Alfabetos, cadenas y lenguajes

Alan Reyes-Figueroa

Teoría de la Computación (Aula 01) 11.julio.2022

Alfabetos

Un **alfabeto** es un conjunto Σ de símbolos, finito y no-vacío. Estos símbolos pueden ser de cualquier tipo: letras, números, caracteres especiales, símbolos.

Ejemplos:

- $\Sigma = \{0, 1\}$: el alfabeto binario.
- $\Sigma = \{a, b, c, \ldots, z\}$: el alfabeto de las letras en el castellano.
- $\Sigma = \{2, 3, \ldots, 10, J, Q, K, A, \clubsuit, \blacklozenge, \spadesuit, \heartsuit, red, black\}$: el alfabeto de una baraja de cartas.

Ejemplo: Nombres de variables en Python.
¿Cuál es el alfabeto para describir los nombres de las variables o identificadores en el lenguaje Python?

Alfabetos

Ejemplo: Nombres de variables en Python.
Para dar nombres a las variables, existen varias reglas:

- Un nombre de variable debe comenzar con una letra o con el caracter *underscore*.
- Los nombres de variable no puede comenzar con un número.
- Los nombres de variable sólo puede contener caracteres alfanuméricos o *underscores* (A-z, 0-9, _).
- Los nombres de variables son *case-sensitive*. Por ejemplo: age, Age y AGE son tres nombres diferentes).
- La longitud máxima para un identificador o nombre de variable es de 79 caracteres.

Así, el alfabeto para identificadores en el lenguaje de Python es:

$$\Sigma = \{A, B, \ldots, Z\} \cup \{a, b, \ldots, z\} \cup \{0, 1, \ldots, 9\} \cup \{\_\}$$

Cadenas

Una **cadena de caracteres** (o **palabra**) es una secuencia finita de símbolos seleccionados de algún alfabeto Σ.

Ejemplos:

- *11010001* es una cadena del alfabeto binario $\Sigma = \{0, 1\}$.
- *class*, *del*, *False*, *lambda*, son cadenas en el alfabeto de Python. De hecho, son palabras reservadas.

Observaciones:

- Usaremos las primeras letras del alfabeto $a, b, c, d, \ldots$, o dígitos $0, 1, \ldots, 9$ para representar a los símbolos o letras de un alfabeto.
- Usaremos las últimas letras del alfabeto $\ldots, w, x, y, z$ para representar cadenas o palabras.
- Existe una cadena especial, la **cadena vacía**, la cual no posee símbolos. La denotamos por ε.

Cadenas

La **longitud** de una cadena es el número de caracteres que contiene. Denotamos la longitud de una cadena w por $|w|$.

Ejemplos:

- La cadena $w = 011000$ tiene longitud 6. $|w| = 6$.
- La cadena $x = autómata$ tiene longitud 8. Esto es $|x| = 8$.
- La cadena vacía tiene longitud 0. $|\varepsilon| = 0$.

Para un alfabeto Σ, y un número natural $k \geq 0$, denotamos por Σ^k el conjunto de todas las cadenas de longitud k en Σ:

$$\begin{aligned} \Sigma^k &= \{\text{cadenas en } \Sigma \text{ de longitud } k\} \\ &= \{w \in \Sigma^* : |w| = k\}. \end{aligned}$$

Cadenas

Ejemplo: En el alfabeto binario, $\Sigma = \{0, 1\}$.

- $\Sigma^0 = \{\varepsilon\}$.
- $\Sigma^1 = \{0, 1\}$.
- $\Sigma^2 = \{00, 01, 10, 11\}$.
- $\Sigma^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$.

En general, si el alfabeto Σ contiene exactamente n símbolos, $|\Sigma| = n$, entonces

$$\left|\Sigma^k\right| = |\Sigma|^k = n^k.$$

Obs: Cuidado! Σ y Σ^1 se ven iguales, pero no son lo mismo. El primero es un alfabeto. El segundo es un conjunto de cadenas.

Definición

Para un alfabeto Σ, definimos la ***suma de Kleene*** *como el conjunto de todas sus cadenas de longitud $k \geq 1$.*

$$\begin{aligned} \Sigma^+ &= \bigcup_{k\geq 1} \Sigma^k = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots \\ &= \{w \in \Sigma^* : |w| \geq 1\} = \{\textit{cadenas de longitud} \geq 1\}. \end{aligned}$$

Definición

Para un alfabeto Σ, definimos la ***estrella de Kleene*** *como el conjunto de todas sus cadenas.*

$$\begin{aligned} \Sigma^* &= \bigcup_{k\geq 0} \Sigma^k = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots \\ &= \{w \in \Sigma^* : |w| \geq 0\} = \{\textit{cadenas en } \Sigma\}. \end{aligned}$$

Cadenas

Siempre vale $\Sigma* = \{\varepsilon\} \cup \Sigma^+$.

Ejemplo: En el alfabeto binario $\Sigma = \{0, 1\}$, tenemos

- $\Sigma^+ = \{0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111, \ldots\}$.
- $\Sigma^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111, \ldots\}$.

Finalmente, para las cadenas tenemos una operación binaria, la **concatenación**. En algunos libros, la concatenación de dos palabras x y y se denota por xy. En otros libros de denota por $x + y$.

Ejemplo: Si $x = 1111$, $y = 001110$, entonces

$$xy = x + y = 1111001110.$$

Propiedades

La concatenación posee las siguiente propiedades:

Propiedades *Para todo $x, y, z, w \in \Sigma^*$, valen*

a) $(xy)z = x(yz)$ *(asociatividad).*

b) $w\varepsilon = w$, *y* $\varepsilon w = w$, *(ε es el elemento neutro)*

Consideremos ahora la siguiente operación entre un número natural $n \in \mathbb{N}$ y una cadena w:

$$w^n = \underbrace{w + w + \ldots + w}_{n \text{ veces}}.$$

Entonces la concatenación posee la propiedad adicional:

c) $w^{n+m} = w^n\, w^m$, *(ley de exponentes).*

Lenguajes

Dado un alfabeto Σ, cualquier subconjunto de cadenas, todas ellas seleccionadas de Σ^*, se llama un **lenguaje**.

Si Σ es un alfabeto y $L \subseteq \Sigma^*$, entonces L es un lenguaje de Σ.

Ejemplos:

- El lenguaje de todas las frases del castellano.
- El lenguaje de todos los programas en Python.
- Todas las cadenas binarias, que consisten de n ceros, seguidos de n unos, para cualquier $n \geq 0$
$$L = \{\varepsilon, 01, 0011, 000111, 00001111, \ldots\}$$
- Todas las cadenas formadas por los símbolos "(" y ")" que producen jerarquías de paréntesis sintácticamente correctas
$$L = \{(), ()(), (()), ()()(), ()(()), (())(), (()()), ((())), \ldots\}$$

Ejemplos:

- El lenguaje vacío $L = \varnothing$
- El lenguaje que consiste sólo de la cadena vacía: $L = \{\varepsilon\}$.

Es habitual describir un lenguaje utilizando una "descripición de conjuntos":

- $\{w : \ w$ consta de un número igual de ceros que de unos$\}$.
- $\{w : \ w$ es un entero binario que es primo$\}$.
- $\{w : \ w$ es un programa C sintácticamente correcto$\}$.

También es habitual reemplazar w por alguna expresión con parámetros y describir las cadenas del lenguaje estableciendo condiciones sobre los parámetros.

- $L = \{0^n 1^n : \ n \geq 1\} = \{01, 0011, 000111, \ldots\}$.
- $L = \{0^i 1^j : \ 0 \leq i \leq j\} = \{\varepsilon, 1, 11, \ldots, 01, 011, \ldots, 0011, 00111, ldots\}$.