예제와 함께하는 정규표현식

?의 신비

자바카페 서동우

INDEX

정규표현식이란?

Why?

Question

1. 대소문자를 구별하지 않고 car 라는 텍스트가 포함된 단어 찾기
2. 블로그에 글을 작성하는데 URL 형식의 내용을 들어오면 자동으로 링크 생성하도록 하기
3. 이메일 주소 체크하기

프로그램으로 작성하려고 하면?

정규표현식을 사용하면?

1. /[Cc][Aa][Rr]/g
2. /(https?:₩/₩/[a-zA-Z0-9.-₩/]+)/g
3. /^[_₩.0-9a-zA-Z-]+@([0-9a-zA-Z][0-9a-zA-Z-]+₩.)+[a-zA-Z]{2,6}$/

간단하게 구현이 가능!(하지만...)

When

When?

1. Text 검색
2. Text 치환
3. Text validation

자 그럼 도대체 멀까?

1. 텍스트를 **찾고**, **조작**하는데 쓰이는 **문자열**
2. 정규 표현 언어를 사용하여 만든다.
3. 하나의 작은 언어로 자바, c#, 자바스크립트 등의 다른 언어에서 라이브러리화 시켜서 지원

식	기능	설명
.	문자	1개의 문자와 일치한다. 단일행 모드에서는 새줄 문자를 제외한다.
₩	이스케이프	특수 문자를 식에 문자 자체로 포함한다.
¦	선택	여러 식 중에서 하나를 선택한다. 예를 들어, "abc¦adc"는 abc와 adc 문자열을 모두 포함한다.
^	부정	문자 클래스 안의 문자를 제외한 나머지를 선택한다. 예를 들면 [^abc]d는 ad, bd, cd는 포함하지 않고 ed, fd 등을 포함한다. [^a-z]는 알파벳 소문자로 시작하지 않는 모든 문자를 의미한다.
[]	문자 클래스	"["과 "]" 사이의 문자 중 하나를 선택한다. "¦"를 여러 개 쓴 것과 같은 의미이다. 예를 들면 [abc]d는 ad, bd, cd를 뜻한다. 또한, "-" 기호와 함께 쓰면 범위를 지정할 수 있다. "[a-z]"는 a부터 z까지 중 하나, "[1-9]"는 1부터 9까지 중의 하나를 의미한다.
()	하위식	여러 식을 하나로 묶을 수 있다. "abc¦adc"와 "a(b¦d)c"는 같은 의미를 가진다.
*	0회 이상	0개 이상의 문자를 포함한다. "a*b"는 "b", "ab", "aab", "aaab"를 포함한다.
+	1회 이상	"a+b"는 "ab", "aab", "aaab"를 포함하지만 "b"는 포함하지 않는다.
?	0 또는 1회	"a?b"는 "b", "ab"를 포함한다.
{m}	m회	"a{3}b"는 "aaab"만 포함한다.
{m,}	m회 이상	"a{2,}b"는 "aab", "aaab", "aaaab"를 포함한다. "ab"는 포함되지 않는다.
{m, n}	m회 이상 n회 이하	"a{1,3}b"는 "ab", "aab", "aaab"를 포함하지만, "b"나 "aaaab"는 포함하지 않는다.

Mission 1

Question 1

문자열 :

자바카페 세미나파이팅 입니다.

목표 :

머냐 이건… -_-?

답변

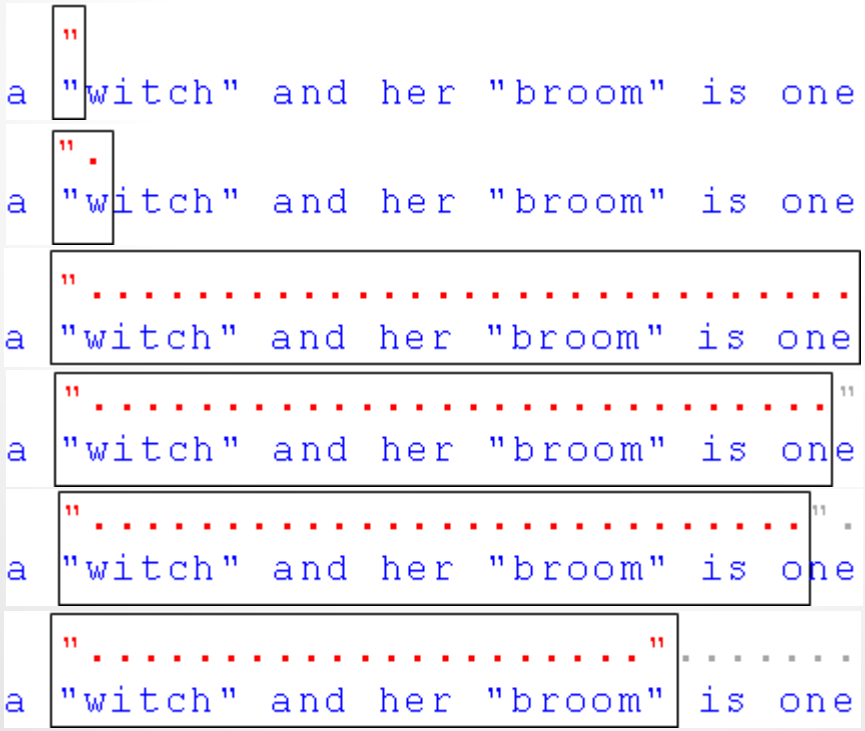
.+

실제결과 :

자바카페세미나파이팅 입니다.

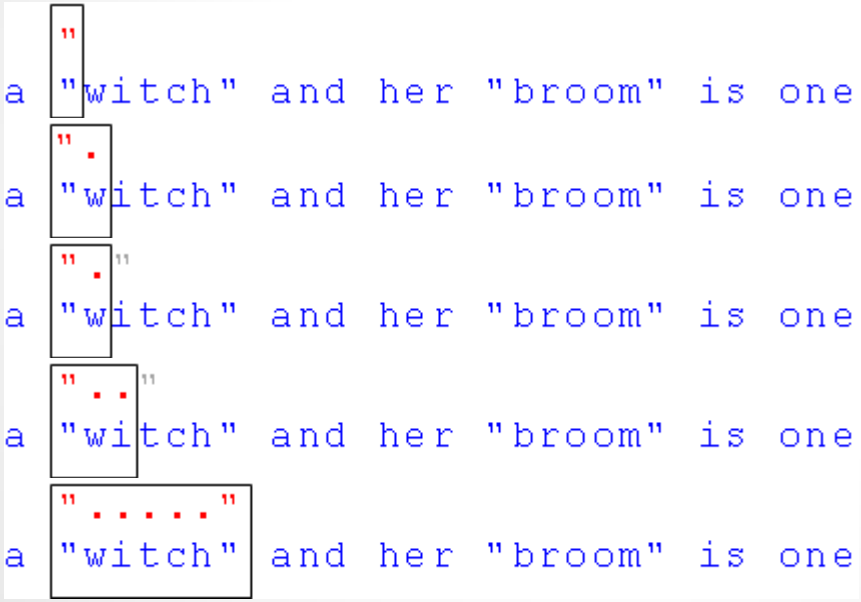
Greedy Quantifier

- 정규표현식 수량자는 기본적으로 Greedy Qnantifier을 가짐
- 이 분은 매우 욕심이 많음.
- ".+"



Lazy Quantifier

- 정규표현식 수량자에 ?를 붙임으로써 표현
- 이 분은 매우 귀찮아함.
- ".+?"



그래서 어떻게?

답변 :

.+?

실제결과 :

자바카페세미나파이팅 입니다.

Mission 2

Question

문자열 :

<h1>자바카페</h1>세미나<h2>파이팅</h2> 입니다.

목표 :

<h1>자바카페</h1>세미나<h2>파이팅</h2> 입니다.

답변 :

<h[1-9]>.+?</h[1-9]>

약간 변형된 Question

문자열 :

<h1>자바카페</h2>세미나<h2>파이팅</h1> 입니다.

목표 :

<h1>자바카페</h2>세미나<h2>파이팅</h1> 입니다.

결과 :

<h1>자바카페</h2>세미나<h2>파이팅</h1> 입니다.

()

- 그룹을 지정한다.
  abc 가 3번 이상 시작하는 문자열을 검색 ➔ ^(abc){3,}.*

- 캡쳐의 기능을 제공한다.
  <h1>자바카페</h1>세미나<h2>파이팅</h2> 입니다.

  ➔ <h([1-9])>.*?<h₩1>

만약 캡쳐 기능이 필요 없다면?

- 데이터 저장을 하는데 리소스 낭비
- 이것을 피할 방법은 없을까?

(?:정규표현식) 을 사용

Mission 3

Question 1

문자열 :

나는 어제 201호 아저씨한테 5000원을 줄래, 500대 맞을래 하고 선택하라는 이야기를 듣고 5000원들 주는 것으로 이야기 했습니다.

목표 :

나는 어제 201호 아저씨한테 5000원을 줄래, 500대 맞을래 하고 선택하라는 이야기를 듣고 5000원들 주는 것으로 이야기 했습니다.

답변 :

[0-9]*?원 ➔ Replace 해서 원을 없앤다....

Lookahead(전방탐색)

- 작성한 패턴과 일치하는 영역이 나오면 그 부분을 제외하고 나머지 부분 나오는 패턴
- 앞에 있는 문자열을 검색
- (?=...) 형태로 사용한다.
- 앞의 예제의 정답은 ➔ [0-9]+?(?=원)

- 그럼 원을 제외한 201(호), 500(대)를 가져올 수 있는 방법은 없을까?

[0-9]+?(?!원)

Question 2

문자열 :

나는 어제 #201 아저씨한테 $5000를 줄래, 500대 맞을래 하고
선택하라는 이야기를 듣고 $5000들 주는 것으로 이야기 했습니다.

목표 :

나는 어제 #201 아저씨한테 $5000를 줄래, 500대 맞을래 하고
선택하라는 이야기를 듣고 $5000들 주는 것으로 이야기 했습니다.

답변 :

$[0-9]*? ➔ Replace 해서 $를 없앤다....

Lookbehind(후방탐색)

- 작성한 패턴과 일치하는 영역이 나오면 그 부분을 제외하고 나머지 부분 나오는 패턴
- 뒤에 있는 문자열을 탐색
- (?<=...) 형태로 사용한다.

- 앞의 예제의 정답은 ➔ (?<=$)[0-9]+?

- 그럼 원을 제외한 (#)201, ( )500를 가져올 수 있는 방법은 없을까?

(?<!$)[0-9]+?

Last Mission

Question

문자열 :

<h1>자바카페</h1> 세미나<h2>화이팅</h2>하하하

목표 :

<h1>자바카페</h1> 세미나<h2>화이팅</h2>하하하

답변 :
(?<=<h([0-9])>).*?(?=</h₩1>)

THANK YOU!!!