

Python Crawling

발표자 : 김동휘

Tool

- Only Chrome(Developer Tools)

http://sports.news.naver.com/wfootball/news/read.nhn?oid=139&aid=0002090884&m_view=1&sort=LIKE

Forthrlux

파리에서 안풀겠다는데 원

35분 전 | 신고

답글 4

83 9

인터넷

사실 네이마르가 기자들 돈주고 퍼뜨림

39분 전 | 신고

답글

80 11

베르나데스키

딱봐도 찌라시 언론이네 ㅋㅋㅋㅋ

41분 전 | 신고

답글 8

54 21

oo

파리에서 떠야지 부합시나 RMC 마르카 아스원소용이나 원래저랬음

40분 전 | 신고

답글

39 16

어떠

다수언론이 짜고 친건가..

32분 전 | 신고

답글

36 6

dvd1

마르카 아스를 믿음? 레알언론 ㅋㅋㅋ

```
<!DOCTYPE html>
<html lang="ko">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta http-equiv="X-UA-Compatible" content="requiresActiveX=true">
    <meta name="viewport" content="width=1080px,maximum-scale=2.0,minimum-scale=0.4,user-scalable=yes">
    <meta property="me:feed:serviceId" content="sports"/>
    <meta property="me:feed:feedGroupId" content="sports"/>
    <meta property="me:feed:feedId" content="" />
    <meta property="me:feed:feedName" content="" />
    <meta property="me:feed:feedGroupName" content="" />
    <meta property="me:feed:url" content="" />
    <meta property="og:update_time" content="20180509084810"/>
    <meta property="article:contentLocation:name" content="" />
    <meta property="article:contentLocation:url" content="" />
    <meta property="og:type" content="article"/>
    <meta property="og:image" content="http://imgnews.naver.net/image/139/2018/05/09/0002090884_001_20180509084810436.jpg"/>
    <meta property="og:title" content="네이마르, 다음 시즌 레알에 합류(다수 매체)"/>
    <meta property="og:description" content="【스포탈코리아】 이현민 기자= 네이마르(26, 파리 생제르맹)의 레알 마드리드행이 가시화되고 있다. 네이마르는 이번 시즌 내내 이적설이 돌았다. 파리
    할에 불만을 품고 있으며 팀 내 분위기를 흐리는 등 잇단 구설에 올랐다. 영국 '익스프레스'는 9일 스페인 'OR 디아리오' 에두아르도 안다 기자의 말을 빌어 "네이마르는..." />
    <meta property="og:url" content="http://sports.news.naver.com/wfootball/news/read.nhn?oid=139&aid=0002090884"/>
    <meta property="og:article:author" content="네이버 스포츠 | 스포탈코리아"/>
    <meta property="og:article:author:url" content="http://sports.news.naver.com"/>
    <meta property="og:article:thumbnailUrl" content="https://imgnews.pstatic.net/image/sports/common/naverme/image-sports.png"/>
<title>네이마르, 다음 시즌 레알에 합류(다수 매체)</title>
```

Can't find

```
req = requests.get('http://sports.news.naver.com/wfootball/news/read.nhn?oid=139&aid=0002090884&m_view=1&sort=LIKE')
html = req.text
print(html)
```

Chrome 파일 수정 보기 방문 기록 북마크 사용자 창 도움말

97% A (수) 오전 9:42

Sports news.naver.com/wfootball/news/read.nhn?oid=139&aid=0002090884&m_view=1&sort=LIKE

기사입력 2018.05.09 오전 08:48 | 최종수정 2018.05.09 오전 08:48 기사원문

div#focusComment.news_comment | 660 x 2962

댓글 83

주제와 무관한 댓글, 악플은 삭제될 수 있습니다.

스포츠 기사 댓글 정렬 및 운영 안내

공감순 최신순

Forthlux
파리에서 안풀겠다는데 뭔
41분 전 신고

답글 6

인터넷
사실 네이마르가 기자들 돈주고 퍼뜨림
44분 전 신고

답글 3

베르나데스키

Elements Console Sources Network Performance Memory

Styles

:hover .cls + element.style {

nsportsCss.css.news_comment { padding-top: 16px; }

user agent s... i v { display: block; }

Inherited from...

nsportsCss.css body, input, textarea, select, button, table { font-family: '나눔고딕', 'nanum gothic', '은고' }

html body #wrap #container #content div div div div#focusComment.news_comment

Console What's New

Highlights from the Chrome 66 update

Pretty-printing in the Preview and Response tabs

The Preview tab now pretty-prints by default, and you can force pretty-printing in the Response tab via the new Format button.

Previewing HTML content in the Preview tab

The Preview tab now always does a basic rendering of HTML content.

Local Overrides with styles defined in HTML

슬라이드 4/4 영어(미국)

99%

Slideshow Note Memo

9 5월 1월 2월 3월 4월 5월 6월 7월 8월 9월 10월 11월 12월 TALK Chrome 워크스페이스 맥북 새 탭

```

<div name="focusComment" id="focusComment" class="news_comment">

<div id="cbox_module"></div>
<script type="text/javascript">
document.domain= 'naver.com';
var isLoginForComment = false;
if (jindo.$Cookie().get('NID_SES') != null & jindo.$Cookie().get('NID_SES') != ''){
    isLoginForComment = true;
}

var sports = sports || {};
sports.common = sports.common || {};
sports.common.comment = sports.common.comment || {};

sports.common.comment.init = function(options) {
    window.__htCboxOption = options;
    var s = document.createElement('script');
    s.type = 'text/javascript';
    s.charset = 'utf-8';
    s.src = options.sDomain + '/js/cbox.core.js?v=' + Math.floor(new Date().getTime());
    (document.head || document.getElementsByTagName("head")[0]).appendChild(s);
};

sports.common.comment.options = {
    sDomain : 'https://ssl.pstatic.net/static.cbox',
    sApiDomain : 'https://apis.naver.com/commentBox/cbox2',
    bLogin : isLoginForComment,
    sTicket : 'sports',
    sTemplateId : 'view',
    sObjectId : 'news139_0002090884',
    sCategoryId : '',
    sLikeItId : 'ne_139_0002090884',
    sCategoryImage : '',
    sGroupId : '',
    sLanguage : 'ko',
    sCountry : 'KR',
    sPageType : 'more',
    sHelp : 'up',
}

```

Name	Headers	Preview	Response	Cookies	Timing
web_naver_list_jsonp.json?tic...					
cc?a=RPS.sym&r=&i=&bw=6...					
	:method: GET :path: /commentBox/cbox/web_naver_list_jsonp.json?ticket=sports&templatelId=view&pool=cbox2&callback=jQuery111308024093469217737_1525826381603&lang=ko&country=KR&objectId=news139%2C0002090884&categoryId=&pageSize=20&indexSize=10&groupId=&listType=OBJECT&pageType=more&page=1&refresh=true&sort=like&_=1525826381606 :scheme: https accept: */* accept-encoding: gzip, deflate, br accept-language: ko-KR,ko;q=0.9,en-US;q=0.8,en;q=0.7 cookie: NNB=LV7HGQW3LQ5FU; npic=WnGjpcV+eaRsfq0GmLGGfxK0rLcTCwWbiD6IbOK1AX0DBonUoFC6781Ip5qNmnnVACA==; ASID=7a2c2a37000001609654ba1b0000004e; _ga=GA1.2.134714237.1514701762; NDMyZTMyZWUt0TIyZi000GZkLWJjNjktYzcx0RhYmUz0WU5_8K34wuZR=true; NDMyZTMyZWUt0TIyZi000GZkLWJjNjktYzcx0RhYmUz0WU5_g4WxgVY=true; nx_ssl=2; nsr_acl=1; nid_iplevel=1; nid_info=390430306; ND_JKL=bjJ+8ISqEGtapa1/5+S4sN6z089+3/aC3e7V5zHSazs=; page_uid=TybZvwpl6IossSsukxhssssssV-254161 referer: http://sports.news.naver.com/wfootball/news/read.nhn?oid=139&aid=0002090884&m_view=1&sort=LIKE user-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_4) AppleWebKit/537.36 (KHTML like Gecko) Chrome/66.0.3359.139 Safari/537.36				
	2 requests 6.2 KB transferred				

https://apis.naver.com/commentBox/cbox/web_naver_list_jsonp.json?ticket=sports&templatelId=view&pool=cbox2&callback=jQuery111308024093469217737_1525826381603&lang=ko&country=KR&objectId=news139%2C0002090884&categoryId=&pageSize=20&indexSize=10&groupId=&listType=OBJECT&pageType=more&page=1&refresh=true&sort=like&_=1525826381606

Why?

```
url = 'https://apis.naver.com/commentBox/cbox/web_naver_list_jsonp.json?ticket=sports&templateId=view&pool=cbox2&callback=j
req = requests.get(url)
html = req.text
print(html)
```

```
jQuery111308024093469217737_1525826381603({"success":false,"code":"3999","message":"잘못된
접근입니다.","lang":"ko","country":"KR","result":{},"date":"2018-05-09T00:48:14+0000"});
[Finished in 1.9s]
```

Header Check

- Referer
- Cookie
- User-Agent

```
url = 'https://apis.naver.com/commentBox/cbox/web_naver_list_jsonp.json?ticket=sports&templateId=view&pool=cbox2&_callback='
headers = {
    'referer' : 'http://sports.news.naver.com/wfootball/news/read.nhn?oid=529&aid=0000023174&m_view=1&sort=LIKE'
}
req = requests.get(url,headers=headers)
html = req.text
print(html)
```

```
jQuery111308024093469217737_1525826381603({"success":true,"code":"1000","message":"요청을 성공적으로 처리하였습니다.", "lang":"ko","country":"KR","result":{"sort":"LIKE","count":{"comment":107,"reply":33,"exposeCount":109,"delCommentByUser":0,"delCommentByMon":0,"blindCommentByUser":0,"blindReplyByUser":0,"total":140}, "exposureConfig":{"reason":null,"status":"COMMENT_ON"}, "bestList":[],"pageModel":{"page":1,"pageSize":20,"indexSize":10,"startRow":1,"endRow":20,"totalRows":107,"startIndex":0,"totalPages":6,"firstPage":1,"prevPage":0,"nextPage":2,"lastPage":6,"current":null,"moveToLastPage":false,"moveToComment":false,"moveToLastPrev":false}, "commentList":[{"ticket":"sports","objectId":"news139,0002090884","categoryId":"*","templateId":"default","commentNo":146497590,"parentCommentNo":146497590,"replyLevel":1,"replyCount":9,"replyAllCount":9,"replyPreviewNo":null,"replyList":null,"imageCount":0,"imageList":null,"imagePathList":null,"imageWidthList":null,"imageHeightList":null,"commentType":"txt","stickerId":null,"sticker":null,"sortValue":1525823916666,"contents":"파리에서 안팔겠다는데 뭔","userIdNo":"28V0d","exposedUserIp":null,"lang":"ko","country":"KR","idType":"naver","idProvider":"naver","userName":"Forthrlux","userProfileImage":"http://profile.phinf.naver.net/20203/f99fba320ed07c6a809beb8bfaa12fcfe59bedf35b3924545332f497ed39d22f.jpg","profileType":"naver","modTime":"2018-05-09T08:58:36+0900","modTimeGmt":"2018-05-08T23:58:36+0000","regTime":"2018-05-09T08:58:36+0900","regTimeGmt":"2018-05-08T23:58:36+0000","sympathyCount":136,"antipathyCount":13,"userBlind":false,"hideReplyButton":false,"status":0,"mine":false,"best":false,"mentions":null,"toUser":null,"userStatus":0,"categoryImage":null,"open":false,"levelCode":"level04","grades":null,"sympathy":false,"antipathy":false,"snsList":null,"metaInfo":null,"extension":null,"audioInfoList":null,"translation":null,"report":null,"middleBlindReport":false,"spamInfo":null,"userHomepageUrl":null,"defamation":false,"deleted":false,"expose":true,"visible":true,"secret":false,"blind":false,"maskedUserId":"qazw****","maskedUserName":"Fo****","validateBanWords":false,"exposeByCountry":false,"virtual":false,"containText":true,"blindReport":false,"manager":false,"shardTableNo":null,"profileUserId":null,"anonymous":false}, {"ticket":"sports","objectId":"news139,0002090884","categoryId":"*","templateId":"default","commentNo":146497231,"parentCommentNo":146497231,"replyLevel":1,"replyCount":3,"replyAllCount":3,"replyPreviewNo":null,"replyList":null,"imageCount":0,"imageList":null,"imagePathList":null,"imageWidthList":null,"imageHeightList":null,"commentType":"txt","stickerId":null,"sticker":null,"sortValue":1525823689600,"contents":"사실 네이마르가 기자들 돈주고 퍼뜨림","userIdNo": "28V0d","userProfileImage": "http://profile.phinf.naver.net/20203/f99fba320ed07c6a809beb8bfaa12fcfe59bedf35b3924545332f497ed39d22f.jpg"}])
```

Split

```
name = "kim, lee, park, choi"
name = name.split(",")
print(name)
```

```
['kim', 'lee', 'park', 'choi']
```

```

2 import requests
3
4
5 url = 'https://apis.naver.com/commentBox/cbox/web_naver_list_jsonp.json?ticket=sports&templateId=view&pool
6 headers = {
7     'referer' : 'http://sports.news.naver.com/wfootball/news/read.nhn?oid=139&aid=0002090884&m_view=1&sort
8 }
9 req = requests.get(url,headers=headers)
10 html = req.text
11 print(html)
12
13
14
15 comment_num = len(html.split('contents":')) - 1
16 for i in range(comment_num):
17     nick = html.split('userName":')[i+1].split('"')[0]
18     comment = html.split('contents":')[i+1].split('"')[0]
19     print(nick + ": " + comment)
20

```

네이마르는 그의 또 다른 면모를 드러냈다.
 어때: 다수언론이 짜고 친건가..
 00: 파리에서 떠야지 부합시나 RMC 마르카 아스원소용이냐 원래저랬음
 dvdI: 마르카 아스를 믿음? 레알언론ㅋㅋㅋ
 수컷연후: 파리가 얼마나 팔까?
 김은찬: ㅋㅋㅋㅋㅋ 그냥 웃지요
 JUST: 벤제마 베일 이런애들 현금이랑 같이주면 트레이드 해준다 파리도. 왜냐하면 네이마르 데리고 있어봤자 팀에 마음도 없는데 케미000 나거든ㅋㅋ네이마
 푸키푸키: 좀 본인이 말좀해라.. 기자들먹여살릴라고 말안하나
 신밧드: 불가능한게 호우가 행복하지않음 재가오면
 레스폰: 왜 호날두 따끼리를 자처해서 가는지 이해가 안가네 카바니 박힌들 빼내는거보다 호날두 재끼는게 더 힘들텐데....
 kier****: 글쎄? 네이마르가 또 호날두 밑닦아줄려고 레알로 간다고? 메시 뒷치닥거리 하기 싫어서 psg로 간건데....만약 네이마르가 레알로 온다면 호날
 찢지마벗을게: 바르샤를 나왔는데... 같은 곳은 딱 한군데 밖에 없지. 살라가 자기 자리 뺏었는데 똥줄타서라도 더 빨리 레알 가야지
 kms5****: 팬심 까심 버리고 네이마르가 만약 진짜 오기로했다면 확실한건 호날두 이적도 유력해진다는 소리다.. 아무리 레알이라도 초고액 연봉자를 한번
 골은 기복있어서 그렇지 잘넣지만) 그나마 요즘같이 좀 가치만큼 해줄때 고액에 빨리 팔아버릴 시점에 있는데 데리고 있을 이유가 없음.. 무엇보다 네이마르가
 내걸었음이 확실하거든.. 레알 입장에서는 호날두를 그나마 품 남아있는 지금 고액에 팔아버리고 돈좀 더 써서 네이마르와 살라 혹은 케인을 데려오는게 훨씬
 메시: 네 다음 개소리
 hans: 오면 레알 트레블 각
 무리뉴DogBaby: 부상회복했다니 슬슬 또 기울라오네 기사가
 탱크: 네이마르 온다면야 바르샤 그리즈만 화력보다 더 쎄지겠네
 jsw7****: 레알 가면 바르샤 잇을때 메시보다 더 위력적 일거 같다 문제는 베일이랑 아센시오 이스코랑 어캐 공존한지가 문제임 날두는 이미 타겟형이어서

BeautifulSoup?

1. 정규표현식 활용

- 가장 빠른 처리가 가능하나, 정규표현식의 룰을 만드는 것이 번거롭고 복잡하여 다양한 처리를 하기 어려운 점이 있습니다.
- 때에 따라 필요할 수 있습니다.

2. HTML Parser 라이브러리를 활용(V)

- DOM Tree를 탐색하는 방식으로 적용이 쉬운 장점이 있습니다.
- Ex) BeautifulSoup4, lxml

To be continued