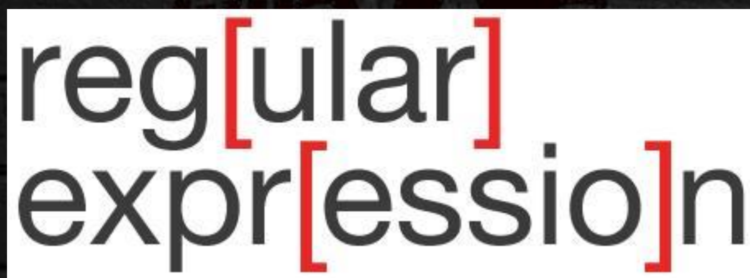
Expresiones regulares



Compiladores
Prof. Luis Enrique Hernández Olvera

Contenido

- Repaso de definiciones.
- Lenguajes Regulares
- Expresiones regulares
- Operadores de las expresiones regulares

Compiladores
Prof. Luis Enrique Hernández Olvera

Repasando definiciones

- Un conjunto no vacío y finito de símbolos se conoce como alfabeto (**Σ**).
- Una secuencia finita de símbolos de un determinado alfabeto se conoce como **palabra** sobre dicho alfabeto (a-z).
- Cada símbolo de un alfabeto es una cadena sobre dicho alfabeto.
- La **cadena vacía**, la cual se denota por el símbolo **ε (λ)**, es una palabra sobre cualquier alfabeto.

Repasando definiciones

- Un **lenguaje** es un conjunto de palabras. Por lo tanto el conjunto {1,12,123,1234,12345} es un lenguaje sobre el alfabeto compuesto por dígitos.
- Si Σ es un alfabeto, también es un lenguaje (El formado por todas las cadenas con un único símbolo).

Repasando definiciones

- Dado que un lenguaje es un conjunto de cadenas, se puede tener el lenguaje compuesto por ninguna cadena (El lenguaje vacío).
- El lenguaje vacío se denota de la misma forma que el conjunto vacío Ø.

Repasando definiciones

- Ø, el **lenguaje vacío**, es un lenguaje de cualquier **alfabeto**.
- {ε}, el lenguaje que consta sólo de la **cadena vacía**, también es un lenguaje de cualquier alfabeto.
- Ø ≠ {ε}; el primero no contiene **ninguna cadena** y el segundo sólo tiene **una cadena**.

ESCOM

Repasando definiciones

- El conjunto de todas las cadenas de un alfabeto Σ, se conoce como **cerradura de Σ** o **lenguaje universal de Σ** y se denota por **Σ*.**
- Por ejemplo, si se tiene el alfabeto Σ={1}, entonces:

$$\Sigma^*=\{\varepsilon,1,11,111,1111,11111,\ldots\}$$

Compiladores
Prof. Luis Enrique Hernández Olvera
7

ESCOM

Repasando definiciones

- El conjunto de todas las cadenas de un alfabeto Σ se designa mediante **Σ***.

$$\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \ldots$$

- El conjunto de cadenas no vacías del alfabeto Σ se designa como **cerradura positiva Σ^+**.
- Por lo tanto las dos equivalencias son:

$$\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \cdots$$
$$\Sigma^* = \Sigma^+ \cup \{\varepsilon\}.$$

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

Compiladores
Prof. Luis Enrique Hernández Olvera
8

Lenguajes Regulares

- Los lenguajes regulares sobre un alfabeto pueden ser usados para especificar la construcción de analizadores léxicos (programas que analizan un texto y extraen los lexemas o unidades léxicas) que hay en el mismo.

Lenguajes Regulares

- Los lenguajes más sencillos que formalmente se consideran son los **lenguajes regulares**.
- Un lenguaje regular se puede generar a partir de lenguajes básicos, con la aplicación de las operaciones de **unión**, **concatenación** y cerradura de Kleene un número finito de veces.

Lenguajes Regulares

- Los lenguajes regulares se llaman así porque sus palabras contienen **"regularidades"** o repeticiones de los mismos componentes.
- Por ejemplo:
  - L_1 = {ab, abab, ababab, abababab, . . .}
    - En este ejemplo se aprecia que las palabras de L_1 son simplemente repeticiones de "ab" cualquier número de veces. Aquí la "regularidad" consiste en que las palabras contienen "ab" algún número determinado de veces.

Lenguajes Regulares

Un **lenguaje regular** es un tipo de **lenguaje formal** que satisface las siguientes propiedades:

- **Puede ser reconocido** por:
  - Un autómata finito determinista
  - Un autómata finito no determinista
  - Un autómata de pila
  - Un autómata finito alterno
  - Una máquina de Turing de solo lectura
- **Es generado por:**
  - Una gramática regular
  - Una gramática de prefijos
- **Es descrito por:**
  - Una expresión regular

ESCOMP

Definición formal de lenguaje regular

- Sea Σ un alfabeto. El conjunto de los lenguajes regulares sobre Σ se define recursivamente como sigue:
  - El lenguaje vacío Ø es un lenguaje regular.
  - El lenguaje cadena vacía {ε} es un lenguaje regular.
  - Para todo a ϵ Σ, {a} es un lenguaje regular.
  - Si A y B son lenguajes regulares, entonces A U B, A·B y A* son lenguajes regulares.
  - Ningún otro lenguaje sobre Σ es regular.

ESCOM

Ejemplo

Dado Σ = {a,b}, las siguientes afirmaciones son ciertas:

- Ø y {ε} son lenguajes regulares
- {a} y {b} son lenguajes regulares
- {a, b} es un lenguaje regular
- {ab} es un lenguaje regular
- {a, ab, b} es un lenguaje regular
- $\{a^i \mid i \geq 0\}$ es un lenguaje regular
- $\{a^i b^j \mid i \geq 0$ y $j \geq 0\}$ es un lenguaje regular
- $\{(ab)^i \mid i \geq 0\}$ es un lenguaje regular

Expresiones regulares

- Una **expresión regular** es una forma abreviada de representar cadenas de caracteres que se ajustan a un determinado patrón. Al **conjunto de cadenas representado por la expresión** *r* se lo llama *lenguaje generado por la expresión regular r* y se escribe *L(r).*
- Una expresión regular se define sobre un alfabeto Σ y **es una cadena formada por caracteres** de dicho alfabeto y por una serie de **operadores**.

Expresiones regulares

- Las expresiones regulares sirven como lenguaje de entrada de muchos sistemas que procesan cadenas por ejemplo generadores de analizadores léxicos, como Lex o Flex.
- Un generador de analizadores léxicos acepta descripciones de las formas de las unidades lógicas, que son principalmente expresiones regulares, y produce un AFD que reconoce qué unidad lógica aparece a continuación en la entrada.

Expresiones regulares

Si Σ es un alfabeto, una expresión regular sobre este alfabeto se define de la siguiente forma:

- Ø es una expresión regular que denota el lenguaje vacío.
- ε es una expresión regular que denota el lenguaje {ε}
- Si a ϵ Σ, a es una expresión regular que denota el lenguaje {a}
- Si r y s son expresiones regulares, entonces r U s, r · s y r* también lo son.

Operadores de las expresiones regulares

- Unión: (r+s) es una expresión regular que denota el lenguaje R U S.
- Concatenación: (rs) es una expresión regular que denota el lenguaje R·S
- Clausura: r* es una expresión regular que denota el lenguaje R*.

Ejemplos

$\Sigma = \{0,1\}$

- 00 El conjunto {00}
- 01* Conjunto de palabras que empiezan por 0 y después tienen una sucesión de unos.
- (1+10)* Conjunto de palabras en las que los ceros están precedidos siempre por unos.
- (0+1)*011 Conjunto de palabras que terminan en 011

Ejemplos

$\Sigma = \{0,1\}$

- 0*1* Conjunto de palabras formadas por una sucesión de ceros seguida de una sucesión de unos. Ambas sucesiones pueden ser vacías.
- 00*11* Conjunto de palabras formadas por una sucesión de ceros seguida de una sucesión de unos. Ninguna de las sucesiones puede ser vacía.

A r*r se le denota como r^+. Por lo que en la ultima expresión regular quedaría como 0^+1^+

Ejemplos

∑ ={a, b}

– a|b denota el lenguaje L(a|b)={a, b}.
– (a|b)(a|b) denota a L((a|b)(a|b)), el lenguaje de todas las cadenas de longitud dos sobre el alfabeto ∑.
– (a|b)(a|b) =aa|ab|ba|bb
– L(a*)={λ, a, aa, aaa, …}, todas las cadenas de cero o más a 's.

Ejemplos

(a)|((b)*(c)) = a|b*c

– Descripción:
  - Conjunto de cadenas que son una sola a o ninguna o más b's seguidas por una c.

– Algunas cadenas del lenguaje:
  - L(a|b*c)={a,c,bc,bbc,bbbc,bbbbc,…}

Ejercicio

Dado Σ = {a,b,c} y la E.R c*(a|bc*)* ¿Describe el lenguaje de todas las cadenas sobre el Σ que no contiene ninguna subcadena ac es regular?

Ejercicio (Solución)

Dado Σ = {a,b,c} y la E.R c*(a|bc*)* ¿Describe el lenguaje de todas las cadenas sobre el Σ que no contiene ninguna subcadena ac es regular?

- {c}, {a}, {b} son L.R. (3)
- {c}* es un L.R. (4, *)
- {b}{c}* es un L.R. (4, ·)
- {a}∪{b}{c}* es un L.R. (4, ∪)
- ({a}∪{b}{c}*)* es un L.R. (4, *)
- {c}*({a}∪{b}{c}*)* es un L.R. (4, ·)

Equivalencias en Expresiones Regulares

- Existen muchas equivalencias con respecto a expresiones regulares basadas en las correspondientes igualdades de lenguajes. Se resumen en el siguiente teorema.
- Sean r,s y t expresiones regulares sobre el mismo alfabeto Σ. Entonces:

Equivalencias en Expresiones Regulares

1. r U s = s U r
2. r U Ø = r = Ø U r
3. r U r = r
4. (r U s) U t = r U (s U t)
5. rε = εr = r
6. r Ø = Ø r = Ø
7. (rs)t = r(st)
8. r(s U t) = rs U rt
9. (r U s)t = rtUst

10. r* = r** = r*r* = (ε U r)* = r*(r U ε) = (r U ε)r* = ε U rr*
11. (r U s)*=(r*U s*)*=(r*s*)*=(r*s)*r*= r*(sr*)*
12. r(sr)*=(rs)*r
13. (r*s)*= ε U (r U s)*s
14. (rs*)*= ε U r(r U s)*
15. s(r U ε)*(r U ε) U s=sr*
16. rr*=r*r

Ejercicios

- Describa los lenguajes generados por las siguientes expresiones regulares y enumere al menos 7 cadenas de cada lenguaje. ∑ ={a, b, c, d, e, …,z}

- (ab)|(cz)|d*
- a+b+c | (b+d)z
- (abc)*z
- (a|b)*|a
- $(ab)^{+}\ (bc)^{+}$