

Análise léxica

Especificação dos *tokens*

Prof. Edson Alves

Faculdade UnB Gama

Definição de alfabeto

Um alfabeto, ou classe de caracteres, é um conjunto finito de símbolos.

Exemplos de alfabetos: ASCII, EBCDIC, a alfabeto binário $\{ 0, 1 \}$, os dígitos decimais, etc.

Definição de cadeia

Uma cadeia sobre um alfabeto \mathcal{A} é uma sequência finita de elementos de \mathcal{A} . Os termos sentença, palavra e string são geralmente usados como sinônimos de cadeia.

Conceitos associados à cadeias

- ▶ O comprimento (número de caracteres) de uma cadeia s é denotado por $|s|$
- ▶ A cadeia vazia ϵ tem comprimento igual a zero
- ▶ Um prefixo de s é uma cadeia obtida pela remoção de zero ou mais caracteres do fim de s
- ▶ Um sufixo de s é uma cadeia obtida pela remoção de zero ou mais caracteres do início de s
- ▶ Uma subcadeia de s é uma cadeia obtida pela remoção de um prefixo e de um sufixo de s
- ▶ Um prefixo, sufixo ou subcadeia de s são ditos próprios se diferem de ϵ e de s
- ▶ Um subsequência de s é uma cadeia obtida pela remoção de zero ou mais símbolos de s , não necessariamente contíguos

Definição de linguagem

Uma linguagem é um conjunto de cadeias sobre algum alfabeto \mathcal{A} fixo.

Esta definição contempla também linguagens abstratas como \emptyset (o conjunto vazio), ou $\{ \epsilon \}$, o conjunto contendo apenas a cadeia vazia.

Operações em cadeias

- ▶ Se x e y são duas cadeias, então a concatenação de x e y , denotada xy , é a cadeia formada pelo acréscimo, ao final de x , de todos os caracteres de y , na mesma ordem
- ▶ Por exemplo, se $x = \text{"rodo"}$ e $y = \text{"via"}$, então $xy = \text{"rodovia"}$
- ▶ A cadeia vazia ϵ é o elemento neutro da concatenação
- ▶ Se a concatenação for visualizada como um produto, é possível definir uma “exponenciação” de cadeias
- ▶ Seja s uma cadeia e n um natural. Então
 1. $s^0 = \epsilon$
 2. $s^n = ss^{n-1}$

Operações em linguagens

Sejam L e M duas linguagens. São definidas as seguintes operações sobre linguagens:

Operação	Notação	Definição
união	$L \cup M$	$L \cup M = \{ s \mid s \in L \vee s \in M \}$
concatenação	LM	$LM = \{ st \mid s \in L \wedge t \in M \}$
fechamento de Kleene	L^*	$L^* = \bigcup_{i=0}^{\infty} L^i$
fechamento positivo	L^+	$L^+ = \bigcup_{i=1}^{\infty} L^i$

Exemplos de operações em linguagens

Seja $L = \{ A, B, C, \dots Z, a, b, c, \dots z \}$ e $M = \{ 0, 1, 2, \dots 9 \}$. Então:

1. $L \cup M$ é o conjunto de letras e dígitos
2. LM é o conjunto de cadeias formadas por uma letra, seguida de um dígito
3. L^4 é o conjunto de todas as cadeias formadas por exatamente quatro letras
4. L^* é o conjunto de todas as cadeias formadas por letras, incluindo a cadeia ϵ
5. $L(L \cup D)^*$ é o conjunto de cadeias de letras e dígitos, que iniciam com uma letra
6. D^+ é o conjunto de cadeias formadas por um ou mais dígitos

Definição de expressão regular

Sejam Σ um alfabeto. As expressões regulares sobre Σ são definidas pelas seguintes regras, onde cada expressão regular define uma linguagem:

1. ϵ é uma expressão regular que denota a linguagem $\{ \epsilon \}$
2. Se $a \in \Sigma$, então a é uma expressão regular que denota a linguagem $\{ a \}$
3. Se r e s são duas expressões regulares que denotam as linguagens $L(r)$ e $L(s)$, então
 - (a) (r) é uma expressão regular que denota $L(r)$
 - (b) $(r)|(s)$ é uma expressão regular que denota $L(r) \cup L(s)$
 - (c) $(r)(s)$ é uma expressão regular que denota $L(r)L(s)$
 - (d) $(r)^*$ é uma expressão regular que denota $(L(r))^*$

Expressões regulares e parêntesis

O uso de parêntesis em expressões regulares pode ser reduzido se forem adotadas as seguintes convenções:

1. o operador unário $*$ possui a maior precedência e é associativo à esquerda
2. a concatenação tem a segunda maior precedência e é associativa à esquerda
3. o operador $|$ tem a menor precedência e é associativo à esquerda

Neste cenário, a expressão regular $(a) | ((b)^* (c))$ equivale a $a | b^* c$.

Exemplos de expressões regulares

Seja $\Sigma = \{ a, b \}$. Então

- ▶ $a \mid b$ denota a linguagem $\{ a, b \}$
- ▶ $(a \mid b)(a \mid b)$ denota $\{ aa, ab, ba, bb \}$
- ▶ a^* denota $\{ \epsilon, a, aa, aaa, \dots \}$
- ▶ $(a \mid b)^*$ denota todas as cadeias formadas por zero ou mais instâncias de a ou de b
- ▶ $a \mid a^* b$ denota a cadeia a e todas as cadeias iniciadas por zero ou mais a 's, seguidos de um b

Propriedades das expressões regulares

Sejam r, s, t expressões regulares. Valem as seguintes propriedades:

Axioma	Descrição
$r s = s r$	$ $ é comutativo
$r (s t) = (r s) t$	$ $ é associativo
$r(st) = (rs)t$	a concatenação é associativa
$r(s t) = rs rt$ $(r s)t = rt st$	a concatenação é distributiva em relação a $ $
$\epsilon r = r$ $r\epsilon = r$	ϵ é o elemento neutro da concatenação
$r^* = (r \epsilon)^*$	relação entre ϵ e $*$
$r^{**} = r^*$	$*$ é idempotente

Definição

Seja Σ um alfabeto. Uma definição regular sobre Σ é uma sequência de definições da forma

$$d_1 \rightarrow r_1$$

$$d_2 \rightarrow r_2$$

\dots

$$d_n \rightarrow r_n$$

onde cada d_i é um nome distinto e r_i uma expressão regular sobre o alfabeto $\Sigma \cup \{ d_1, d_2, \dots, d_{i-1} \}$.

Exemplo de definição regular

Os identificadores de Pascal, e em muitas outras linguagens, são formados por cadeias de caracteres e dígitos, começando com uma letra.

Abaixo segue a definição regular para o conjunto de todos os identificadores válidos em Pascal:

$$\begin{aligned}\text{letra} &\rightarrow A \mid B \mid \dots \mid Z \mid a \mid b \mid \dots \mid z \\ \text{digito} &\rightarrow 0 \mid 1 \mid 2 \mid \dots \mid 9 \\ \text{id} &\rightarrow \text{letra} (\text{letra} \mid \text{digito})^*\end{aligned}$$

Simplificações notacionais

As seguintes notações podem simplificar as expressões regulares:

1. *Uma ou mais ocorrências.* Se r é uma expressão regular, então $(r)^+$ denota $(L(r))^+$. O operador $+$ tem a mesma associatividade e precedência do operador $*$. Vale que $r^* = r^+ | \epsilon$ e que $r^+ = rr^*$.
2. *Zero ou uma.* Se r é uma expressão regular, então $r?$ denota $L(r) \cup \epsilon$. O operador $?$ é posfixo e unário, e $r? = r | \epsilon$.
3. *Classes de caracteres.* A notação $[abc]$, onde a, b, c são símbolos do alfabeto, denota a expressão regular $a | b | c$. A notação $[a-z]$ abrevia a expressão regular $a | b | \dots | z$.

Limitações das expressões regulares

- ▶ Existem linguagens que não podem ser descritas por meio de expressões regulares
- ▶ Por exemplo, não é possível descrever o conjunto \mathcal{P} de todas as cadeias de parêntesis balanceados por meio de expressões regulares
- ▶ Contudo, o conjunto \mathcal{P} pode ser descrito por meio de uma gramática livre de contexto
- ▶ Existem linguagens que não podem ser descritas nem mesmo por meio de uma gramática livre de contexto
- ▶ Por exemplo, o conjunto

$$\mathcal{C} = \{wcw \mid w \text{ é uma cadeia de } a\text{'s e } b\text{'s}\}$$

não pode ser descrito nem por expressões regulares e nem por meio de uma gramática livre de contexto

Referências

1. **AHO**, Alfred V, **SETHI**, Ravi, **ULLMAN**, Jeffrey D. *Compiladores: Princípios, Técnicas e Ferramentas*, LTC Editora, 1995.
2. GeeksForGeeks. [Flex \(Fast Lexical Analyzer Generator\)](#), acesso em 04/06/2022.