


| 카카오톡 봇 강좌 | >

[중급] (라이노 엔진) 자바스크립트 기초 강의 (4 - 3) - 정규식과 Jsoup select 보충

마른얼음 BOT 1:1 채팅

2019.07.16. 20:42 조회 945

💬 댓글 32

🔗 URL 복사

⋮

많은 분들이 이 부분에 대해 어려워하시는 것 같아서, 보충 강좌를 준비했습니다

일단 정규식부터 알아보겠습니다.

정규식은 /식/(패턴변경자) 으로 생겼는데, 가장 많이 쓰이는 패턴변경자가 g입니다.
이 정규식의 용도는 크게 두가지 입니다.

- 1. 문자열.replace(정규식, 바뀔문자열)
- 2. 정규식.test(문자열)

첫번째는 정규식에 해당되는 문자열을 바뀔문자열로 바꿔주고, 두번째는 문자열이 정규식에 해당하냐를 논리값(true, false)로 알려줍니다.
그럼 이제 본격적으로 정규식을 배워보겠습니다.

/문자열/g

정규식의 기본형입니다. 예를 한번 들어보죠

"안녕안녕".replace(/안녕/g, "하이")
//결과 : 하이하이

/안녕/g.test("안녕하세요")
//결과 : true

네. 하지만 이렇게 정규식을 쓴다면 정규식을 왜 쓰나 싶죠? 정규식에서만 해당되는 특수 기호들을 소개해드리겠습니다.

- (문자열1)+ : 문자열1이 한 개 또는 그 이상 있는 것
- (문자열1)* : 문자열1이 0개 또는 그 이상 있는 것
- ^(문자열1) : 대상이 문자열1로 시작하는 것
- (문자열1)\$: 대상이 문자열1로 끝나는 것
- (문자열1)? : 문자열 1이 한개 또는 없는 것
- . : 임의의 한 문자
- [] : 단일 문자들의 그룹. ([abc]는 a, b, c를 찾음)
- (문자열1){num} : 문자열1이 num개 만큼 있는 것
- (문자열1){num,} : 문자열1이 num개 또는 그 이상 있는 것
- (문자열1){onum,tnum} : 문자열1이 onum개 이상 있지만 그 개수가 tnum을 넘지는 않는 것

뭔 싹소린가 싶죠? 예를 한번 들어봅시다.

"1233333345".replace(/3+/g,"3")
//결과 : 12345

"사과 배 사과 포도".replace(/^사과/g,"복숭아")
//결과 : 복숭아 배 사과 포도

//결과 : 중간에 매 시리 포조

```
/47$/g.test("10293847")
```

//결과 : true

```
"자스조아".replace(/[조아]/g,"시리")
```

//결과 : 자스시리시리

```
"1233334512345".replace(/123{2,4}/g,"3")
```

//결과 : 1234512345

이제 이해가 되실겁니다. (저런 특수기호들을 정규식 안에서 메타문자라고 부릅니다.)

그리고 이 메타문자들을 문자열 한개만 인식시키지 않고 여러개를 인식시키고 싶다면, 괄호로 묶어주시면 됩니다. 예를 들어보자면,

```
"안녕하하하하세요".replace(/안녕하+/g,"하")
```

//결과 : 하세요

```
"안녕하안녕하안녕하하세요".replace(/(안녕하)+/g,"하")
```

/결과 : 하세요

둘다 결과값이 똑같긴 하지만, 괄호를 묶음으로서 어떻게 변하는지 보여드렸습니다.

또 정규식에서, 대괄호 안에서 특정 메타문자는 뜻이 바뀌어버리는 메타문자가 있습니다.

대괄호([]) 안에서

^(문자열1) : 문자열1이 아닌 것

(문자열1)-(문자열2) : 문자열1에서 문자열2까지의 문자인 것

(문자열1)-(문자열2)가 잘 이해되지 않으실텐데, 예제를 통해 알아보시다.

```
"가나다라abcdABCD1234".replace(/[a-z]/gi,"")
```

//결과 : 가나다라1234

```
"가나다라abcdABCD1234".replace(/[0-9]/g,"")
```

//결과 : 가나다라abcdABCD

정규식 타입에서 i는, 대문자와 소문자를 구분하지 않는다는 것입니다.

여기서, 우리가 예제를 통해 [0-9]를 하면 모든 숫자를 감지한다는 것을 알았지만, 사실 더 적합한 메타문자가 있습니다.

바로 \d인데 \d는 모든 숫자를 감지합니다. \d와 같은 메타문자를 몇개 더 알아보겠습니다.

\d : 숫자

\s : 공백문자

\b : 문자와 공백 사이

\w : 단어 ([0-9a-zA-Z_]와 같은 의미임)

이런 것들이 있습니다.

마지막으로, 가끔씩 문자를 인식시키고 싶을 때가 있는데 그 문자가 메타문자로 인식되어서 다른 의미가 되어버리는 경우가 있습니다. 그럴때는 \ (역슬래시)를 붙여주면 됩니다. 예를 들어봅시다

```
"안녕^^".replace(/^안녕\w^/g,"^")
```

//결과 : ^^

이런식으로 말이죠

여러분이 정규식을 잘 이해했나 테스트 하기 위해 HTML 태그를 곱살시켜주는 정규식을 분석해보겠습니다.

/<[^>]+>/g

자, 이 정규식을 분석해봅시다. 혹시 혼자 분석해보고 싶다면 아래로 내리지 말아주세요!

먼저, < 는 말 그대로 문자 < 를 의미합니다. 그리고 [^>]+는 > 가 아닌 문자가 한개 이상 있다는걸 의미하고, 마지막 > 는 문자 > 를 의미합니다.

결과적으로 정규식을 해석하자면,

<(>가 아닌 문자들)>

가 되겠네요!

이제는 Jsoup select를 조금 더 알아보도록 하겠습니다.

일단 지난 시간에 배운 selector(select 뒤 괄호에 있는 것들이 selector입니다)를 정리해보겠습니다.

예제 HTML)

```
<body>
<div class="abc">닥토봇</div>
<div class="abc">메봇</div>
<span id="abcd">새카봇</span>
<a href="abcde">마얼봇(?)</a>
</body>
<div class="abc">함정이닷!</div>
```

1. 마얼봇(?)을 가져오고 싶을때

```
select("a[href=abcde]")
```

2. 새카봇을 가져오고 싶을때

```
select("span#abcd")
```

3. 닥토봇과 메봇을 가져오고 싶을때

```
select("body > div.abc")
```

4. 메봇만 가져오고 싶을때

```
select("div.abc").get(1)
```

여기까지 배워봤던 것 같네요!

이제는 selector를 좀 더 알아봅시다.

일단 selector에 또 어떤것들이 있는지 알아보겠습니다.

select("div[class]") : 속성이 class인 div 태그를 가져옴

select("div[^cl]") : 속성이 cl로 시작하는 div 태그를 가져옴

select("div[class^=a]") : 클래스명이 a로 시작하는 div 태그를 가져옴

select("div[class\$=z]") : 클래스명이 z로 끝나는 div 태그를 가져옴

select("div[class*=g]") : 클래스명이 g를 포함하는 div 태그를 가져옴

select("div[class~=식]") : 식(정규식)에 해당되는 클래스명의 div 태그를 가져옴

select("div:nth-child(2n+5)") : 여러개의 div 태그들 중 2n+5번째의 태그를 가져옴. 예) 7번째, 9번째, 11번째 ...

흠.. 무슨 개소리냐 싶죠? 차근차근 예제를 들며 설명하겠습니다.

(편의를 위해 앞으로 나올 예제 HTML이 변수 doc에 담겨있다고 가정합니다)

예 1)

```
<div class="abcd">닥토봇</div>
```

```
<div class= abcd >카톡봇 조아1</div>
<div class="abde">카톡봇 조아2</div>
<div class="bcde">카톡봇 시러</div>
```

```
doc.select("div[class^=ab]").text()
//결과 : 카톡봇 조아1 카톡봇 조아2
//클래스명이 ab로 시작하는것을 select해서 abcd, abde를 가져옴
```

```
doc.select("div[class$=de]").text()
//결과 : 카톡봇 조아2 카톡봇 시러
//클래스명이 de로 끝나는것을 select해서 abde, bcde를 가져옴
```

```
doc.select("div[class~=a[bcd]]").text()
//결과 : 카톡봇 조아1 카톡봇 조아2
//클래스명이 정규식 a[bcd]에 해당되는것을 select해서 abcd, abde를 가져옴
```

예 2)

```
<div>1</div>
<div>2</div>
<div>3</div>
<div>4</div>
<div>5</div>
<div>6</div>
<div>7</div>
<div>8</div>
```

```
doc.select("div:nth-child(2n)").text()
//결과 : 2 4 6 8
//2n에 해당하는 2번째, 4번째, 6번째, 8번째를 가져옴
//아이러니하게도 nth-child 쓸때는 순서가 1부터 시작
```

```
doc.select("div:nth-child(3n-2)").text()
//결과 : 1 4 7
//3n-2에 해당하는 1번째, 4번째, 7번째를 가져옴
```

```
doc.select("div:nth-child(5)").text()
//결과 : 5
//n에 상관없이 5번째를 가져옴
```

이 예제들로 이해가 되셨길 바랍니다.

강좌 꼬을

크롤링 후 파싱에 대해 어려워하시는 분들이 많아서 강의를 써봅니다. 다음 강좌는 여러 파싱 예제를 올리도록 하겠습니다
당
그럼 이만



마른얼음 BOT님의 게시글 더보기 >

❤ 좋아요 11 💬 댓글 32

🔗 공유 | 신고



인디벨
정성추

2019.07.16. 20:45 답글 쓰기



마른얼음 BOT 작성자
헤헷

2019.07.16. 20:52 답글 쓰기



큐브마인

[]는 문자열이 아니라 문자 또는 문자입니다.
문자열 또는은 |를 사용해야 합니다.

예시) abc 또는 def
(abc|def)

2019.07.16. 20:48 답글 쓰기



마른얼음 BOT 작성자

[(문자열1)(문자열2)]
여기서 괄호를 쳐주면서 문자열1이라고 해놓았으니 여기서는 문자열1 또는 문자열2가 맞습니다.
그리고 또는은 |를 사용해도 되지만 []를 사용하면 더 쉽습니다.

2019.07.16. 20:50 답글 쓰기



마른얼음 BOT 작성자

마른얼음 BOT 아하 문자 또는 문자라는 뜻을 제가 잘못 이해했네요 죄송합니다..

2019.07.16. 20:51 답글 쓰기



평범한 사람

초보는 읍니다(?)흠

2019.07.16. 20:49 답글 쓰기



마른얼음 BOT 작성자

ㅠㅠ.. 노력하면 언젠가 고일 수 있어요 힘내세요!

2019.07.16. 20:52 답글 쓰기



평범한 사람

흠흠

2019.07.16. 20:53 답글 쓰기



엘지

정——성

2019.07.17. 00:24 답글 쓰기



doami2005

a=12345 에서
결과를 ㄱ ㄴ ㄷ ㄹ ㅁ 이런식으로 하나하나를 한번에 바꾸는 법이 있나요?
숫자들을 모두
간지나는 특수문자로 바꾸려는데

2019.07.17. 13:38 답글 쓰기



큐브마인

replace에다가 함수 쓰세요

2019.07.17. 16:12 답글 쓰기



doami2005

큐브마인 ..?

2019.07.17. 17:41 답글 쓰기



큐브마인

doami2005 시간 생기면 단편 강좌 하나 쓸게여

2019.07.17. 18:00 답글 쓰기



마른얼음 BOT 작성자

2,3,4번째를 갖고오는 방법은 강 노가다밖에 없는것 같아요..
get(1)
get(2)
get(3)
이런식으로

2019.07.20. 23:00 답글쓰기

삭제된 댓글입니다.



마른얼음 BOT 작성자

selector가 잘못되었네요
select("a[href=/stats/나오는 코드]")

2019.07.22. 14:57 답글쓰기



마른얼음 BOT 작성자

냥냥이 selector에 정규식을 쓰면 됩니당

2019.07.22. 15:28 답글쓰기



마른얼음 BOT 작성자

냥냥이 a[href~?=^₩/stats₩/]

2019.07.22. 15:35 답글쓰기



또르륵

<dc:date>Mon, 29 Jul 2019 20:17:56 GMT</dc:date>

dc:date 에 있는 걸 가져오려고 items.get(i).select("date").text() 어떻게 하면 될까요 ㅠ

2019.07.30. 13:15 답글쓰기



마른얼음 BOT 작성자

select("dc:date") 하면 되죠

2019.07.30. 13:16 답글쓰기



또르륵

마른얼음 BOT 그렇게하니 에러가 나요 ㅠ
javaException 이라고 ㅠㅠ

2019.07.30. 17:08 답글쓰기



마른얼음 BOT 작성자

또르륵 헐.. select("dc") 해보세요
태그에 :이 들어간건 한번도 못 봐서....

2019.07.30. 20:51 답글쓰기



또르륵

마른얼음 BOT responseMsg += "- 일자 : " + items.get(i).select("dc").text() + "₩n₩n";
해봤는데 실패했어요 ㅠㅠ

2019.07.31. 11:12 답글쓰기



또르륵

마른얼음 BOT http://rss.nocutnews.co.kr/nocutnews.xml
여기 한번 봐 주시면 ^^;;

2019.07.31. 11:13 답글쓰기



마른얼음 BOT 작성자

또르륵 String으로 바꾼다음 split하는방법밖에 없는것 같아요

2019.07.31. 11:31 답글쓰기



bass90301

Aㅏ select모르고 방금전에 split하고 replace로 노가다 하다가 정규식 사이에 변수가 안넣어져서 이 게시글 이어서 보고 있는
데.. select가 있어..(내 시간!!!!)



2019.08.24. 19:25 답글 쓰기



마른얼음 BOT 작성자

select는 Jsoup 안에서 쓸 수 있는 것으로, Utils.getWebText 방법에는 못써요

2019.08.24. 19:35 답글 쓰기



bass90301

마른얼음 BOT 아;;그렇군요..(머쓱)

2019.08.24. 19:36 답글 쓰기



마른얼음 BOT 작성자

bass90301 대신 Utils.getWebText(링크) 를 org.jsoup.Jsoup.connect(링크).get() 으로 치환한다면 select를 쓸 수 있지요. 대신 split이랑 replace는 못 쓰지만 (형변환 해준다면 물론 쓸수는 있죠)

2019.08.24. 19:43 답글 쓰기



bass90301

마른얼음 BOT 감사합니다.

들어가기 전에 실례가 안된다면 질문하나만 하겠습니다.

split(/cursor:pointer">/g)[1]

이곳에서 a라는 변수를 끝에 추가하는 방법이 있나요?

split(/cursor:pointer">/g+a)[1]

로 하니 전에 직접 써넣었을때 잘되던게 a라는 변수에 넣어서 저런식으로 테스트 하니 안되더라고요.

인터넷에 검색해보니 new RegExp를 이용해 하라고 하지만 테스트 결과 안됐습니다.

방법이 있다면 알려주시면 감사하겠습니다.

그럼 좋은밤 되세요!

2019.08.24. 20:01 답글 쓰기



마른얼음 BOT 작성자

bass90301 a가 어느 위치에 들어가야 하는건가요?

2019.08.24. 20:08 답글 쓰기



bass90301

마른얼음 BOT split(/cursor:pointer">(여기요)/g)[1]

2019.08.24. 20:10 답글 쓰기



bass90301

bass90301 과거의 나를 보았다..!!

2020.03.09. 18:41 답글 쓰기

Hibot

댓글을 남겨보세요



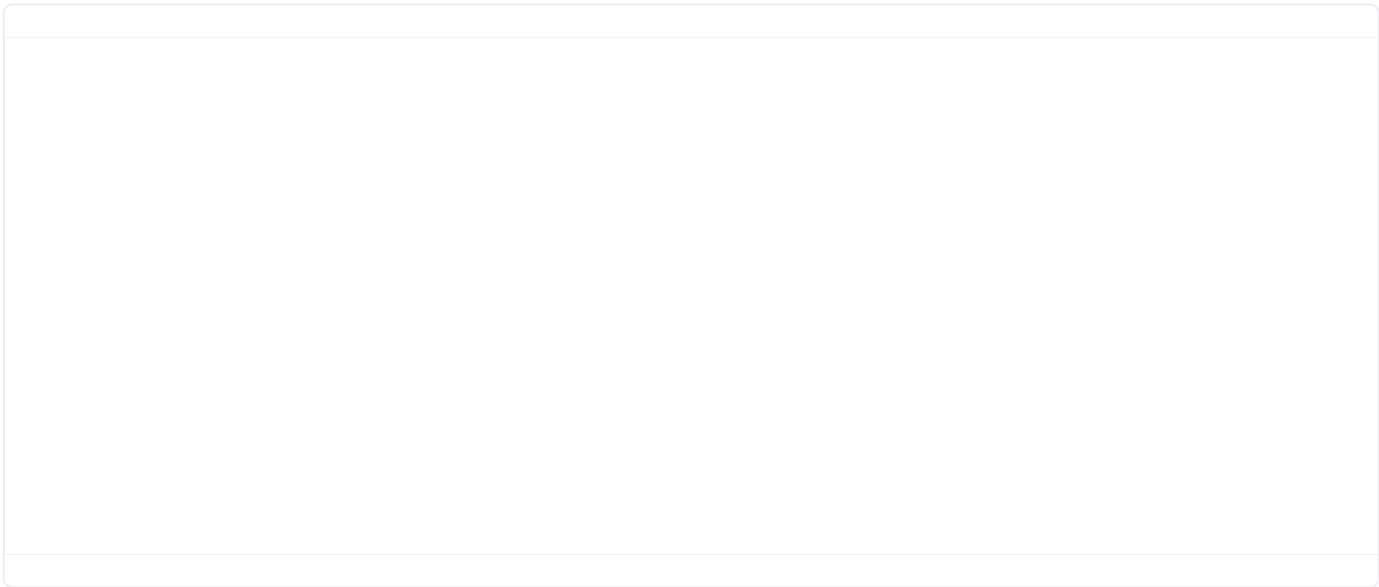
등록

글쓰기

답글

목록

▲ TOP



'| 카카오톡 봇 강좌 |' 게시판 글

전체 [중급] 말머리 글

이 게시판 새글 구독하기 ☐

[중급] 여러 개의 replace 1개로 만들기 [9]	큐브마인	2019.07.17.
[중급] Jsoup 사용 시 개행 태그() 살리기 [1]	인디벨	2019.07.16.
[중급] (라이노 엔진) 자바스크립트 기초 강의 (4 - 3) - 정규식과 Jsoup select 보충 [32]	마른얼음 BOT	2019.07.16.
[기타] 크롬에서 모바일 페이지로 보기 📱	인디벨	2019.07.15.
[중급] (라이노 엔진) 자바스크립트 기초 강의 (4 - 2) - 파싱 (Jsoup) 🤖 [21]	마른얼음 BOT	2019.07.06.

1 2 3

전체보기

이 카페 인기글

복불

하하하하
♡0 💬5

가르치기 리로드..

Milk2
♡0 💬9



자동응답봇으로 RPG봇 만들었습니다

간편 자동응답과 한글코딩을 메인으로 제공
하는 카톡봇앱 개발 예정

성빈
♡1 💬14

[봇드게임] 게이트 v0.9.0

ZUMP
♡0 💬4



오류뜨는데 해결법 알려주세요ㅠㅠ

1 2 3 4

반가워요.

천방지축하연
♡0 💬4



틱택토 (Player vs Player)



태양, 달 정보 구현 완료