

7. Expresiones Regulares

El metalenguaje para especificar una expresión regular en este contexto es más completo de lo que a continuación se detalla, sin embargo será suficiente para nuestro propósito (ver el manual de JFlex incluido dentro del paquete).

Una secuencia de uno o más caracteres es una expresión regular, dicha secuencia puede también ir encerrada entre comillas dobles. Por ejemplo:

a aa "a" "xxx" "****"

Los siguientes son símbolos del metalenguaje llamados caracteres especiales. Los mismos tendrán un determinado efecto excepto que estén encerrados entre [] o "", o bien precedidos de la barra invertida, en donde pierden su significado especial.

'.', '*', '+', '?', '|', '[', '\', ']', '^', '\$', '{', '}'

7.1. Símbolos del Metalenguaje

7.1.1. El punto (.)

Es un metasímbolo que concuerda con todos los caracteres, menos con '\n'.

7.1.2. Clases de caracteres

Una cadena de caracteres no vacía entre corchetes es una expresión regular de un caracter que concuerda con cualquier caracter en esa cadena por ejemplo:

[0123456789]

concuera con sólo uno de los números del intervalo 0 a 9.

Otra forma de expresar lo anterior, es introduciendo el signo '-' a la manera de un subrango:

[0-9]

En el caso de que deseemos reconocer el signo '-' dentro de los corchetes, el mismo deberá ser el primer caracter. Así la siguiente expresión denota todos los números de 0 al 9, más el signo '-':

[-0-9]

7.1.3. El asterisco (*)

La expresión regular seguida de un asterisco denota cero o más ocurrencias de dicha expresión regular. Por ejemplo:

a* denota { λ , a, aa, aaa, ...}

[A-Za-z][A-Za-z_0-9]* podría denotar un identificador en un lenguaje de programación.

7.1.4. El signo +

La expresión regular seguida de un +, denota una o más ocurrencias de dicha expresión regular. Por ejemplo:

n^+ denota $\{n, nn, nnn, nnnn, \dots\}$

7.1.5. El signo de interrogación (?)

Una expresión regular seguida por ? denota 0 o 1 ocurrencia de la expresión regular. Por ejemplo:

$aa^?b^?$ denota $\{aab, aa, ab, a\}$

7.1.6. La Unión (|)

Una expresión regular separada por '|' concuerda con cadenas de caracteres denotada por cualquiera de las expresiones regulares. Por ejemplo:

$aaa | bbb | ccc$ denota $\{aaa, bbb, ccc\}$

7.1.7. Las llaves ({ })

Cuando asignamos un nombre a una expresión regular es necesario encerrarla entre llaves para expandirla y que sea reemplazada por la expresión regular que éste denota. Por ejemplo, si se define la expresión regular llamada **pal** de la siguiente manera:

pal $[a-zA-Z]^+$.

Luego se puede utilizar

{pal}

que es equivalente a $[a-zA-Z]^+$

7.2. Precedencia de los operadores

La precedencia de los operadores en orden de mayor a menor es la siguiente:

1. $[]$
2. $^?^+$
3. concatenación, que si bien no se mencionó, es simplemente poner una expresión regular seguida de la otra para formar una nueva.
4. $|$

De manera tal que si deseamos alterarla se hace uso de los paréntesis. Por ejemplo en el siguiente caso:

abc^+ denota $\{abc, abcc, abccc, \dots\}$

mientras que con paréntesis

$(abc)^+$ denota $\{abc, abcab, abcabcab, \dots\}$

aquí hace primero la concatenación y luego aplica el operador +

7.3. La barra invertida

Anula el efecto de un caracter especial. Por ejemplo, si se desea construir una expresión regular que denote una secuencia de uno o más asteriscos se deberá escribir:

`*+`

7.4. Otros ejemplos

`\.*` denota una secuencia de caracteres, que no incluye '`\n`', la cual comienza con un punto '.' (uso de la barra invertida).

`a(bb|cc)*d` denota {`ad`, `abbd`, `accd`, `abbccd`, `accbbd`, ...}

`x[*?+.]y` denota {`x*y` , `x+y` , `x?y` , `x.y`}
Aquí el efecto de los caracteres especiales está anulado por estar dentro de `[]`

`[-+]?[0-9]+` denota una constante entera con signo.

Otros símbolos especiales son: '^' y '\$', los cuales indican que una expresión regular debe ser encontrada al principio y al final de la línea respectivamente. Algunos ejemplo son:

`^[0-9]+` Constante entera al principio de línea.

`^\\\\` Doble barra invertida al principio de la línea (comentario en C++ o Java).

`A$` Una 'a' al final de la línea

`^a$` Una línea que contiene sólo una 'a'