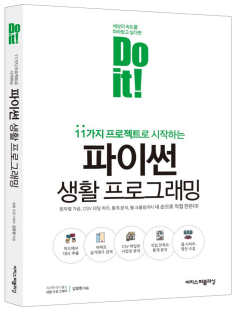
세상의 속도를
따라잡고 싶다면

Do it!

파이썬
생활 프로그래밍



이지스퍼블리싱(주)

06

Web Crawling으로 정보 모으기

06-3 Portal Site에서 기사 crawling하기

06-3 Portal Site에서 기사 crawling 하기

이런 상황이라면?

- Portal Site에서 머리기사 정보 모으기

06-3 Portal Site에서 기사 crawling 하기

Web Crawling 기본 환경 준비하기

- 필요한 모듈 임포트

```
>>> import os, re
>>> import urllib.request as ur
>>> from bs4 import BeautifulSoup as bs
```

06-3 Portal Site에서 기사 crawling 하기

Web Crawling 기본 환경 준비하기

- 저장할 파일 위치와 접속할 url 저장

```
>>> os.chdir(r'C:\Users\user\python')

>>> news = 'https://news.daum.net/'
```

06-3 Portal Site에서 기사 crawling 하기

Web Crawling 기본 환경 준비하기

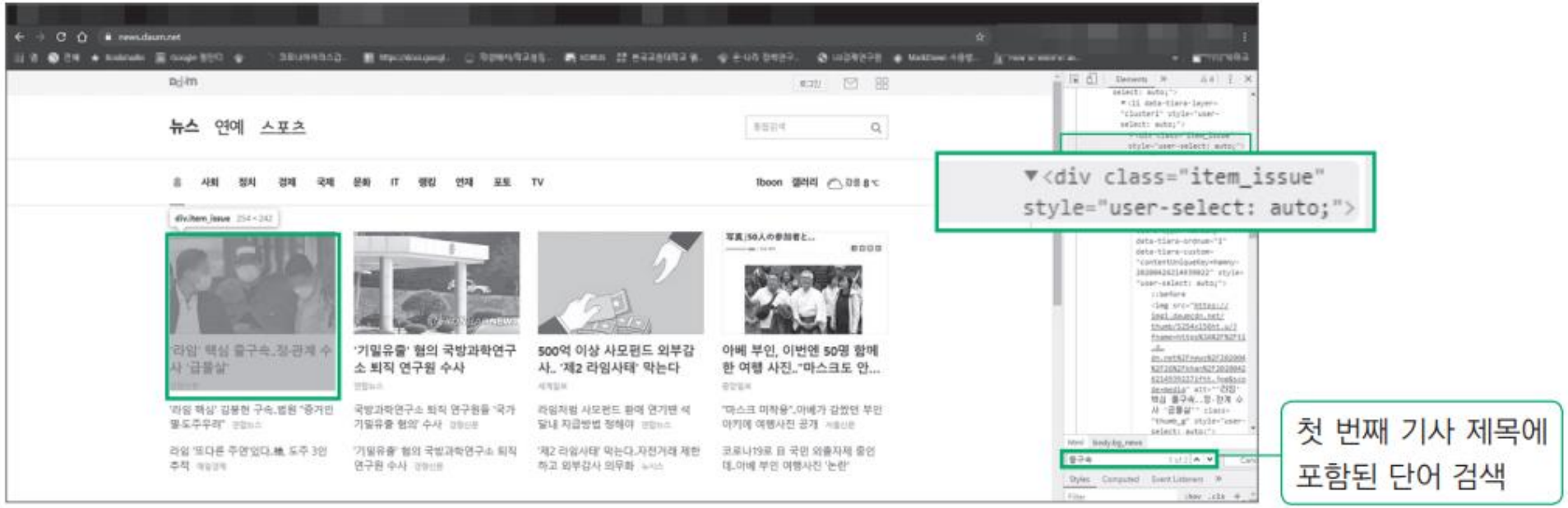
- html 파싱

```
>>> soup = bs(ur.urlopen(news).read(), 'html.parser')
```

06-3 Portal Site에서 기사 crawling 하기

머리기사 제목 추출하기

- 기사 제목에 들어있는 단어 검색



06-3 Portal Site에서 기사 crawling 하기

머리기사 제목 추출하기

- find_all로 내용 추출

```
soup.find_all('div', {"class" : "item_issue"})
```

06-3 Portal Site에서 기사 crawling 하기

반복문으로 기사 제목 모두 추출하기

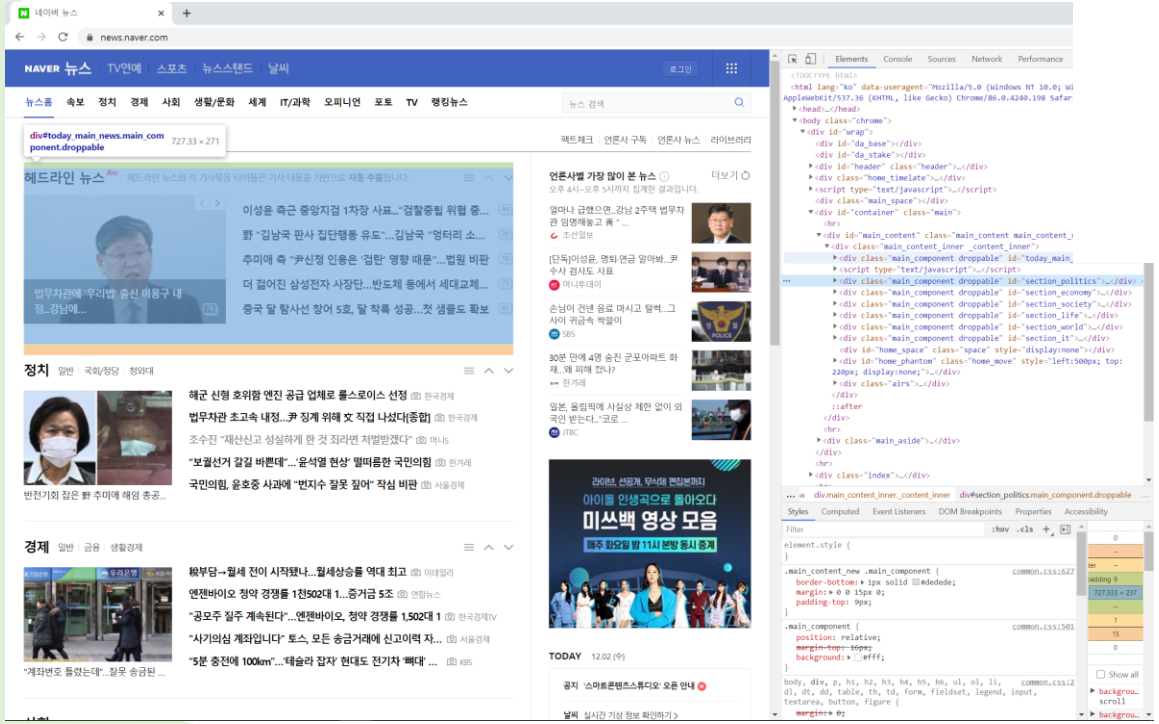
- 추출한 div 안의 텍스트를 모두 출력

```
>>> for i in soup.find_all('div', {"class":"item_issue"}):
        i.text

'\n\'라임 핵심\' 김봉현 구속..법원 "증거인멸·도주우려"\n연합뉴스\n'
"\n라임 '또다른 주연'있다..檢, 도주 3인 추적\n매일경제\n"
"\n국방과학연구소 퇴직 연구원들 '국가 기밀유출 혐의' 수사\n경향신문\n"
"\n'기밀유출' 혐의 국방과학연구소 퇴직 연구원 수사\n경향신문\n"
```

06-3 Portal Site에서 기사 crawling 하기

머리기사 제목 추출하기

- 기사 제목에 들어있는 단어 검색




```
▼<div id="container" class="main">
  <hr>
 ▼<div id="main_content" class="main_content main_content_new">
  ▼<div class="main_content_inner _content_inner">
   ▶<div class="main_component droppable" id="today_main_news">…</div>
   ▶<script type="text/javascript">…</script>
   ▶<div class="main_component droppable" id="section_politics">…</div>
   ▶<div class="main_component droppable" id="section_economy">…</div>
   ▶<div class="main_component droppable" id="section_society">…</div>
   ▶<div class="main_component droppable" id="section_life">…</div>
   ▶<div class="main_component droppable" id="section_world">…</div>
   ▶<div class="main_component droppable" id="section_it">…</div>
    <div id="home_space" class="space" style="display:none"></div>
   ▶<div id="home_phantom" class="home_move" style="left:500px; top:
   220px; display:none;">…</div>
   ▶<div class="airs">…</div>
```

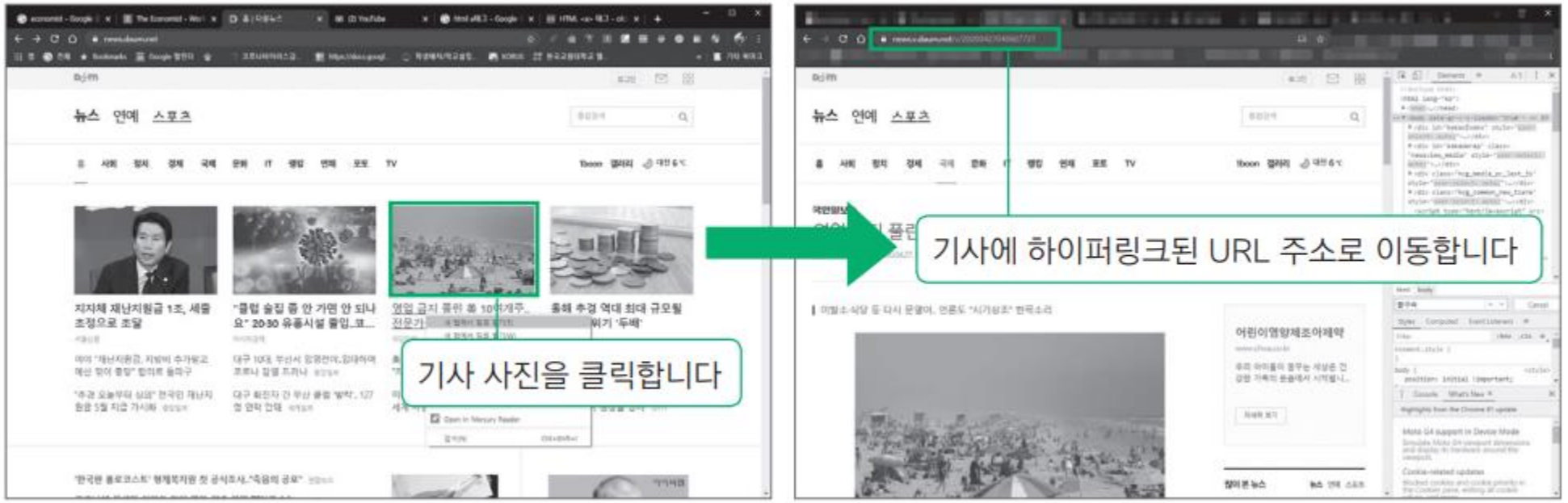
```
soup.find_all('div', {"class" : "main_component droppable"})
→
```

```
>>> Warning : ('Connection aborted.', RemoteDisconnected('Remote end closed connection wit
hout response'))
>>>
```

06-3 Portal Site에서 기사 crawling 하기

hyperlink 주소 추출하기

- 어떤 요소를 누를 경우 다른 링크로 이동시키는 것이 hyperlink



06-3 Portal Site에서 기사 crawling 하기

hyperlink 주소 추출하기

- <a> tag 사용법 알아보기

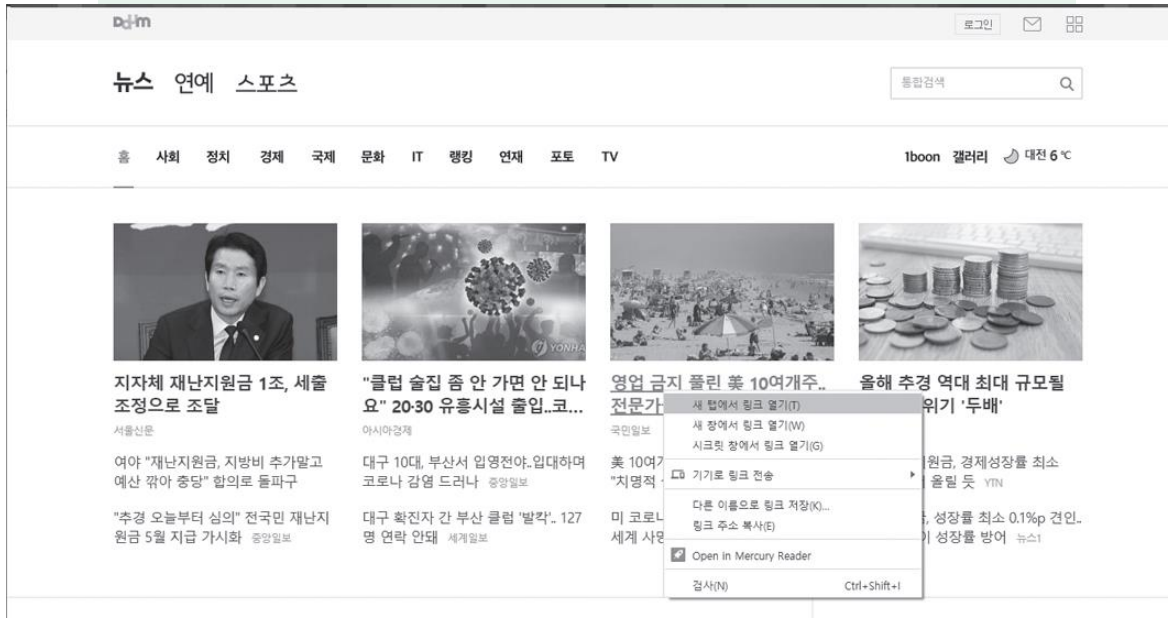
```
<a href = "링크할 URL 주소">하이퍼링크 텍스트</a>
```

```
<HTML>
  <body>
    <a href="http://www.naver.com"> 네이버 바로가기 </a>
  </body>
</HTML>
```

06-3 Portal Site에서 기사 crawling 하기

hyperlink 주소 추출하기

- 새 탭에서 링크 열기가 있으면 <a> tag를 사용한 것



06-3 Portal Site에서 기사 crawling 하기

hyperlink 주소 추출하기

- <a> tag만 추출하기

```
soup.find_all('a')
```

06-3 Portal Site에서 기사 crawling 하기

hyperlink 주소 추출하기

- href 속성값 추출하기
- 이동할 대상 url을 지정하는 속성
- get() 함수를 통하여 속성 얻기 가능

```
a.get('속성')
```

06-3 Portal Site에서 기사 crawling 하기

hyperlink 주소 추출하기

- href 속성값 추출하기
- 이동할 대상 url을 지정하는 속성
- get() 함수를 통하여 속성 얻기 가능

```
>>> for i in soup.find_all('a')[:5]:
        i.get('href')

'#kakaoBody'
'#kakaoGnb'
'https://news.daum.net/'
'https://entertaindaum.net'
'https://sports.media.daum.net/sports'
```

06-3 Portal Site에서 기사 crawling 하기

원하는 영역에서 hyperlink 모두 추출하기

- 파싱 가능하게 준비

```
>>> news = 'https://news.daum.net/'
>>> soup = bs(ur.urlopen(news).read(), 'html.parser')
```

06-3 Portal Site에서 기사 crawling 하기

원하는 영역에서 hyperlink 모두 추출하기

- find_all로 <a> tag 추출하기

```
>>> for i in soup.find_all('div', {"class":"item_issue"}):
        i.find_all('a')

[<a class="link_thumb" data-tiara-custom="contentUniqueKey=hamny-20200427060512538"
data-tiara-id="20200427060512538" data-tiara-layer="article_thumb" data-tiara-ord-
num="1" data-tiara-type="harmony" href="https://news.v.daum.net/v/20200427060512538">
<img alt="정부, 무급휴직 주한미군 근로자 임금 선지급 방침" class="thumb_g" src="https://
img1.daumcdn.net/thumb/S254x156ht.u/?fname=https%3A%2F%2Ft1.daumcdn.net%2Fnews%2F202
004%2F27%2Fsegye%2F20200427060512905mrmi.jpg&amp;scode=media"/>
</a>, <a class="link_txt" data-tiara-custom="contentUniqueKey=hamny-20200427060512538"
data-tiara-id="20200427060512538" data-tiara-layer="article_main" data-tiara-ord-
num="1" data-tiara-type="harmony" href="https://news.v.daum.net/v/20200427060512538">
정부, 무급휴직 주한미군 근로자 임금 선지급 방침</a>]
[<a class="link_thumb" data-tiara-custom="contentUniqueKey=hamny-20200427060438523"
data-tiara-id="20200427060438523" data-tiara-layer="article_thumb" data-tiara-ord-
num="1" data-tiara-type="harmony" href="https://news.v.daum.net/v/20200427060438523">
...(생략)...
```

06-3 Portal Site에서 기사 crawling 하기

원하는 영역에서 hyperlink 모두 추출하기

- find_all로 <a> tag 추출 후 바로 get은 list 형식이기에 불가능

```
>>> for i in soup.find_all('div', {"class":"item_issue"}):
        i.find_all('a').get('href')

Traceback (most recent call last):
  File "<pyshell#85>", line 2, in <module>
    i.find_all('a').get('href')
  File "C:\Users\user\Anaconda3\lib\site-packages\bs4\element.py", line 1884, in __
getattr__
    "ResultSet object has no attribute '%s'. You're probably treating a list of items
like a single item. Did you call find_all() when you meant to call find()?" % key
AttributeError: ResultSet object has no attribute 'get'. You're probably treating a
list of items like a single item. Did you call find_all() when you meant to call find()?
```

06-3 Portal Site에서 기사 crawling 하기

원하는 영역에서 hyperlink 모두 추출하기

- 인덱스를 지정 후 get 사용 가능

```
>>> for i in soup.find_all('div', {"class":"item_issue"}):
        i.find_all('a')[0]


<a  class="link_thumb"  data-tiara-custom="contentUniqueKey=hamny-20200427060512538"
data-tiara-id="20200427060512538"  data-tiara-layer="article_thumb"  data-tiara-ord-
num="1" data-tiara-type="harmony" href="https://news.v.daum.net/v/20200427060512538">
<img  alt="정부, 무급휴직 주한미군 근로자 임금 선지급 방침"  class="thumb_g"  src="https://
img1.daumcdn.net/thumb/S254x156ht.u/?fname=https%3A%2F%2Ft1.daumcdn.net%2Fnews%2F202
004%2F27%2Fsegye%2F20200427060512905mrmi.jpg&amp;scode=media"/>
</a>
...(생략)...
```

06-3 Portal Site에서 기사 crawling 하기

원하는 영역에서 hyperlink 모두 추출하기

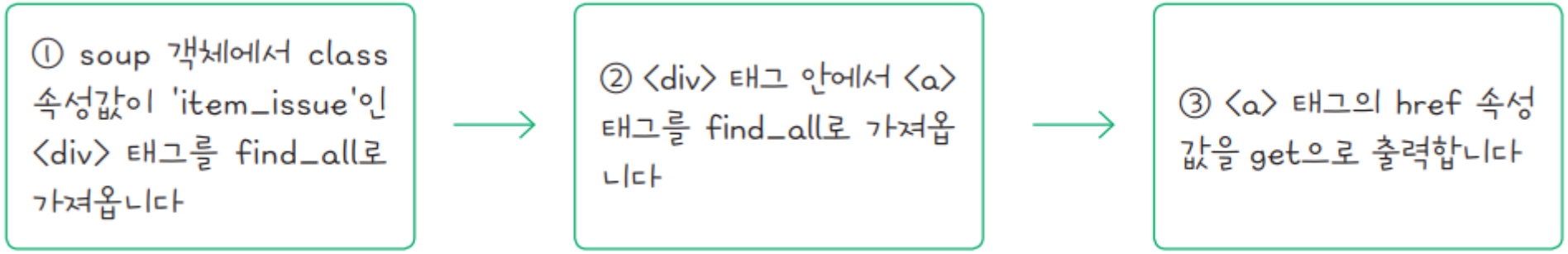
- 인덱스를 지정 후 get 사용 가능

```
>>> for i in soup.find_all('div',{"class":"item_issue"}):
        i.find_all('a')[0].get('href')


'https://news.v.daum.net/v/20200426214939022'
'https://news.v.daum.net/v/20200426205218584'
'https://news.v.daum.net/v/20200426201134187'
'https://news.v.daum.net/v/20200426180251978'
```

06-3 Portal Site에서 기사 crawling 하기

원하는 영역에서 hyperlink 모두 추출하기

- hyperlink를 추출하는 과정



06-3 Portal Site에서 기사 crawling 하기

기사 제목과 내용 한번에 추출하기

- 기사 URL 얻기

06-3 Portal Site에서 기사 crawling 하기

기사 제목과 내용 한번에 추출하기

- URL 저장 후 BeautifulSoup로 열기

```
>>> article1 = 'https://go.seoul.co.kr/news/newsView.php?id=20200427004004&wlog_
tag3=daum'
```

```
>>> soup2 = bs(ur.urlopen(article1).read(), 'html.parser')
```

06-3 Portal Site에서 기사 crawling 하기

기사 제목과 내용 한번에 추출하기

- 기사 내용 가져오기
- <p> tag 안만 확인하면 확인 가능

```
>>> for i in soup2.find_all('p'):
        print(i.text)

오늘 여야 예결특위서 2차 추경 심사…전 국민 지원 확대로 4조 6000억 증액

여야가 27일부터 국회 상임위원회와 예산결산특별위원회를 열고 코로나19 대응 긴급재난지원금 지급을 위한 2차 추가경정예산(추경)안 심사에 들어가기로 했다. 또 전 국민에게 재난지원금을 지급하기 위해 추가로 필요한 재원 가운데 지방정부가 부담할 예정이었던 1조원을 세출 조정을 통해 조달키로 했다. 더불어민주당 이인영, 미래통합당 심재철 원내대표는 26일 국회에서 이 같은 입장을 밝혔다.이 원내대표는 “심 원내대표가 ‘지방정부가 당초 부담하기로 했던 1조원 규모라도 세출 조정을 통해 마련하면 어떻겠느냐’고 요청했다”면서 “긴급하게 기획재정부 담당자를 불러 그게 가능한지를 상의했고 최종적으로 가능하게 하기로 정리했다”고 말했다.

...(생략)...
```

06-3 Portal Site에서 기사 crawling 하기

기사 제목과 내용 한번에 추출하기

- 기사 제목 가져오기
- class 속성이 ‘item_issue’인 div 안에 존재

```
>>> headline = soup.find_all('div', {"class" : "item_issue"})
```

```
>>> print(headline[0].text)

지자체 재난지원금 1조, 세출 조정으로 조달

서울신문
```

06-3 Portal Site에서 기사 crawling 하기

hyperlink된 모든 기사의 제목과 본문 추출하기

- 머리 기사의 제목 추출

```
>>> for i in headline:
        print(i.text, '\n')

지자체 재난지원금 1조, 세출 조정으로 조달
서울신문

"클럽 술집 좀 안 가면 안 되나요" 20·30 유흥시설 출입…코로나19 확산 우려"
아시아경제

영업 금지 풀린 美 10여개주… 전문가들 “치명적 실수”
국민일보

올해 추경 역대 최대 규모될 듯… 금융위기 ‘두배’
한국일보
```

06-3 Portal Site에서 기사 crawling 하기

hyperlink된 모든 기사의 제목과 본문 추출하기

- 기사 제목 출력 후 연결된 링크를 통하여 기사 본문 URL 정보 얻기

```
>>> for i in headline:

        print(i.text, '\n')

        soup3 = bs(ur.urlopen(i.find_all('a')[0].get('href')).read(), 'html.parser')
        for j in soup3.find_all('p'):
            j.text
```

해당 기사의 URL 주소

web_crawling_2.py

File Edit Format Run Options Window Help

```python
#
# 참조 : http://hleecaster.com/python-web-crawling-with-beautifulsoup/
# "11가지 프로젝트로 시작하는 생활프로그래밍", 이창현 저, 이지스퍼블리싱, 2020
#

D = True
D_1 = False

import os, re
import usecsv
import requests
import urllib.request as ur

from bs4 import BeautifulSoup as bs

try:
    os.chdir(r'C:₩과소사₩과소사-강의예제₩web_crawling')

    '''
    news = "https://news.yahoo.com/"
    news = "https://www.chosun.com/"
    news = "https://news.naver.com/"
    '''
    news = "https://news.daum.net/"

    webpage = requests.get(news)
    if D:
        print("₩n1) >> webpage : ", webpage)

    soup = bs(webpage.content, 'html.parser')
    if D:
       print("₩n2) >> soup : ", soup)

    # 기사 제목 추출하기
    # find_all로 <div> 내용 추출
    # class 속성이 'item_issue'인 div 안에 존재
    if D:
        print("₩n4) >> 기사 제목 추출하기")

    headline = soup.find_all('div', {"class" : "item_issue"})

    for i in headline:
        print(i.text, "₩n")

    # find_all로 <a> tag 추출하기
    if D:
        print("₩n5) >>  find_all로 <a> tag 추출하기 : ")

    for i in soup.find_all('a')[:5]:
        print(i.get('href'))

    # 원하는 영역에서 하이퍼링크 모두 추출하기
    # 인덱스를 지정 후 get 사용 가능
    if D:
        print("₩n6) >> 원하는 영역에서 하이퍼링크 모두 추출하기")

    for i in headline:
        print(i.find_all('a')[0].get('href'))


    # 기사 제목 출력 후 연결된 링크를 통하여 기사 본문 URL 정보 얻기
    if D:
        print("₩n7) >> 기사 제목 출력 후 연결된 링크를 통하여 기사 본문 URL 정보 얻기")
    for i in headline:
        if D:
            print("₩n7-1) >> 상위기사 제목 : ")
        print(i.text, "₩n")

        new_link = i.find_all('a')[0].get('href')
        link_webpage = requests.get(new_link)
        if D:
            print("₩n7-2) >> new_link : ", new_link)
            print("₩n7-3) >> link_webpage : ", link_webpage)

        soup_link = bs(link_webpage.content, 'html.parser')
        for j in soup_link.find_all('p'):
            print(j.text, "₩n")

    # Portal Site에서 기사 crawling 하기
    # 기사 제목과 내용 한번에 추출하기
    if D:
        print("₩n8) >> Portal Site에서 기사내용 crawling 하기")
        print("₩n       기사 제목과 내용 한번에 추출하기")

    article_link = 'http://go.seoul.co.kr/news/newsView.php?id=20201109014005'

    article = requests.get(article_link)
    if D:
        print("₩n9) >> article : ", article)

    soup_article = bs(article.content, 'html.parser')
    if D:
       print("₩n10) >> soup_article : ", soup_article)
       print("₩n11) >> 전체기사 내용")

    for i in soup_article.find_all('p'):
        print(i.text)


except Exception as e:
    print("₩n>>> Warning : ", e)
```

Python 3.8.5 Shell

File Edit Shell Debug Options Window Help

```
Python 3.8.5 (tags/v3.8.5:580fbb0, Jul 20 2020, 15:43:08) [MSC v.1926 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:₩과소사₩과소사-강의예제₩web_crawling₩web_crawling_2.py ==========

1) >> webpage :  <Response [200]>

2) >> soup :  Squeezed text (1204 lines).

4) >> 기사 제목 추출하기



문대통령 "코로나 극복의지 담은 예산 통과.. 여야에 감사"

국민일보



복귀한 윤석열, 월성 원전 영장부터 승인..파장

머니투데이



조슈아 웡 13.5개월..'홍콩 민주화' 청년 3명 실형 선고

한겨레
```

Ln: 3 Col: 4

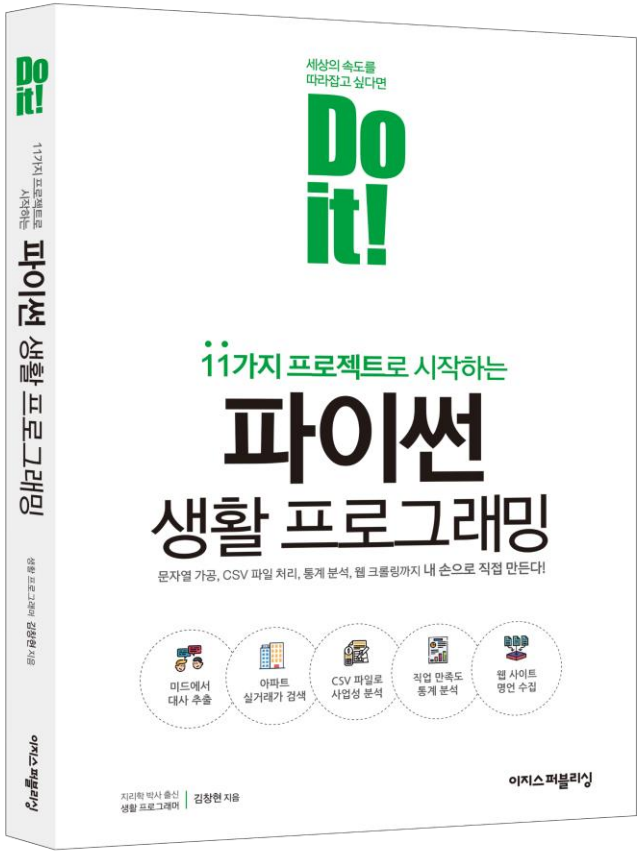
Python 3.8.5 Shell

File Edit Shell Debug Options Window Help

```
로써 금연 정책에 대한 변혁이 시작될 것으로 기대하고 있다"면서 "앞으로 방배동, 서초동, 반포동, 잠
원동 등으로 확대하겠다"고 밝혔다.이민영 기자 min@seoul.co.kr

2020-11-09 14면



김인호 서울시의회 의장, 온라인 청소년
김인호 서울시의회 의장, 안전한 수능 실
서울특별시의회 김인호 의장(더불어민주당, 동대문3)은 오는 12월 3일에 치러지는 2021학년도 대학수
학능력시험의…
정윤경 경기도의회 교육위원장,도교육청
장현국 경기도의회 의장, 일일 소원이로


조은희 서초구청장 서울시장 출마 "여성가산점 안 받고 실력
"지금은 남성·여성보다 일 잘하는 일꾼 필요"
정무부시장·구청장 등 서울행정 10년 경험
내일 부동산·세금 문제 등 입장 발표 예정
김종인 "文정부 비판보다 시민 마음 얻길"

"공공원룸 베란다는 주거인권… 국유지에 주택 공급"
쪽방촌 재개발하는 김영종 종로구청장

수험생 지원!… 광진, 고3 1인당 마스크 10장씩
학원·교습소 등 815곳도 16만장 전달
수능 당일 수험생 수송 상황실 운영

"장애인 배려·주민 편의 윈윈 복지관"
[현장 행정] 은평 2호 '우리장애인복지관' 개관

최신 장비 시설로 장애인들 복지 향상
주민 편의시설 체력단련실·카페 갖춰
초기 주민들 반대 어려움 딛고 문열어
김미경 구청장 "장애인 행복한 삶 기여"

은평, 문체부 등 평가·공모사업 성적 탁월
금천구의 자랑, 청소년상담복지센터 '안전
'라이브 관악' 구독·인증 땐 추첨 통해
코로나 블루 날리는 강동 '희망의 빛'…
수험생 지원!… 광진, 고3 1인당 마스크 10
중구 구민 여러분, 마음 안녕하신가요

자료 제공 : 정책브리핑 korea.kr
주소 : 100-745 서울시 중구 세종대로 124 (태평로1가  25번지) 서울신문사빌딩 I 대표전화 : (02) 20
00-9000

>>>
```

Ln: 643 Col: 20

감사합니다


세상의 속도를
따라잡고 싶다면
Do it!
11가지 프로젝트로 시작하는
파이썬
생활 프로그래밍
문자열 가공, CSV 파일 처리, 통계 분석, 웹 크롤링까지 내 손으로 직접 만든다!
미드에서 대사 추출
아파트 실거래가 검색
CSV 파일로 사업성 분석
직업 만족도 통계 분석
웹 사이트 명언 수집
지리학 박사 출신 생활 프로그래머 | 김창현 지음
이지스퍼블리싱