정규표현식

정규표현식(正規表現式, Regular Expression)은 문자열을 처리하는 방법 중의 하나로 특정한 조건의 문자를 '검색'하거나 '치환'하는 과정을 매우 간편하게 처리 할 수 있도록 하는 수단이다.

정규 표현식의 용어들

정규 표현식에서 사용되는 기호를 Meta문자라고 표현한다. 표현식에서 내부적으로 특정 의미를 가지는 문자를 말하며 간단하게 정리하면 아래의 표와 같다.

표현식	의미
^x	문자열의 시작을 표현하며 x 문자로 시작됨을 의미한다.
x$	문자열의 종료를 표현하며 x 문자로 종료됨을 의미한다.
.x	임의의 한 문자의 자리수를 표현하며 문자열이 x 로 끝난다는 것을 의미한다.
x+	반복을 표현하며 x 문자가 한번 이상 반복됨을 의미한다.
x?	존재여부를 표현하며 x 문자가 존재할 수도, 존재하지 않을 수도 있음을 의미한다.
x*	반복여부를 표현하며 x 문자가 0번 또는 그 이상 반복됨을 의미한다.
x|y	or 를 표현하며 x 또는 y 문자가 존재함을 의미한다.
(x)	그룹을 표현하며 x 를 그룹으로 처리함을 의미한다.
(x)(y)	그룹들의 집합을 표현하며 앞에서 부터 순서대로 번호를 부여하여 관리하고 x, y 는 각 그룹의 데이터로 관리된다.
(x)(?:y)	그룹들의 집합에 대한 예외를 표현하며 그룹 집합으로 관리되지 않음을 의미한다.
x{n}	반복을 표현하며 x 문자가 n번 반복됨을 의미한다.
x{n,}	반복을 표현하며 x 문자가 n번 이상 반복됨을 의미한다.
x{n,m}	반복을 표현하며 x 문자가 최소 n번 이상 최대 m 번 이하로 반복됨을 의미한다.

Meta 문자들 중에서 좀 더 특수하게 사용되는 문자들이 존재한다. '[]'는 내부에 지정된 문자열의 범위 중에서 한 문자만을 선택하다는 특수한 의미를 가진다. 그리고 내부에서 Meta문자를 사용하면 다른 의미를 가지고 동작할 수 있으므로 잘 확인하고 사용해야 한다. 좀 더 특별한 용도로 사용되는 것들은 아래의 표와 같다.

표현식	의미
[xy]	문자 선택을 표현하며 x 와 y 중에 하나를 의미한다.
[^xy]	not 을 표현하며 x 및 y 를 제외한 문자를 의미한다.
[x-z]	range를 표현하며 x ~ z 사이의 문자를 의미한다.
₩^	escape 를 표현하며 ^ 를 문자로 사용함을 의미한다.
₩b	word boundary를 표현하며 문자와 공백사이의 문자를 의미한다.
₩B	non word boundary를 표현하며 문자와 공백사이가 아닌 문자를 의미한다.
₩d	digit 를 표현하며 숫자를 의미한다.
₩D	non digit 를 표현하며 숫자가 아닌 것을 의미한다.
₩s	space 를 표현하며 공백 문자를 의미한다.
₩S	non space를 표현하며 공백 문자가 아닌 것을 의미한다.
₩t	tab 을 표현하며 탭 문자를 의미한다.
₩v	vertical tab을 표현하며 수직 탭(?) 문자를 의미한다.
₩w	word 를 표현하며 알파벳 + 숫자 + _ 중의 한 문자임을 의미한다.
₩W	non word를 표현하며 알파벳 + 숫자 + _ 가 아닌 문자를 의미한다.

정규표현식을 사용할 때 Flag 라는 것이 존재하는데 Flag를 사용하지 않으면 문자열에 대해서 검색을 한번만 처리하고 종료하게 된다. Flag는 다음과 같은 것들이 존재한다.

Flag	의미
g	Global 의 표현하며 대상 문자열내에 모든 패턴들을 검색하는 것을 의미한다.
i	Ignore case 를 표현하며 대상 문자열에 대해서 대/소문자를 식별하지 않는 것을 의미한다.
m	Multi line을 표현하며 대상 문자열이 다중 라인의 문자열인 경우에도 검색하는 것을 의미한다.