

Compiladores

Aula 2

Expressões Regulares

Prof. Dr. Luiz Eduardo G. Martins

UNIFESP



Expressões Regulares

- No processo de análise léxica, os *tokens* válidos da linguagem normalmente são especificados por *expressões regulares*
- Expressões regulares representam *padrões de cadeias de caracteres*
- Uma expressão regular r é completamente definida pelo conjunto de cadeias de caracteres com as quais ela “casa”
- Esse conjunto é chamado de *linguagem gerada pela expressão regular*, e é denotado como $L(r)$

Expressões Regulares

- Expressões regulares básicas
 - São os caracteres em separado do alfabeto
 - **Alfabeto**: conjunto de símbolos legais da linguagem, usualmente denotado pelo símbolo grego Σ
 - Dado um caractere a do alfabeto Σ indicamos que a expressão regular **a** casa com o caractere a escrevendo $L(\mathbf{a}) = \{a\}$

Expressões Regulares

- Expressões regulares básicas
 - Precisamos de símbolos adicionais para situações especiais
 - Utilizamos o meta-símbolo ϵ para denotar a **cadeia vazia** (uma cadeia sem caracteres)
 - Utilizamos o meta-símbolo Φ para denotar a linguagem que não casa com nenhuma cadeia de caracteres (linguagem vazia)

$$L(\epsilon) = \{\epsilon\}$$

$$L(\Phi) = \{\}$$

Expressões Regulares

- Operações básicas de expressões regulares
 - Escolha entre **alternativas**
 - Se r e s são expressões regulares, então $r \mid s$ é uma expressão regular que casa com qualquer cadeia que case com r ou com s
 - Em termos de linguagem, a linguagem de $r \mid s$ é a união de r e s

$$L(r \mid s) = L(r) \cup L(s)$$

Exemplos:

Para a expressão regular $a \mid b$,

$$L(a \mid b) = L(a) \cup L(b) = \{a\} \cup \{b\} = \{a, b\}$$

Para a expressão regular $a \mid \epsilon$

$$L(a \mid \epsilon) = \{a, \epsilon\}$$

Para a expressão regular $a \mid b \mid c \mid d$

$$L(a \mid b \mid c \mid d) = \{a, b, c, d\}$$

Expressões Regulares

- Operações básicas de expressões regulares

- Concatenação

- A concatenação de duas expressões regulares r e s é denotada por rs , e casa com qualquer cadeia de caracteres que seja a concatenação de duas cadeias, desde que a 1ª case com r e a 2ª case com s

$$L(rs) = L(r)L(s)$$

Exemplos:

Para a expressão regular $(a|b)c$,

$$L((a|b)c) = L(a|b)L(c) = \{a, b\}\{c\} = \{ac, bc\}$$

Para a expressão regular rs , em que $r = (a|b)$ $s = (c|d)$, temos

$$L(rs) = L((a|b)(c|d)) = L(a|b)L(c|d) = \{a, b\}\{c, d\} = \{ac, ad, bc, bd\}$$

Expressões Regulares

- Operações básicas de expressões regulares

- Repetição

- Seja uma expressão regular r , a operação de repetição é denotada por r^*
 - Essa operação é também chamada de fecho (de *Kleene*)
 - A expressão r^* casa com qualquer concatenação finita de cadeias de caracteres (inclusive a cadeia vazia), desde que cada cadeia case com r

Seja a expressão regular a^* ,

$$L(a^*) = L(a)^* = \{a\}^* = \{\epsilon, a, aa, aaa, \dots\}$$

Exemplo

Considere a expressão regular $(a|bb)^*$

$$L((a|bb)^*) = L(a|bb)^* = \{a, bb\}^* = \{\epsilon, a, bb, aa, abb, bba, bbbb, aaa, aabb, abba, bbba, abbabb, \dots\}$$

Expressões Regulares

- Operações básicas de expressões regulares
 - Precedência de operações
 - Repetição
 - Concatenação
 - Escolha (alternativas)
 - Uso de parênteses
 - Quando queremos indicar uma precedência diferente, devemos utilizar parênteses
 - Exemplos:
 - $(a|b)c$ a escolha terá precedência sobre a concatenação
 - $(a|b)^*$ a escolha terá precedência sobre a repetição

Expressões Regulares

- Operações básicas de expressões regulares
 - Nomes para expressões regulares
 - É útil simplificar a notação com nomes significativos para expressões regulares

Exemplo:

dígito dígito*

onde

dígito = 0|1|2|...|9

dizemos que o nome *dígito* é uma **definição regular**

Expressões Regulares

- Definição de expressão regular
 - Uma expressão regular é uma das seguintes:
 1. Uma expressão regular **básica**, composta por um único caractere **a**, onde *a* pertence a um alfabeto Σ de caracteres legais; o metacaractere ϵ ; ou o metacaractere Φ
 - No 1º caso, $L(a) = \{a\}$
 - No 2º caso $L(\epsilon) = \{\epsilon\}$
 - No 3º caso $L(\Phi) = \{\}$
 2. Uma expressão da forma $r|s$, onde r e s são expressões regulares
 $L(r|s) = L(r) \cup L(s)$
 3. Uma expressão da forma rs , onde r e s são expressões regulares
 $L(rs) = L(r)L(s)$

Expressões Regulares

- Definição de expressão regular

4. Uma expressão da forma r^* , onde r é uma expressão regular

$$L(r^*) = L(r)^*$$

5. Uma expressão da forma (r) , onde r é uma expressão regular

$$L((r)) = L(r)$$

Os parênteses são utilizados apenas para ajustar a precedência dos operadores

Expressões Regulares

- Operações estendidas de Expressões Regulares
 - Operações que estendem o conjunto básico
 - Uma ou mais repetições
 - Seja r uma expressão regular, r^+ indica uma ou mais repetições de r
 - Intervalo de caracteres
 - Utiliza-se os metacaracteres colchetes e hífen
 - Exemplos: $[a - z]$ $[0 - 9]$ $[a-zA-Z]$ $[abc]$
 - Essa notação com colchetes é chamada de **classe de caracteres**
 - Qualquer caractere fora de um conjunto
 - Utiliza-se os metacaracteres til (\sim) ou circunflexo ($^$), que indica “não” ou o complemento
 - Exemplos:
 $[^a]$ expressão regular para um caractere no alfabeto que não seja a
 $\sim(a|b|c)$ ou abc qualquer caractere que não seja a , b ou c

Expressões Regulares

- Operações estendidas de Expressões Regulares
 - Operações que estendem o conjunto básico
 - Subexpressões opcionais
 - Utiliza-se o metacaractere interrogação (?), seja r uma expressão regular, $r?$ indica que as cadeias que casam com r são opcionais
 - Exemplo:
 - $natural = [0-9]^+$
 - $naturalComSinal = (+|-)? natural$
 - Qualquer caractere
 - Utiliza-se o metacaractere ponto "." para representar o casamento com qualquer caractere do alfabeto
 - Exemplo: todas as cadeias de caractere que contêm pelo menos um m
 - $.^*m.^*$

Expressões Regulares

- Operações estendidas de Expressões Regulares
 - Operações que estendem o conjunto básico
 - Caractere de escape
 - Utiliza-se o metacaractere barra invertida “\” seguido do caractere literal que pretende-se especificar
 - O caractere de escape permite que um metacaractere seja interpretado de forma literal
 - Exemplo: $r = [0-9]\.[0-9]$
 - $L(r) = \{0.0, 0.1, 0.2..., 1.0, 1.1, 1.2...9.9\}$
- *Testador de expressão regular*
 - *Regex Tester*
 - <https://regex101.com>

Expressões Regulares

- Ambiguidade
 - Ocorre quando uma cadeia pode casar com mais do que uma expressão regular
 - Formas de resolver
 - Palavras reservadas
 - Deve haver uma expressão regular para cada palavra reservada
 - Exemplos: **if else do while**
 - Princípio da subcadeia mais longa
 - A expressão regular que casa com a cadeia mais longa deve ser adotada
 - Exemplos: **>= == <= <> !=**

Expressões Regulares

- Delimitadores de Lexemas
 - Espaço em branco
 - Caractere de tabulação
 - Caractere de mudança de linha
 - Caracteres que não casam com a expressão regular em análise
 - Exemplos
 - $x = 10$ (delimitados por espaços em branco)
 - $a = b + c$

Expressões Regulares

- Exercícios

1) Apresente o conjunto de cadeias de caracteres (com pelo menos 5 elementos, quando couber) que casam com as expressões regulares

a) a

$$L(a) = \{a\}$$

b) a^*

$$L(a^*) = L(a)^* = \{\epsilon, a, aa, aaa, aaaa, \dots\}$$

c) a^+

$$L(a^+) = L(a)^+ = \{a, aa, aaa, aaaa, aaaaa, \dots\}$$

d) $(a|b|c)^+$

e) $(ab)^*$

f) $(0|1)^+$

g) $ab^*(c|\epsilon)$

h) $(ab)^*(cd)^*$

Expressões Regulares

- Exercícios

1) Apresente o conjunto de cadeias de caracteres (com pelo menos 5 elementos, quando couber) que casam com as expressões regulares

a) a

$$L(a) = \{a\}$$

b) a^*

$$L(a^*) = L(a)^* = \{\epsilon, a, aa, aaa, aaaa, \dots\}$$

c) a^+

$$L(a^+) = L(a)^+ = \{a, aa, aaa, aaaa, aaaaa, \dots\}$$

d) $(a|b|c)^+$

$$L((a|b|c)^+) = \{a, b, c, aa, ab, bc, abc, aabbcc, \dots\}$$

e) $(ab)^*$

$$L((ab)^*) = \{\epsilon, ab, abab, ababab, abababab, \dots\}$$

f) $(0|1)^+$

$$L((0|1)^+) = \{0, 1, 01, 11, 011, \dots\}$$

g) $ab^*(c|\epsilon)$

$$L(ab^*(c|\epsilon)) = \{a, ab, ac, abb, abc, \dots\}$$

h) $(ab)^*(cd)^*$

$$L((ab)^*(cd)^*) = \{\epsilon, ab, cd, abab, cdcd, abcd, \dots\}$$

Expressões Regulares

- Exercícios

2) A partir das cadeias de caracteres apresentadas, encontre as expressões regulares que as definem:

a) *{a, b, ab, abba, aaab, baaa, ...}*

b) *{abc, aabc, aaabc, aaaabc, ...}*

c) *{1, 2, 20, 1234, 56789, 803459, ...}*

d) *{if, else, do, while}*

e) *{casa, nota, mEdia, Peso, zap, IRPF, ...}*

f) *{nota01, p1, xyz, xyz123, cod2f1,...}*

g) *números hexadecimais em linguagem C*

h) *números reais*

Expressões Regulares

- Exercícios

2) A partir das cadeias de caracteres apresentadas, encontre as expressões regulares que as definem:

- a) $\{a, b, ab, abba, aaab, baaa, \dots\}$ $(a|b)^+$
- b) $\{abc, aabc, aaabc, aaaabc, \dots\}$ a^*abc
- c) $\{1, 2, 20, 1234, 56789, 803459, \dots\}$ $[0-9]^+$
- d) $\{if, else, do, while\}$ $if | else | do | while$
- e) $\{casa, nota, media, Peso, zap, IRPF, \dots\}$ $[a-zA-Z]^+$
- f) $\{nota01, p1, xyz, xyz123, cod2f1, \dots\}$ $[a-z]([a-z]|[0-9])^*$
- g) *números hexadecimais em linguagem C* $0x([0-9]|[A-F])^+$
- h) *números reais* $(+|-)?[0-9]^+(\.[0-9]^+)?$

Expressões Regulares

- Exercícios

3) Considere o alfabeto $\Sigma = \{a,b,c,d...,z\}$

a) *Apresente uma expressão regular que aceite cadeias de caracteres que contêm exatamente um z*

$[\wedge z]^* z [\wedge z]^*$

b) *Apresente uma expressão regular que aceite cadeias de caracteres que contêm no máximo um z*

$[\wedge z]^* [z | \epsilon] [\wedge z]^*$ ou $[\wedge z]^* z ? [\wedge z]^*$

4) *Considere o alfabeto $\Sigma = \{a,b,c\}$, e apresente uma expressão regular que aceite cadeias que não contêm dois bs consecutivos*

$(a | c | ba | bc)^* (b | \epsilon)$

Expressões Regulares

- Bibliografia consultada

LOUDEN, K. C. **Compiladores: princípios e práticas.**

São Paulo: Pioneira Thompson Learning, 2004

AHO, A. V.; LAM, M. S.; SETHI, R. e ULLMAN, J. D.

Compiladores: princípios, técnicas e ferramentas.

2ª edição – São Paulo: Pearson Addison-Wesley,
2008