GRAMATICAS REGULARES - EXPRESIONES REGULARES

Gramáticas

Las gramáticas formales definen un lenguaje describiendo cómo se pueden generar las cadenas del lenguaje.

Una gramática formal es una cuadrupla G = <N, T, P, S> donde

- N es un conjunto finito de símbolos no terminales
- T es un conjunto finito de símbolos terminales $\quad N \cap T = \varnothing$
- P es un conjunto finito de producciones

  Cada producción de P tiene la forma

  $\alpha \to \beta, \ \alpha = \varphi A \rho \quad y \quad \beta = \varphi \omega \rho$

  $\varphi, \omega, \rho \in (N \cup T)^* \quad$ y A es S ó $A \in N$
- S es el símbolo distinguido o axioma $\quad S \notin (N \cup T)$

Restringiendo los formatos de producciones permitidas en una gramática, se pueden especificar cuatro tipos de gramáticas (tipo 0, 1, 2 y 3) y sus correspondientes clases de lenguajes.

Gramáticas regulares (Tipo 3)

Generan los lenguajes regulares (aquellos reconocidos por un autómata finito). Son las gramáticas más restrictivas. El lado derecho de una producción debe contener un símbolo terminal y, como máximo, un símbolo no terminal. Estas gramáticas pueden ser:

- Lineales a derecha, si todas las producciones son de la forma

$$A \to aB \quad ó \quad A \to a \qquad \begin{cases} A \in N \cup \{S\} \\ B \in N \\ a \in T \end{cases}$$

(en el lado derecho de las producciones el símbolo no terminal aparece a la derecha del símbolo terminal)

- Lineales a izquierda, si todas las producciones son de la forma

$$A \to Ba \quad ó \quad A \to a \qquad \begin{cases} A \in N \cup \{S\} \\ B \in N \\ a \in T \end{cases}$$

(en el lado derecho de las producciones el símbolo no terminal aparece a la izquierda del símbolo terminal)

En ambos casos, se puede incluir la producción $S \to \varepsilon$, si el lenguaje que se quiere generar contiene la cadena vacía.

Por ejemplo las siguientes gramáticas G_1 y G_2, son gramáticas regulares lineales a derecha y lineales a izquierda respectivamente, que generan el lenguaje $L = \{a^{2n} / n \geq 0\}$

$G_1 = <\{A, B\}, \{a\}, P_1, S_1>$	$G_2 = <\{C, D\}, \{a\}, P_2, S_2>$
donde P_1 es el cjto.	donde P_2 es el cjto.
$S_1 \to \varepsilon$	$S_2 \to \varepsilon$
$S_1 \to aA$	$S_2 \to Ca$
$A \to aB$	$C \to Da$
$A \to a$	$C \to a$
$B \to aA$	$D \to Ca$

<u>Algoritmo para obtener la gramática regular desde el autómata finito</u>

Existe un algoritmo que permite obtener una gramática regular que genera un lenguaje regular dado a partir del autómata finito que reconoce ese lenguaje. Los pasos a seguir son los siguientes:

1) Asociar al estado inicial el símbolo distinguido S.
2) Asociar a cada estado del autómata (menos el estado inicial) un símbolo no terminal. Si al estado inicial llega algún arco asociar también un símbolo no terminal (además del símbolo distinguido). No asociar símbolo no terminal a aquellos estados finales de los que no salen arcos.
3) Para cada transición definida δ (e_i, a) = e_j, agregar al conjunto de producciones, la producción A $\rightarrow$ aB, siendo A y B los símbolos no terminales asociados a e_i y e_j respectivamente. Si e_j es un estado final, agregar también la producción A $\rightarrow$ a. Si e_j es el estado inicial (tiene dos símbolos asociados, el distinguido y un no terminal), utilizar el símbolo no terminal (de esta manera se evita que el símbolo distinguido aparezca a la derecha de una producción).
4) Si el estado inicial es también final agregar la producción S $\rightarrow$ ε.

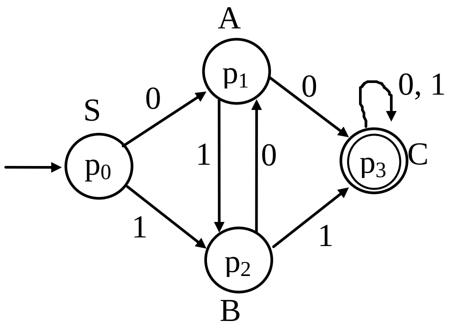
<u>Ejemplo 1</u>:

Derivación de la gramática correspondiente al lenguaje del ej. 4 del apunte de autómatas finitos

L_4 = { x / x $\in$ $\{0, 1\}^*$ y x contiene la subcadena 00 ó x contiene la subcadena 11}

$L_4 = L(M_{4Dmin})$, M_{4Dmin} = < {p_0, p_1, p_2, p_3}, {0, 1}, δ, p_0, {p_3}>

δ está definida por el siguiente diagrama de transición de estados



Como al estado inicial no entran arcos, se asocia únicamente el símbolo distinguido S.

La gramática correspondiente a este lenguaje es

G = <{A, B, C}, {0, 1}, P, S>, siendo P el siguiente conjunto:

S $\rightarrow$ 0A ya que δ (p_o, 0) = p_1 y S y A están asociado a p_0 y p_1 respectivamente.

S $\rightarrow$ 1B ya que δ (p_o, 1) = p_2 y S y B están asociado a p_0 y p_2 respectivamente.

A $\rightarrow$ 0C

A $\rightarrow$ 0

A $\rightarrow$ 1B

B $\rightarrow$ 0A

B $\rightarrow$ 1C

B $\rightarrow$ 1

C $\rightarrow$ 0C

C $\rightarrow$ 0

C $\rightarrow$ 1C

C $\rightarrow$ 1

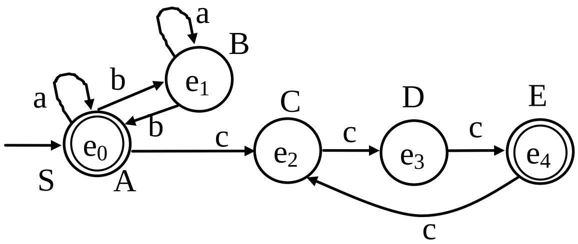
<u>Ejemplo 2</u>:

Derivación de la gramática correspondiente al lenguaje del ej. 3 del apunte de autómatas finitos.

$L_3 = \{xc^{3m} / x \in \{a, b\}^*$ y la cantidad de b's es par y $m \geq 0\}$, siendo $L_3 = L(M_{3D})$

$M_{3D} = < \{e_0, e_1, e_2, e_3, e_4\}, \{a, b, c\}, \delta_{3D}, e_0, \{e_0, e_4\}>$

δ_{3D} está definida por el siguiente diagrama de transición de estados



Como al estado inicial entran arcos, se asocia el símbolo distinguido S y además un símbolo no terminal A.

La gramática correspondiente a este lenguaje es

G = <{A, B, C, D, E}, {a, b, c}, P, S>, siendo P el siguiente conjunto:

$S \rightarrow \varepsilon$ (el estado inicial es también final)	$A \rightarrow cC$
$S \rightarrow aA$	$B \rightarrow aB$
$S \rightarrow a$	$B \rightarrow bA$ (se usa el símbolo no terminal asociado al estado inicial)
$S \rightarrow bB$	$B \rightarrow b$
$S \rightarrow cC$	$C \rightarrow cD$
$A \rightarrow aA$	$D \rightarrow cE$
$A \rightarrow a$	$D \rightarrow c$
$A \rightarrow bB$	$E \rightarrow cC$

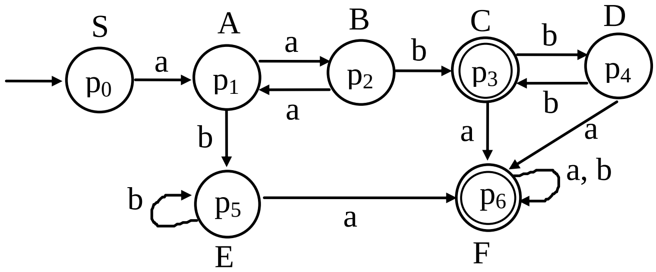
<u>Ejemplo 3</u>:

Derivación de la gramática correspondiente al lenguaje del ej. 7 del apunte de autómatas finitos.

$L_7 = \{ a^{2n}b^{2k+1} / n \geq 1$ y $k \geq 0\} \cup \{ax / x \in \{a, b\}^*$ y x contiene la subcadena ba}

siendo $L_7 = L(M_{7Dmin})$, $M_{7Dmin} = < \{p_0, p_1, p_2, p_3, p_4, p_5, p_6\}, \{a, b\}, \delta, p_0, \{p_3, p_6\}>$

δ está definida por el siguiente diagrama de transición de estados



La gramática correspondiente a este lenguaje es

G = <{A, B, C, D, E, F}, {a, b}, P, S>, siendo P el siguiente conjunto:

$S \rightarrow aA$	$B \rightarrow bC$	$C \rightarrow a$	$D \rightarrow a$	$F \rightarrow aF$
$A \rightarrow aB$	$B \rightarrow b$	$D \rightarrow bC$	$E \rightarrow bE$	$F \rightarrow a$
$A \rightarrow bE$	$C \rightarrow bD$	$D \rightarrow b$	$E \rightarrow aF$	$F \rightarrow bF$
$B \rightarrow aA$	$C \rightarrow aF$	$D \rightarrow aF$	$E \rightarrow a$	$F \rightarrow b$

Expresiones regulares

Se denominan expresiones regulares sobre un alfabeto A, a las expresiones que se pueden construir a partir de las siguientes reglas:

- $\emptyset$ es una expresión regular que describe el lenguaje vacío;
- ε es una expresión regular que describe el lenguaje $\{\varepsilon\}$, esto es el lenguaje que contiene únicamente la cadena vacía;
- Para cada símbolo $a \in A$, a es una expresión regular que describe el lenguaje $\{a\}$, esto es el lenguaje que contiene únicamente la cadena a;
- Si r y s son expresiones regulares que describen los lenguajes L(r) y L(s) respectivamente:

i) $r + s$ es una expresión regular que describe el lenguaje $L(r) \cup L(s)$

ii) $r \,.\, s$ es una expresión regular que describe el lenguaje $L(r) \,.\, L(s)$

iii) r^* es una expresión regular que describe el lenguaje $L(r)^*$.

El operador de clausura es el que tiene mayor precedencia, seguido por el operador de concatenación y por último el operador de unión.

Las expresiones regulares describen los lenguajes regulares (aquellos reconocidos por autómatas finitos).

Por ejemplo las siguientes son expresiones regulares válidas:

- $a^* \,.\, b$ que describe el lenguaje $L = \{a^n b \,/\, n \geq 0\}$
- $(a + b)^*$ que describe el lenguaje $L = \{ x \,/\, x \in \{a, b\}^* \}$
- $(aa)^* \, c \, b^* b$ que describe el lenguaje $L = \{ a^{2n} cb^k \,/\, n \geq 0 \text{ y } k > 0\}$
- $aa^* \,.b$ que describe el lenguaje $L = \{a^n b \,/\, n > 0\}$

Leyes algebraicas para expresiones regulares

Dos expresiones regulares r y s son equivalentes ($r \equiv s$) si $L(r) = L(s)$

Sean r, s y t expresiones regulares:

1) $r + \emptyset \equiv \emptyset + r \equiv r$

2) $r \,.\, \varepsilon \equiv \varepsilon \,.\, r \equiv r$

3) $r \,.\, \emptyset \equiv \emptyset \,.\, r \equiv \emptyset$

4) $r + s \equiv s + r$

5) $(r + s) + t \equiv r + (s + t)$

6) $(r \,.\, s) \,.\, t \equiv r \,.\, (s \,.\, t)$

7) $r \,.\, (s + t) \equiv r \,.\, s + r \,.\, t$

8) $(s + t) \,.\, r \equiv s \,.\, r + t \,.\, r$

9) $r + r \equiv r$

10) $\emptyset^* \equiv \varepsilon$

11) $r \,.\, r^* \equiv r^* \,.\, r$

12) $r \,.\, r^* + \varepsilon \equiv r^*$

13) $(r^* \,.\, s^*)^* \equiv (r + s)^*$

14) $(r^*)^* \equiv r^*$

Ejemplo

- $a^* + a$ describe el lenguaje $L = \{a^n \,/\, n \geq 0\}$
- a^* describe el lenguaje $L = \{a^n \,/\, n \geq 0\}$

Luego $a^* + a = a^*$ son expresiones regulares equivalentes porque describen los mismos lenguajes