

파이썬 웹 크롤링

강사장철원

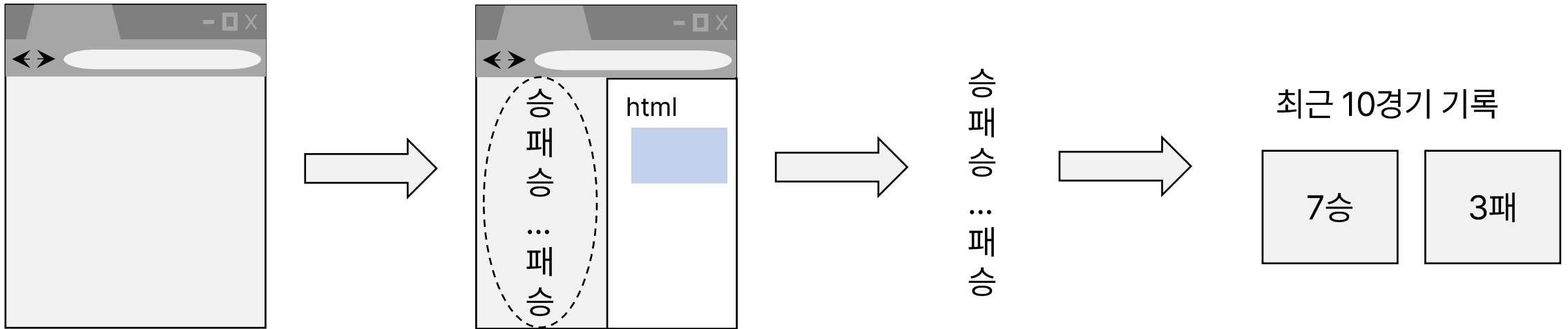
□ 동적 크롤링 활용해서 해외 축구 데이터 크롤링하기

파이썬 웹 크롤링

Section 1. 해외 축구 데이터 크롤링

Section 1-1. 크롤링 웹 사이트 확인

크롤링 개요

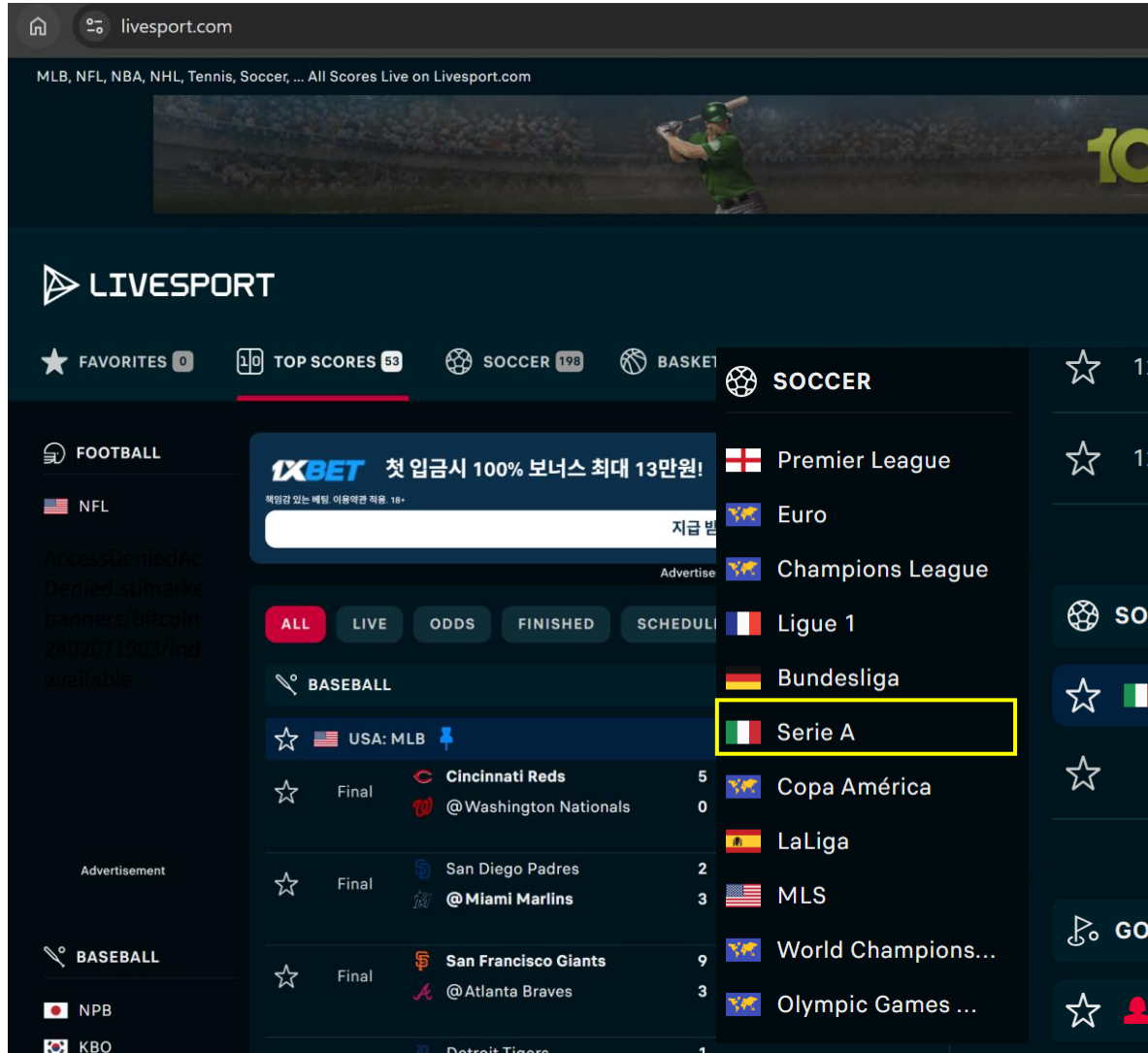


Section

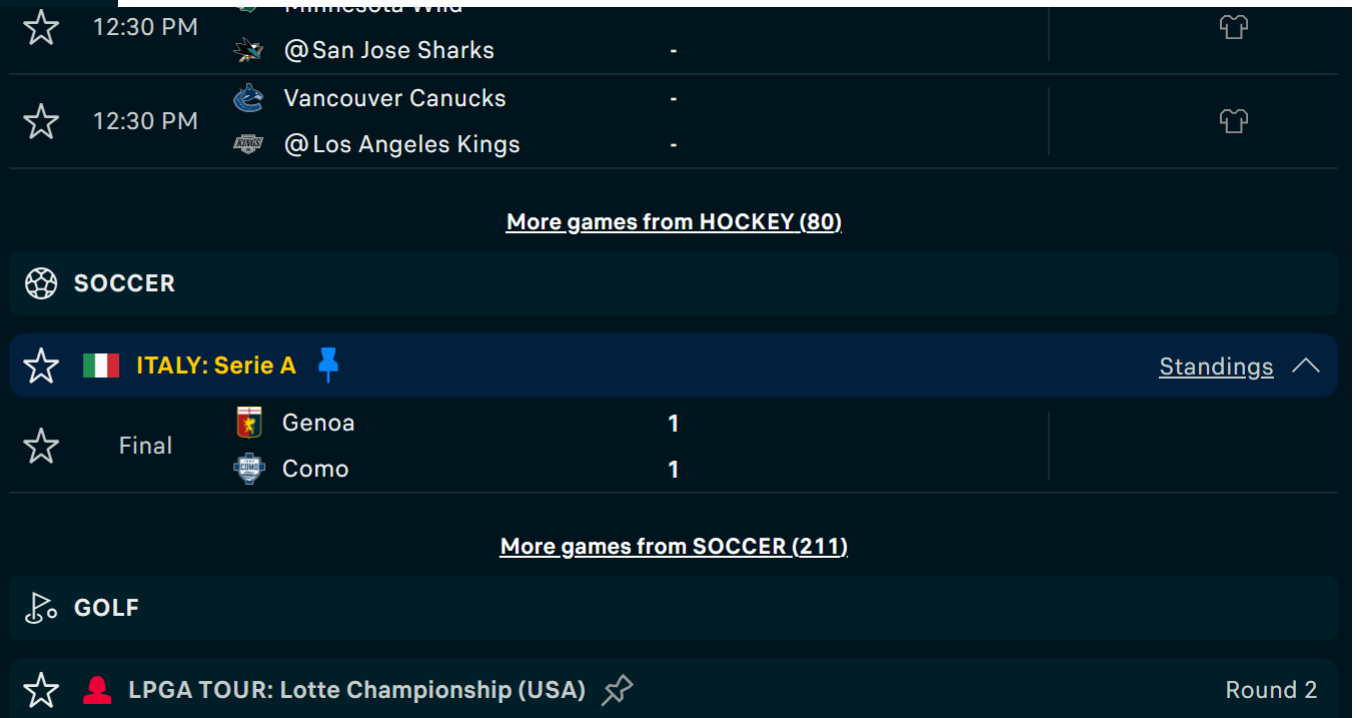
크롤링 웹사이트 확인

해외 스포츠 기록 사이트 확인

<https://www.livesport.com/>




스크롤 내려서 SOCCER 메뉴의 Serie A 확인



해외 스포츠 기록 사이트 확인

<https://www.livesport.com/>



Advertisement

Azerbaijan

Bahrain

Bangladesh

Barbados

Belarus

Belgium

Benin

Bermuda

Bhutan

OVERALL		HOME	AWAY											
# ▲	TEAM	GP	W	T	L	G	GD	PTS	FORM					
1.	Napoli	11	8	1	2	18:8	10	25	?	L	W	W	W	W
2.	Inter	11	7	3	1	25:13	12	24	?	W	W	T	W	W
3.	Atalanta	11	7	1	3	29:14	15	22	?	W	W	W	W	W
4.	Fiorentina	11	6	4	1	22:9	13	22	?	W	W	W	W	W
5.	Lazio	11	7	1	3	24:14	10	22	?	W	W	W	L	W
6.	Juventus	11	5	6	0	19:7	12	21	?	W	T	T	W	T
7.	AC Milan	10	5	2	3	17:11	6	17	?	W	L	W	L	W
8.	Udinese	11	5	1	5	14:16	-2	16	?	L	L	W	L	W
9.	Bologna	10	3	6	1	12:11	1	15	?	W	W	T	T	T
10.	Torino	11	4	2	5	15:16	-1	14	?	L	L	W	L	L

해외 스포츠 기록 사이트 확인

livesport.com/team/ac-milan/8Sa8HInO/

LIVESPORT

FAVORITES 0

TOP SCORES 22

SOCCER 211

BASKETBALL 199

TENNIS 265

HOCKEY 80

ITALY

Serie A

Serie B

Serie C - Group A

Serie C - Group B

Serie C - Group C

Serie C - Promotion - Play Offs

Serie C - Play Out

Serie D - Group A

Serie D - Group B

Serie D - Group C

Show more (21)

1XBET

첫 입금시 100% 보너스 최대 13만원!

책임감 있는 베팅. 이용약관 적용. 18+

지금 받기

Advertisement

SOCCER > ITALY

AC Milan ☆

Stadium: Stadio Giuseppe Meazza (San Siro) (Milan)
Capacity: 80 018

SUMMARY

NEWS

RESULTS


FIXTURES

STANDINGS


TRANSFERS

SQUAD


The latest news



REVEALED: Real Madrid teammates frustrated with Mbappe



France coach Deschamps explains omitting Real Madrid striker



Struggling captain Mbappe once again left out of France squad

최근 전적 확인

Antigua & Barbuda	Latest Scores			
Argentina	EUROPE: Champions League - League phase			Standings
Armenia	Nov 06 05:00 AM	Real Madrid	1	W
Aruba		AC Milan	3	
Australia	ITALY: Serie A			Standings
Austria	Nov 03 04:45 AM	Monza	0	W
		AC Milan	1	
	Oct 30 04:45 AM	AC Milan	0	L
		Napoli	2	
	EUROPE: Champions League - League phase			Standings
	Oct 23 01:45 AM	AC Milan	3	W
		Club Brugge KV	1	
	ITALY: Serie A			Standings
	Oct 20 01:00 AM	AC Milan	1	W
		Udinese	0	
	Oct 07 03:45 AM	Fiorentina	2	L
		AC Milan	1	
Azerbaijan	EUROPE: Champions League - League phase			Standings
Bahrain	Oct 02 04:00 AM	Bayer Leverkusen	1	L
Bangladesh		AC Milan	0	
Barbados	ITALY: Serie A			Standings
Belarus	Sep 28 03:45 AM	AC Milan	3	W
Belgium		Lecce	0	
Benin	Sep 23 03:45 AM	Inter	1	W
Bermuda		AC Milan	2	
Bhutan	EUROPE: Champions League - League phase			Standings
Bolivia	Sep 18 04:00 AM	AC Milan	1	L
Bosnia and Herzegovina		Liverpool	3	

크롤링 웹사이트 확인

<https://www.livesport.com/>

The screenshot shows the 'Elements' tab in a web browser's developer tools. The DOM tree on the left shows a hierarchy starting with 'div' (id='g_1_hvJWFBf5'), followed by 'div' (id='g_1_h6u4EbJO'), and then 'div' (class='wcl-header_uBhYi'). The selected element is a 'span' (class='wcl-simpleText_Asp-0') with data-testid='wcl-scores-simpleText-01'. The main pane shows the HTML code for this element and its surrounding context, including a 'button' and several 'div' elements. The code is as follows:

```

<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">W
</span>
</button>
</div>
<div id="g_1_hvJWFBf5" class="event__match event__match--withRowLink event__match--static event__match--twoLine" data-event-row="true">...</div>
<div id="g_1_h6u4EbJO" class="event__match event__match--withRowLink event__match--static event__match--last event__match--twoLine" data-event-row="true">...</div>
<div class="wcl-header_uBhYi wclLeagueHeader wclLeagueHeader--collapsed wclLeagueHeader--noCheckBox wclLeagueHeader--indent"

```


해외 스포츠 기록 사이트 확인

<https://www.livesport.com/>

```
▼<button class="wcl-badgeform_yYFgV wcl-win_9cMnW wcl-clickable_BbSlo  
formIcon__lastMatches" title="Win" data-testid="wcl-badgeForm-win">  
  flex  
  <span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk"  
    data-testid="wcl-scores-simpleText-01">W</span> == $0  
</button>  
</div>
```

span 태그 사용 및
wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk
클래스 사용 확인

해외 스포츠 기록 사이트 확인

<https://www.livesport.com/>

```
[1]: from urllib.request import urlopen  
     from bs4 import BeautifulSoup
```

```
[2]: url = "https://www.livesport.com/team/ac-milan/8Sa8HIn0/"  
     html = urlopen(url)  
     bs_obj01 = BeautifulSoup(html, "html.parser")  
     win01 = bs_obj01.find_all("span", {"class": "wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk"})
```

```
[3]: print(win01)
```

[] 아무것도 안나온다?

해외 스포츠 기록 사이트 확인

<https://www.livesport.com/>

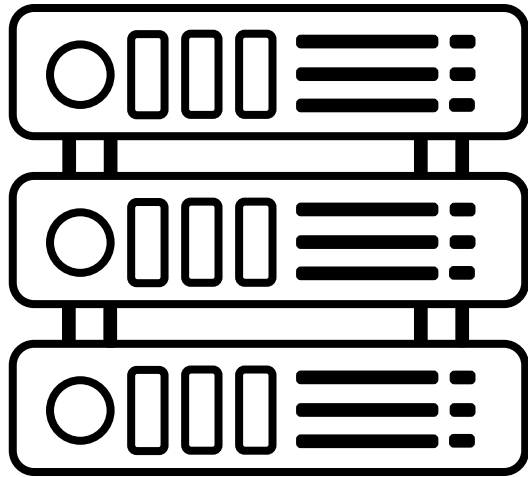
[4]: `print(bs_obj01)` **bs_obj01은 잘 나오는거 같은데?**

```
<!DOCTYPE html>

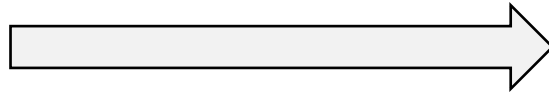
<html lang="en-US">
<head>
<meta charset="utf-8"/>
<title>Livesport: AC Milan - results, fixtures, Liverpool - AC Milan live</title>
<script charset="UTF-8" data-document-language="true" data-domain-script="be4cd6ae-414e-4784-a9f8-29c9eb574714" defer="" src="https://cdn.cookie law.org/scripts/templates/otSDKStub.js" type="text/javascript">
</script>
<link crossorigin="" href="https://500.ds.lsapp.eu/pq_graphql" rel="preconnect"/>
<link crossorigin="" href="https://500.flashscore.ninja" rel="preconnect"/>
<link as="font" crossorigin="" href="https://static.flashscore.com/res/font/LivesportFinderLatin-Regular_Static.woff2" rel="preload" type="font/woff2"/>
<link as="font" crossorigin="" href="https://static.flashscore.com/res/font/LivesportFinderLatin-Bold_Static.woff2" rel="preload" type="font/woff2"/>
<noscript>
<meta content="0;url=https://www.flashscore.mobi/" http-equiv="refresh">
</meta></noscript>
<meta content="AC Milan scores and fixtures - follow AC Milan livescore, results, fixtures and match details on Livesport.com." name="description"/>
<meta content="Copyright (c) 2012-2025 Livesport s.r.o." name="copyright"/>
<meta content="index,follow" name="robots">
<meta content="width=device-width, initial-scale=1" name="viewport"/>
```

wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk 클래스 검색하면 안나옴

정적 페이지 vs 동적 페이지

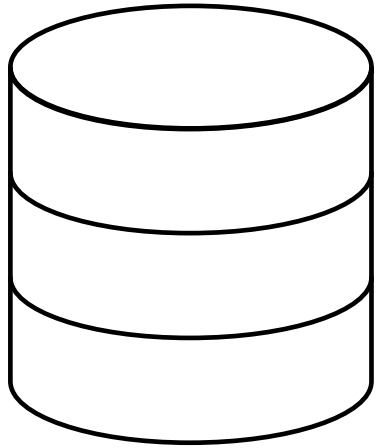


서버

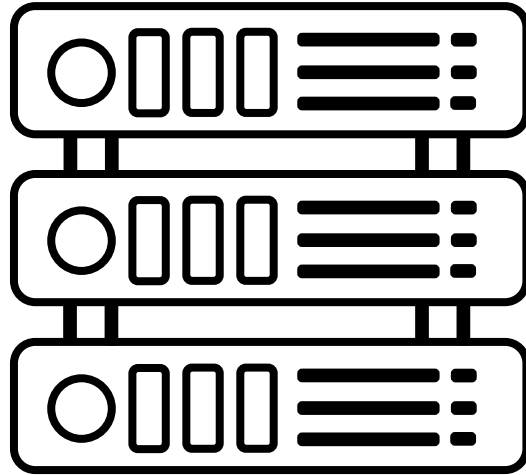
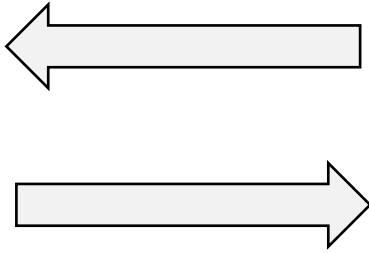


클라이언트

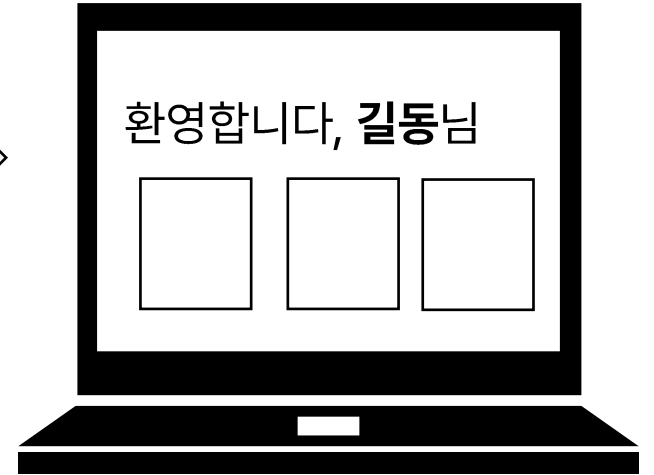
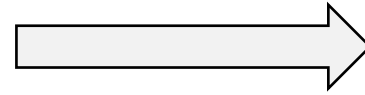
정적 페이지 vs 동적 페이지



데이터베이스



서버

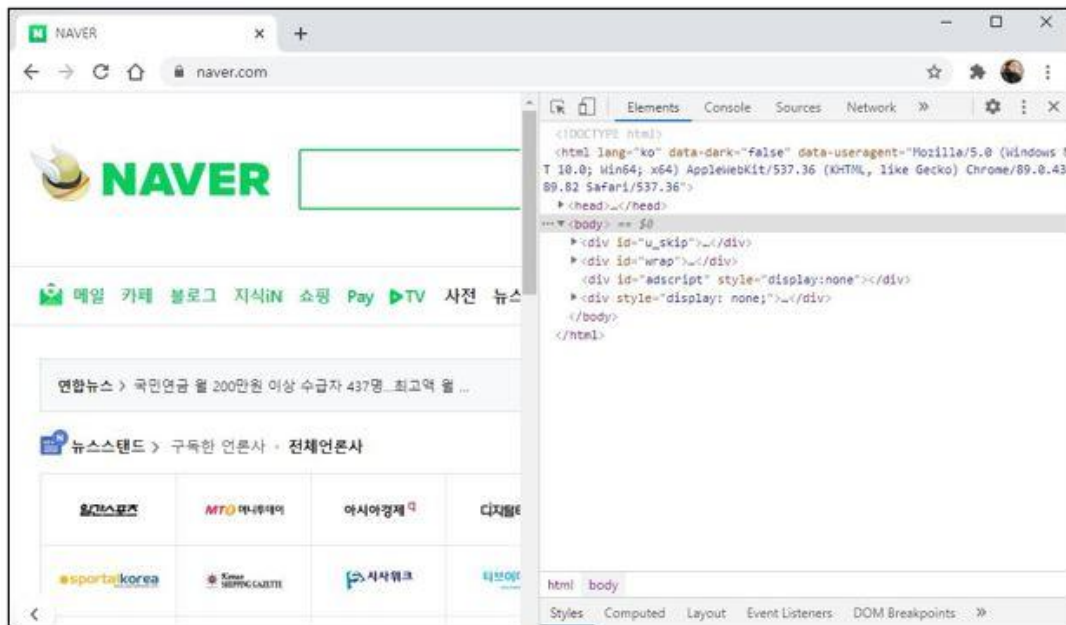


클라이언트

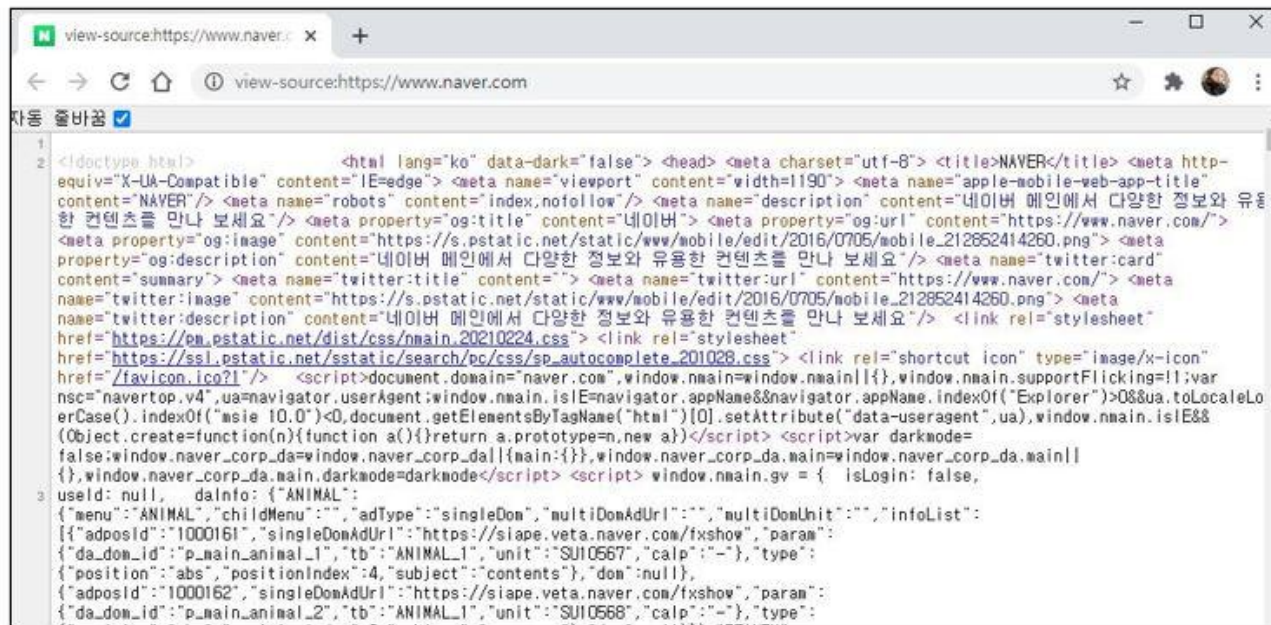
개발자 도구 vs 페이지 소스

개발자 도구 - 웹 서버에서 전달 받은 이후 가공한 후의 최종적으로 브라우저가 그린 결과물

페이지 소스 보기 - 웹 서버에서 최초로 전달 받은 내용이 그대로 나타남. beautifulsoup을 이용해 파싱할 때는 페이지 소스를 파싱하므로 가공 이후의 내용은 표시되지 않음(자바스크립트 실행되기 이전 원본 HTML)



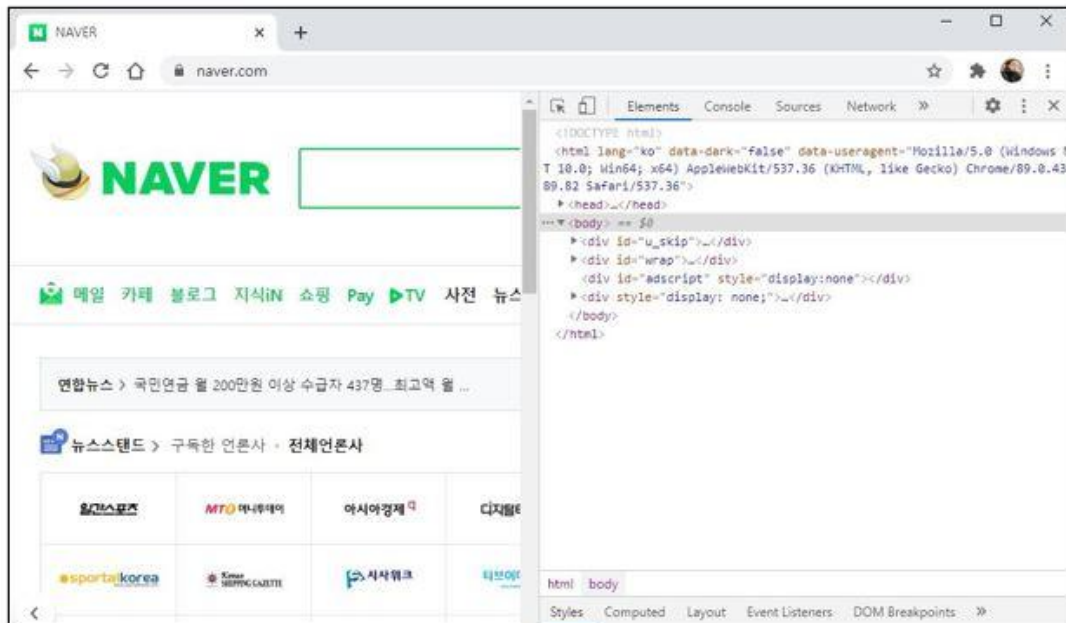
개발자 도구



페이지 소스

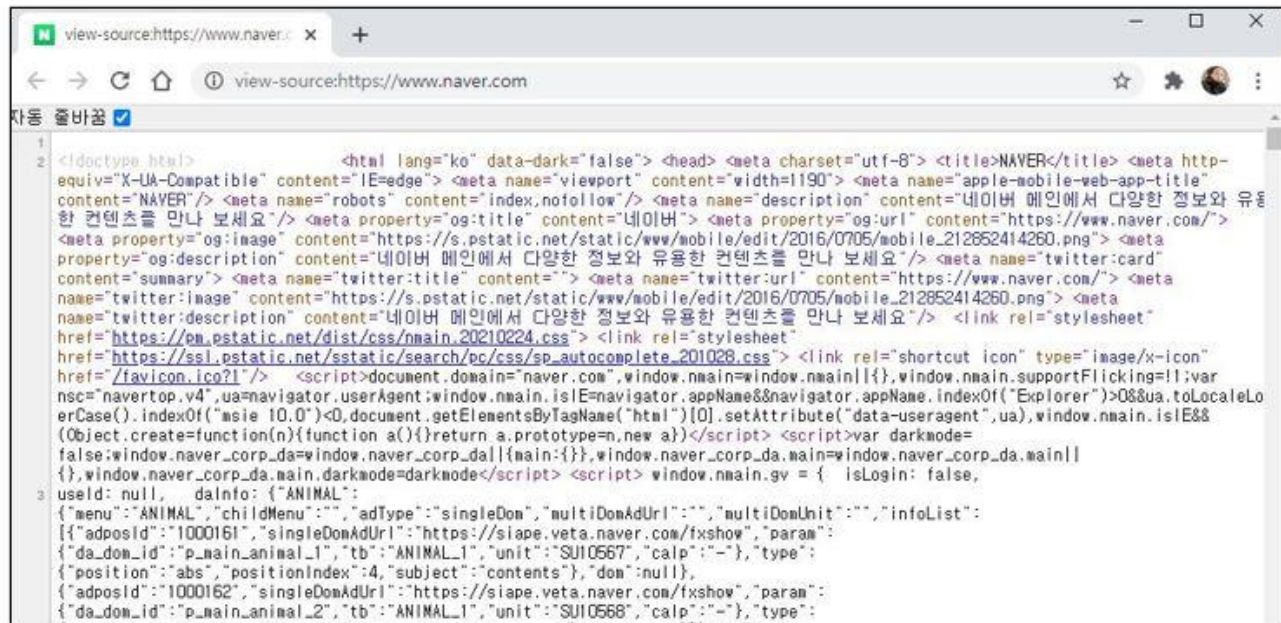
개발자 도구 vs 페이지 소스

개발자 도구



웹 브라우저가 문서를 읽고,
자바스크립트도 실행해서 화면에 보여준 최종 상태

페이지 소스



원본 스크립트만 존재

개발자 도구 - 키워드 나옴

← → ↺ 🏠 🌐 naver.com

☰ (pay)

N

📧 메일

☕ 카페

🗨️ 블로그

🛒 스토어


📰 뉴스


📈 증권

🏠 부동산

📍 지도

XEXYMIX





젝시믹스 쿨캉스 페스타 ~69%

입자마자 뛰고 싶어지는 여름 BEST 쇼츠

지금 구매하기

뉴스스탠드 · 언론사편집 / 엔터 / 스포츠 LIVE / 경제 / 쇼핑투데이

국내증시 · 해외증시 · 시장지표 · 관심종목

증권홈 | 부동산홈

Elements Console Sources Network Performance >> 2 ⚙️ ⋮ ✕

🔍

▼ #document (https://www.naver.com/)

<!DOCTYPE html>

▼ <html>

▶ <head> ⋮ </head>

▼ <body id="gfp_sf_body" class="banner_ad_wrapper type_step2_3x2 type_min_w437 type_min_h129">

▼ <div id="gfp_sf_align">

<script src="https://ssl.pstatic.net/tveta/libs/glad/prod/3.3.1/gfp-display-safeiframe.js"></script>

▼ <div id="ad-element" class="timeboard_basic">

▼ ⋮

▶ ⋮

▶ ⋮

▼ <div class="copy">

<strong class="title">젝시믹스 쿨캉스 페스타 ~69%

 == \$0

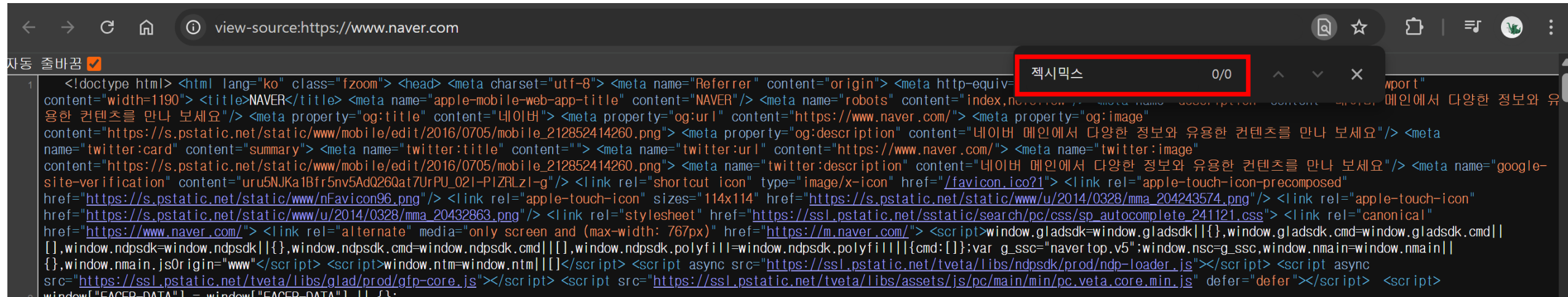
입자마자 뛰고 싶어지는 여름 BEST 쇼츠

</div>

▶ <div class="cta" style="background-color: rgb(123, 139, 156);"> ⋮ </div> ⋮ </div>

▶ <i class="icon_ad" data-gfp-role="adMarkLegacy" style="display:none;"> ⋮ </i>

페이지 소스 - 키워드 안나옴



```
<!doctype html> <html lang="ko" class="fzoom"> <head> <meta charset="utf-8"> <meta name="Referrer" content="origin"> <meta http-equiv=
content="width=1190"> <title>NAVER</title> <meta name="apple-mobile-web-app-title" content="NAVER"/> <meta name="robots" content="index, nofollow"> <meta name="description" content="네이버 메인에서 다양한 정보와 유
용한 콘텐츠를 만나 보세요"/> <meta property="og:title" content="네이버"> <meta property="og:url" content="https://www.naver.com/"> <meta property="og:image"
content="https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png"> <meta property="og:description" content="네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요"/> <meta
name="twitter:card" content="summary"> <meta name="twitter:title" content=""> <meta name="twitter:url" content="https://www.naver.com/"> <meta name="twitter:image"
content="https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png"> <meta name="twitter:description" content="네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요"/> <meta name="google-
site-verification" content="uru5NJKa1Bfr5nv5AdQ26Qat7UrPU_02l-PlZRLzl-g"/> <link rel="shortcut icon" type="image/x-icon" href="/favicon.ico?1"> <link rel="apple-touch-icon-precomposed"
href="https://s.pstatic.net/static/www/nFavicon96.png"/> <link rel="apple-touch-icon" sizes="114x114" href="https://s.pstatic.net/static/www/u/2014/0328/mma_204243574.png"/> <link rel="apple-touch-icon"
href="https://s.pstatic.net/static/www/u/2014/0328/mma_20432863.png"/> <link rel="stylesheet" href="https://ssl.pstatic.net/sstatic/search/pc/css/sp_autocomplete_241121.css"> <link rel="canonical"
href="https://www.naver.com/"> <link rel="alternate" media="only screen and (max-width: 767px)" href="https://m.naver.com/"> <script>window.gladsdk=window.gladsdk||{};window.gladsdk.cmd=window.gladsdk.cmd||
[],window.ndpsdk=window.ndpsdk||{};window.ndpsdk.cmd=window.ndpsdk.cmd||[],window.ndpsdk.polyfill=window.ndpsdk.polyfill||{};var g_ssc="navertop.v5";window.nsc=g_ssc;window.nmain=window.nmain||
{};window.nmain.jsOrigin="www"</script> <script>window.ntm=window.ntm||{}</script> <script async src="https://ssl.pstatic.net/tveta/libs/ndpsdk/prod/ndp-loader.js"></script> <script async
src="https://ssl.pstatic.net/tveta/libs/glad/prod/gfp-core.js"></script> <script src="https://ssl.pstatic.net/tveta/libs/assets/js/pc/main/min/pc.veta.core.min.js" defer="defer"></script> <script>
window["FACEBOOK_DATA"] = window["FACEBOOK_DATA"] || {};
```

해외 스포츠 기록 사이트 확인

<https://www.livesport.com/>

- [5]: `from selenium import webdriver` webdriver: 브라우저를 제어하기 위한 다양한 도구를 포함하고 있는 모듈
- []: `#options = webdriver.ChromeOptions()` 크롬 브라우저 설정 변경을 위한 객체 생성
`#options.add_argument("--headless")` headless 옵션: 브라우저 창이 열리지 않고 백그라운드에서 실행하는 모드
- [6]: `driver = webdriver.Chrome()` 크롬 브라우저 제어를 위한 객체 생성
- [7]: `driver.implicitly_wait(3)` 웹 페이지가 로딩될때까지 기다리는 시간, 단위는 초(second)
- [8]: `driver.get("https://www.livesport.com/team/ac-milan/8Sa8HIn0/")` 해당 웹 페이지로 이동

해외 스포츠 기록 사이트 확인 <https://www.livesport.com/>

```
[9]: from bs4 import BeautifulSoup
```

```
[10]: page = driver.page_source  현재 브라우저에서 로드된 전체 HTML 소스 코드 변환
```

```
[11]: bs_obj = BeautifulSoup(page, "html.parser")
      print(bs_obj)
```

```
<html lang="en-US"><head>
<meta charset="utf-8"/>
<title>Livesport: AC Milan - results, fixtures, Liverpool - AC Milan live</title>
<script async="" src="https://www.googletagmanager.com/gtag/destination?id=AW-999393459&ai
1509157~103116026~103200004~103233427~104684208~104684211" type="text/javascript"></scrip
web-vitals.iife.js" type="text/javascript"></script><script async="" src="https://www.goo
7m0v71314122za200&tag_exp=101509157~103116026~103200004~103233427~104684208~104684211
ogletagmanager.com/gtm.js?id=GTM-PWJ3NQ"></script><script charset="UTF-8" data-document-1
b574714" defer="" src="https://cdn.cookie law.org/scripttemplates/otSDKStub.js" type="text
</script>
<link crossorigin="" href="https://500.ds.lsapp.eu/pq_graphql" rel="preconnect"/>
```

<https://www.livesport.com/>

[14]: 19

해외 스포츠 기록 사이트 확인

<https://www.livesport.com/>

```
for i in range(0, len(win)):  
    print(win[i])
```

```
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01"></span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">Bracket</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">Standings</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01"></span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">L</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">Standings</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">W</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">L</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">Bracket</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">L</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">Standings</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">W</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">W</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">W</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">Bracket</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">W</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">W</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">Standings</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">L</span>  
<span class="wcl-simpleText_Asp-0 wcl-scores-simpleText-01_pV2Wk" data-testid="wcl-scores-simpleText-01">W</span>
```

```
for i in range(0, len(win)):  
    print(win[i].text)
```

Bracket
Standings

L
Standings
W
L
Bracket
L
Standings
W
W
W
Bracket
W
Standings
L
W

해외 스포츠 기록 사이트 확인

<https://www.livesport.com/>

```
[17]: result = {'win':0, 'draw':0, 'lose':0}
```

```
[18]: keys = list(result.keys())  
print(keys)
```

```
['win', 'draw', 'lose']
```

```
[19]: for i in range(0, len(win)):  
        if win[i].text == 'W':  
            result['win'] += 1  
        elif win[i].text == 'D':  
            result['draw'] += 1  
        elif win[i].text == 'L':  
            result['lose'] += 1  
    print(result)
```

```
{'win': 0, 'draw': 0, 'lose': 0}  
{'win': 0, 'draw': 0, 'lose': 0}  
{'win': 0, 'draw': 0, 'lose': 0}  
{'win': 0, 'draw': 0, 'lose': 0}  
{'win': 0, 'draw': 0, 'lose': 1}  
{'win': 0, 'draw': 0, 'lose': 1}  
{'win': 1, 'draw': 0, 'lose': 1}  
{'win': 1, 'draw': 0, 'lose': 2}  
{'win': 1, 'draw': 0, 'lose': 2}  
{'win': 1, 'draw': 0, 'lose': 3}  
{'win': 1, 'draw': 0, 'lose': 3}  
{'win': 2, 'draw': 0, 'lose': 3}  
{'win': 3, 'draw': 0, 'lose': 3}  
{'win': 4, 'draw': 0, 'lose': 3}  
{'win': 4, 'draw': 0, 'lose': 3}  
{'win': 5, 'draw': 0, 'lose': 3}  
{'win': 5, 'draw': 0, 'lose': 3}  
{'win': 5, 'draw': 0, 'lose': 4}  
{'win': 6, 'draw': 0, 'lose': 4}
```

```
[20]: print(result)
```

```
{'win': 6, 'draw': 0, 'lose': 4}
```

감사합니다.

Q & A