

**크롤링(crawling)**

- 크롤링의 뜻과 사용하는 이유를 설명할 수 있다.
- 웹 페이지의 구조를 설명할 수 있다.
- 크롤링을 통해 웹 페이지의 정보를 리스트에 저장할 수 있다.

**크롤링이란?**

# 크롤링이란

4

영어사전

crawl



고급 검색

전체 | 단어·속어 | 뜻풀이 | 예문 | 유의어 | 영영사전

T

T

T

연관검색어 creep crawling gasp knock over gratitude trip stack keep to oneself  
turn out walk along crawl snap end up walk up to sneak hang out jealous crouch  
encouragement creep

단어·속어 79

글로벌 발음듣기 ☒

crawl ★ +

1. 동사 (엎드려) 기다
2. 동사 (곤충이) 기어가다
3. 명사 기어가기, 서행 (→pub crawl)
4. 명사 (수영의) 크롤법

**크롤링(crawling)**은 '기다'라는 뜻의 crawl의 명사형이다.

소프트웨어와 같은 무언가가 인터넷을 돌아다니며 정보를 수집해 오는 작업을 의미하고, 그러한 작업을 하는 소프트웨어를 **크롤러(crawler)**라고 한다.

월드와이드웹에서 웹페이지의 데이터를 '긁어' 오는 행위를 **스크레이핑(scraping)**이라고도 하는데 대체로 유사한 의미이다.

**크롤링하는 이유**

**크롤링(crawling)**은 웹 사이트(web site), 하이퍼링크(hyperlink), 데이터(data), 정보 자원을 자동화된 방법으로 수집, 분류, 저장하는 것을 의미함

**웹 크롤링(web crawling)** 또는 **데이터 크롤링(data crawling)**이라고도 함

웹 상의 다양한 정보 자원을 자동화된 방법으로 수집해서  
분류 및 저장하기 위함

# 크롤링하는 이유

8

N

bts



통합 뉴스 이미지 VIEW 지식iN 인플루언서 동영상 쇼핑 어학사전 지도 ...

강원도민일보 PICK | 6시간 전 | 네이버뉴스

## 새단장한 삼척 BTS '버터비치', 일본인 관광객으로 복적

BTS)이 앨범 재킷 사진 촬영을 한 삼척 맹방해변을 찾는 국내·외 관광객들이 급증하고 있어 새로운 관광자원화 가능성을 높이고 있다. 삼척시에 따르면 일본인 관...



삼척 맹방해변 'BTS 포토존' 새단장... 日 관... 노컷뉴스 | 9시간 전 | 네이버뉴스

삼척 맹방해변 BTS앨범 촬영지 새단장 강원일보 | 10시간 전 | 네이버뉴스

관련뉴스 전체보기 >

엑스포츠뉴스 | 11시간 전 | 네이버뉴스

## 방탄소년단(BTS) 진, 첫 솔로 앨범 발매 후 한터 주간차트 2관왕 ...

그룹 방탄소년단(BTS) 멤버 진과 가수 원호(WONHO)가 한터차트 10월 5주 주간차트 1위에 등극했다. 지난 10월 31일 한터차트는 2022년 10월 5주 차 주간 월드차...



방탄소년단(BTS) 진, 솔로 앨범 'The Astronaut'...한... 뉴스인사이드 | 14시간 전

'한터차트' 스트레이키즈 2관왕, 블랙핑크·BTS 제쳤다 국제뉴스 | 4시간 전

스포츠조선 | 8시간 전 | 네이버뉴스

## BTS 진 "슈퍼스타는 한정판 못 참아→돈 쓴 티 난 옷 좋아" 필수...





# 크롤링하는 이유

9

N

bts



통합 뉴스 이미지 VIEW 지식iN 인플루언서 동영상 쇼핑 어학사전 지도 ...

강원도민일보 PICK | 6시간 전 | 네이버뉴스

새단장한 삼척 **BTS** '버터비치', 일본인 관광객으로 복적

**BTS**이 앨범 재킷 사진 촬영을 한 삼척 맹방해변을 찾는 국내·외 관광객들이 급증하고 있어 새로운 관광자원화 가능성을 높이고 있다. 삼척시에 따르면 일본인 관...



삼척 맹방해변 '**BTS** 포토존' 새단장...[더 관...](#) | 노컷뉴스 | 9시간 전 | 네이버뉴스

삼척 맹방해변 **BTS** 앨범 촬영지 새단장 | 강원일보 | 10시간 전 | 네이버뉴스

[관련뉴스 전체보기 >](#)

엑스포츠뉴스 | 11시간 전 | 네이버뉴스

방탄소년단(**BTS**) 진, 첫 솔로 앨범 발매 후 한터 주간차트 2관왕 ...

그룹 방탄소년단(**BTS**) 멤버 진과 가수 원호(WONHO)가 한터차트 10월 5주 주간차트 1위에 등극했다. 지난 10월 31일 한터차트는 2022년 10월 5주 차 주간 월드차...



방탄소년단(**BTS**) 진, 솔로 앨범 'The Astronaut'...한... | 뉴스인사이드 | 14시간 전

'한터차트' 스트레이키즈 2관왕, 블랙핑크·**BTS** 제쳤다 | 국제뉴스 | 4시간 전

스포츠조선 | 8시간 전 | 네이버뉴스

**BTS** 진 "슈퍼스타는 한정판 못 참아→돈 쓴 티 난 옷 좋아" 필수...



# 웹 페이지 구조

# 웹 페이지 구조(html)

11

N | bts

통합 뉴스 이미지 VIEW 지식iN 인플루언서 동영상 쇼핑 어학사전 지도 ...

강원도민일보 PICK | 6시간 전 | 네이버뉴스

새단장한 삼척 **BTS** '버터비치', 일본인 관광객으로 북적

**BTS**이 앨범 재킷 사진 촬영을 한 삼척 맹방해변을 찾는 국내·외 관광객들이 급증하고 있어 새로운 관광자원화 가능성을 높이고 있다. 삼척시에 따르면 일본인 관...



삼척 맹방해변 '**BTS** 포토존' 새단장...[더 관...](#) | 노컷뉴스 | 9시간 전 | 네이버뉴스  
삼척 맹방해변 **BTS**앨범 촬영지 새단장 강원일보 | 10시간 전 | 네이버뉴스

관련뉴스 전체보기 >

엑스포츠뉴스 | 11시간 전 | 네이버뉴스

방탄소년단(**BTS**) 진, 첫 솔로 앨범 발매 후 한터 주간차트 2관왕 ...

그룹 방탄소년단(**BTS**) 멤버 진과 가수 원호(WONHO)가 한터차트 10월 5주 주간차트 1위에 등극했다. 지난 10월 31일 한터차트는 2022년 10월 5주 차 주간 월드차...



방탄소년단(**BTS**) 진, 솔로 앨범 'The Astronaut'...[한...](#) | 뉴스인사이드 | 14시간 전  
'한터차트' 스트레이키즈 2관왕, 블랙핑크·**BTS** 제쳤다 국제뉴스 | 4시간 전

스포츠조선 | 8시간 전 | 네이버뉴스

**BTS** 진 "슈퍼스타는 한정판 못 참아→돈 쓴 티 난 옷 좋아" 필수...



뒤로 Alt+왼쪽 화살표

앞으로 Alt+오른쪽 화살표

새로고침 Ctrl+R

다른 이름으로 저장... Ctrl+S

인쇄 Ctrl+P

전송...

Google Lens로 이미지 검색

기기로 전송

이 페이지의 QR 코드 생성

한국어(으)로 번역

Google에서 이미지 설명 가져오기 ▶

페이지 소스 보기 Ctrl+U

검사



# 웹 페이지 구조(html)

13

```
<html>
<head>
  <title>div example</title>

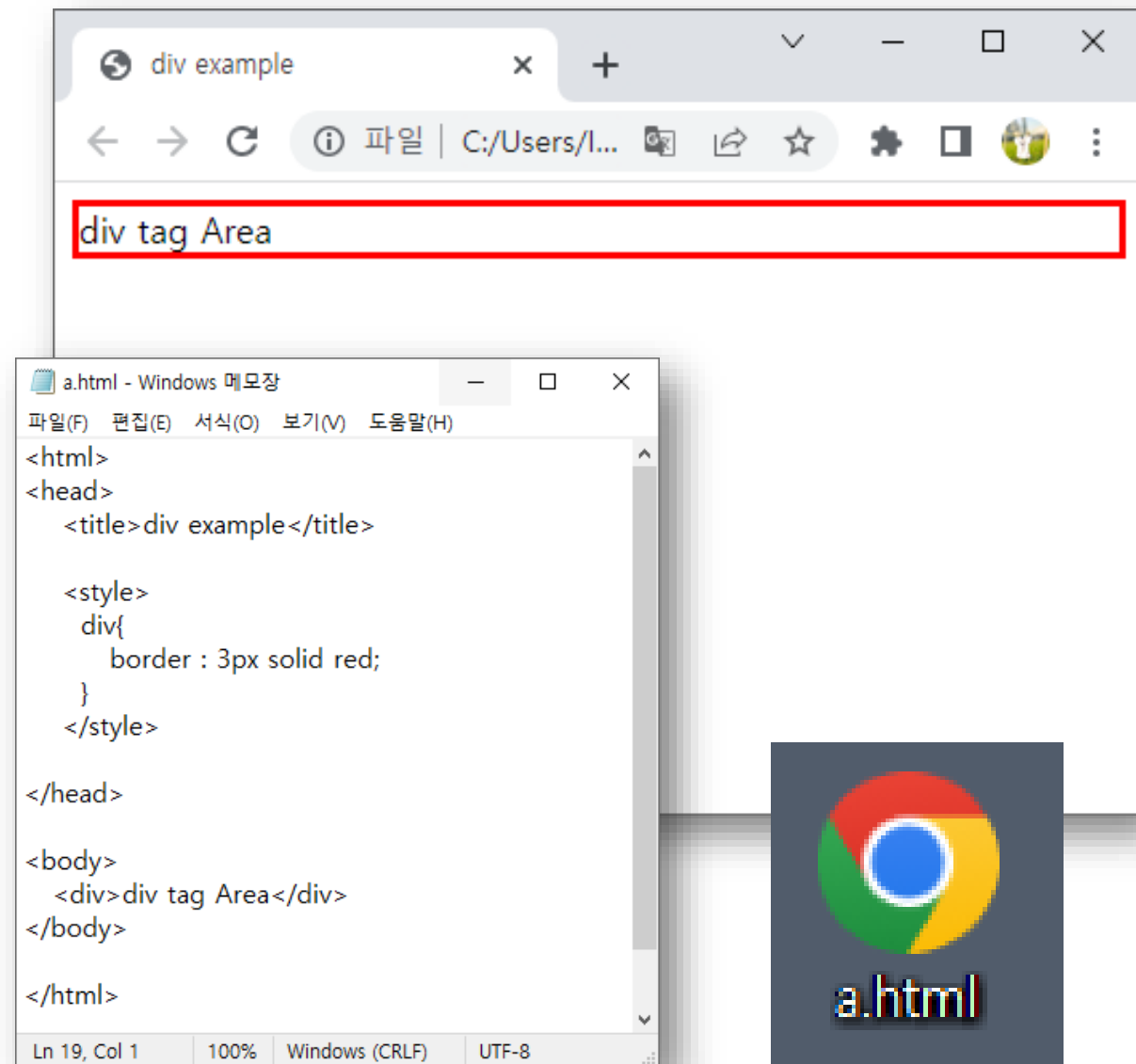
  <style>
    div{
      border : 3px solid red;
    }
  </style>

</head>

<body>
  <div>div tag Area</div>
</body>

</html>
```

<출처> <https://yaegreena.tistory.com/52>



# 웹 페이지 구조(html)

14

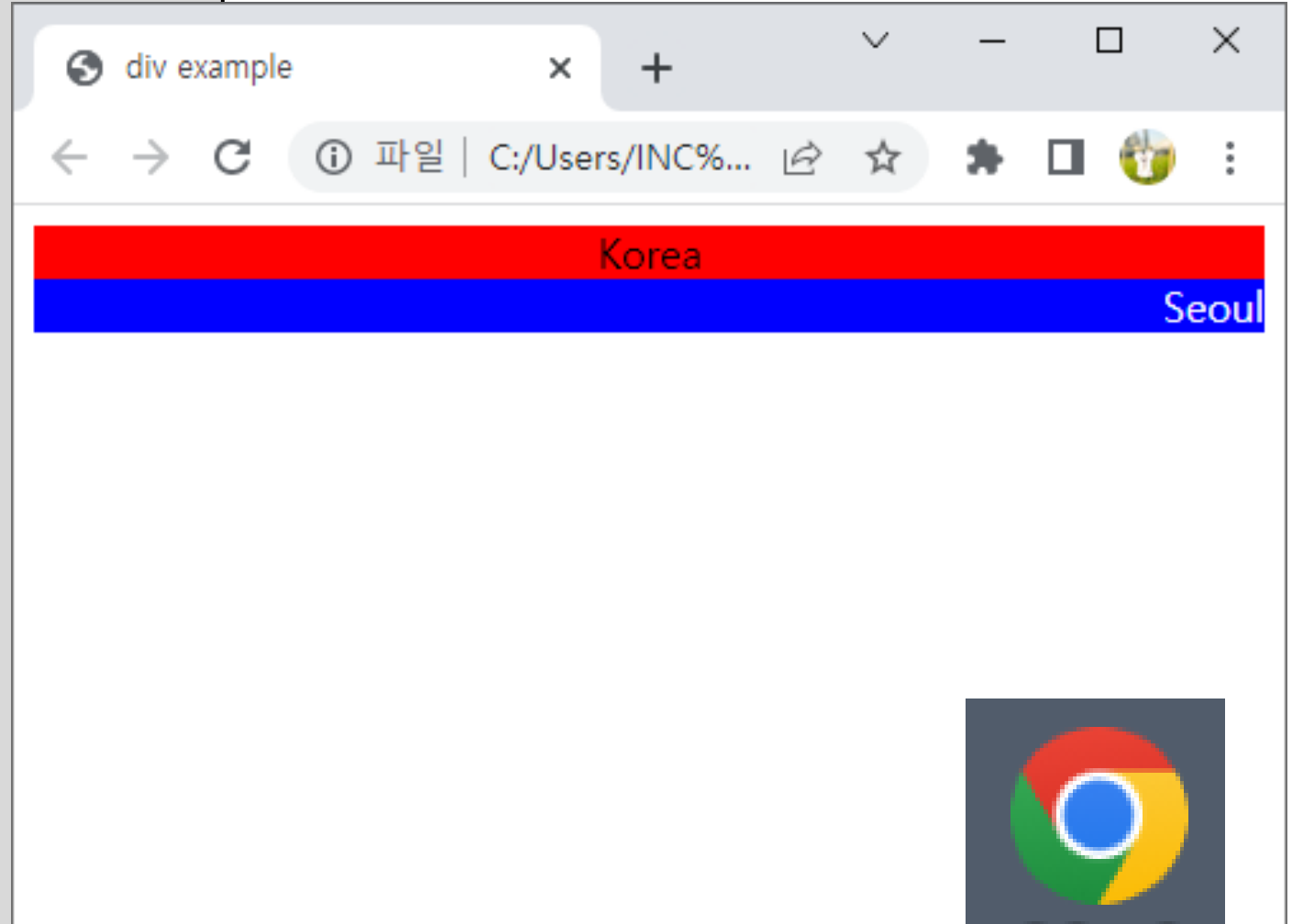
```
<html>
<head>
  <title>div example</title>

  <style>
    .nation {
      background-color : red;
      text-align: center;
    }

    .city {
      background-color : blue;
      text-align: right;
      color:white;
    }
  </style>
</head>

<body>
  <div class="nation">Korea</div>
  <div class="city">Seoul</div>
</body>

</html>
```



<출처> <https://balmostory.tistory.com/100>

# 웹 페이지 구조(html)

15

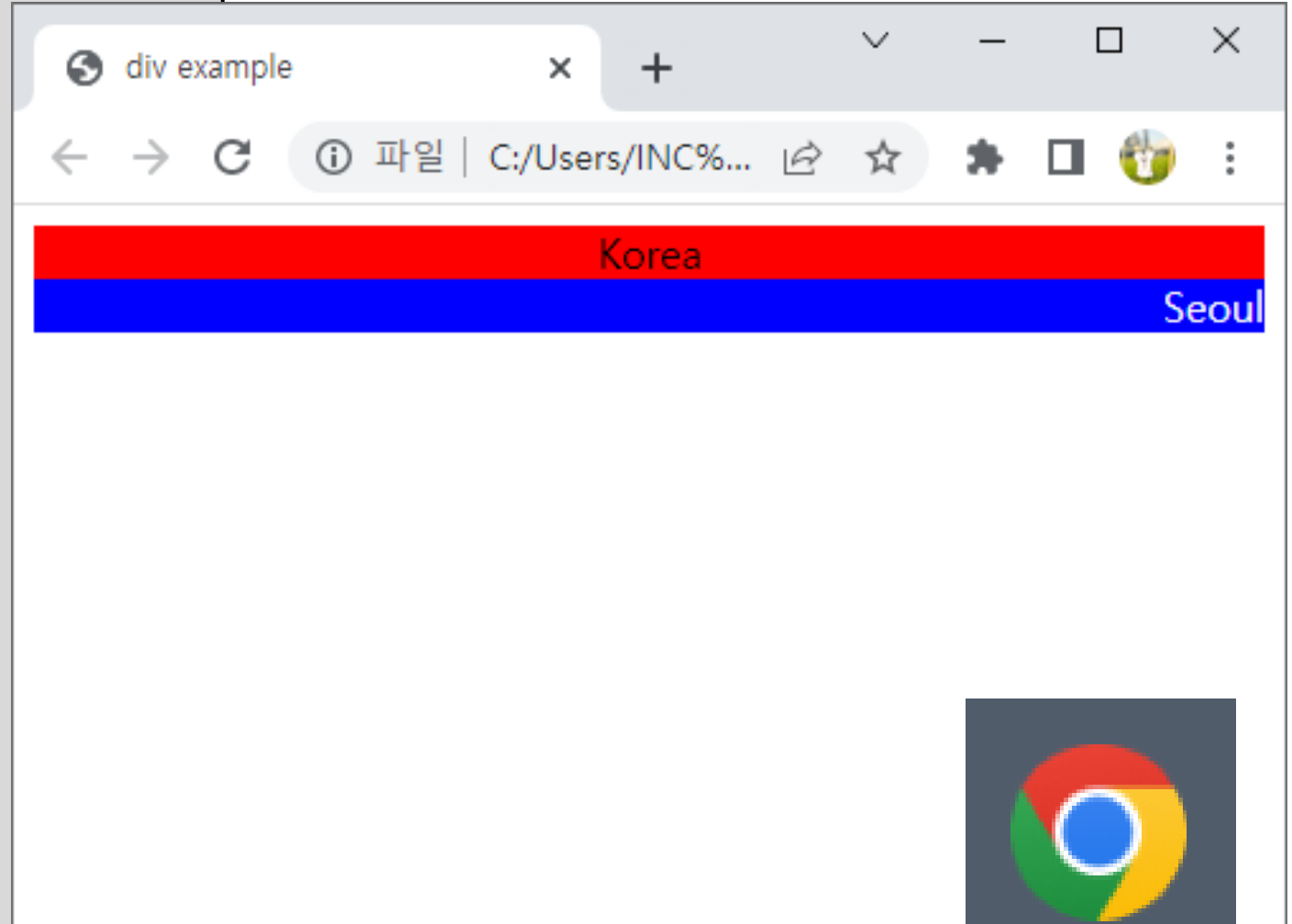
```
<html>
<head>
  <title>div example</title>

  <style>
    #nation {
      background-color : red;
      text-align: center;
    }

    #city {
      background-color : blue;
      text-align: right;
      color:white;
    }
  </style>
</head>

<body>
  <div id="nation">Korea</div>
  <div id="city">Seoul</div>
</body>

</html>
```



<출처> <https://balmostory.tistory.com/100>

```
<html>
<head>
  <title>div example</title>

  <style>
    .nation {
      background-color : red;
      text-align: center;
    }

    .city {
      background-color : blue;
      text-align: right;
      color:white;
    }
  </style>

</head>

<body>
  <div class="nation">Korea</div>
  <div class="city">Seoul</div>
</body>

</html>
```

**.클래스 이름**  
(일반화된 내용 적용시)

```
<html>
<head>
  <title>div example</title>

  <style>
    #nation {
      background-color : red;
      text-align: center;
    }

    #city {
      background-color : blue;
      text-align: right;
      color:white;
    }
  </style>

</head>

<body>
  <div id="nation">Korea</div>
  <div id="city">Seoul</div>
</body>

</html>
```

**#아이디 이름**  
(한 요소에 하나만 적용)



`<ul> ... </ul>` : **u**nordered **l**ist(순서가 없는 목록 태그)

`<ol> ... </ol>` : **o**rdered **l**ist(순서가 있는 목록 태그)

`<li> ... </li>` : **l**ist **i**tem(목록에 해당하는 항목 태그)

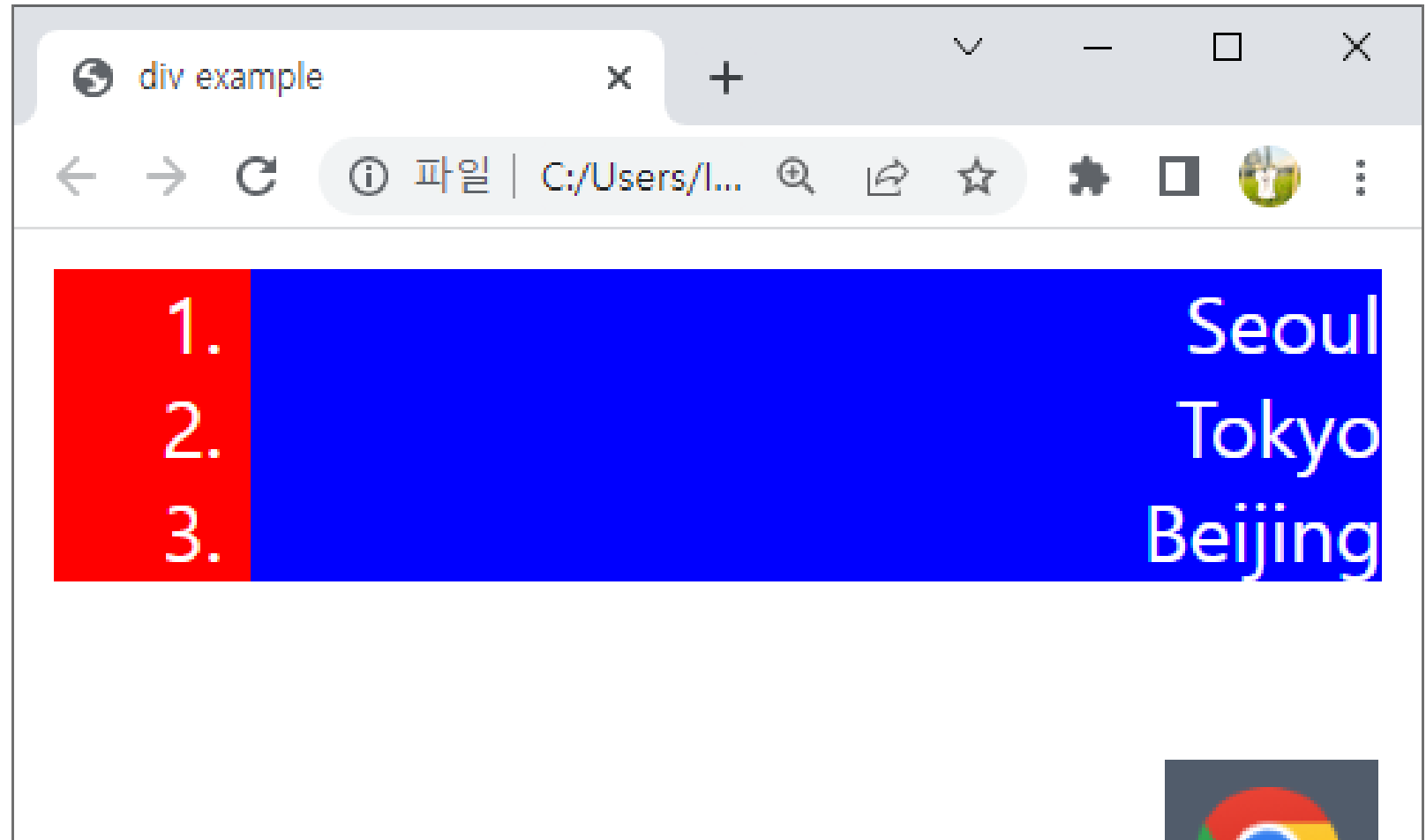
※`<ul>..</ul>`, `<ol>...</ol>` 사이에 사용됨

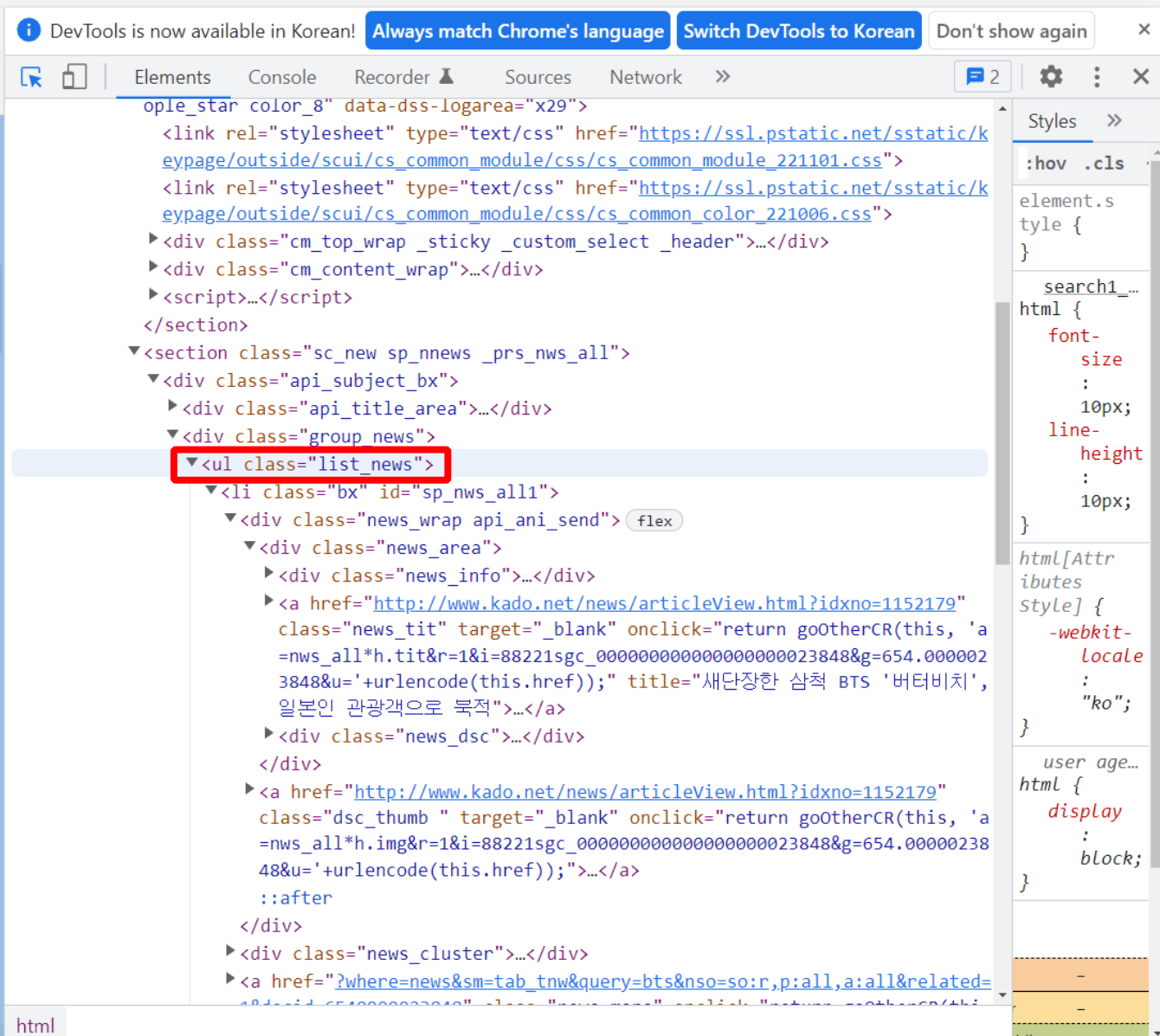
```
<html>
<head>
  <title>div example</title>

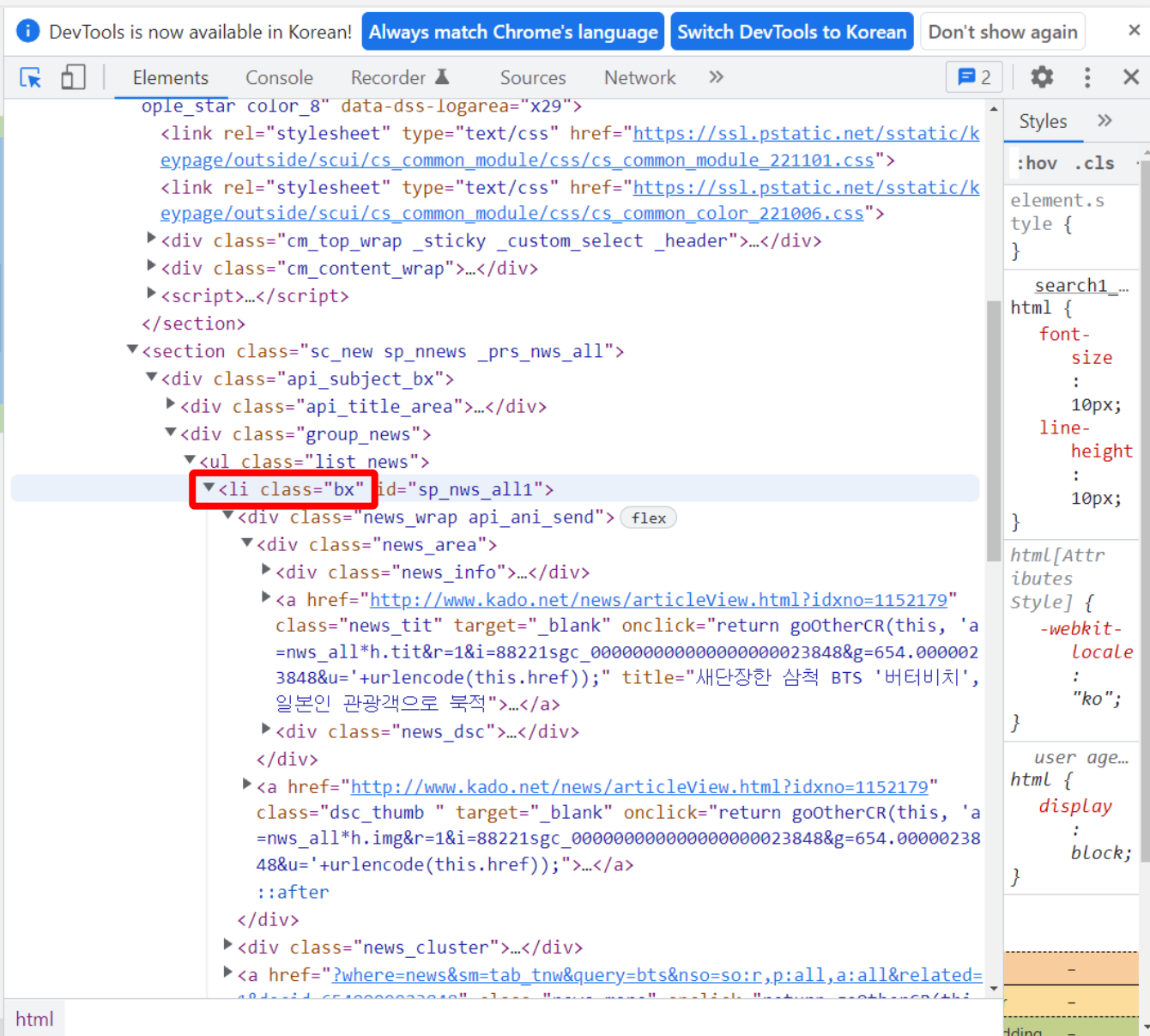
  <style>
    .nation {
      background-color : red;
      text-align: center;
    }

    .city {
      background-color : blue;
      text-align: right;
      color:white;
    }
  </style>
</head>

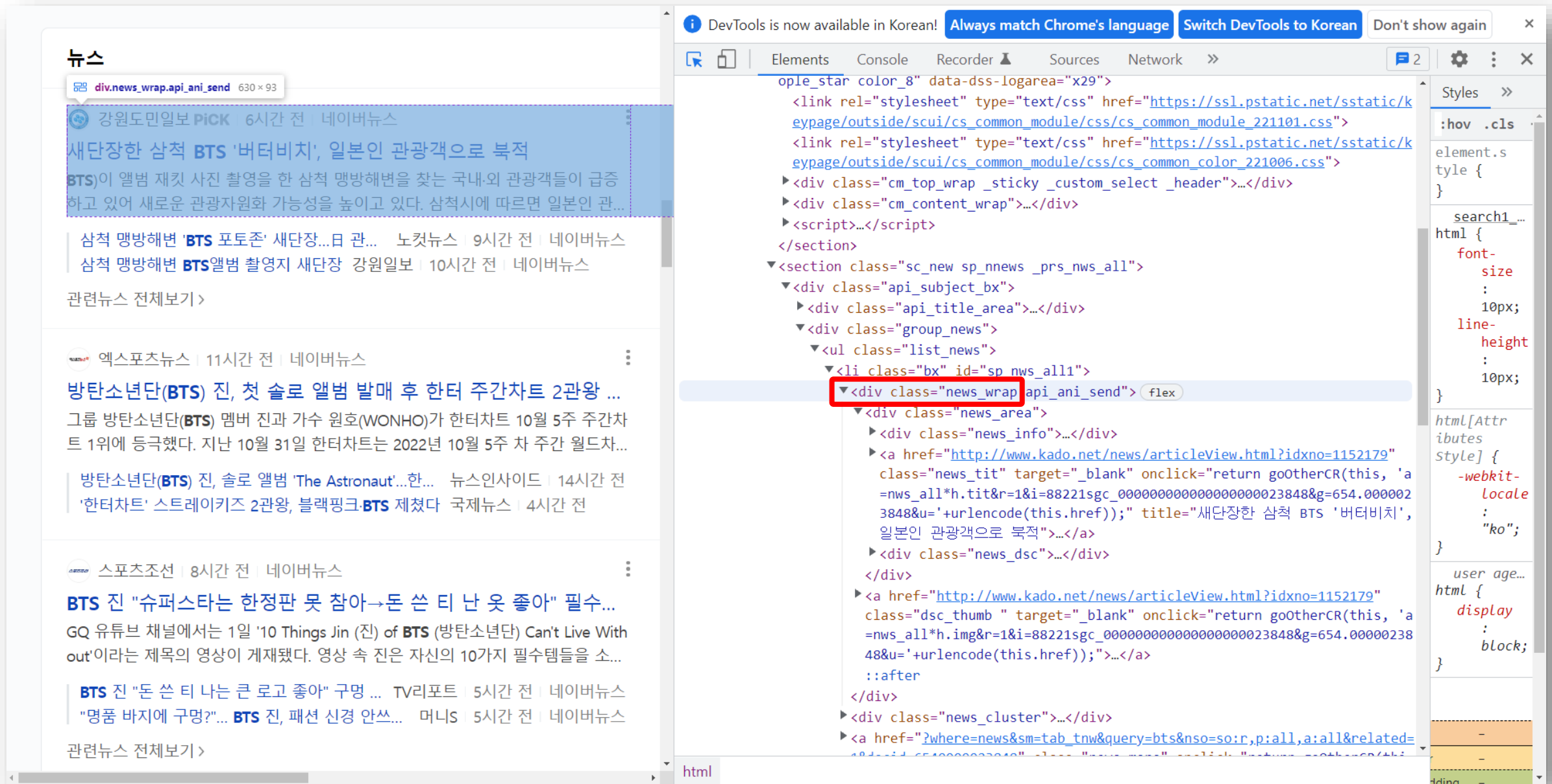
<body>
  <ol class = "nation">
    <li class = "city">Seoul</li>
    <li class = "city">Tokyo</li>
    <li class = "city">Beijing</li>
  </ol>
</body>
</html>
```







21



div.news area 468 × 93

강원도민일보 PiCK 6시간 전 네이버뉴스

새단장한 삼척 **BTS** '버터비치', 일본인 관광객으로 북적

**BTS**이 앨범 재킷 사진 촬영을 한 삼척 맹방해변을 찾는 국내·외 관광객들이 급증하고 있어 새로운 관광자원화 가능성을 높이고 있다. 삼척시에 따르면 일본인 관

삼척 맹방해변 'BTS 포토존' 새단장...目 関... 노컷뉴스 9시간 전 네이버뉴스

삼척 맹방해변 **BTS** 앨범 촬영지 새단장 강원일보 10시간 전 네이버뉴스

[관련뉴스 전체보기](#) >

엑스포츠뉴스 | 11시간 전 | 네이버뉴스

방탄소년단(BTS) 진, 첫 솔로 앨범 발매 후 한터 주간차트 2관왕 ...

그룹 방탄소년단(BTS) 멤버 진과 가수 윈호(WONHO)가 한터차트 10월 5주 주간차트 1위에 등극했다. 지난 10월 31일 한터차트는 2022년 10월 5주 차 주간 월드차...

방탄소년단(BTS) 진, 솔로 앨범 'The Astronaut'...한... 뉴스인사이드 | 14시간 전  
'한터차트' 스트레이키즈 2관왕, 블랙핑크·BTS 제쳤다 국제뉴스 | 4시간 전

스포츠조선 8시간 전 네이버뉴스

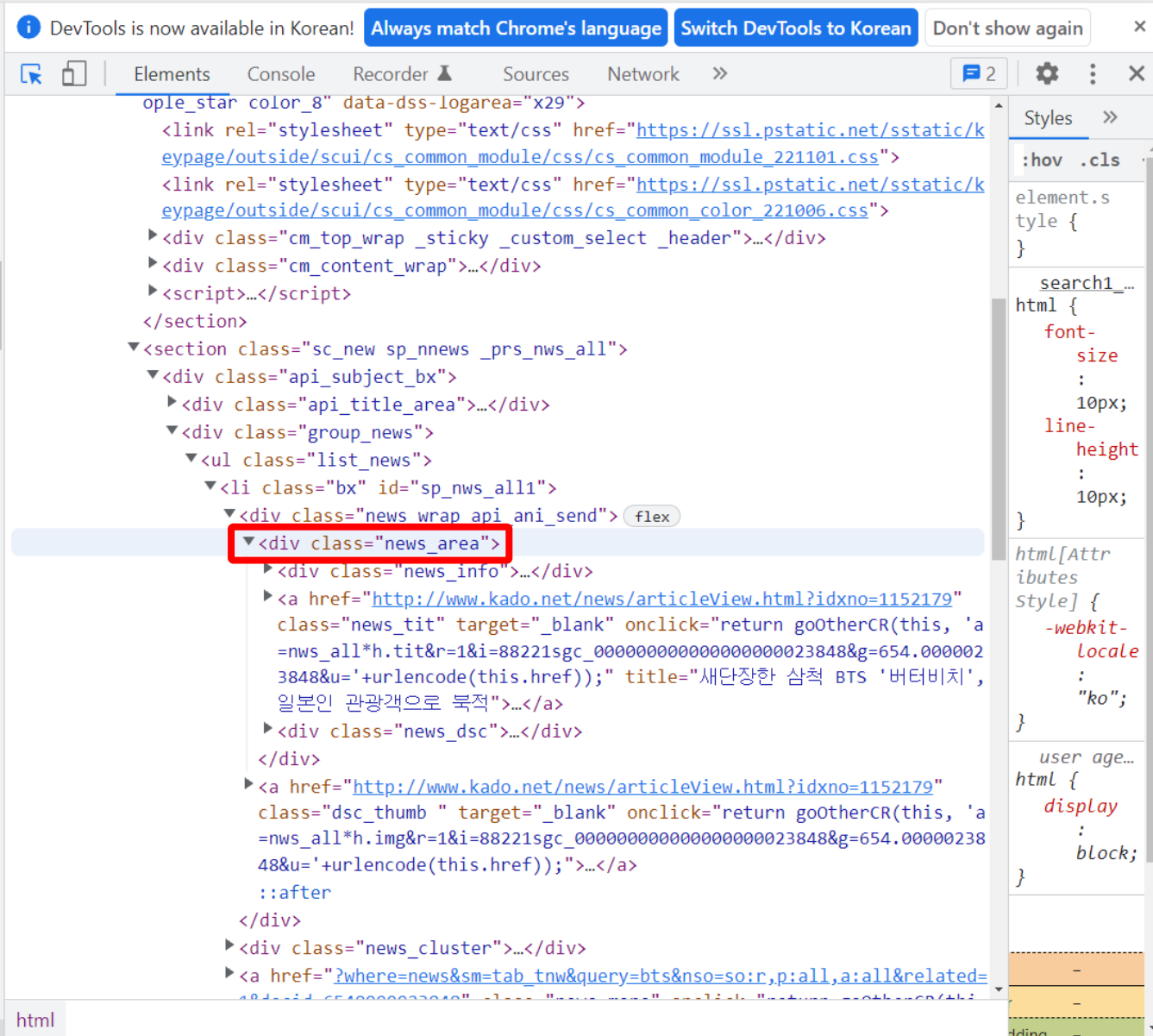
## BTS 진 "슈퍼스타는 한정판 못 참아→돈 쓴 티 난 옷 좋아" 필수...

GQ 유튜브 채널에서는 1일 '10 Things Jin (진) of **BTS** (방탄소년단) Can't Live Without'이라는 제목의 영상이 게재됐다. 영상 속 진은 자신의 10가지 필수템들을 소...

BTS 진 "돈 쓴 티 나는 큰 로고 좋아" 구멍 ... TV리포트 5시간 전 네이버뉴스

"명품 바지에 구멍?"... **BTS** 진, 패션 신경 안쓰... 머니S 5시간 전 | 네이버뉴스

[관련뉴스 전체보기 >](#)





[관련뉴스 전체보기 >](#)

`<a href=?where=news&sm=tab_tnw&query=bts&nso=s:r,p:all,a:all&related=`

... api\_animation div#wrap div#container div#content.pack\_group div#main\_pack.main\_pack section. ..

**크롤링하기**



# 웹페이지 크롤링하기

25

**새단장한 삼척 BTS '버터비치', 일본인 관광객으로 북적**  
BTS가 앨범 재킷 사진 촬영을 한 삼척 맹방해변을 찾는 국내·외 관광객들이 급증하고 있어 새로운 관광자원화 가능성을 높이고 있다. 삼척시에 따르면 일본인 관...

삼척 맹방해변 'BTS 포토존' 새단장... 노컷뉴스 | 9시간 전 | 네이버뉴스  
삼척 맹방해변 BTS 앨범 촬영지 새단장 강원일보 | 10시간 전 | 네이버뉴스  
관련뉴스 전체보기 >

엑스포츠뉴스 | 11시간 전 | 네이버뉴스  
**방탄소년단(BTS) 진, 첫 솔로 앨범 발매 후 한터 주간차트 2관왕...**  
그룹 방탄소년단(BTS) 멤버 진과 가수 원호(WONHO)가 한터차트 10월 5주 주간차트 1위에 등극했다. 지난 10월 31일 한터차트는 2022년 10월 5주 차 주간 월드차...

방탄소년단(BTS) 진, 솔로 앨범 'The Astronaut'...한... 뉴스인사이드 | 14시간 전  
'한터차트' 스트레이키즈 2관왕, 블랙핑크 BTS 제쳤다 국제뉴스 | 4시간 전

스포츠조선 | 8시간 전 | 네이버뉴스  
**BTS 진 "슈퍼스타는 한정판 못 참아→돈 쓴 티 난 옷 좋아" 필수...**

```
[2] import requests
from bs4 import BeautifulSoup

titles = []
search_word = 'bts'
url = f'https://search.naver.com/search.naver?where=nexearch&sm=top_hy&fbm=1&ie=utf8&query={search_word}'
#print(url)
req = requests.get(url)
html = req.text
#print(html)

soup = BeautifulSoup(html, 'html.parser')
search_result = soup.select_one('.list_news')
news_links = search_result.select('.bx > .news_wrap > .news_area > a')
#print(news_links)

for i in news_links:
    titles.append(i.get_text())
print(titles)
```

["새단장한 삼척 BTS '버터비치', 일본인 관광객으로 북적", '방탄소년단(BTS) 진, 첫 솔로 앨범 발매 후 한터 주간차트 2관왕...', 'BTS 진 "슈퍼스타는 한정판 못 참아→돈 쓴 티 난 옷 좋아" 필수...']

```
import requests  
from bs4 import BeautifulSoup
```

**requests** 모듈 : Python에서 HTTP 요청을 보내기 위해 사용

## BeautifulSoup 라이브러리

- 루이스 캐럴의 『이상한 나라의 앨리스』에서 이름을 따옴
- 이야기 속 '모조 거북'이 이 노래를 부름
- BeautifulSoup는 잘못된 HTML을 수정하여 쉽게 탐색할 수 있는 XML 형식의 **파이썬 객체로 변환**하므로 골치 아픈 웹을 탐색할 때 유용함
- (지저분한 구조를 가진 웹문서를 두고 마구 뒤섞인 수프와 같다고 표현함)

아름다운 수프, 풍부한 녹색,  
그릇에서 기다리거라!  
누가 이 맛있는 것에 속이지 않으리?  
저녁 수프, 아름다운 수프!



```
#추출한 제목을 저장할 리스트  
titles = []
```

```
#검색할 단어  
search_word = 'bts'
```

```
#검색할 웹페이지 주소  
url = f'https://search.naver.com/search.naver?where=nexearch&sm=top_h ty&fbm=1&ie=utf8&query={search_word}'
```

```
#검색어가 포함된 웹페이지 주소 확인  
print(url)
```

#requests 라이브러리를 통해 웹페이지 접근

```
req = requests.get(url)
```

```
html = req.text
```

```
print(html)
```

## API

requests 라이브러리는 매우 직관적인 API를 제공하는데요. 어떤 방식(method)의 HTTP 요청을 하느냐에 따라서 해당하는 이름의 함수를 사용하면 됩니다.

- GET 방식: `requests.get()`
- POST 방식: `requests.post()`
- PUT 방식: `requests.put()`
- DELETE 방식: `requests.delete()`

```
[2] import requests
from bs4 import BeautifulSoup

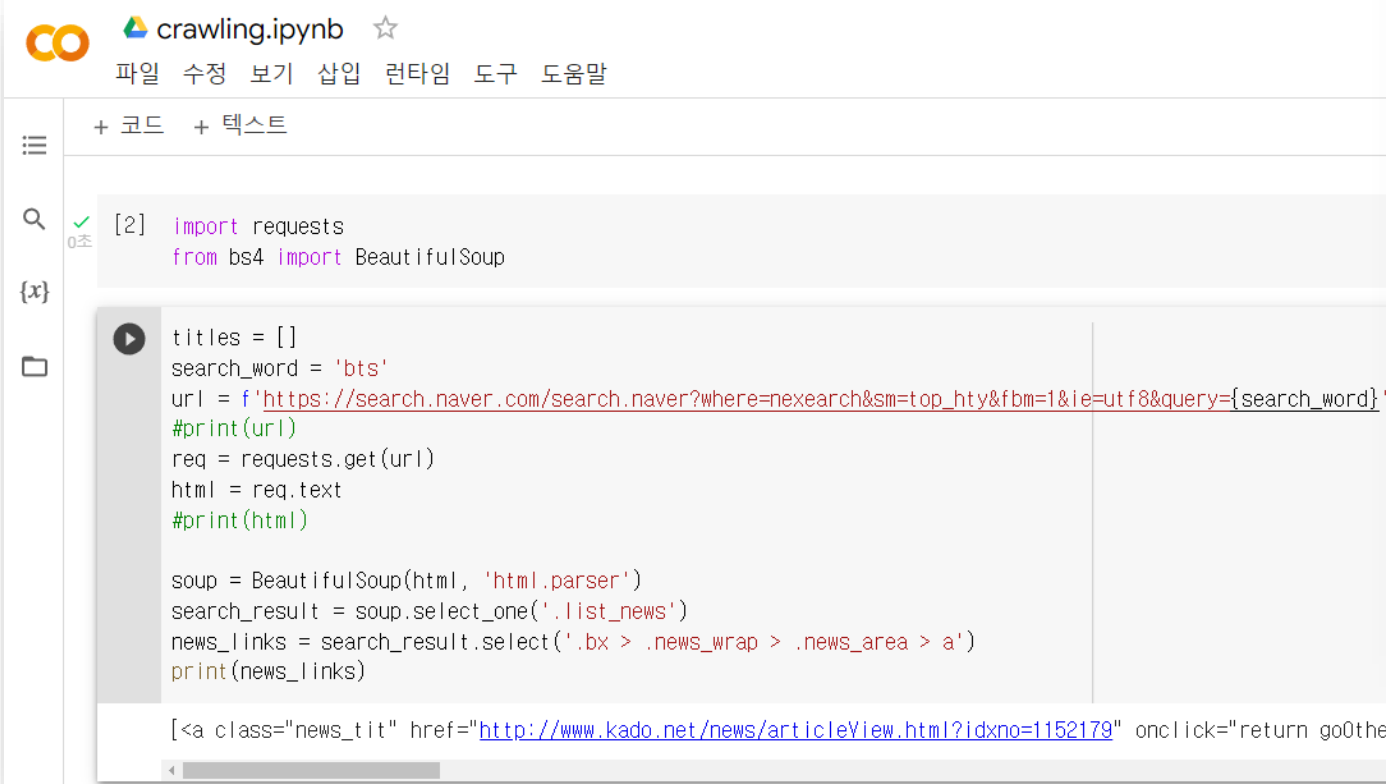
titles = []
search_word = 'bts'
url = f'https://search.naver.com/search.naver?where=nexearch&sm=top_hly&fbm=1&ie=utf8&query={search_word}'
#print(url)
req = requests.get(url)
html = req.text
print(html)
```

```
<!doctype html> <html lang="ko"> <head> <meta charset="utf-8"> <meta name="referrer" content="always"> <meta name="format-detection" content="telephone=no,address=no,email=no"> <meta name="viewport"
  <span class="area_page" data-kgs-page-item></span>
</script> <script>(function(){var startApplication=function(){var require=window.require.config({"context":"'노출ID'로 변경해주세요.", "paths":{"@lapin-plus/flicking-page":"https://ssl.pstatic.r

<script> /* [AU] 인라인 스크립트 */ (function () { var CLASS_SECTION = "_sp_ntotal"; var CLASS_MARGIN = "type_margin"; var sectionEls = Array.prototype.slice.call(document.querySelectorAll("sectio
<span class="etc">드라마 장르 전문</span>
<span class="etc">팬 <span class="_fan_count">1.3만+</span></span></div></div><div class="my_keyword_area"><div class="my_type_overview"><div class="keyword_list"><span class="bx"><span class="keyw
<span class="bx"><span class="keyword">한국 드라마-500편 이상 리뷰</span></span></div><div class="my_type_more"><a href="#" role="button" aria-expanded="false" aria-haspopup="true" class="btn_more_btn_expand"
<a target="_blank" href="https://in.naver.com/shyhjin/contents/internal/495113136773504?areacode=ink*A&query=bts" class="dsc_link_cross_trigger_foryou_trigger" data-cr-gdid="a0209r14_nblog_post_2
<span class="etc">캐주얼 스타일</span>
<span class="etc">팬 <span class="_fan_count">5,681</span></span></div></div><div class="my_keyword_area"><div class="my_type_overview"><div class="keyword_list"><span class="bx"><span class="keyw
<span class="bx"><span class="keyword">국내 브랜드 선호</span></span></div>
<span class="bx"><span class="keyword">날씬한 체형</span></span></div>
```

<출처> <https://www.daleseo.com/python-requests/>

```
soup = BeautifulSoup(html, 'html.parser')
search_result = soup.select_one('.list_news')
news_links = search_result.select('.bx > .news_wrap > .news_area > a')
print(news_links)
```



```

▼<section class="sc_new sp_nnews _prs_nws_all">
  ▼<div class="api_subject_bx">
    ▶<div class="api_title_area">...</div>
    ▼<div class="group_news">
      ▼<ul class="list_news">
        ▼<li class="bx" id="sp_nws_all1">
          ▼<div class="news_wrap" api_ani_send"> flex
            ▼<div class="news_area">
              ▶<div class="news_info">...</div>
              ▼<a href="http://www.kado.net/news/articlev
                class="news_tit" target="_blank" onclick="
                =nws_all*h.tit&r=1&i=88221sgc_0000000000000
                3848&u='+urlencode(this.href));" title="새
                일본인 관광객으로 북적">...</a> == $0
              ▶<div class="news_dsc">...</div>

```

```
for i in news_links:  
    titles.append(i.get_text())  
print(titles)
```

for 변수 in 리스트:  
 반복할 코드

```
a = [38, 21, 53, 62, 19]  
for i in a :  
    print(i)
```

```
38  
21  
53  
62  
19
```

# 웹페이지 크롤링하기

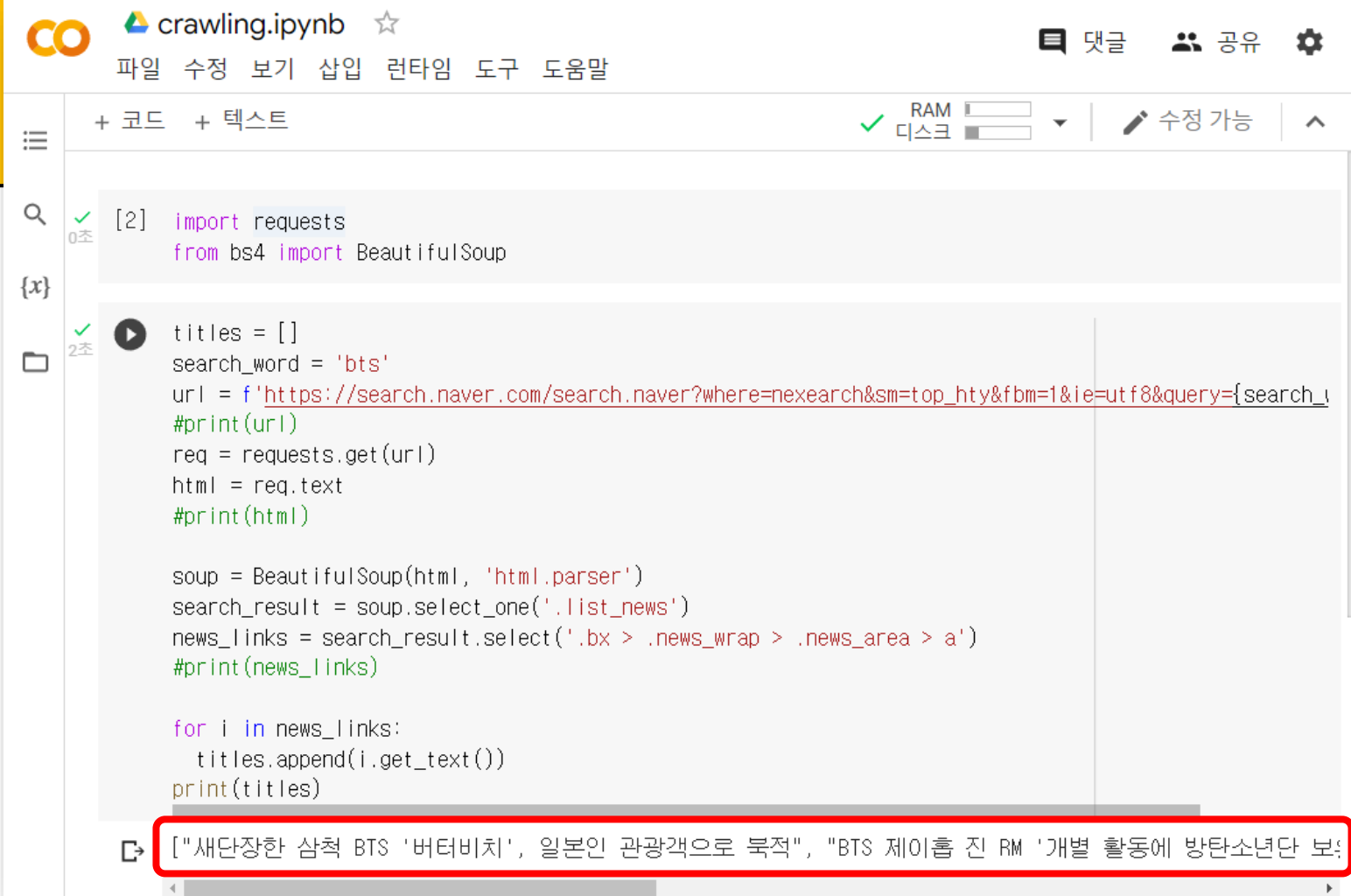
31

```
for i in news_links:  
    titles.append(i.get_text())  
print(titles)
```

for 변수 in 리스트:  
 반복할 코드

```
a = [38, 21, 53, 62, 19]  
for i in a:  
    print(i)
```

38  
21  
53  
62  
19



The screenshot shows a Jupyter Notebook titled 'crawling.ipynb'. The code cell contains the following Python code:

```
[2] import requests  
from bs4 import BeautifulSoup  
  
titles = []  
search_word = 'bts'  
url = f'https://search.naver.com/search.naver?where=nexearch&sm=top_hy&fbm=1&ie=utf8&query={search_u'  
#print(url)  
req = requests.get(url)  
html = req.text  
#print(html)  
  
soup = BeautifulSoup(html, 'html.parser')  
search_result = soup.select_one('.list_news')  
news_links = search_result.select('.bx > .news_wrap > .news_area > a')  
#print(news_links)  
  
for i in news_links:  
    titles.append(i.get_text())  
print(titles)
```

The output cell shows the result of the script execution, which is a list of news titles. The first title is highlighted with a red box:

```
["새단장한 삼척 BTS '버터비치', 일본인 관광객으로 북적", "BTS 제이홉 진 RM '개별 활동에 방탄소년단 보"]
```

- 크롤링의 뜻과 사용하는 이유를 설명할 수 있다.
  - 크롤링(crawling)은 '기다'라는 뜻의 crawl의 명사형
  - 인터넷을 돌아다니며 정보를 수집해 오는 작업. 그러한 작업을 하는 소프트웨어 : 크롤러(crawler)
  - 월드와이드웹에서 웹페이지의 데이터를 '긁어' 오는 행위 : 스크레이핑(scraping)
- 웹 페이지의 구조를 설명할 수 있다.
  - div : division
  - .클래스 이름 : 일반화된 내용 적용시
  - #아이디 이름 : 한 요소에 하나만 적용
- 크롤링을 통해 웹 페이지의 정보를 리스트에 저장할 수 있다.
  - requests 모듈 : Python에서 HTTP 요청을 보내기 위해 사용
  - BeautifulSoup 라이브러리 : XML 형식의 파이썬 객체로 변환



**Q & A**

**감사합니다**