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.

ex) ^

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.

(1) .

임의의 한 문자

s.e
s 와 e 사이에 임의의 한 글자를 갖는 문자열과 일치함.
sae, sbe, sce, sde, ...

.ce
ce 앞에 임의의 한 글자를 갖는 문자열과 일치함.
ace, kce, dce, ...

(2) *

바로 앞의 문자가 없거나 하나 이상 ← 문자열

s*e
e 앞에 s 가 없거나 하나 이상 존재하는 모든 문자열을 나타냄.
e, se, see, ssse, ...

abc*
ab 다음에 c 가 없거나 하나 이상 존재하는 모든 문자열을 나타냄.
ab, abc, abcc, abccc, ...

h*im
im 앞에 h 가 없거나 하나 이상 존재하는 모든 문자열을 나타냄.
im, him, hhim, hhhim, ...

(3) +

바로 앞의 문자가 하나 이상 ← 문자열

s+e
문자 e 앞에 s 가 최소한 하나 이상 존재하는 모든 문자열을 나타냄.
se, sse, sssse, settle, ...

(4) ?

바로 앞의 문자가 없거나 하나

th?e
t 와 e 사이에 h 가 하나 있거나 혹은 없는 문자열과 일치함.
te, the, their, lotte, ...

(5) ^

바로 뒤의 문자열로 시작

```
^The
The 로 시작하는 모든 문자열을 나타냄.
The girl is beautiful, Theater, ... (뒷부분부터 공백까지 검사)

^a?bc
bc 나 abc 로 시작하는 모든 문자열을 나타냄.
bc++ 3.0, abcdef), ...
^.e
e 앞에 한 글자가 존재하는 문자열로 시작하는 모든 문자열을 나타냄.
he, me, request, settle, ...
^s.e?
s 와 임의의 한 문자로 시작하고 그 뒤에 문자 e 가 하나 있거나 혹은 없는 문자열을 나타냄.
sa, sae, sb, sbe, ... (e는 나와도 되고 안나와도 되고)
```

(6) $

바로 앞의 문자열로 종료

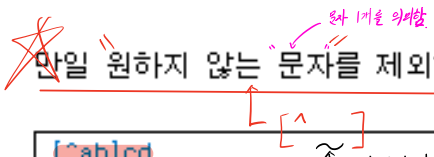
```
a?bc$
bc 로 끝나는 문자열 앞에 문자 a 가 없거나 하나 존재하는 문자열과 일치함.
eeabe, seebc, bc, ...

t.e$
t 와 임의의 한 글자, 그리고 그 다음에 e 로 연결되어 끝나는 문자열과 일치함.
onthetoe, bctae, appetitte, ...
s?e+$
첫 글자는 s 나 e 로 시작하고 하나 이상의 e 로 끝나는 문자열과 일치함.
e, se, ee, eee, seee, seee, ...

^the$
the 로 시작해서 the 로 끝나는 문자열과 일치함.
즉, 이 경우는 문자열 자체가 the 뿐인 경우에만 일치함.
the
```

(7) []

[] 안에 있는 문자 중 하나, 범위는 '-'로 지정함.
[a-d] 는 [abcd] 와 동일하며 [0-9] 는 [0123456789] 와 같은 의미임.

```
[ab]cd
acd 또는 bcd 를 포함하는 문자열과 일치함.
acd, tacde, "bcd", "tbcde", ...

^[ab]cd
acd 또는 bcd 로 시작하는 문자열과 일치함.
acds, bcdt, acdsee32, ...

[a-z]
영문 소문자 한 글자를 포함하는 문자열과 일치함.
a0c2ds, ta123cde, Student, ...

[a-zA-Z]
영문 소문자나 대문자 한 글자를 포함하는 문자열과 일치함.
LINUX, 386AT, ...

[0-9]
십진수 한 자를 포함하는 문자열과 일치함.
a0c2ds, ta123cde, 386, ...

ga[a-z]
하나의 영문 소문자 앞에 ga를 갖는 문자열과 일치함.
LINgazUX, gazzett, ...

^ab[cd]ef
abcef 또는 abdef로 시작하는 문자열과 일치함.
abcef0z, abdef386, ...

^[a-zA-Z]
영문자로 시작하는 모든 문자열과 일치함.
LINgazUX, abcef0z, ...

[a-z]+
영문소문자 한 자 이상을 갖는 문자열과 일치함.
tgabcabcef, MySQL, ...

[aA][bB]
ab, aB, Ab, AB를 포함하는 문자열과 일치함.
386ABIT, abcef0z, tgabcabcef, ...

.[a-zA-Z0-9]$
점마와 하나와 영문자 또는 숫자로 끝나는 문자열과 일치함.
abdef38,6, 199,2, ...
```

만일 원하지 않는 문자를 제외한 문자를 가리킬 때에는 [] 안의 첫 문자로 ^ 를 사용함.



[^ab]cd
acd 와 bcd 를 제외하고는 패턴 .cd와 같음.
즉, cd 앞에 a 나 b 를 제외한 하나의 문자를 포함하는 문자열과 일치함.
ccd, scd, 018cd, tgcdcdabcef, gazcd18, ...

s[^ab]t
sat 와 sbt 를 제외하고는 패턴 s.t 와 같음.
즉, t 앞에 a 나 b 를 제외한 임의의 한 문자와 그 앞에 s 가 있는 문자열과 일치함.
sct, sdt, tgcdsctda28, settle, ...

[^a-z]
영문 소문자를 제외한 한 글자를 포함한 문자열과 일치함.
MySQL, 386sAB,IT, abcef0z, 199,2, ...

[^a-zA-Z]
영문자를 제외한 한 글자를 포함하는 문자열과 일치함.
286sAB,IT, gazscd18, abcef0z, 199,2, ...

[^0-9]
숫자를 제외한 한 글자를 포함하는 문자열과 일치함.
settle, gazscd18, LINUX, ...

(8) {}

{} 앞의 문자나 문자열 출현 횟수

a{2}b
aab 를 가진 문자열과 일치함.
즉, {2} 는 {} 앞에 있는 문자 a 의 개수가 2 개임을 의미함.
aab, ...

a{2,}b
a 의 개수가 최소한 2 개 이상인 문자열을 포함하는 문자열과 일치함.
aab, aaab, aaaab, ...

a{1,3}b
b 앞에 1 개부터 3 개까지의 a 를 갖는 문자열을 포함하는 문자열과 일치함.
ab, aab, aaab, ...

(9) ()

괄호안의 문자열을 '한 문자'로 취급

()안에 있는 문자를 그룹화

a(bc){2}
a 뒤에 bc 의 개수가 두 개인 문자열 abcbc 를 포함하는 모든 문자열과 일치함.
[bc] 가 b 또는 c 중 하나를 의미하는 것에 비해 (bc) 는 bc 를 하나의 그룹으로 처리함.
docabcbctor, tabcbc++, ...
a(bc)*
a 뒤에 bc 가 없거나 하나 이상인 문자열과 일치함.
sea, abcd, abcbcbcbc, ...

(10) |

or 연산자

he|she
he 나 she 를 포함하는 문자열과 일치함.
he is handsome, she's gone, ...

(he|she)is
is 앞에 he 나 she 를 포함하는 문자열과 일치함.
즉, heis 나 sheis 를 포함하는 모든 문자열과 일치함.
heis, sheis, ...

(le|li)*ft
ft 앞에 le 나 li 가 없거나 하나 이상인 문자열과 일치함.
mlefto, lft, lelift, fclelelilefte, ...

mo(no)+
mo 뒤에 no 가 하나 이상인 문자열과 일치함.
mono, monono, mononono, acmonoe, ...

(11) 특수 문자 사용

^ [] $ () | * + ? {} ₩
앞에 ₩ 붙여서 사용해야 함.

₩++ : + 가 하나 이상 포함된 패턴
₩d : 순수한 숫자, 정수값, 0-9
₩d{2,3}-₩d{3,4}-₩d{4} : 전화번호 정규식. -? 하이퍼뒤에 물음표가 있으면 하이퍼가 있어도 되고 없어도 된다는 뜻임.
₩D : ₩d와 반대 (숫자를 제외한 나머지)
₩w : [a-zA-Z0-9] 의 줄임표현
₩W : [^a-zA-Z0-9] 영문자와 숫자만 아니면 된다는 뜻.
₩s : 공백 문자
₩S : ₩s와 반대 (공백 문자를 제외한 나머지)