

| 카카오톡 봇 강좌 | >

[응용] Jsoup & map 강좌



도미 doami2005 챗봇 창시자 🌞 1:1 채팅

2020.07.06. 18:02 조회 259

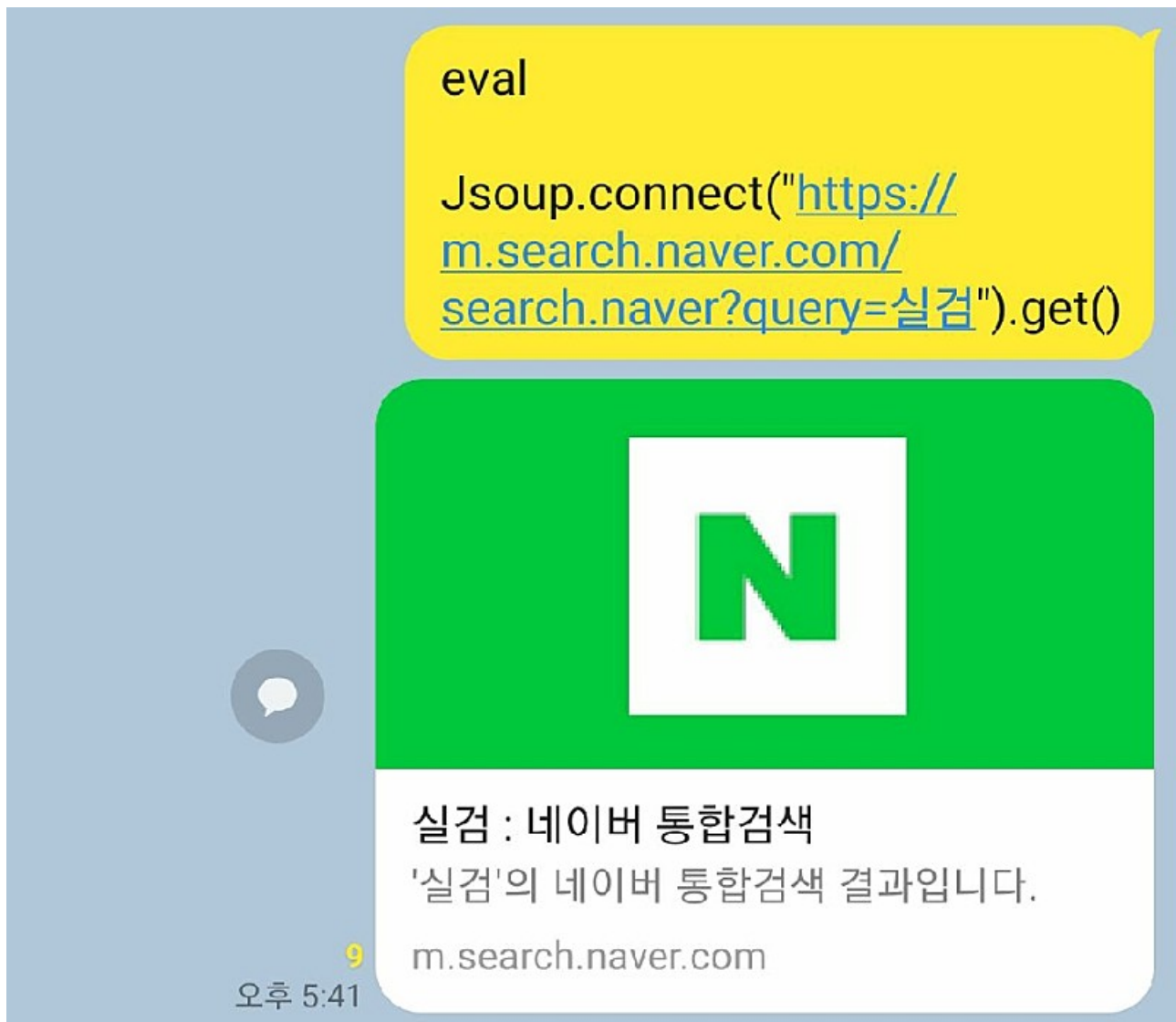
💬 댓글 11 URL 복사 ⋮

전에 강좌 신청 글 올렸을 때
map 강좌, 파싱 강좌 해달라는 분들이 있길래 둘이 같이 합니다.

한번 간단한 예제로
네이버 실시간 검색어를 파싱해 봅시다.

<https://m.search.naver.com/search.naver?query=실검>

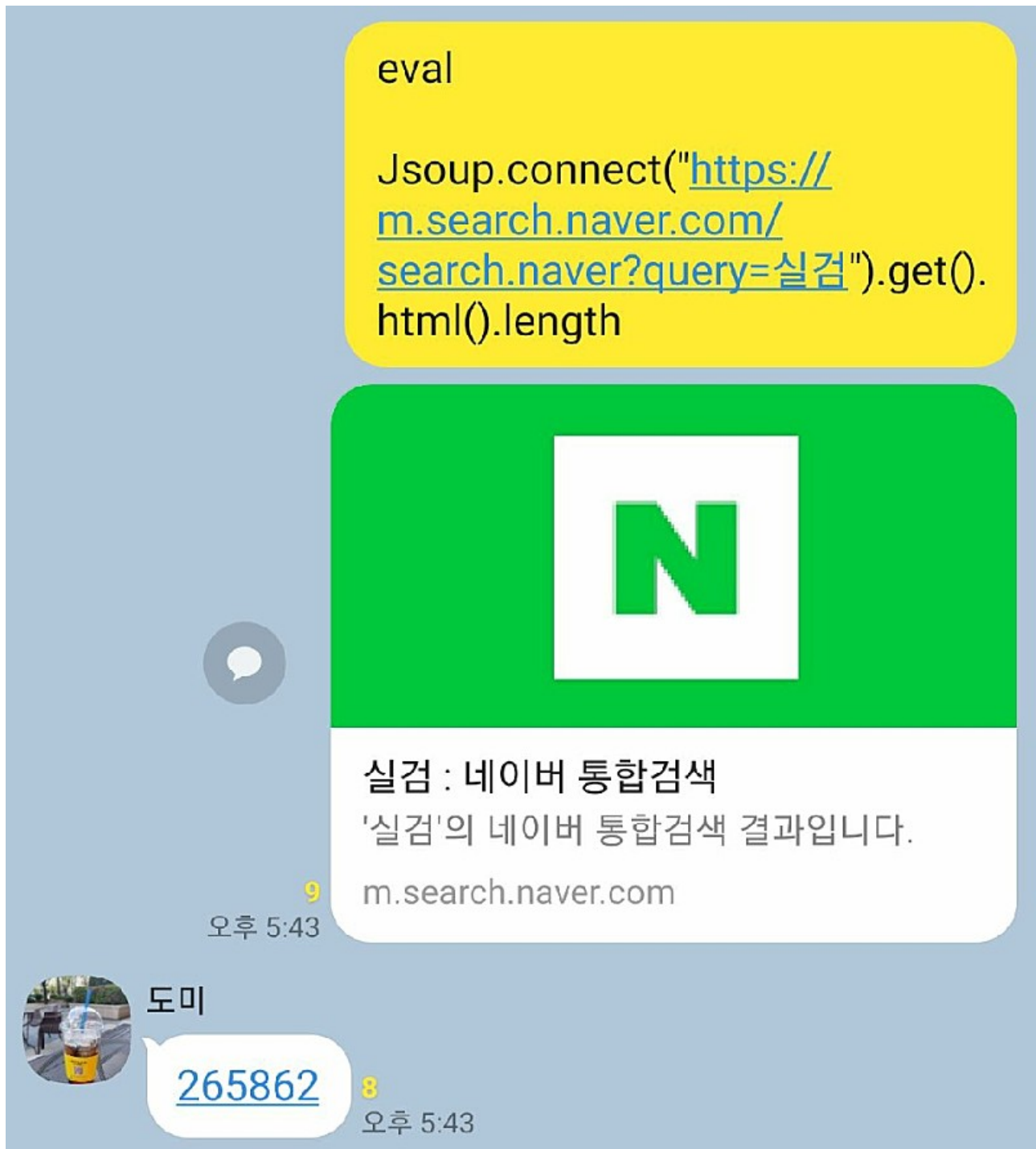
네이버에 실검 을 검색한 링크입니다.
한번 html 을 훑어보죠.



어째선지 결과가 나오지 않습니다.

그 이유는 카카오톡에선 약 26만자 이상의 단어는 보내지지 않기 때문입니다.

따라서 저 html 을 문자열로 만든 후 길이를 가져와 봅시다.

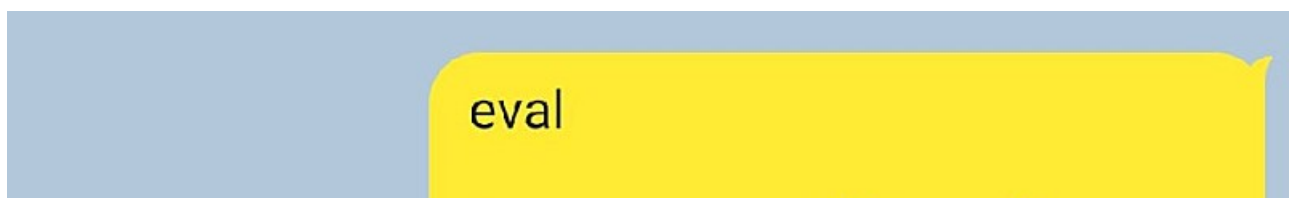


TMI) html() 붙이면 문자열 됩니다.

26 만자가 넘네요.

따라서 글자를 좀 잘라서 봐야 할 것 같습니다.

그런데 앞에서부터 찢르면 실시간 검색어가 잘릴 수도 있어서 뒤에서부터 찢르겠습니다.



```
Jsoup.connect("https://  
m.search.naver.com/  
search.naver?query=실검").get().  
html().slice(0, 150000)
```



실검 : 네이버 통합검색
'실검'의 네이버 통합검색 결과입니다.
m.search.naver.com

9
오후 5:45



도미

```
<!doctype html>  
<html lang="ko" class="">  
<head>  
  <meta http-equiv="Content-  
Type" content="text/html;  
charset=UTF-8">  
  <meta http-equiv="Content-  
Script-Type" content="text/  
javascript">  
  <meta name="referrer"  
content="always">  
  <meta name="viewport"  
content="width=device-width,  
initial-scale=1.0, maximum-  
scale=2.0, minimum-scale=1.0,
```

```
user-scalable=yes">
  <meta name="format-
detection"
content="telephone=no,address
=no,email=no">
  <meta property="og:title"
content="실검 : 네이버 통합검색">
  ...
```

전체보기



8

오후 5:45

문자열 slice 메서드로 앞에서부터 15 만자만 떼왔습니다.

← 전체보기



```
A0+%EC%9E%A5%EC%9C%A4%EC%A0%95"
class="_link keyword" onclick="return
goOtherCR(this,
'a=crk_old.list&r=&i=&u='+urlencod
e(urlexpand(this.href));"> <em class="num">1</
em> <span class="tit _keyword">트라이애슬론
장윤정</span> </a>
</div> </li>
<li>
<div class="kwd_bx"> <a
href="?where=m&sm=mtb_crk.agallgrpnam
si0en0sp0&query=%EC%86%90%EC%A0%95
%EC%9A%B0" class="_link keyword"
onclick="return goOtherCR(this,
'a=crk_old.list&r=&i=&u='+urlencod
e(urlexpand(this.href));"> <em class="num">2</
em> <span class="tit _keyword">손정우</span> </
a>
```



```

</div> </li>
<li>
  <div class="kwd_bx"> <a
href="?where=m&sm=mtb_crk.agallgrpnam
si0en0sp0&query=%ED%9D%91%EC%82%AC
%EB%B3%91" class="_link keyword"
onclick="return goOtherCR(this,
'a=crk_old.list&r=&i=&u='+urlencod
e(urlexpand(this.href)));"> <em class="num">3</
em> <span class="tit _keyword">흑사병</span> </
a>

```

```

</div> </li>

```

```

<li>

```

```

  <div class="kwd_bx"> <a
href="?where=m&sm=mtb_crk.agallgrpnam
si0en0sp0&query=%EC%9E%A5%EC%9C%A4
%EC%A0%95+%EC%84%A0%EC%88%98"
class="_link keyword" onclick="return
goOtherCR(this,
'a=crk_old.list&r=&i=&u='+urlencod

```

한참 내려다보니 실검 부분이 있네요.

여기서 중요합니다.

저기 써있는 '트라이애슬론', '손정우' 등이 키워드입니다.

그리고 그 앞에 동일한 태그가 존재합니다.

```

  <div class="kwd_bx"> <a
href="?where=m&sm=mtb_crk.agallgrpnam
si0en0sp0&query=%ED%8A%B8%EB%9D%BC
%EC%9D%B4%EC%95%A0%EC%8A%AC%EB%A1%
A0+%EC%9E%A5%EC%9C%A4%EC%A0%95"
class="_link keyword" onclick="return
goOtherCR(this,
'a=crk_old.list&r=&i=&u='+urlencod

```

```

e(urlencode(this.href));"> <em class="num">1</
em> <span class="tit_keyword">트라이애슬론
장원</span> </a>
</div> </li>
<li>
<div class="kwd_bx"> <a
href="?where=m&amp;sm=mtb_crk.agallgrpnam
si0en0sp0&amp;query=%EC%86%90%EC%A0%95
%EC%9A%B0" class="_link keyword"
onclick="return goOtherCR(this,
'a=crk_old.list&amp;r=&amp;i=&amp;u='+urlencod
e(urlencode(this.href));"> <em class="num">2</
em> <span class="tit_keyword">문정우</span> </
a>
</div> </li>
<li>
<div class="kwd_bx"> <a
href="?where=m&amp;sm=mtb_crk.agallgrpnam
si0en0sp0&amp;query=%ED%9D%91%EC%82%AC
%EB%B3%91" class="_link keyword"
onclick="return goOtherCR(this,
'a=crk_old.list&amp;r=&amp;i=&amp;u='+urlencod
e(urlencode(this.href));"> <em class="num">3</
em> <span class="tit_keyword">득남병</span> </
a>

```

 키워드

이런 형식이네요.

Jsoup select 메서드를 이용해 이것들을 꺼내와봅시다.

```
eval
```

```
Jsoup.connect("https://
m.search.naver.com/
```

```
search.naver?query=실검").get().  
select("span[class=tit  
_keyword]")
```



실검 : 네이버 통합검색
'실검'의 네이버 통합검색 결과입니다.
m.search.naver.com

8
오후 5:50



도미

```
<span class="tit  
_keyword">트라이애슬론 장윤정</  
span>  
<span class="tit  
_keyword">손정우</span>  
<span class="tit  
_keyword">흑사병</span>  
<span class="tit  
_keyword">함영준</span>  
<span class="tit  
_keyword">장윤정 선수</span>  
<span class="tit  
_keyword">메가로돈</span>  
<span class="tit  
_keyword">정용진</span>  
<span class="tit
```



```
_keyword">장윤정</span>
<span class="tit
_keyword">노현정</span>
<span class="tit
_keyword">안성훈</span>
<span class="tit _keyword">중국
흑사병</span>
<span class="tit
_keyword">임오경</span>
<span class="tit
_keyword">모범...
```

전체보기



8

오후 5:50

간단하죠?

이제 저기서 원소를 뜯어올 땐 get 을 써주면 됩니다.

eval

```
Jsoup.connect("https://
m.search.naver.com/
search.naver?query=실검").get().
select("span[class=tit
_keyword]").get(0)
```



실검 : 네이버 통합검색

'실검'의 네이버 통합검색 결과입니다.
m.search.naver.com

9
오후 5:51



도미

```
<span class="tit  
_keyword">트라이애슬론 장윤정</  
span>
```

8
오후 5:51

배열과 원리가 같습니다.
순서가 0부터 시작하죠.

이제 저 태그를 없애야 하니 text() 로 제거해줍니다.

eval

```
Jsoup.connect("https://  
m.search.naver.com/  
search.naver?query=실검").get().  
select("span[class=tit  
_keyword]").get(0).text()
```



실검 : 네이버 통합검색
'실검'의 네이버 통합검색 결과입니다.
m.search.naver.com

9
오후 5:52



도미



잘 나오네요.

이제 저걸 20번 반복해줍니다.

```
Jsoup.connect("https://m.search.naver.com/  
search.naver?query=실검").get().select("span[class=tit_keyword]").get(0).text()  
+  
Jsoup.connect("https://m.search.naver.com/  
search.naver?query=실검").get().select("span[class=tit_keyword]").get(1).text()  
+  
Jsoup.connect("https://m.search.naver.com/  
search.naver?query=실검").get().select("span[class=tit_keyword]").get(2).text()  
+  
Jsoup.connect("https://m.search.naver.com/  
search.naver?query=실검").get().select("span[class=tit_keyword]").get(3).text()  
+  
Jsoup.connect("https://m.search.naver.com/  
search.naver?query=실검").get().select("span[class=tit_keyword]").get(4).text()  
+  
Jsoup.connect("https://m.search.naver.com/  
search.naver?query=실검").get().select("span[class=tit_keyword]").get(5).text()  
+  
Jsoup.connect("https://m.search.naver.com/  
search.naver?query=실검").get().select("span[class=tit_keyword]").get(6).text()  
+
```

```
Jsoup.connect("https://m.search.naver.com/  
search.naver?query=실검").get().select("span[clas  
s=tit_keyword]").get(7).text()
```

+

하란다고 이대로 하시는 분은 없겠죠?
반복문을 사용해야 합니다.

그래서 사용하는 게 map!

```
Jsoup.connect("https://m.search.naver.com/search.naver?query=실검")  
.get()  
.select("span[class=tit_keyword]")
```

여기서 반환되는 elements 객체에는 toArray() 라는 메서드를 사용할 수 있습니다.

```
eval  
Jsoup.connect("https://  
m.search.naver.com/  
search.naver?query=실검")  
.get()  
.select("span[class=tit  
_keyword]").toArray()
```



실검 : 네이버 통합검색
'실검'의 네이버 통합검색 결과입니다.
m.search.naver.com

오후 5:58



도미



[Ljava.lang.Object;@e2058b9

8

오후 5:58

결과가 저렇게 나오는 이유는 자바 배열이기 때문이지만,
map 과 join 을 쓸 수 있습니다.

```
eval  
Jsoup.connect("https://  
m.search.naver.com/  
search.naver?query=실검")  
.get()  
.select("span[class=tit  
_keyword]").toArray()  
.map(e => e.text())
```



실검 : 네이버 통합검색

'실검'의 네이버 통합검색 결과입니다.

m.search.naver.com

9

오후 5:58



도미

트라이애슬론

장윤정,손정우,흑사병,정용진,함영준,
장윤정 선수,메가로돈,엔니오
모리꼬네,노현정,안성훈,중국
흑사병,임오경,모범형사,박상면,서예
지

허리,안희정,페스트,강영수,파맙신,라 그나로크 오리진

8

오후 5:58

그냥 간단히 `map(e => e.text())`
를 해주었습니다.

모든 원소들에 `text()` 를 한겁니다.
실검 배열이 나왔죠?

하지만 등수도 표기해야 해야 하니
`map` 메서드의 두번째 인자로 제공되는 인덱스를 사용해 보겠습니다.

```
eval  
Jsoup.connect("https://  
m.search.naver.com/  
search.naver?query=실검")  
.get()  
.select("span[class=tit  
_keyword]").toArray()  
.map((e, i) => (i+1)+"위 "+e.text())
```

A green square with a white 'N' inside, representing the Naver logo.

실검 : 네이버 통합검색
'실검'의 네이버 통합검색 결과입니다.
m.search.naver.com

9

오후 6:00



도미

1위 트라이애슬론 장윤정,2위

손정우,3위 흑사병,4위 정용진,5위
함영준,6위 장윤정 선수,7위
메가로돈,8위 엔니오 모리꼬네,9위
안성훈,10위 노현정,11위 중국
흑사병,12위 임오경,13위
모범형사,14위 김택진,15위 서예지
허리,16위 박상면,17위 안희정,18위
신도기연,19위 강영수,20위 파맙신

8

오후 6:00

이제 join 만 치면 되네요.

```
eval  
Jsoup.connect("https://  
m.search.naver.com/  
search.naver?query=실검")  
.get()  
.select("span[class=tit  
_keyword]")  
.toArray()  
.map((e, i) => (i+1)+"위 "+e.text())  
.join("\n")
```

A green square with a white 'N' in the center, representing the Naver logo.

실검 : 네이버 통합검색
'실검'의 네이버 통합검색 결과입니다.

9

m.search.naver.com

오후 6:01



도미

- 1위 트라이애슬론 장윤정
- 2위 손정우
- 3위 흑사병
- 4위 정용진
- 5위 함영준
- 6위 장윤정 선수
- 7위 메가로돈
- 8위 엔니오 모리꼬네
- 9위 안성훈
- 10위 노현정
- 11위 중국 흑사병
- 12위 임오경
- 13위 모범형사
- 14위 김택진
- 15위 서예지 허리
- 16위 박상면
- 17위 안희정
- 18위 신도기연
- 19위 강영수
- 20위 파맴신

8

오후 6:01

깔끔하쥬 ?

팁을 드리자면 (i+1) 는 ++i 로 줄여쓰실 수도 있습니다.

이번 강좌엔 jsoup 조금, 그리고 map 응용이 들어갔네요

댓글 피드백 환영합니다



도미 doami2005님의 게시글 더보기 >

❤ 좋아요 11 💬 댓글 11

🔗 공유 | 신고



도미 doami2005 작성자

하트 눌러주세요
아니 눌러요

2020.07.06. 18:05 답글쓰기



SP청정

파싱 드디어

2020.07.06. 18:16 답글쓰기



hyun123

하트 드리겠습니다



2020.07.06. 18:17 답글쓰기



도미 doami2005 작성자

아주 칭찬합니다 께께

2020.07.06. 18:17 답글쓰기



Steve28

근데 태그가 없는 text 는 파싱을 못하나요?
어쩔수 없이 텍스트가 있는 곳을 다 뜯어내서 잘라써야 하는건가용

2020.07.06. 19:38 답글쓰기



핀밀크

Jsoup.parse(텍스트)하면 org.jsoup.nodes.Document(맞나) 형이 됩니다

2020.07.06. 19:45 답글쓰기



Steve28

핀밀크 죄송하지만 혹시 .select(tag) 까지 한 부분 에서 마지막 요소만 빼내오려면 어떻게 해야하나요?

2020.07.06. 20:37 답글쓰기



핀밀크

Steve28 .last()

2020.07.06. 20:45 답글쓰기



핀밀크

org.jsoup.nodes.Document(맞나)는 java.lang.ArrayList를 상속했고 .toArray()를 하면 java.lang.Array가 되는 걸로 알고 있음
니당

2020.07.06. 19:44 답글쓰기



승준

toArray 알아갑니다 ☺

2020.07.06. 22:39 답글쓰기



NotreDame

와 정말 어렵네요 ㄸ

2020.07.07. 16:16 답글쓰기

Hibot

댓글을 남겨보세요



등록

'| 카카오톡 봇 강좌 '| 게시판 글

전체 [응용] 말머리 글

이 게시판 새글 구독하기 ☐

[응용] 카톡봇 반복문으로 자연 로그 값 구해보기 🤖 [2]	Steve28	2020.07.07.
[중급] 함수지향과 객체지향(11) - prototype은 무엇인가	choi0108	2020.07.07.
[응용] Jsoup & map 강좌 🤖 [11]	도미 doami2005	2020.07.06.
[중급] 함수지향과 객체지향(10) - 생성자와 new	choi0108	2020.07.06.
[중급] [카카오링크 강좌] 5. 도미님 포켓몬 검색 사용방법 🤖 [10]	그린스크린	2020.07.04.

1 2 3

전체보기

이 카페 인기글

getSchoolFood

도미 doami2005

♡0 💬6

안녕하십니까.

JSR

♡0 💬8



(수정5)카카오톡 오픈채팅 공식 봇관련 정

자동읽음소스

tomohong

♡0 💬9



메신저봇 글씨 색

카카오링크 소스 공유 및 다운로드금지

♡0 💬6

지민

♡3 💬8

[봇드게임] 세븐 다이스 v1.1.0

ZUMP

♡0 💬2



카카오톡이 이런걸 추천하다니

바른 누르면 intent가 함수를 실행하게하려
면

ACTIVE

♡2 💬5

jomin398

♡0 💬18

1 2 3 4 5