BBM401-Lecture 1: Strings, Languages, and Regular Expressions

Lecturer: Lale Özkahya

Resources for the presentation:
https://courses.engr.illinois.edu/cs498374/lectures.html

Definitions for strings

- Σ = finite alphabet of symbols
  Σ = {0,1}, or Σ={*a,b,c,…,z*}, or Σ=all ascii characters
- *string* or *word* = **finite** sequence of symbols of Σ
- *length* of a string *w* is denoted |*w*|. |cat|=3
- the *empty string* is denoted "ε". |ε| = 0.

Conventions

a, b, c, ...   denote strings of length 1; elements of Σ
w, x, y, z, ... denote strings of length 0 or more
A, B, C,...   denote sets of strings

Much ado about nothing

- ε is a *string* containing no symbols. It is not a set.
- {ε} is a *set* containing one string: the empty string ε. It is a *set*, not a string.
- Ø is the *empty set.* It contains no strings.
- {Ø} is a *set* containing one element, which itself is a set with no elements.

Concatenation & its properties

- If x and y are strings, xy denotes the concatenation
- associative: $(uv)w = u(vw)$ and we write uvw.
- NOT commutative: $ab \neq ba$
- identity element ε: $\varepsilon w = w\varepsilon = w$
- length (can be defined inductively)

  $|\varepsilon| = 0$

  $|a| = 1$

  $|au| = 1 + |u|$

Substrings, prefix, suffix, exponents

- v is a *substring* of w iff there exist strings x, y, such that $w=xvy$.
  - If $x=\varepsilon$ then v is a *prefix* of w.
  - If $y=\varepsilon$ then v is a *suffix* of w.
- If w is a string, then w^i is defined inductively by:

  $w^i = \varepsilon$ if $i=0$

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

  e.g. $(\text{blah})^4$ = blahblahblahblah

Set Concatenation

- If *X* and *Y* are sets of strings, then

  $XY = \{xy \mid x \text{ in } X \text{ and } y \text{ in } Y\}$

  e.g. *X* = {fido, rover, spot}, *Y* = {fluffy, tabby}

  then *XY* ={fidofluffy, fidotabby, roverfluffy, ...}

Σ^n, Σ^*, *and* Σ^+

- Σ^n defined as all strings over Σ of length *n* inductively:

  $\Sigma^0 = \{\varepsilon\}$

  $\Sigma^n = \Sigma\Sigma^{n-1}$ if $n > 0$

- Σ^* is the set of *all finite length strings:*

  $$\Sigma^* = \bigcup_{n \geq 0} \Sigma^n$$

- Σ^+ is the set of *all nonempty finite length strings:*

  $$\Sigma^+ = \bigcup_{n \geq 1} \Sigma^n = \Sigma\Sigma^*$$

Σ^n, Σ^*, *and* Σ^+

Examples

- $\Sigma=\{0,1\}$. Then $\Sigma^2=\{00,01,10,11\}$. $\Sigma^0 = \{\varepsilon\}$
- $\Sigma=\{a,b,c,..z, A,...,Z,\_,-,+,...$ *<other symbols>*$\}$.
  - $\cup_{n\leq 100} \Sigma^n$ contains all English words (and more)
  - Σ^* contains all books sold by Amazon (and more)
- $\Sigma=\emptyset$. Then $\Sigma^1=\Sigma^2=...= \Sigma^{100} = \emptyset$

  $\Sigma^0 = \{\varepsilon\}$

$$\Sigma^* = \cup_{n \geq 0} \Sigma^n$$

- What is the cardinality of Σ^n ?

  $|\Sigma^n| = |\Sigma|^n$

- What is the cardinality of Σ^* ?

  $|\Sigma^*| = \aleph_0 = |N|$ (provided that Σ is nonempty)

- What is the length of the longest element of Σ^* ?

  there is no longest element

- Are there any infinitely long strings in Σ^* ?

  NO! Σ^* has strings of arbitrary size, but no single unbounded (infinite) string

Canonical Order

- Enumerate Σ^* in order of increasing length strings
  and for strings of same length, in dictionary order

e.g. $\{0,1\}^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, ...\}$

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

Inductive Definitions

- Often strings and functions on strings are defined inductively.
- Example: w^R, the reverse of word w is defined:

  if $|w| = 0$, then $w = \varepsilon$, and $w^R = \varepsilon$

  if $|w| > 0$, then $w = au$ for some a in Σ and u in Σ^* with $|u| < |w|$

  and then $w^R = u^R a$

$(abc)^R = (bc)^R a = (c^R b)a = ((c\varepsilon)^R b)a = ((\varepsilon^R c)b)a = cba$

Inductive proofs follow inductive defs

Theorem: For any strings u and v, $(uv)^R = v^R u^R$

e.g. $(dogcat)^R = (cat)^R(dog)^R = tacgod$

Proof: by induction.

On what??

$|uv| = |u| + |v|$?

$|u|$?

$|v|$?

$|u|$ and in induction, do an *inner induction* on $|v|$?

Induction on $|u|$

Proof: by induction on $|u|$ is most natural

Base case: If $|u| = 0$, then $u = \varepsilon$, and for any v,

$$(uv)^R = (\varepsilon v)^R = v^R = v^R\varepsilon = v^R\varepsilon^R = v^R u^R$$

Inductive Step

- Assume for any u of length $< n$ that:

  for all v, $(uv)^R = v^R u^R$

- Let u be an arbitrary string of length n.

  Then $u = ay$ for some a in Σ and $|y| < n$

Then

$(uv)^R$	$= ((ay)v)^R$	because $u = ay$
	$= (a(yv))^R$	because concatenation is associative
	$= (yv)^R a$	by inductive definition of reverse
	$= (v^R y^R)a$	applying inductive hypothesis $(\lvert y \rvert < n)$
	$= v^R(y^R a)$	because concatenation is associative
	$= v^R(ay)^R$	by inductive definition of reverse
	$= v^R u^R$	because $u = ay$

Induction on |v|

- Base cases need $|v| = 0$ *or* 1.
- Assume for any v of length $< n$ that:

  for all v, $(uv)^R = v^R u^R$
- Let v be an arbitrary string of length $n > 1$.

  Then $v = ax$ for some a in Σ and $|x| < n$

Then

$(uv)^R$	$= (u(ax))^R$	because $v = ax$
	$= ((ua)x)^R$	because concatenation is associative
	$= x^R(ua)^R$	applying inductive hypothesis ($|x| < n$)
	$= x^R(a^R u^R)$	applying inductive hypothesis ($|a| < n$)
	$= x^R(au^R)$	($a=a^R$ via definition of reverse)
	$= (x^R a)u^R$	because concatenation is associative
	$= (ax)^R u^R$	by inductive definition of reverse
	$= v^R u^R$	because $v = ax$

Languages

- If Σ is a (finite) alphabet, then a *language* is *any subset of* Σ^* [often, Σ is clear from context]
- Thus, a language is just a set of strings (words)

  Examples
  - $\{\varepsilon\}$
  - $\{w : |w| > 5\}$
  - {*w*: *w* is a syntactically correct Python program}
  - {*w*: *w* is the text of a book in the Library of Congress}
  - $\emptyset$

- The *complement* of a language L is $\overline{L} = \Sigma^* - L$
  (where $A - B$ is set subtraction)

- L^n, L^*, and L^+ defined as were Σ^n, Σ^*, and Σ^+
  note that a word in L^n is the
  concatenation of n *possibly different* words in L.

Boundary conditions: what is $\{\varepsilon\}^*$?
what is $\emptyset^*$?

The Study of Languages is Important

- A fundamental computing problem:

  Given (some description of) L, and w, is w in L?

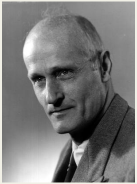
- Examples
  - H = {<G> | <G> encodes a graph that contains a Hamiltonian cycle}
  - G = {*n* | *n is even and is the sum of two primes*}

    Goldbach's conjecture: G = {all even numbers > 2}
  - P = {*p* | *p* is a Python program that for any input will terminate properly}

Regular Expressions

Regular Expressions

- a way to denote the regular languages
- simple *patterns* to describe related strings
- useful in
  - text search (editors, Unix/grep)
  - compilers: lexical analysis
  - compact way of representing sets of strings
- dates back to 50's: Stephen Kleene, who has a star named after him*



* The star named after him is the Kleene star "*"

Inductive Definition

A regular expression ***r*** over alphabet Σ is one of the following:

Base cases

- **∅** denotes the language L(**∅**) = ∅ = { }
- **ε** denotes the language L(**ε**) = {ε}
- ***a*** for *a* in Σ, denotes the language L(***a***) = {*a*}

Inductive Definition

A regular expression **r** over alphabet Σ is one of the following:

Inductively defined cases

If $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ are regular expressions denoting languages R_1 and R_2, then

$(\boldsymbol{r_1 + r_2})$ is a regular expression denoting $R_1 \cup R_2$

$(\boldsymbol{r_1 r_2})$ is a regular expression denoting $R_1 R_2$

$(\boldsymbol{r_1})^*$ is a regular expression denoting $(R_1)^*$

Compare with regular languages

REGULAR LANGUAGES

- Ø regular
- {ε} regular
- {*a*} regular for *a* in Σ
- R_1 U R_2 regular if both are
- R_1R_2 is regular if both are
- R^* is regular if *R* is.

REGULAR EXPRESSIONS

- **Ø** denotes Ø
- **ε** denotes {ε}
- ***a*** denotes {*a*}
- $\mathbf{r_1+r_2}$ denotes R_1 U R_2
- $\mathbf{r_1r_2}$ denotes R_1R_2
- $\mathbf{r}^*$ denotes R^*

Regular expressions denote regular languages

(they show the operations used to form the language)

Parentheses

- Omit parentheses by adopting precedence order: *, concat, +. E.g., $r^* s + t = ((r^*)s)+t$
- Omit parentheses by associativity of each of these operations. E.g., $rst = (rs)t = r(st)$

Superscript +

- For convenience, define $\boldsymbol{r^+ = rr^*}$

  so if $\boldsymbol{r}$ denotes language R, then $\boldsymbol{r^+}$ denotes R^+

Other notation

- $r + s$, $r \cup s$, and $r|s$ all denote the "or" or union
- rs is sometimes written $r \bullet s$

Examples

- $(0+1)^*001(0+1)^*$
  - strings with 001 as a substring
- $0^* + (0^*10^*10^*10^*)^*$
  - strings with a number of 1's divisible by 3
- $\emptyset 0$
  - concatenation of anything in here { } with anything in here {0}, so = { } = Ø (no strings may be so formed)
- $(\varepsilon+1)(01)^*(\varepsilon+0)$
  - alternating 0s and 1s
- $(\varepsilon+0)(1+10)^*$
  - strings without two consecutive 0s

Challenge: create regular expressions

- bitstrings with either the pattern 001 or the pattern 100 occurring somewhere

  one answer: $(0+1)^*001(0+1)^* + (0+1)^*100(0+1)^*$

- bitstrings with an odd number of 1s

  one answer: $0^*10^*(0^*10^*10^*)^*$

Real challenge: bitstrings with an odd number of 1s AND an odd number of 0s

Regular Expression Identities

- $r^*r^* = r^*$
- $(r^*)^* = r^*$
- $rr^* = r^*r$
- $(rs)^*r = r(sr)^*$
- $(r+s)^* = (r^*s^*)^* = (r^* + s^*)^* = (r+s^*)^* = \ldots$

An inductively defined language

Define *L* over $\{0,1\}^*$ by:

- ε is in *L*
- if *w* is in *L*, *then* 0*w*1 is in *L*

What do strings in *L* look like?

Give a characterization of *L* and prove it correct.

Can you find a regular expression for *L* ?

Conjecture: $L = \{0^i 1^i : i \geq 0\}$

How can we prove this is correct?

Prove (by induction) that

(a) $L \subseteq \{0^i 1^i : i \geq 0\}$

(b) $L \supseteq \{0^i 1^i : i \geq 0\}$

$$L \subseteq \{0^i 1^i : i \geq 0\}$$

Show by induction on $|w|$, that if w is in L, then w is of the form 0^i1^i.

Base case: $|w| = 0$.

Then $w = \varepsilon = 0^0 1^0$

Let $n > 0$, and assume for all $k < n$ that

for any w in L with $|w| = k$, w is of form 0^i1^i

Inductive step

Now consider arbitrary w in L, with $|w| = n$.

Then $w=0u1$ where u in L *has size* $n-2 < n$
(by definition of L)

By induction, u is of form 0^i1^i.

Then $w = 0u1 = 00^i1^i1 = 0^{i+1}1^{i+1}$, *the required form*

$$L \supseteq \{0^i 1^i : i \geq 0\}$$

Show by induction on $|w|$, that if w is of the form $0^i 1^i$, then w is in L.

Base case: $|w| = 0$.

Then $w = 0^0 1^0 = \varepsilon$, which is in L by definition

Inductive step:

Let $n > 0$, and assume for all $k < n$ that $0^k 1^k$ in L

$0^n 1^n = 0 0^{n-1} 1^{n-1} 1 = 0u1$, with u in L by induction

Since u in L, so is $0u1 = 0^n 1^n$ by definition of L