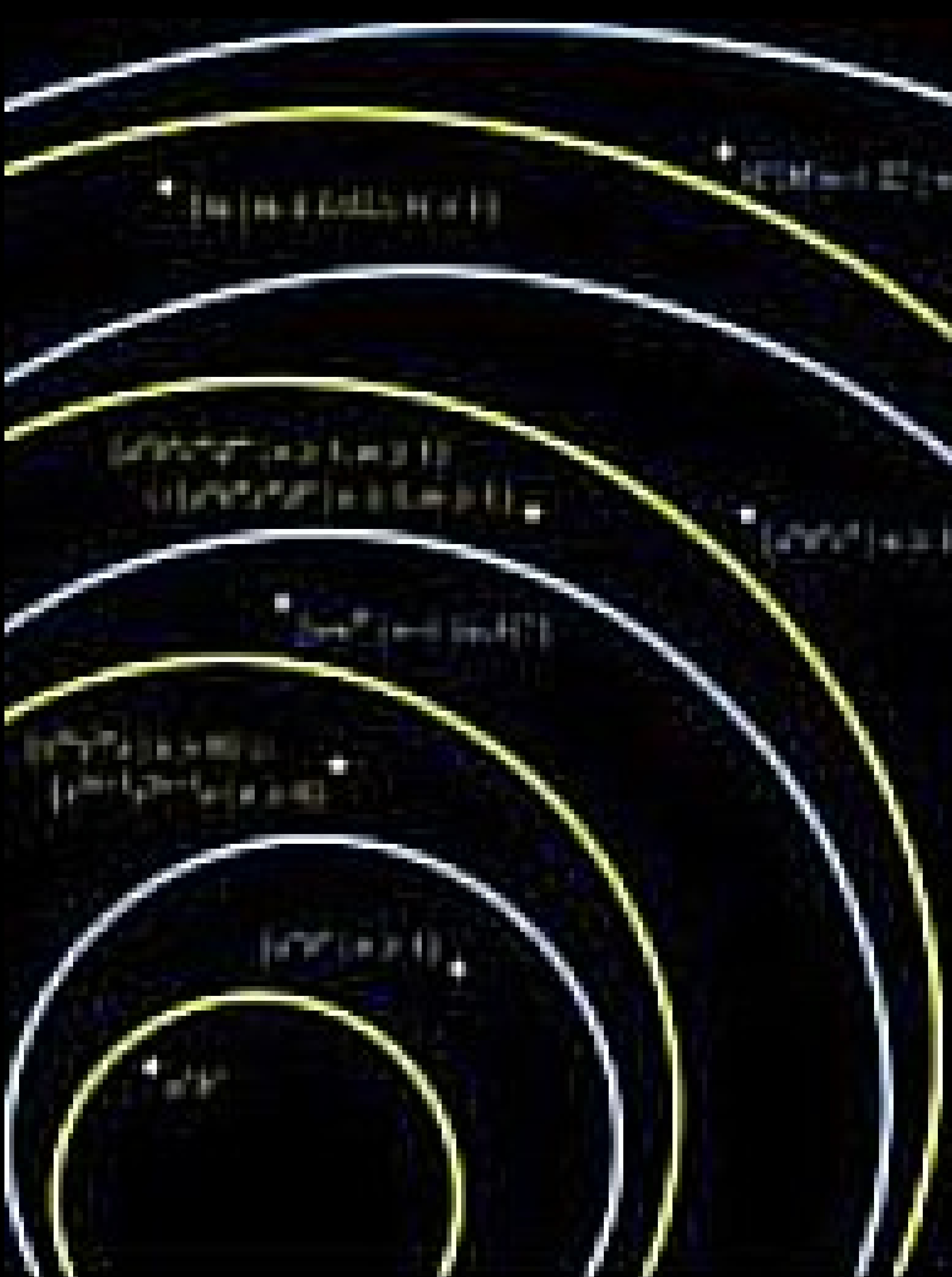
Linguagens Formais e Autômatos

Aula 3:

Alfabetos, Palavras,
Linguagens e Gramática



Prof. Dr. Rodrigo Xavier de Almeida Leão

Cientista de Dados e Big Data

ALFABETO

Um *Alfabeto* é um conjunto finito de *Símbolos*. ❑

Portanto, um conjunto vazio também é considerado um alfabeto. Um *símbolo* (ou *caractere*) é uma entidade abstrata básica a qual não é definida formalmente. Letras e dígitos são exemplos de símbolos freqüentemente usados.

ALFABETO

Assimile

Um alfabeto (chamado também de vocabulário) é um conjunto finito não vazio de símbolos.

ALFABETO

O alfabeto latino moderno é o seguinte conjunto de 26 símbolos: {A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z}. É comum representarmos o alfabeto pela letra $\sum$. Outros exemplos de alfabeto são:

$$\Sigma_1 = \{\alpha, \beta, \gamma, \delta, \ldots, \omega\}$$
$$\Sigma_2 = \{0,1\}$$

Com o primeiro você pode escrever palavras gregas e com o segundo você pode escrever palavras binárias (números na base 2).

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Uma *Palavra*, *Cadeia de Caracteres* ou *Sentença* sobre um alfabeto é uma seqüência finita de símbolos (do alfabeto) justapostos. ❑

A *palavra vazia*, representada pelo símbolo ε, é uma palavra sem símbolo. Se Σ representa um alfabeto, então Σ^* denota o conjunto de todas as palavras possíveis sobre Σ. Analogamente, Σ^+ representa o conjunto de todas as palavras sobre Σ excetuando-se a palavra vazia, ou seja, $\Sigma^+ = \Sigma^* - \{\varepsilon\}$.

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Assimile

Uma cadeia de símbolos de um dado alfabeto (também chamada de *string*, palavra ou sentença) é uma sequência finita de símbolos deste alfabeto.

Para o alfabeto $\Sigma_1 = \{\alpha, \beta, \gamma, \delta, \ldots, \omega\}$ podemos escrever a cadeia "$\psi\omega$".

Para o alfabeto $\Sigma_2 = \{0,1\}$ podemos escrever a cadeia "10001".

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

A cadeia formada por uma sequência com nenhum símbolo é conhecida como a cadeia vazia. Representamos a cadeia vazia com o símbolo ϵ. Note que a cadeia vazia é uma cadeia, ou palavra, sobre qualquer alfabeto. Cadeias de símbolos, ou palavras, sendo sempre uma sequência finita de símbolos, possuem comprimento, que é a quantidade de símbolos que ocorrem na mesma.

A *palavra vazia*, representada pelo símbolo ε, é uma palavra sem símbolo. Se Σ representa um alfabeto, então Σ^* denota o conjunto de todas as palavras possíveis sobre Σ. Analogamente, Σ^+ representa o conjunto de todas as palavras sobre Σ excetuando-se a palavra vazia, ou seja, $\Sigma^+ = \Sigma^* - \{\varepsilon\}$.

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Qualquer cadeia de símbolos tem um comprimento. Por exemplo, a cadeia "10001" tem comprimento 5. A cadeia vazia é normalmente representada em linguagens de programação como "".

O *Tamanho* ou *Comprimento* de uma palavra w, representado por $|w|$, é o número de símbolos que compõem a palavra. ❑

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

A *Concatenação* é uma operação binária, definida sobre uma linguagem, a qual associa a cada par de palavras uma palavra formada pela justaposição da primeira com a segunda. Uma concatenação é denotada pela justaposição dos símbolos que representam as palavras componentes. A operação de concatenação satisfaz às seguintes propriedades (suponha v, w, t palavras):

a) *Associatividade.*

$$v(wt) = (vw)t$$

b) *Elemento Neutro à Esquerda e à Direita.*

$$\varepsilon w = w = w\varepsilon$$

❑

Uma operação de concatenação definida sobre uma linguagem L não é, necessariamente, fechada sobre L, ou seja, a concatenação de duas palavras de L não é, necessariamente, uma palavra de L.

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Considere a linguagem L de palíndromos sobre { a, b }. A concatenação das palavras aba e bbb resulta na palavra ababbb a qual não é palíndromo. Portanto, a operação de concatenação não é fechada sobre L. ❑

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

A *Concatenação Sucessiva* de uma palavra (com ela mesma), representada na forma de um expoente w^n onde w é uma palavra e n indica o número de concatenações sucessivas, é definida indutivamente a partir da concatenação binária, como segue:

a) *Caso 1.* $w \neq \varepsilon$

$w^0 = \varepsilon$

$w^n = w^{n-1}w$, para $n > 0$

b) *Caso 2.* $w = \varepsilon$

$w^n = \varepsilon$, para $n > 0$

w^n é indefinida para $n = 0$ ❑

Note-se que a concatenação sucessiva é indefinida para ε^0.

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

EXEMPLO 21 Concatenação Sucessiva.

Sejam w uma palavra e a um símbolo. Então:

$w^3 = www$

$w^1 = w$

$a^5 = aaaaa$

$a^n = aaa...a$ (o símbolo a repetido n vezes)

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Assimile

Dadas duas cadeias, definimos sua concatenação como a justaposição de seus valores. Por exemplo, se ω_1 ="101" e ω_2 ="000", sua concatenação é "101000". Representamos a concatenação como $\omega_1 \circ \omega_2$ ou simplesmente $\omega_1\omega_2$.

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Definição 1.23 Prefixo, Sufixo, Subpalavra.

Um *Prefixo* (respectivamente, *Sufixo*) de uma palavra é qualquer seqüência de símbolos inicial (respectivamente, final) da palavra. Uma *Subpalavra* de uma palavra é qualquer seqüência de símbolos contígüa da palavra. ❑

EXEMPLO 18 Palavra, Prefixo, Sufixo.

a) abcb é uma palavra sobre o alfabeto { a, b, c }

b) Se $\Sigma = \{ a, b \}$, então $\Sigma^+ = \{ a, b, aa, ab, ba, bb, aaa, \ldots \}$ e $\Sigma^* = \{ \varepsilon, a, b, aa, ab, ba, bb, aaa, \ldots \}$

c) $|abcb| = 4$ e $|\varepsilon| = 0$

d) ε, a, ab, abc, abcb são os prefixos da palavra abcb e ε, b, cb, bcb, abcb são os respectivos sufixos;

e) Qualquer prefixo ou sufixo de uma palavra é uma subpalavra. ❑

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Assimile

Dadas duas cadeias, ω_1 e ω_2, dizemos que ω_1 é prefixo de ω_2 se existe uma cadeia ω_1 tal que $\omega_1 \circ \omega_3 = \omega_2$.

A cadeia "101" possui os seguintes prefixos: ϵ, "1", "10" e "101".

Os sufixos de uma cadeia são definidos de forma análoga, porém tomando as subsequências do final da cadeia. Deixamos a definição como exercício para o leitor. A cadeia "100" possui os seguintes sufixos: ϵ, "0", "00" e "100".

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Assimile

Dadas duas cadeias, ω_1 e ω_2, dizemos que ω_1 é prefixo de ω_2 se existe uma cadeia ω_1 tal que $\omega_1 \circ \omega_3 = \omega_2$.

A cadeia "101" possui os seguintes prefixos: ϵ, "1", "10" e "101".

Os sufixos de uma cadeia são definidos de forma análoga, porém tomando as subsequências do final da cadeia. Deixamos a definição como exercício para o leitor. A cadeia "100" possui os seguintes sufixos: ϵ, "0", "00" e "100".

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Uma *Linguagem Formal* é um conjunto de palavras sobre um alfabeto. ❑

EXEMPLO 19 Linguagem.

Suponha o alfabeto $\Sigma = \{a, b\}$. Então:

a) O conjunto vazio e o conjunto formado pela palavra vazia são linguagens sobre Σ (obviamente $\{\ \} \neq \{\varepsilon\}$);

b) O conjunto de palíndromos (palavras que têm a mesma leitura da esquerda para a direita e vice-versa) sobre Σ é um exemplo de linguagem infinita. Assim, ε, a, b, aa, bb, aaa, aba, bab, bbb, aaaa,... são palavras desta linguagem. ❑

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Assimile

Dado um alfabeto definimos uma linguagem sobre este alfabeto como um conjunto de cadeias sobre este alfabeto.

Para o alfabeto $\Sigma_2 = \{0,1\}$ temos infinitas linguagens possíveis, entre elas:

$$L_1 = \varnothing$$

$$L_2 = \{\epsilon\}$$

$$L_3 = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\}$$

A primeira linguagem não possui cadeia. A segunda linguagem possui apenas a cadeia vazia, enquanto a última possui todas as cadeias possíveis com símbolos do alfabeto Σ_2. Observe que a linguagem vazia, $\varnothing$, e a linguagem que só tem a palavra vazia, $\{\epsilon\}$, são diferentes, por quê?

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Sabemos que podemos concatenar duas cadeias. Esta operação pode ser estendida para uma linguagem. Definimos a concatenação de duas linguagens como a linguagem cujas cadeias são todas as possíveis concatenações entre cadeias da primeira linguagem com cadeias da segunda linguagem.

Assimile

Dadas as linguagens L_1 e L_2, definimos sua concatenação como a linguagem $L_1 \circ L_2 = \{\omega_1 \circ \omega_2 \mid \omega_1 \in L_1$ e $\omega_2 \in L_2\}$.

Se L_1 é uma linguagem finita com n cadeias e L_2 é uma linguagem finita com m cadeias, então quantos elementos possui $L_1 \circ L_2$?

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Quando repetimos a operação de concatenação com a mesma linguagem usamos a notação de potência. Por exemplo, $L_1^2 = L_1 \circ L_1$ e $L_1^3 = L_1 \circ L_1 \circ L_1$.

Definimos $L^0 = \varnothing$ e $L^{n+1} = L^n \circ L$.

Se fizermos a união de todas as potências de L, de L^0 em diante, obtemos o fecho de Kleene da linguagem L, representado por L^*.

Assimile

Definimos $L^* = L^0 \cup L^1 \cup L^2 \cup \ldots$

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Para o alfabeto $L = \{0,1\}$ temos:

$$L^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\}$$

Usando a concatenação de conjuntos, a união e o fecho de Kleene, podemos especificar algumas linguagens simples.

$L =$ { números na base 2 que são múltiplos de 4 (terminam com 00) } =
$\{0,1\}^\circ\{00\}$

$L =$ { todos os números na base 2, sem permitir 0's desnecessários à esquerda } = $(\{1\}^\circ\{0,1\}^*) \cup \{0\}$

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Uma variação do fecho de Kleene é usar o símbolo +. Definimos $L^+ = L^1 \cup L^2 \cup L^3 \cup \ldots$

Podemos definir o operador * usando o operador + e vice-versa. Podemos definir L^* como a união de L^0 com L^+, enquanto que L^+, por sua vez, pode ser definida como $L \circ L^*$.

PALAVRA, CADEIA DE CARACTER OU SENTENÇA

Vamos pensar no alfabeto $\Sigma = \{n, +, \times\}$. Vamos entender n como representando um número qualquer, $+$ como a soma, e $\times$ como a multiplicação. Queremos definir uma linguagem L sobre Σ como sendo a linguagem de todas as 'expressões' bem formadas usando-se essas duas operações:

$$L = \{n, n+n, n\times n, n+n+n, n+n\times n, n\times n+n, n\times n\times n, \ldots\}$$

Descrição da situação-problema

O sistema de numeração originário na Roma antiga, aproximadamente no século VIII a.C., é aquele baseado nas letras I, V, X, L, C, D e M. Este sistema foi amplamente utilizado desde a sua criação até o século XIV d.C. Ainda hoje existem usos modernos deste sistema para representar quantidades ou itens em uma ordenação, como a denominação dos séculos no ocidente. Sabe-se que o sistema caiu em desuso e foi substituído pelo sistema posicional com zero (Hindu-Arábico) por este ser uma representação que facilita em muito a aplicação de algoritmos de adição e multiplicação. No sistema romano, a justaposição de símbolos é mais complexa que no sistema decimal hindu-arábico. No sistema decimal, os símbolos 0,1,2,3,4,5,6,7,8 e 9 podem ser justapostos lado a lado em qualquer ordem e livres de quaisquer restrições, a exceção dos zeros à esquerda, que devem ser evitados, por razões de ordem prática. Qualquer sequência de algarismos é um número decimal. Outra propriedade interessante

dos numerais decimais é a sua capacidade de representar qualquer número. O mesmo não acontece com os numerais romanos. Em primeiro lugar não é qualquer sequência de letras I, V, X, L, C, D e M que é um numeral romano válido. Por exemplo, a sequência IIIV não é um numeral romano válido. Além disso, é sabido que os numerais romanos tradicionais não conseguem representar mais que MMMCMXLIX números naturais distintos. Existiram extensões do sistema romano que conseguiam passar disso, mas não chegavam a representação de bilhões. Neste livro vamos nos limitar ao número MMMCMXLIX mesmo.

Nesta seção, você aprendeu que qualquer conjunto de palavras sobre um alfabeto é uma linguagem formal. Assim, tanto a linguagem dos numerais decimais quanto a dos numerais romanos são linguagens formais. Uma é uma linguagem infinita e outra é uma linguagem finita. Se não levarmos em consideração os "zeros à esquerda" que devem ser evitados na notação decimal, podemos dizer que os numerais decimais são a linguagem definida pelo conjunto $\{0,1,2,3,4,5,6,7,8,9,\}^+$. Ou seja, usamos um dos operadores aprendidos, o chamado fecho de Kleene, para definir o conjunto de todos os numerais decimais de uma forma compacta. Lembre-se que a linguagem em questão é infinita.

Você consegue representar a restrição de não haver zeros à esquerda através de conjuntos de símbolos e as operações entre linguagens formais apresentadas nesta seção?

$$DECIMAIS = (\{1,2,3,4,5,6,7,8,9\}^\circ\{0,1,2,3,4,5,6,7,8,9,\}^+) \cup (\{1,2,3,4,5,6,7,8,9\})$$

E temos então incluído a restrição de não haver zeros à esquerda. Note que o conjunto acima também pode ser especificado como

$$DECIMAIS = \{1,2,3,4,5,6,7,8,9\}^\circ\{0,1,2,3,4,5,6,7,8,9,\}^*$$

Como relação aos numerais romanos, vamos estruturar nosso conhecimento: (1) as palavras I, II e III são as únicas que podem ser escritas só com I; (2) imediatamente à esquerda de um V só pode ocorrer um I; (3) À direita de um V podem ocorrer até no máximo 3 Is; (4) À direita de um X podem ocorrer no máximo 3 Is e à sua esquerda somente um I; (5) a regra 4 também vale em relação a L, C, D e M. Em resumo, todo numeral romano tem um núcleo de maior valor, por exemplo o núcleo de MMXVII é MM, o núcleo de CDXXIV é CD. Antes do núcleo pode ocorrer um, e somente um, símbolo de menor valor e depois do núcleo de maior valor pode aparecer um núcleo de valor menor.

Observe que a explicação em linguagem natural, mesmo organizada, fica complicada. Vamos fazer usando as operações entre conjuntos. Para facilitar a leitura vamos denotar cada novo conjunto especificado.

$$NI = \{I, II, III, IV, V, VI, VII, VIII, IX\}$$

$$AX = \{X, XX, XXX, XL, L, LX, LXX, LXXX, XC\}$$

$$AC = \{C, CC, CCC, CD, D, DC, DCC, DCCC, CM$$

$$NX = AX \cup (AX \circ NI)$$

$$NC = AC \cup (AC \circ NX)$$

$$AM = \{M, MM, MMM\}$$

$$NM = AM \cup (AM \circ NC)$$

O conjunto NM é a linguagem das cadeias que são numerais romanos até a numeração de 3999.

1. Considere a linguagem $L = \{aab, a\}$ sobre o alfabeto $\Sigma = \{a, b\}$.
Marque a alternativa correta:

a) $L^0 = \varnothing$

b) $L^2 = \{aabaab, aa\}$

c) $L^3 = \{aaa, aaaab, aaabaab, aabaabaab, aabaa, aabaaba\}$

d) $L^4 \subset L^5$

e) $L^4 \subset L^*$

2. Suponha que L_1 e L_2 sejam linguagens sobre o alfabeto $\Sigma = \{a, b\}$.
Assinale a alternativa verdadeira:

a) Se $L_1 \circ L_2 = \varnothing$, então $L_1 = \varnothing$.

b) Se $L_1 \circ L_1 = \varnothing$, então $L_1 = \varnothing$.

c) Se $L_1 = \{\epsilon\}$, então $L_1 \circ L_2 = \{\epsilon\}$.

d) Se $L_1 = \{\epsilon\}$, então $L_1 \circ L_2 = \varnothing$.

e) Se $L_1 \subseteq L_2$, então $L_1^* = L_2^*$.

3. Considere a cadeia $\omega = ababa$.

Assinale a cadeia que pode ser formada concatenando-se dois prefixos de ω:

a) $abba$

b) $abaabb$

c) bb

d) ba

e) a