^ 이전글

| 카카오톡 봇 강좌 | >

[초급] 파싱 시 Utils 와 Jsoup 의 장단점



doami2005 챗봇 창시자 🌞 1:1 채팅

2019.08.08. 04:31 조회 591

제가 파싱할 때 주로 html에서 태그와 띄어쓰기를

 $replace(/(<([^>]+)>)/g, "").replace(/[\lambda n\lambda s]{2,}/g,"\lambda n")$

이 방법으로 싹 소멸시켜버리는데

결과 자체는 되게 깔끔하여 보기 쉬우며 split 을 쓰기 쉽다는 이점이 있습니다.

하지만, 특정 사이트들은 태그에 필요한 정보가 있어서 태그를 제거해버리면 정보가 사라집니다.

(태그란? 웹사이트 html에 <와 > 사이에 써있는 것예) <div id=~~> <div class~~> 등등)

그래서 파싱할 때 태그 제거는 미리 제거 한 뒤 결괏값을 보고,

필요한 정보가 쓰여있지 않다면 태그를 보면서 하시는 것을 추천드립니다.

태그 제거 전 -

SKT

▲ 🦠 세 13% 🗎 오전 4:29





muuen- mase -

<h3 class="blind">급상승 검색어 검색어</h3>

<div class="ah_roll_area

PM_CL_realtimeKeyword_rolling">

ul class="ah_l">

class="ah_item"> <a href="#" class="ah_a"
data-clk="lve.keyword"> 1

span> 위메프88특가

class="ah_item"> <a href="#" class="ah_a"
data-clk="lve.keyword"> 2</
span> 백종원 골목식당

<a href="#" class="ah_a"
data-clk="lve.keyword"> 3</
span> 김민주

class="ah_item"> <a href="#" class="ah_a"
data-clk="lve.keyword"> 4</
span> 입추

class="ah_item"> <a href="#" class="ah_a"
data-clk="lve.keyword"> 5</
span> 골목식당

class="ah_item"> <a href="#" class="ah_a"
data-clk="lve.keyword"> 6</
span> 로켓펀치

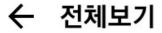
<a href="#" class="ah_a"
data-clk="lve.keyword"> 7</
span> 송병준

ali alacc="ah itam"> <a hraf="#" alacc="ah a"</pre>

태그 제거 후 -

SKT 🔤

▲ 🦠 세 13% 🗎 오전 4:29





규드 증권(금융) 부동산 지도 영화 뮤직

겍 만화 / 웬툰 더보기 급상승 검색어 검색어 1 위메프88특가 2 백종원 골목식당 3 김민주 4 입추 5골목식당 6 송병준 7로켓펀치 8 기생충 해석 9 케빈 나 10 기문둔갑 11 한끼줍쇼 12 이진혁 13 워너원 14 김희철 모모 15 포방터 돈까스 16 이대 백반집 17 꽁지 18 봉오동 전투 19 아드레날린24 20 주옥순 급상승 검색어 DataLab.급상승 트래킹 1 - 1 1 0 1

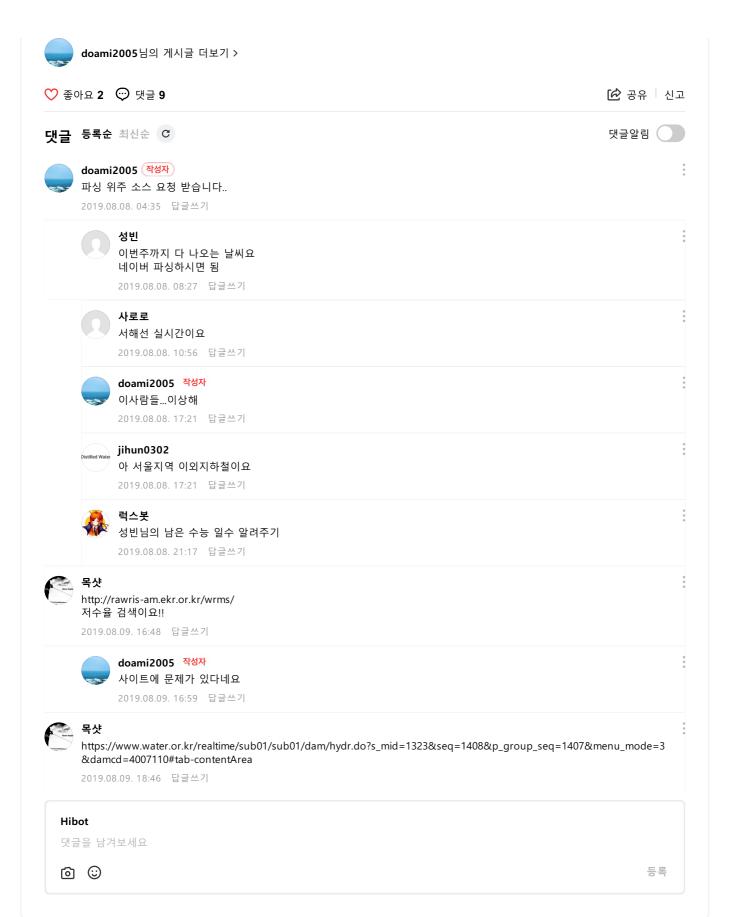
네이버는 필요한 정보가 태그에 저장되어 있지 않아 태그 제거 시에 보기가 매우 깔끔해집니다

대표적인 정규식 말고도 필요에 따라 본인이 개발해서 쓰는것이 중요합니다

/* 도미의 강좌 시리즈 */

[Tip]

- *eval 의 위험성과 사용법*
- https://cafe.naver.com/nameyee/4341
- *파싱 시 Utils 와 Jsoup 의 장단점*
- https://cafe.naver.com/nameyee/4342

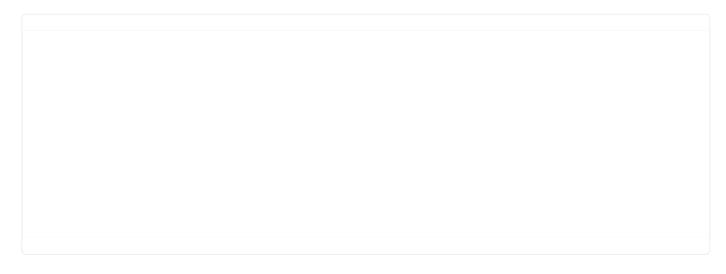


🖍 글쓰기

답글

모로

▲ TOP



'| 카카오톡 봇 강좌 |' 게시판 글

전체 [초급] 말머리 글		이 게시판 새글	이 게시판 새글 구독하기	
 옾챗뚫으시려는 분들을 위한 자그마한 팁 [28]		아이곰	2019.08.15.	
[초급] 오류를 잡아주는 try문 사용법 [7]		doami2005	2019.08.15.	
[초급] 파싱 시 Utils 와 Jsoup 의 장단점 ❷ [9]		doami2005	2019.08.08.	
[초급] eval의 위험성/사용법 🔐 [11]		doami2005	2019.08.08.	
[초급] 봇 기초 강좌 1-1 <mark>[3]</mark>		럭스봇	2019.08.01.	
	1 2 3		전체보기	

이 카페 인기글

반가워요.

천방지축하연 ♡0 ⊙4



전에 말했던 마피아 계열 게임 소스 올렸습니다. 카카오톡 단순 자동응답봇 파일 ZUMP 도록움 ①1 ♡0 ②2



시간별/날짜별 날씨 구현

요층 대라는 새로운 홍보 수법 중 하나가 Dark Tornado ♡0 ⊙7

중수가 초보에게 하는 강좌 [2] 하하하하 ♡1 ⓒ11

1 2 3 4

가르치기 리로드..

Milk2 ♥0 ⊕9

[치***, 한**] 님 강제 탈퇴

AlphaDo ♥1 ⊕2

JPG 개발 일지 #7

재승 ♡0 ♀4