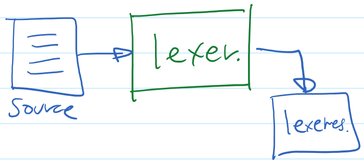
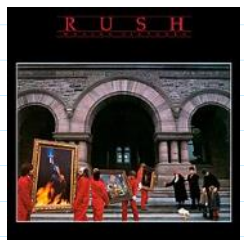
4 Regular Expressions

Wednesday, August 30, 2023 10:56 AM



Linguistics.-

- Orthography.- How to write words?
  apple manzana pomme [illegible]
- Syntax.- How to write sentences?
  "has jump apple dog green"
- Semantics.- What do sentences mean?
  "colorless green ideas sleep fouriously"
  "the chickens are ready to eat"
  "moving pictures" ←



while
for
int

- HOW TO WRITE WORDS?
  FOR
  WHILE
  IF
  - user defined names: "identifiers"
  - "Literals"
    - numerical values
      - integers
      - real numbers.
      - complex numbers
      - scientific numbers.
    - strings.
    - arrays,
    - dictionaries
    - tuples

- MATHEMATICAL LANGUAGES.
  - S. Kleeny
  - E. Post
  - Alphabet.- a finite set of symbols.
  - Language.- a set of sequences/words made from symbols in an alphabet.

e.g. $\Sigma = \{a, b\}$

$L = \{a, aa, aba, bbaa\}$

- Specification problem
  How to precisely descrive a Language

  $\Sigma = \{a, b\}$

  $L = \{a, ab, bab, abb, \ldots\}$?

- Recognition Problem.

  given a word w and Language L,

  $w \in L$?

SPECIFICATION: REGULAR EXPRESSIONS

Σ, | * ( )

Regular Expressions:

- a word w is a regular expression. $\{w\}$
- if r_1 and r_2 are regular expressions
  - $r_1 r_2$ $\{w_1 w_2 \mid w_1 \in r_1 \wedge w_2 \in r_2\}$
  - $r_1 \mid r_2$ $\{w \mid w \in r_1 \vee w \in r_2\}$
  - $r*$ $\{w \mid w$ is zero or more repetitions of a word in $r\}$

eg. $(a \mid b)b = \{ab, bb\}$

$a : \{a\}$ $\quad a|b = \{a, b\}$

$b : \{b\}$

$ab* = \{a, ab, abb, abbb, abbbb, \ldots\}$

$a* = \{\lambda, a, aa, aaa, aaaa, \ldots\}$

$(aa \mid ab)* = \{\lambda, aa, ab, aaab, abaa, aaaa, abab, \ldots\}$ ← aa aa aaaa ab

$((aa)*)|((ab)*) =$ ← aaaaaaaa ... abababab ...

Shorthands

$r? \equiv (r \mid \lambda)$

$r+ \equiv r(r*)$

$\Sigma = \{a, b\}$

$ab?a = \{aa, aba\}$

$l+ = \{\lambda, l, ll, lll, llll, \ldots\}$

$r+ \equiv r(r*)$

$ab?a = \{aa, aba\}$

$b^+ = \{b, bb, bbb, bbbb, \ldots\}$

- REGULAR EXPRESSIONS & PROGRAMMING LANGUAGES
  - note: we cannot use regex to specify the "Language of all valid programs"

[¿ what is the regex to validate HTML? closed]

Why?

eg. $\{a^n b^n \mid n \in \mathbb{Z}\} = \{ab, aabb, aaabbb, \ldots\}$

{([])} ✓ {([)]} ✗ <a> .... <h3> .... </h3>

- Good enough to describe "tokens"

  Token:- a category of atomic strings.

  lexeme :- instance of a token.

E.G

```
while ( flag == 62.3 else if ) 14 7 != x + var_1
```

Keyword | Ident | oper | lit real | Keyword | lit int | oper | Ident.

Keywords = (while | if | else | (|) | do | return | ....)

Some tokens are infinite. ⊏ integer literals, identifiers.

Σ = keyboard symbols (ASCII)

Short-hands

$[0\text{-}9] \equiv (0|1|2|\ldots|9)$

$[A\text{-}Z] \equiv (A|B|C|D|\ldots|Z)$

$[a\text{-}z] \equiv (a|b|c|d|\ldots|z)$

Regex for integer literals.

$(+|-)?([0-9])^+$ { -1, 123, -9, -00, 000, +3, +000 }

Regex for identifiers:

Fortran $([A-Z]|[a-z])([A-Z]|[a-z]|[0-9])^*$

Pascal $([A-Z]|[a-z])([A-Z]|[a-z]|[0-9]|'_')^*$ { A, J, App13, ~~3d~~, }

Python $([A-Z]|[a-z]|'_')([A-Z]|[a-z]|[0-9]|'_')^*$ { app_3, a_1_x ... }

{ __init__, __name__, ... ------, }

we can write reg-ex for: phone numbers
url's
serial no
licence plates
file names.

Regex for scientific notation

1239e-5

$(+|-)?[0-9]^+(.[0-9]^+)?\ 'e'\ (+|-)?[0-9]^+$

0.3e5

— EOF —