Expresiones Regulares

Una **expresión regular** o **regex** es una serie de caracteres especiales que permiten describir un texto que queremos buscar. Por ejemplo, si quisiéramos buscar la palabra «linux» bastaría poner esa palabra en el programa que estemos usando. La propia palabra es una expresión regular. Pero, ¿y si queremos buscar todos los números que hay en un determinado archivo? ¿O todas las líneas que empiezan por una letra mayúscula? En esos casos ya no se puede poner una simple palabra. La solución es usar una expresión regular.

Podemos decir que una expresión regular es una cadena de caracteres que se utiliza para describir o encontrar patrones de texto. Son de gran utilidad y cuando aprendas a aplicarlas, te darás cuenta de lo poderosas que son, la gran cantidad de problemas que se pueden resolver, y lo mucho que aumentará tu productividad al programar.

Un punto importante a tener claro es la diferencia entre una **expresión regular** y los **patrones de archivos** que colocamos como parámetro en comandos como ls, rm, y que sirven para referenciar a varios archivos que se encuentran almacenados en el sistema de archivos de Linux. Por ejemplo, el patrón aba*.txt describe el conjunto de nombres de archivos que comienzan con **aba**, contienen cualquier otro grupo de caracteres, luego un punto, y finalmente la cadena **txt**. El símbolo * se interpreta como «0, 1 o más caracteres cualesquiera». En este capítulo, analizaremos esos operadores y las reglas de construcción de expresiones regulares.

Las expresiones regulares, aunque se parecen en que usan algunos caracteres comunes, son diferentes. Un patrón de archivo se lanza contra los archivos que hay en el disco duro y devuelve los que encajan completamente con el patrón, mientras que una expresión regular se lanza contra un texto y devuelve las líneas que contienen el texto buscado.

Tipos de expresiones regulares

Existen varios tipos de expresiones regulares más o menos estándar, pero hay programas que cambian ligeramente la sintaxis, que incluyen sus propias extensiones o incluso que utilizan unos caracteres completamente diferentes. Existen dos tipos principales de expresiones regulares, que están recogidas en el estándar POSIX, que es el que usan las herramientas de Linux. Son las expresiones regulares básicas y las extendidas. Muchos de los comandos que trabajan con expresiones regulares, como grep o sed, permiten usar estos dos tipos.

Caracteres y Metacaracteres

El patrón puede estar formado por un conjunto de caracteres (letras, números o signos) acompañado de metacaracteres que representan a otros caracteres o permiten una búsqueda contextual. Los metacaracteres reciben este nombre porque no se representan a ellos mismos, sino que son interpretados de una manera especial.

Los metacaracteres más usados son: **. * ? + [ ] ( ) { } ^ $ | **

Metacaracteres de Posicionamiento o Anclas

Los signos **^** y **$** sirven para indicar donde debe estar situado el patrón dentro de la cadena para considerar que existe una coincidencia. El signo **^** indica que el patrón debe aparecer al principio de la línea o cadena. Cuando usamos el signo **$** estamos indicando que el patrón debe aparecer al final del conjunto de caracteres, es decir antes de un carácter de nueva línea.

Expresión	Coincidencia en el texto
^el	**el** signo ^ indica que el patrón debe estar al inicio de la cadena o línea
$ila	Nos dijeron que formáramos una fila y nadie se puso en f**ila**
^$	se puede utilizar para encontrar líneas vacías, donde el inicio de una línea es inmediatamente seguido por el final de ésta

Carácter de Escape

Puede suceder que necesitemos incluir en nuestro patrón algún metacarácter como signo literal, es decir, por sí mismo y no por lo que representa. Para indicar esta finalidad usaremos como carácter de escape a la barra invertida ****. Como regla general, la barra invertida **** convierte en normales a los caracteres especiales.

Expresión	Coincidencia en el texto
\$12	El precio del menú es de **$12**,50

Comodín

El metacarácter **.** (punto) es un comodín. Un punto en el patrón representa cualquier carácter excepto nueva línea.

Expresión	Coincidencia en el texto
.l	**El** precio d**el** menú es de $12,50, así que aproveche la oferta d**el** día (Noten en este ejemplo que también incluye la ocurrencia de un espacio en blanco seguido de la letra l)

Rango de caracteres

Los corchetes **[ ]** definen una clase de caracteres y permiten encontrar cualquiera de los caracteres dentro del grupo especificado. Por ejemplo, queremos encontrar la palabra «niño», pero también queremos encontrar en caso de que hayan escrito la palabra con n en lugar de ñ. Podemos lograr esto con una clase de carácter, de forma que la expresión regular **ni[ñn]o** se interpretaría como «n, seguida de i, seguida ya sea de ñ o n, seguida de o».

La mayoría de los metacaracteres pierden su significado al ser utilizados dentro de clases de caracteres. Es así que la expresión **[a.]** se refiere literalmente a la letra a y al carácter punto.

Un caso especial es el carácter **^**, que al ser utilizado al comienzo de una clase de caracteres significa negación. Es decir que la expresión **[^a]** se refiere a cualquier cadena que NO contenga la letra a. Este significado se pierde si el **^** no está al inicio de la clase de caracteres, así que la expresión **[a^]** se refiere a la letra a y al carácter **^**.

El carácter **–** (guion) al ser utilizado dentro de corchetes adquiere un significado especial e indica que se trata de un rango. Por ejemplo, si queremos referirnos a cualquier letra del alfabeto ya sea en minúscula como en mayúscula utilizaríamos [a-zA-Z].

Expresión	Coincidencia en el texto
[abc]	El p**a**trón **c**oin**c**ide **c**on l**a** **ca**den**a** si en est**a** h**a**y un**a** **a**, un**a** **b** o un**a** **c**
[a-c]	Equiv**a**lente **a**l **ca**so **a**nterior
c[aeo]sa	Coincide con las palabras **casa**, **cesa**, **cosa**
[^abc]	El patrón coincide con la cadena si no hay ninguna a, b y c
[0-9]	Coincide con una cadena que contenga cualquiera de los dígitos
[^0-9]	**Coincide con una cadena que no contenga ningún dígito**

Clases de Caracteres

A continuación, se listan una serie de clases predefinidas y que pueden ser usadas dentro de los corchetes.

Clase	Caracteres	Significado
[:alnum:]	[A-Za-z0-9]	Caracteres alfanuméricos
[:word:]	[A-Za-z0-9_]	Caracteres alfanuméricos y _
[:alpha:]	[A-Za-z]	Caracteres alfabéticos
[:blank:]	[\t]	Espacio y tabulador
[:space:]	[\t\r\n\v\f]	Espacios
[:digit:]	[0-9]	Dígitos
[:lower:]	[a-z]	Letras minúsculas
[:upper:]	[A-Z]	Letras mayúsculas
[:punct:]	[][!"#$%&'()*+,./:;<=>?@\^_`{|}~-]	Caracteres de puntuación

Cuantificadores

También podemos indicar cuantas veces deben aparecer los caracteres. Por defecto se asume que el carácter debe aparecer una, pero podemos estar interesados en que aparezca un número distinto de veces:

Cuantificador	Significado
?	El carácter aparece **cero** o **una** vez
*	El carácter aparece **cero**, **una** o **varias** veces
+	El carácter aparece al **menos una** vez
{n}	El carácter aparece **n** veces
{n,m}	El carácter aparece entre **n** y **m** veces

Expresiones alternativas

La barra vertical se utiliza para alternativas que son evaluadas de izquierda a derecha. Por ejemplo, "Inglés|English". La barra vertical (|) separa las expresiones alternativas. En este caso significa que la palabra encontrada puede ser Inglés o English.

En esta tabla pueden observar una recopilación de los metacaracteres más utilizados y su significancia.

Carácter	Texto buscado
^	Principio de entrada o línea
$	Fin de entrada o línea
*	El carácter anterior 0 o más veces
+	El carácter anterior 1 o más veces
?	El carácter anterior una vez como máximo (es decir, indica que el carácter anterior es opcional)
.	Cualquier carácter individual, salvo el de salto de línea
x|y	x o y
{n}	Exactamente n apariciones del carácter anterior
{n,m}	Como mínimo n y como máximo m apariciones del carácter anterior
[abc]	Cualquiera de los caracteres entre corchetes. Especifique un rango de caracteres con un guión (por ejemplo, [a-f] es equivalente a [abcdef])
[^abc]	Cualquier carácter que no esté entre corchetes. Especifique un rango de caracteres con un guión (por ejemplo, [^a-f] es equivalente a [^abcdef])
\b	Límite de palabra (como un espacio o un retorno de carro)
\B	Cualquiera que no sea un límite de palabra.
\d	Cualquier carácter de dígito. Equivalente a [0-9]
\D	Cualquier carácter que no sea de dígito. Equivalente a [^0-9]
\f	Salto de página
\n	Salto de línea
\r	Retorno de carro
\s	Cualquier carácter individual de espacio en blanco (espacios, tabulaciones, saltos de página o saltos de línea)
\S	Cualquier carácter individual que no sea un espacio en blanco.
\t	Tabulación

\w	Cualquier carácter alfanumérico, incluido el de subrayado. Equivalente a [A-Za-z0-9_]
\W	Cualquier carácter que no sea alfanumérico. Equivalente a [^A-Za-z0-9_]