is a string, not a set.

Operations on Strings

The operation of *concatenation* takes two strings x and y and makes a new string xy by putting them together end to end. The string xy is called the *concatenation* of x and y. Note that xy and yx are different in general. Here are some useful properties of concatenation.

- concatenation is *associative*: $(xy)z = x(yz)$;
- the null string ϵ is an *identity* for concatenation: $\epsilon x = x\epsilon = x$;
- $|xy| = |x| + |y|$.

A special case of the last equation is $a^m a^n = a^{m+n}$ for all $m, n \geq 0$.

A *monoid* is any algebraic structure consisting of a set with an associative binary operation and an identity for that operation. By our definitions above, the set Σ^* with string concatenation as the binary operation and ϵ as the identity is a monoid. We will see some other examples later in the course.

Definition 2.2

- We write x^n for the string obtained by concatenating n copies of x. For example, $(aab)^5 = aabaabaabaabaab$, $(aab)^1 = aab$, and $(aab)^0 = \epsilon$. Formally, x^n is defined inductively:

$$x^0 \stackrel{\text{def}}{=} \epsilon,$$
$$x^{n+1} \stackrel{\text{def}}{=} x^n x.$$

- If $a \in \Sigma$ and $x \in \Sigma^*$, we write $\#a(x)$ for the number of a's in x. For example, $\#0(001101001000) = 8$ and $\#1(00000) = 0$.

- A *prefix* of a string x is an initial substring of x; that is, a string y for which there exists a string z such that $x = yz$. For example, $abaab$ is a prefix of $abaababa$. The null string is a prefix of every string, and every string is a prefix of itself. A prefix y of x is a *proper* prefix of x if $y \neq \epsilon$ and $y \neq x$. □

Operations on Sets

We usually denote sets of strings (subsets of Σ^*) by $A, B, C, \ldots$. The *cardinality* (number of elements) of set A is denoted $|A|$. The empty set $\varnothing$ is the unique set of cardinality 0.

Let's define some useful operations on sets. Some of these you have probably seen before, some probably not.

- *Set union:*

  $$A \cup B \stackrel{\text{def}}{=} \{x \mid x \in A \text{ or } x \in B\}.$$

  In other words, x is in the union of A and B iff[1] either x is in A or x is in B. For example, $\{a, ab\} \cup \{ab, aab\} = \{a, ab, aab\}$.

- *Set intersection:*

  $$A \cap B \stackrel{\text{def}}{=} \{x \mid x \in A \text{ and } x \in B\}.$$

  In other words, x is in the intersection of A and B iff x is in both A and B. For example, $\{a, ab\} \cap \{ab, aab\} = \{ab\}$.

- *Complement in Σ^*:*

  $$\sim\!A \stackrel{\text{def}}{=} \{x \in \Sigma^* \mid x \notin A\}.$$

  For example,

  $$\sim\{\text{strings in } \Sigma^* \text{ of even length}\} = \{\text{strings in } \Sigma^* \text{ of odd length}\}.$$

  Unlike $\cup$ and $\cap$, the definition of $\sim$ depends on Σ^*. The set $\sim\!A$ is sometimes denoted $\Sigma^* - A$ to emphasize this dependence.

- *Set concatenation:*

  $$AB \stackrel{\text{def}}{=} \{xy \mid x \in A \text{ and } y \in B\}.$$

  In other words, z is in AB iff z can be written as a concatenation of two strings x and y, where $x \in A$ and $y \in B$. For example, $\{a, ab\}\{b, ba\} = \{ab, aba, abb, abba\}$. When forming a set concatenation, you include *all* strings that can be obtained in this way. Note that AB and BA are different sets in general. For example, $\{b, ba\}\{a, ab\} = \{ba, bab, baa, baab\}$.

[1] iff = if and only if.

- The *powers* A^n of a set A are defined inductively as follows:

  $$\begin{aligned} A^0 &\stackrel{\text{def}}{=} \{\epsilon\}, \\ A^{n+1} &\stackrel{\text{def}}{=} AA^n. \end{aligned}$$

  In other words, A^n is formed by concatenating n copies of A together. Taking $A^0 = \{\epsilon\}$ makes the property $A^{m+n} = A^m A^n$ hold, even when one of m or n is 0. For example,

  $$\begin{aligned} \{ab, aab\}^0 &= \{\epsilon\}, \\ \{ab, aab\}^1 &= \{ab, aab\}, \\ \{ab, aab\}^2 &= \{abab, abaab, aabab, aabaab\}, \\ \{ab, aab\}^3 &= \{ababab, ababaab, abaabab, aababab, \\ &\qquad abaabaab, aababaab, aabaabab, aabaabaab\}. \end{aligned}$$

  Also,

  $$\begin{aligned} \{a, b\}^n &= \{x \in \{a, b\}^* \mid |x| = n\} \\ &= \{\text{strings over } \{a, b\} \text{ of length } n\}. \end{aligned}$$

- The *asterate* A^* of a set A is the union of all finite powers of A:

  $$\begin{aligned} A^* &\stackrel{\text{def}}{=} \bigcup_{n \geq 0} A^n \\ &= A^0 \cup A^1 \cup A^2 \cup A^3 \cup \cdots. \end{aligned}$$

  Another way to say this is

  $$A^* = \{x_1 x_2 \cdots x_n \mid n \geq 0 \text{ and } x_i \in A,\ 1 \leq i \leq n\}.$$

  Note that n can be 0; thus the null string ϵ is in A^* for any A.

  We previously defined Σ^* to be the set of all finite-length strings over the alphabet Σ. This is exactly the asterate of the set Σ, so our notation is consistent.

- We define A^+ to be the union of all *nonzero* powers of A:

  $$A^+ \stackrel{\text{def}}{=} AA^* = \bigcup_{n \geq 1} A^n.$$

Here are some useful properties of these set operations:

- Set union, set intersection, and set concatenation are *associative*:

  $$\begin{aligned} (A \cup B) \cup C &= A \cup (B \cup C), \\ (A \cap B) \cap C &= A \cap (B \cap C), \\ (AB)C &= A(BC). \end{aligned}$$

- Set union and set intersection are *commutative*:

  $$A \cup B = B \cup A,$$
  $$A \cap B = B \cap A.$$

  As noted above, set concatenation is not.

- The null set $\varnothing$ is an *identity* for $\cup$:

  $$A \cup \varnothing = \varnothing \cup A = A.$$

- The set $\{\epsilon\}$ is an identity for set concatenation:

  $$\{\epsilon\}A = A\{\epsilon\} = A.$$

- The null set $\varnothing$ is an *annihilator* for set concatenation:

  $$A\varnothing = \varnothing A = \varnothing.$$

- Set union and intersection *distribute* over each other:

  $$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$
  $$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

- Set concatenation distributes over union:

  $$A(B \cup C) = AB \cup AC,$$
  $$(A \cup B)C = AC \cup BC.$$

  In fact, concatenation distributes over the union of any family of sets. If $\{B_i \mid i \in I\}$ is any family of sets indexed by another set I, finite or infinite, then

  $$A(\bigcup_{i \in I} B_i) = \bigcup_{i \in I} AB_i,$$
  $$(\bigcup_{i \in I} B_i)A = \bigcup_{i \in I} B_iA.$$

  Here $\bigcup_{i \in I} B_i$ denotes the union of all the sets B_i for $i \in I$. An element x is in this union iff it is in one of the B_i.

  Set concatenation does *not* distribute over intersection. For example, take $A = \{a, ab\}$, $B = \{b\}$, $C = \{\epsilon\}$, and see what you get when you compute $A(B \cap C)$ and $AB \cap AC$.

- The *De Morgan laws* hold:

  $$\sim\!(A \cup B) = \sim\!A \cap \sim\!B,$$
  $$\sim\!(A \cap B) = \sim\!A \cup \sim\!B.$$

- The asterate operation * satisfies the following properties:

$$\begin{aligned} A^*A^* &= A^*, \\ A^{**} &= A^*, \\ A^* &= \{\epsilon\} \cup AA^* = \{\epsilon\} \cup A^*A, \\ \varnothing^* &= \{\epsilon\}. \end{aligned}$$