Expresiones regulares

Contenido

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

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 vacia**, la cual se denota por el símbolo **ε**, es una palabra sobre cualquier alfabeto.

- 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).

- 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.

- 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

- 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:

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

ESCOM

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

ESCOM

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.
- Para un alfabeto Σ dado, los lenguajes regulares sobre Σ constituyen el menor conjunto de lenguajes sobre Σ que es cerrado con respecto a las operaciones de concatenación, la cerradura de Kleene y la unión de lenguajes y además contienen el lenguaje vacío Ø y los lenguajes unitarios {a} para a ϵ Σ.

ESCOM

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.

ESCOM

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 de veces.

ESCOM

- 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

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 regulas.
  - Si A y B son lenguajes regulares, entonces A U B, A·B y A* son lenguajes regulares.

Un lenguaje L es regular **si y solo si se cumple al menos** una de las condiciones siguientes:

i. L es finito *(Estos son lenguajes obviamente regulares y uno podría crear expresiones regulares que serían la unión de todas las palabras del lenguaje que definirían dicho lenguaje.)*

ii. L es la unión o la concatenación de otros lenguajes regulares R_1 y R_2, $L = R_1 \cup R_2$ o $L = R_1R_2$ respectivamente

iii. L es la cerradura de Kleene de algún lenguaje regular, L=R*

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
- $\{a^ib^j \mid i \geq 0$ y $j \geq 0\}$ es un lenguaje regular
- $\{(ab)^i \mid i \geq 0\}$ es un lenguaje regular

Dado Σ = {a,b,c}, ¿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, ·)

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**.

- 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.

ESCOM

Si A 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 A, 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.

ESCOM

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*.

ESCOM

Ejemplos

A={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

ESCOM

Ejemplos

A={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^+

ESCOM

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.

ESCOM

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,…}

- 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:

1. **r + s = s + r**
2. **r + Ø = r = Ø + r**
3. **r + r = r**
4. **(r + s) + t = r + (s + t)**
5. **rε = εr = r**
6. **r Ø = Ø r = Ø**
7. **(rs)t = r(st)**
8. **r(s+t) = rs+rt , y (r+s)t = rt+st**
9. **r*= r**= r*r*=(ε+r)*=r*(r+ε)=(r+ε)r*= ε+rr***
10. **(r+s)*=(r*+s*)*=(r*s*)*=(r*s)*r*=r*(sr*)***
11. **r(sr)*=(rs)*r**
12. **(r*s)*= ε+(r+s)*s**
13. **(rs*)*= ε+r(r+s)***
14. **s(r+ ε)*(r+ ε)+s=sr***
15. **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
- $(ab)^+ (bc)^+$