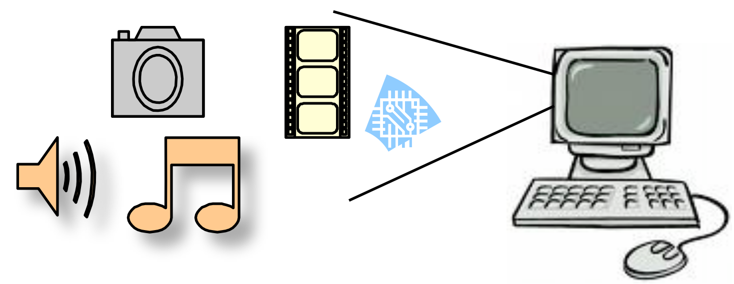
Words and Languages

Discrete Mathematics
Evgeny Skvortsov

Why Strings?

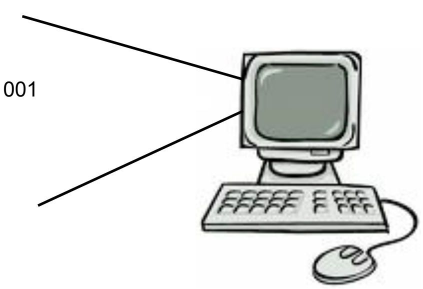
- Computer data is very diverse



- However, before it can be processed it must be converted into
- … strings

0010110101010101010101001010101010101010110101010010010101010101010101010101010101001

110010110101010101010101001010101010101010110101010010010101010101010101010101010101010



Alphabets and Strings

- An alphabet is any finite set. Σ

  Its elements are called symbols or letters

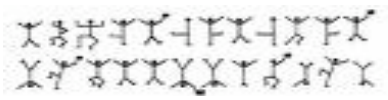
  {0,1} a binary alphabet

  {0,1,2,3,4,5,6,7,8,9}

  {a,b,…,x,y,z} Latin alphabet

  {а, б, ..., э, ю, я} Cyrillic

  .....

  {  } the dancing men alphabet

- A string over alphabet Σ is a sequence of symbols from Σ connected by means of juxtaposition or concatenation

  0010110101010101001010

Powers and Empty String

- Strings that are obtained by concatenation of the same number of symbols are grouped into powers of the alphabet.
- An induction definition:

  $\Sigma^1 = \Sigma$

  $\Sigma^{n+1} = \{xy \mid x \in \Sigma, y \in \Sigma^n\}$ where xy denotes the concatenation of x and y
- If $\Sigma = \{0,1\}$, then $\Sigma^2 = \{00,01,10,11\}$
- If $\Sigma = \{a,b,c,\ldots,z\}$, then $\Sigma^3 = \{act, bad, cat, den, \ldots\}$

  Note that fre, aat, lkj $\in \Sigma^3$
- Empty string λ is the string containing no symbols.

  λ is not the blank symbol, and λ cannot be a symbol in an alphabet
- $\Sigma^0 = \{\lambda\}$

More Powers

- If a string x is obtained by concatenation of n symbols, we say that it has length n, denoted $||x|| = n$
- $||\lambda|| = 0$
- $$\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots = \bigcup_{n=1}^{\infty} \Sigma^n$$
- $\Sigma^* = \Sigma^+ \cup \{\lambda\}$ Kleene star
- Examples:

$$\{0,1\}^+ = \{0,1,00,01,10,11,000,001,010,100,011,101,110,111,\ldots\}$$

$$\{a,b\}^* = \{\lambda, a, b, aa, ab, ba, bb, aaa, aab, aba, \ldots\}$$

Equality and Concatenation

- Two strings $x = x_1 x_2 \ldots x_k$ and $y = y_1 y_2 \ldots y_l$ are equal if k = l and $x_1 = y_1, x_2 = y_2, x_3 = y_3, \ldots, x_k = y_k$

- Concatenation of strings $x = x_1 x_2 \ldots x_k$ and $y = y_1 y_2 \ldots y_l$ is the string $xy = x_1 x_2 \ldots x_k y_1 y_2 \ldots y_l$
- Observe that ||xy|| = ||x|| + ||y||
- For the empty string we define

$$\lambda x = \lambda x_1 x_2 \ldots x_k = x_1 x_2 \ldots x_k = x$$
$$x\lambda = x_1 x_2 \ldots x_k \lambda = x_1 x_2 \ldots x_k = x$$
$$\lambda\lambda = \lambda$$

- Power of a string: $x^0 = \lambda, \quad x^1 = x, \quad x^2 = xx, \quad x^3 = xxx, \quad \ldots$

Prefixes and Suffixes

- For any strings x, w such that w = xy for some string y, string x is called a prefix of w.
- For any strings y, w such that w = xy for some string x, string y is called a suffix of w.
- Note that the empty string is a prefix and suffix of any string
- Examples:

  x =abbabaab Then its prefixes are λ,a,ab,abb,abba,abbab,…

  and suffixes are λ,b,ab,aab,baab,abaab,…

- A substring of w is a prefix of a suffix (or a suffix of a prefix)

substring

abbabaabbaababbaabbabaabbaababba

prefix suffix

Languages

- A (formal) language over alphabet Σ is a subset of Σ^*
  The empty language $\varnothing$ is also a language

- {001,010,100} is a language over alphabet {0,1}
- {0,10,100,110,1000,1010,1100,1110,10000,…}
  the language of binary expansions of even numbers
- The English language = the set of grammatically correct English texts
  over alphabet {a,…,z,A,…,Z,0,…,9,.,,,;,:,!,?,-, }
- Thue-Morse language {a,b,ab,ba,abba,baab,abbabaab,baababba,…}
- The language of properly placed parenthesis
  {(), (()),()(),((())),(())(),(()()),()()(),…} Dyck language

Representing Languages: Language Operations

- Finite languages can be described by a list of their elements:
  {if, then, while, for, to,…}
- Set theoretic operations:
  Union of languages $A \cup B$
  Intersection of languages $A \cap B$
  Complement of a language $\overline{A} = \Sigma^{*} - A$
- Examples
  A is the language of binary expansions of even numbers
  B is the language of binary expansions of odd numbers
  $A \cap B$, $A \cup B$, $\overline{A}$

Representing Languages: Concatenation

- Let A and B be languages over alphabet Σ. Then

  $AB = \{xy \mid x \in A,\ y \in B\}$

- Let
  $$A = \{a^k \mid k \in \mathbb{N}\}$$
  $$B = \{b^k \mid k \in \mathbb{N}\}$$
  $$C = \{c\}$$

  Then $ACB = \{a^k c b^l \mid k, l \in \mathbb{N}\}$

- Concatenation is not commutative

  Let $A = \{a,ab,c\}$ and $B = \{\lambda,b\}$. Then $AB = \{a,ab,abb,c,cb\}$ and $BA = \{a,ab,c,ba,bab,bc\}$.

  In particular, $|AB| = 5 \neq 6 = |BA|$

Properties of Concatenation

Theorem

Let Σ be an alphabet, let A, B, C be languages over Σ.

(1) $A\{\lambda\} = \{\lambda\}A = A$,

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

(3) $A(B \cap C) \subseteq AB \cap AC$,

(4) $(AB)C = A(BC)$

(5) $(B \cup C)A = BA \cup CA$

(6) $(B \cap C)A \subseteq BA \cap CA$

Proof

(4) Take $w \in (AB)C$. Then $w = xy$ such that $x \in AB$ and $y \in C$.

Then $x = uv$, where $u \in A$ and $v \in B$. If $u = u_1u_2\ldots u_k$,

$v = v_1v_2\ldots v_l$, and $y = y_1y_2\ldots y_m$, then

$$w = (u_1u_2\ldots u_kv_1v_2\ldots v_l)y_1y_2\ldots y_m = u_1u_2\ldots u_kv_1v_2\ldots v_ly_1y_2\ldots y_m$$
$$= u_1u_2\ldots u_k(v_1v_2\ldots v_ly_1y_2\ldots y_m) = u(vy) \in A(BC).$$

The reverse inclusion is similar.

Properties of Concatenation (cntd)

(6) Take $w \in (B \cap C)\,A$. Then $w = xy$, where $x \in B \cap C$ and $y \in A$. Thus $x \in B$, and so $w = xy \in BA$. Similarly, $x \in C$, hence $w = xy \in CA$. We conclude $w \in BA \cap CA$.

Q. E. D.

- Example

  Let $B = \{a\}$, $C = \{ab\}$, and $A = \{c,bc\}$. Then $BA = \{ac,abc\}$, $CA = \{abc,abbc\}$. Therefore $BA \cap CA = \{abc\}$, while $B \cap C = \varnothing$, and so $(B \cap C)A = \varnothing$.

Kleene Star

- For a language A over an alphabet Σ:

$$A^0 = \{\lambda\},\ A^1 = A, \quad \text{and for } n \in N,\ A^n = A^{n+1}A = \underbrace{AA\ldots A}_{n\ \text{times}}$$

$$A^+ = A \cup A^2 \cup A^3 \cup \ldots = \bigcup_{n=1}^{\infty} A^n \quad \text{the positive closure of } A$$

$$A^* = \{\lambda\} \cup A^+ = \bigcup_{n=0}^{\infty} A^n \quad \text{the Kleene closure or Kleene star}$$

- Let A = {aa,ab,ba,bb}. Then A^* is the language of all strings of even length.
- If B = {a,b} then BA is the language of all strings of odd length
- What are $\{a\}\{ba\}^*$ and $\{ab\}^*\{b\}$?

Regular Expressions

- An atomic language is a language that contains only one string, and this string has length 1. {a}

  For short we denote such a language simply by a
- Every language that contains only one string can be represented as a concatenation of atomic languages. A = {abba} = abba

  (Careful!!! The abba in the parenthesis is a string, while the abba in the end is a concatenation of languages.)
- Any finite language is a union of concatenations of atomic languages. $A = \{ab,ba,abba\} = ab \cup ba \cup abba$
- An expression constructed from atomic languages by means of concatenation, union, intersection, complementation, and Kleene star is called a regular expression

Regular Expressions: Examples

- What the languages a^*ba^*b, $(a \cup b)^* c^*$, $((ab \cup ba)^* c)^*$ are?

- Write a regular expression for the language over $\{a,b,c\}$ that contains string with exactly one occurrence of c
- with exactly two occurrences of c
- over $\{a,b,c,d\}$ with one occurrence of c and one occurrence of d
- with as many occurrences of c as you wish, but each such occurrence should be followed by an occurrence of d

Homework

Exercises from the Book:
No. 1, 7, 11, 13, 15 (page 317 – 318)