

| 카카오톡 봇 강좌 | >

[중급] (라이노 엔진) 자바스크립트 기초 강의 (4 - 2) - 파싱 (Jsoup)



마른얼음 BOT 1:1 채팅

2019.07.06. 16:55 조회 1,161

💬 댓글 21 URL 복사 ⋮

저번 시간에는 단순히 문자열로 파싱을 하는 법을 알아봤지만, 이번 시간엔 Jsoup 이라는걸로 한번 파싱을 해 보겠습니다.

(참고로 이 글은 겁나게 길 수 있습니다.)

Jsoup이란 무엇일까요?

Jsoup은 우리가 파싱을 더욱 쉽게 할 수 있도록 만들어진 것 입니다. Utils.getWebText는 단순히 HTML만 가져왔다면, Jsoup은 HTML을 객체로 가져와서 더 쉽게 파싱을 할 수 있습니다.

일단 Jsoup의 사용법을 알아보겠습니다.

org.jsoup.Jsoup.connect(링크)

이 뒤엔 get() 이나 post() 로 서버 접속 방법을 설정할 수 있습니다.

org.jsoup.Jsoup.connect(링크).get()
org.jsoup.Jsoup.connect(링크).post()

일단 처음 배워보는것이니 get을 써보도록 하겠습니다.

먼저, 기본 파싱에 replace와 split이 있듯이, 여기서도 자주 쓰이는게 있습니다.

.select(태그)
.get(숫자)
.text()
.html()

등등 여러가지가 있습니다. 이 중에서 제일 많이 쓰이는 것은 select인데, 예제를 통해 사용 방법을 알아보시다.

예제 HTML)

```
<div class="abc">  
<div class="abcde">안녕하세요</div>  
</div>
```

우리가 HTML에서 이런걸 너무나 많이 보았는데 무시하고 지나갔죠? 이제는 애를 해석해야합니다.

일단 중간에 <div class="abcde">를 주목해주세요!

앞에 있는 div는 태그명이고, 뒤에 있는 class는 태그의 속성입니다. 태그 속성에는 class나 id 등등 여러가지가 있습니다. 그리고 속성 뒤에 있는 "abcde"는 속성값입니다.

그리고 뒤에 있는 </div>는 div 태그를 끝낸다 이 말입니다. 우리가 자바스크립트에서 중괄호 닫기랑 비슷한 역할입니다.

아무튼, 우리가 select를 하려면 태그명과 속성 그리고 속성값을 알아야 합니다. 저 예시 HTML에서 안녕하세요를 가져와 봅시다.

일단 저기서 div 태그의 속성값이 abcde인 태그를 가져오려면 이렇게 하면 됩니다. 저 HTML이 저장된 변수의 값을 doc 라고 하면,

```
doc.select("div[class=abcde]")
```

이렇게 하시면 속성값이 abcde인 div 태그만 뽑아올 수 있습니다. 뽑아온 결과는,

```
<div class="abcde">안녕하세요</div>
```

이렇게 됩니다.

여기서 우리가 안녕하세요만 뽑아올건데, text를 이용하면 됩니다. text는 HTML에 있는 태그를 다 지워줍니다.

```
doc.text()
```

이렇게 하면 결과가 안녕하세요 이겠네요.

결과적으로 최종 소스는,

```
doc.select("div[class=abcde]").text()
```

이렇게 됩니다.

예제 HTML을 한번 더 만들어 보겠습니다.

예제 HTML2)

```
<span id="뽕">뿌앵</span>
<div class="abcd">
  <span id="뽕">와! 샌즈!</span>
  <span id="뽕">언더테일 아시는구나!</span>
</div>
```

여기서 우리는 와! 샌즈! 를 가져오는게 목표입니다.

일단 저 HTML이 저장된 변수 이름도 doc라 해보고, 한번 뽑아와봅시다.

```
doc.select("span[id=뽕]")
```

근데 이러면 문제점이, 저기 보이는 모든 span 태그를 가져와버려서 문제입니다. 이럴땐 div 태그를 먼저 뽑아온다음 span을 뽑아옵시다.

```
doc.select("div[class=abcd]").select("span[id=뽕]")
```

이러면 확실하게 할 순 있지만 너무 기네요! 줄이기 위해 > 를 이용해봅시다.

```
doc.select("div[class=abcd] > span[id=뽕]")
```

이러면 div 안의 id가 뽕인 span 태그를 뽑아올 수 있습니다. 여기서 또 문제점은, 그래도 안에 2개를 다 가져온단 말이죠? 우리는 첫 번째것만 가져오기 위해 get을 사용해볼겁니다.

```
doc.select("div[class=abcd] > span[id=뽕]").get(0)
```

조습니다. 이제 text로 태그를 전멸시키면 끝입니다!

```
doc.select("div[class=abcd] > span[id=뽕]").get(0).text()
```

완벽합니다. 그럼 우리는 이제 와! 샌즈! 를 가져올 수 있게 되었습니다.

추가로 팁을 하나 더 드리자면, class와 id는 굳이 저렇게 대괄호를 쓰지 않아도 됩니다.

```
doc.select("div.abcd > span#뽕").get(0).text()
```

```
doc.select( div.abcd > span#책 ).get(0).text()
```

눈치 채셨나요? class는 점(.) 으로, id는 샵(#) 으로 나타낼 수 있습니다.
참고로 속성값에 띄어쓰기가 있을땐, 점으로 채워주시면 됩니다.

이제 기본적인건 다 배웠으니, 본격적으로 파싱을 해 봅시다. 이번에도 실검입니다!
저번시간에 썼던 실검 사이트를 불러와서 HTML을 가져와봅시다.

```
var doc = org.jsoup.Jsoup.connect("https://m.search.naver.com/search.naver?where=m&sm=mtb_jum&query=%EC%8B%A4%EC%8B%9C%EA%B0%84%EA%B2%80%EC%83%89%EC%96%B4").get()
```

저번에도 변수 부분만 썼듯이 이번에도 변수 부분인 doc만 쓰겠습니다.

SKT  ...

  50%  오후 1:22

← 전체보기



```
<li role="presentation" class="item"> <a href="#"
onclick="tCR('a=crk_old.lefttab&r=&i=&m=0&am
p;u=javascript'); return false;" role="tab" class="_tab tab
selected" aria-selected="true"> <span
class="tab_inner">1~10위</span> </a> </li>
<li role="presentation" class="item"> <a href="#"
onclick="tCR('a=crk_old.righttab&r=&i=&m=0&a
mp;u=javascript'); return false;" role="tab" class="_tab tab">
<span class="tab_inner">11~20위</span> </a> </li>
</ul>
</div>
<div class="flick-wrap _flicking">
<div>
<div class="group_kwd">
<ol class="lst_kwd">
<li>
<div class="kwd_bx">
<a
href="?wh...y=%ED%8
C%94%EB...rk%3A1"
class="_lir...
'a=crk_old.list&r=&i=&u='+urlencode(urlexpan
d(this.href)))"><em class="num">1</em><span class="tit
_keyword">팔라우</span></a>
<a href="https://m.datalab.naver.com/
realtimeDetail.naver?query=%ED%8C%94%EB%9D%BC%EC%9
A%B0&datetime=2019-07-06T13%3A21%3A00&age=
all&period=now&where=search" class="_hlink
datalab" onclick="return goOtherCR(this,
'a=crk_old.numv&r=&i=&u='+urlencode(urlexpan
d(this.href)))"><span class="ico_graph spnew">검색추이</
```

```

span></a>
</div></li>
<li>
<div class="kwd_bx">
<a
href="?where=m&amp;sm=mtb_crk.allnow&amp;query=%EB%8
D%94%EC%BD%9C2&amp;x_nxpr-front=crk%3A2" class="_link
keyword" onclick="return goOtherCR(this,
'a=crk_old.list&amp;r=&amp;i=&amp;u='+urlencode(urlexpand(t
his href)));"><em class="num">?</em><span class="tit

```

보니까 태그 사이에 실검이 있네요? 그러면 select로 뽑아옵시다.

```
doc = doc.select("span.tit_keyword")
```

속성값에 점이 있는 이유는 띄어쓰기가 있어서 입니다.

```

[봇] <span class="tit_keyword">팔라우</span>
<span class="tit_keyword">더콜2</span>
<span class="tit_keyword">강식당3</span>
<span class="tit_keyword">la 지진</span>
<span class="tit_keyword">미국 지진</span>
<span class="tit_keyword">지진</span>
<span class="tit_keyword">이열음</span>
<span class="tit_keyword">히든 피겨스</span>
<span class="tit_keyword">메이비</span>
<span class="tit_keyword">평일 오후 세시의 연인</span>
span>
<span class="tit_keyword">허규</span>
<span class="tit_keyword">신동미</span>
<span class="tit_keyword">잠원동 붕괴</span>
<span class="tit_keyword">현아</span>
<span class="tit_keyword">허재</span>
<span class="tit_keyword">윤상현</span>
<span class="tit_keyword">일본 불매운동</span>
<span class="tit_keyword">현아 입술</span>
<span class="tit_keyword">신촌 물총축제</span>
<span class="tit_keyword">더콜</span>

```

조습니다. 이제 깔끔하게 번호만 붙여봅시다. for문을 이용하도록 하죠

```

var keyWord = [];
for (a = 0; a < doc.size(); a++) {
keyWord.push(a + 1 + ". " + doc.get(a).text());
}
replier.reply(keyWord.join("\n"));
size()는 몇 개가 select되었는지 가져옵니다. 배열의 length랑 비슷한 느낌이에요

```

- [봇] 1. 더콜2
- 2. 팔라우
- 3. 강식당3
- 4. 미국 지진
- 5. la 지진
- 6. 지진
- 7. 히든 피겨스
- 8. 허규
- 9. 메이비
- 10. 신동미
- 11. 이열음
- 12. 평일 오후 세시의 연인
- 13. 허재
- 14. 현아
- 15. 잠원동 붕괴
- 16. 더콜
- 17. 윤상현
- 18. 현아 입술
- 19. 신촌 물총축제
- 20. 일본 불매운동

완벽하네요! 최종 소스는 이렇게 됩니다.

```
var doc = org.jsoup.Jsoup.connect("https://m.search.naver.com/search.naver?where=m&sm=mtb_jum&query=%EC%8B%A4%EC%8B%9C%EA%B0%84%EA%B2%80%EC%83%89%EC%96%B4").get();
doc = doc.select("span.tit_keyword");
var keyWord = [];
for (a = 0; a < doc.size(); a++) {
    keyWord.push(a + 1 + ". " + doc.get(a).text());
}
replier.reply(keyWord.join("\n"));
```

어떤가요? 기본 파싱법은 복잡했지만, Jsoup을 배우고 나니 간편하지 않나요?

뿐만 아니라, Jsoup은 서버에 접속할 때 여러가지 설정을 해 줄 수 있습니다! 이걸 get() 이나 post() 전에서 쓸 수 있는데, 대표적으로 header와 data가 있습니다.

header는 Content Type을 정하거나 기타 등등 여러가지로 쓰입니다.

```
org.jsoup.Jsoup.connect(url).header("Content-Type","application/json") //Content Type을 application/json 으로 바꿈
```

```
org.jsoup.Jsoup.connect(url).header("Authorization","Bearer "+key) // 키 입력
```

```
org.jsoup.Jsoup.connect(url).header("Accpet","*.*)
```

이런식으로 써먹을 수 있습니다. 참고로 ignoreContentType으로 Content Type을 무시해버릴 수 있습니다.

```
org.jsoup.Jsoup.connect(url).ignoreContentType(true)
```

이렇게 써먹을 수 있습니다.

data는 웹사이트에 data를 보낼때 사용합니다.

```
org.jsoup.Jsoup.connect(url).data("email",email).data("password",password) // email과 password를 이용한 로그인
```

```
org.jsoup.Jsoup.connect("https://m.search.naver.com/search.naver?").data("query",검색어).get() // data를 이용한 검색
```

그리고 userAgent를 요구하는 웹사이트의 경우는,

```
org.jsoup.Jsoup.connect(url).userAgent("Mozilla")
```

이렇게 하면 됩니다.

강좌 끝

header부터는 제가 알고있는 정보가 많이 없어서 부족한 부분이 있을 수 있는데 혹시 그런 부분이 있으면 지적 부탁드립니다.

시험기간부터 쓰기 시작해서 이제 올렸네요!

다음에는 Math 객체를 다뤄볼 예정입니다.

그럼 이만



마른얼음 BOT님의 게시글 더보기 >

❤️ 좋아요 13 💬 댓글 21

🔗 공유 | 신고

댓글 등록순 최신순 C

댓글알림 ☐



doami2005

오..Jsoup

2019.07.06. 17:20 답글쓰기



마른얼음 BOT 작성자

제이수프!

2019.07.08. 21:05 답글쓰기



사로로

정확히는 몰랐는데 흥미

2019.07.06. 18:06 답글쓰기



마른얼음 BOT 작성자

앗 도움이 되셨다면 감사합니다

2019.07.08. 21:06 답글쓰기



성빈

봐도 모르겠다

2019.07.06. 18:23 답글쓰기



마른얼음 BOT 작성자

웹 역시 너무 어렵게 설명했군요

2019.07.08. 21:05 답글쓰기

삭제된 댓글입니다.



마른얼음 BOT 작성자

허형 역시 너무 어렵게 설명했네요

2019.07.08. 21:06 답글쓰기



마른얼음 BOT 작성자

냥냥이 아아.... 실검이 제일 무난할 것 같았는데.. 조만간 파싱 예제를 하나 더 올려봐야겠네요

2019.07.08. 21:08 답글쓰기



마른얼음 BOT 작성자

냥냥이 DLC가 무엇이죠

2019.07.08. 21:11 답글쓰기



마른얼음 BOT 작성자

냥냥이 얼 찾아보니 그런뜻이네용

추가 실검예제는 노래가사, IP뜯기(?), 영화정보(또는 게임정보) 생각하고 있습니다.

2019.07.08. 21:13 답글쓰기



AlphaDo

왜 저는 이제야 size() 를 안...(멍청)

2019.07.07. 02:11 답글쓰기



마른얼음 BOT 작성자

저도 처음에 length로 해버린..

2019.07.08. 21:06 답글쓰기



AlphaDo

마른얼음 BOT 크흠 ㅋㅋ

2019.07.08. 21:27 답글쓰기

삭제된 댓글입니다.



마른얼음 BOT 작성자

태그삭제식 파싱부터 차근차근 배워나가는게..

2019.07.08. 21:07 답글쓰기



Semicolon

오오...!

2019.07.07. 17:17 답글쓰기



마른얼음 BOT 작성자

헤헬

2019.07.08. 21:07 답글쓰기

삭제된 댓글입니다.



마른얼음 BOT 작성자

이미 충분히 정리되어 있는데요?

2019.07.12. 15:56 답글쓰기



대기만성

감사합니다! 스푸가 맛있다는걸 처음 알았습니다.

2019.10.26. 23:35 답글쓰기



웹빅탄

한 div에 p가 여러개 있을때 그 p만 빼낼려면 어떻게 해야해요?

2019.11.28. 15:38 답글쓰기



마른얼음 BOT 작성자

일단 모든 p를 뽑아서 원하는 div가 어디 있는지 인덱스를 구한 뒤, get(num) 이용해서 빼내면 됩니다

2019.11.28. 22:16 답글쓰기



choi0108

지금 봐도 좋은 강의네요!

감사합니다

2020.03.17. 01:55 답글쓰기

Hibot

댓글을 남겨보세요



등록

글쓰기

답글

목록

▲ TOP

'| 카카오톡 봇 강좌 |' 게시판 글

전체 [중급] 말머리 글

이 게시판 새글 구독하기 ☐

- | | | |
|--|----------|-------------|
| [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. |
| [기타] [PC] Jsoup 파싱 팁 ☹️ [21] | 인디벨 | 2019.06.25. |
| [초급] 배열의 기초지식???? [8] | 사로로 | 2019.06.19. |

1 2 3

전체보기

이 카페 인기글

복붙

하하하하

♡0 💬5



틱택토 (Player vs Player)

도루묵 2005

[봇드게임] 게이트 v0.9.0

♡0 💬5

ZUMP

♡0 💬4



태양, 달 정보 구현 완료



혹시 여기서 변수를 저장하는 부분이 어딘지
아시는분 계신가요??
간편 자동응답과 한글코딩을 메인으로 제공
하는 카톡봇앱 개발 예정

성빈
♡1 💬14

eval 질문

Kiri

♡0 💬5

카카오톡 단순 자동응답봇 파일

도루묵

♡0 💬2



틱택토 (Player vs Bot)

1 2 3 4