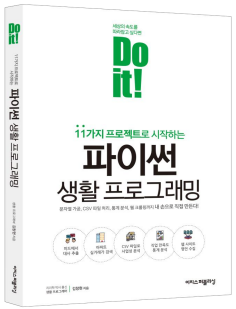
세상의 속도를
따라잡고 싶다면

Do it!

파이썬
생활 프로그래밍



이지스퍼블리싱(주)

06

Web Crawling으로 정보 모으기

06-4 Program 실행 file 만들기

06-4 Program 실행 file 만들기

URL 주소 저장하기

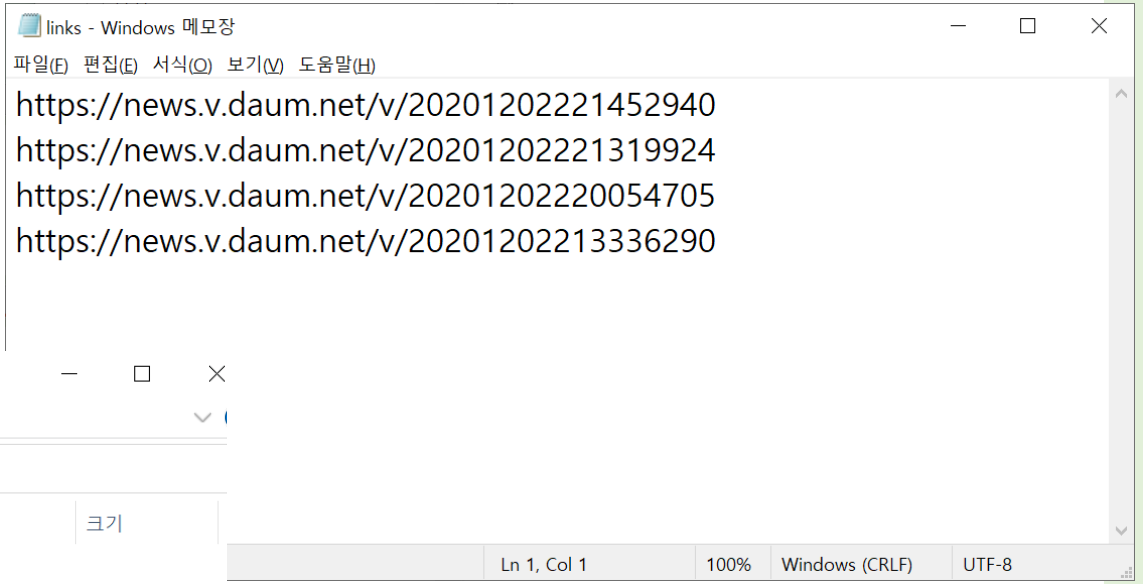
- 출력한 기사의 URL 주소 저장하기

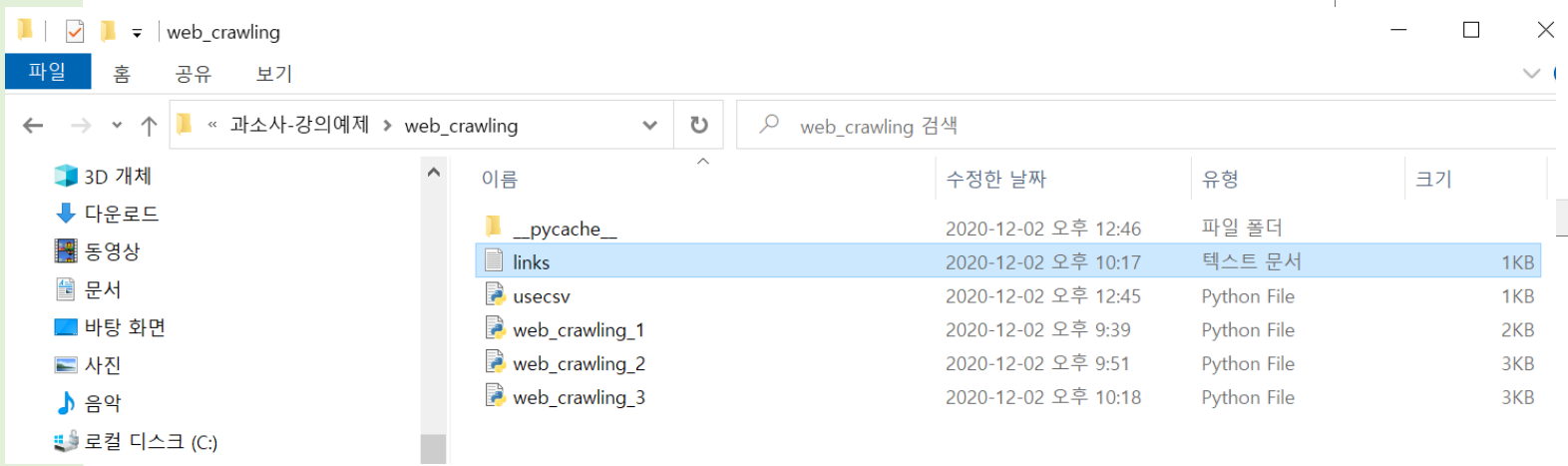
```
>>> os.chdir(r'C:\Users\user\python')
>>> f = open('links.txt', 'w')
>>> for i in soup.find_all('div', {"class":"item_issue"}):
        f.write(i.find_all('a')[0].get('href')+'\n' )


44
44
44
44
>>> f.close()
```

06-4 Program 실행 file 만들기

URL 주소 저장하기

- 저장한 file 확인





06-4 Program 실행 file 만들기

기사 본문을 file로 저장하기

- Python으로 기사의 내용 저장

```
>>> article1 = 'https://news.v.daum.net/v/20200430135751773'

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

>>> f = open('article_1.txt', 'w')

>>> for i in soup.find_all('p'):
        f.write(i.text)

85
100
21
0
203
249
109
177
183
112
26
>>> f.close()
```

06-4 Program 실행 file 만들기

기사 본문을 file로 저장하기

- 저장한 file 확인
- → 다음 페이지 예제와 통합해서 실행

06-4 Program 실행 file 만들기

기사 제목, 본문, hyperlink를 file로 저장하기

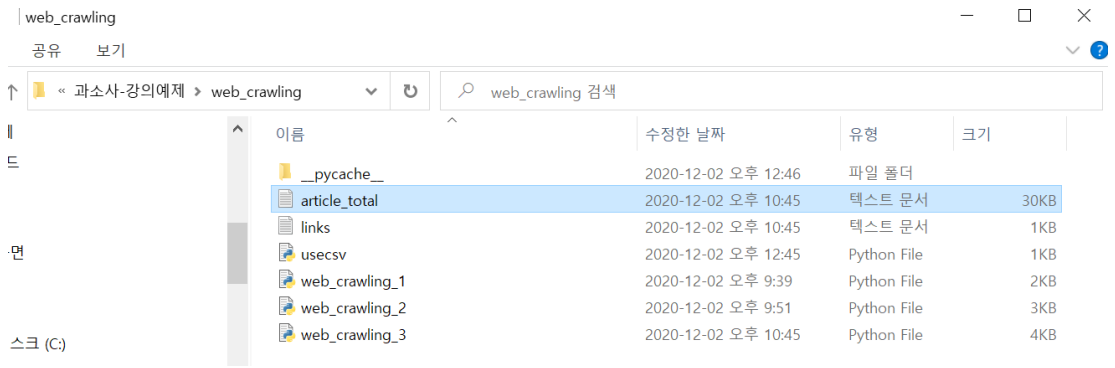
- 제목, 링크, 내용 순으로 저장
- File open 시 ‘w’ 또는 ‘a’ mode

```
>>> url = 'https://news.daum.net/'
>>> soup = bs(ur.urlopen(url).read(),'html.parser')
>>> f = open('article_total.txt', 'w')
>>> for i in soup.find_all('div', {"class":"item_issue"}):
        try:

            f.write(i.text + '\n')

            f.write(i.find_all('a')[0].get('href') + '\n')

            soup2 = bs(ur.urlopen(i.find_all('a')[0].get('href')).read(),'html.parser')

            for j in soup2.find_all('p'):
                f.write(j.text)

            except:
            pass


>>> f.close()
```

06-4 Program 실행 file 만들기

기사 제목, 본문, hyperlink를 file로 저장하기

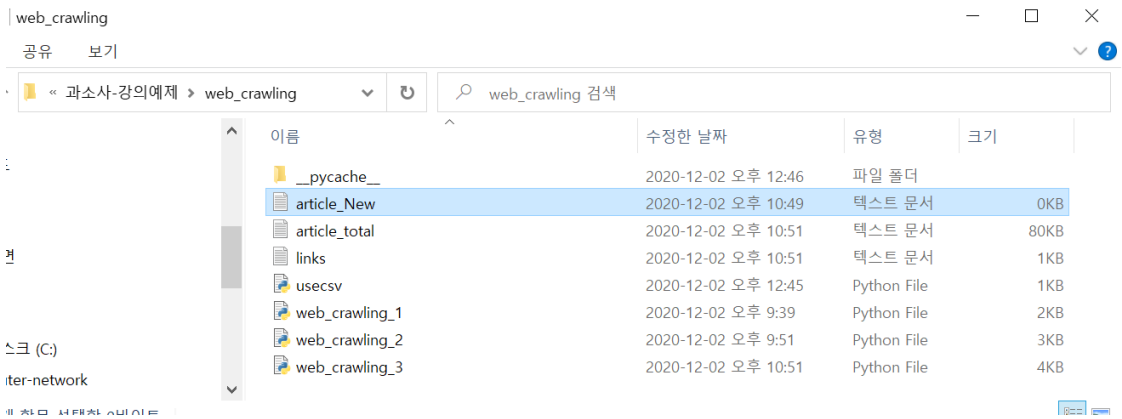
- 저장한 file 확인

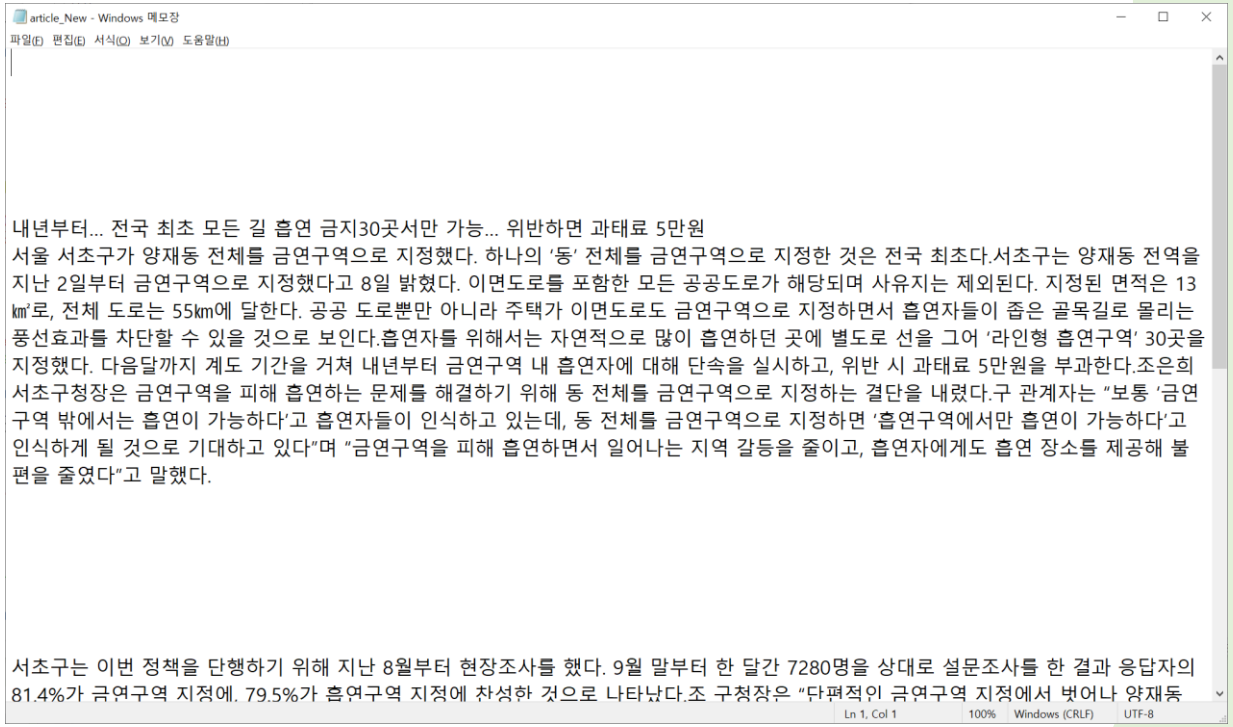


06-4 Program 실행 file 만들기

기사 제목, 본문, hyperlink를 file로 저장하기

- 저장한 file 확인





web_crawling_3.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.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
    # 추출한 hyperlink들을 file f에 저장
    if D:
        print("₩n6) >> 원하는 영역에서 하이퍼링크 모두 추출하여 file로 저장")

    # 한글 처리시 error 발생가능. UNICODE로 처리
    # -1 은  buffer
    f = open('links.txt', 'w', -1, "utf-8")

    for i in headline:
        print(i.find_all('a')[0].get('href'))
        f.write(i.find_all('a')[0].get('href')+'₩n')

    f.close()

    # 기사 제목 출력 후 연결된 링크를 통하여 기사 본문 URL 정보 얻은 후 file에 저장
    if D:
        print("₩n7) >> 기사 제목 출력 후 연결된 링크를 통하여 기사 본문 URL 정보 얻은 후 ₩
file에 저장")

    f_1 = open('article_total.txt', 'a', -1, "utf-8")

    for i in headline:
        if D:
            print("₩n7-1) >> 상위기사 제목 : ")
        print(i.text, "₩n")
        f_1.write(i.text)

        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")
            f_1.write(j.text)

    f_1.close()
```

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

    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")

    f_2 = open('article_New.txt', 'w', -1, "utf-8")

    for i in soup_article.find_all('p'):
        print(i.text)
        f_2.write(i.text)

    f_2.close()


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

Python 3.8.5 Shell

File Edit Shell Debug Options Window Help

```
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 원윈 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
인터넷서울신문에 게재된 콘텐츠의 무단 전재/복사/배포 행위는 저작권법에 저촉되며 위반 시 법적 제
재를 받을 수 있습니다.
Copyright © 서울신문사 All rights reserved.
>>>
```

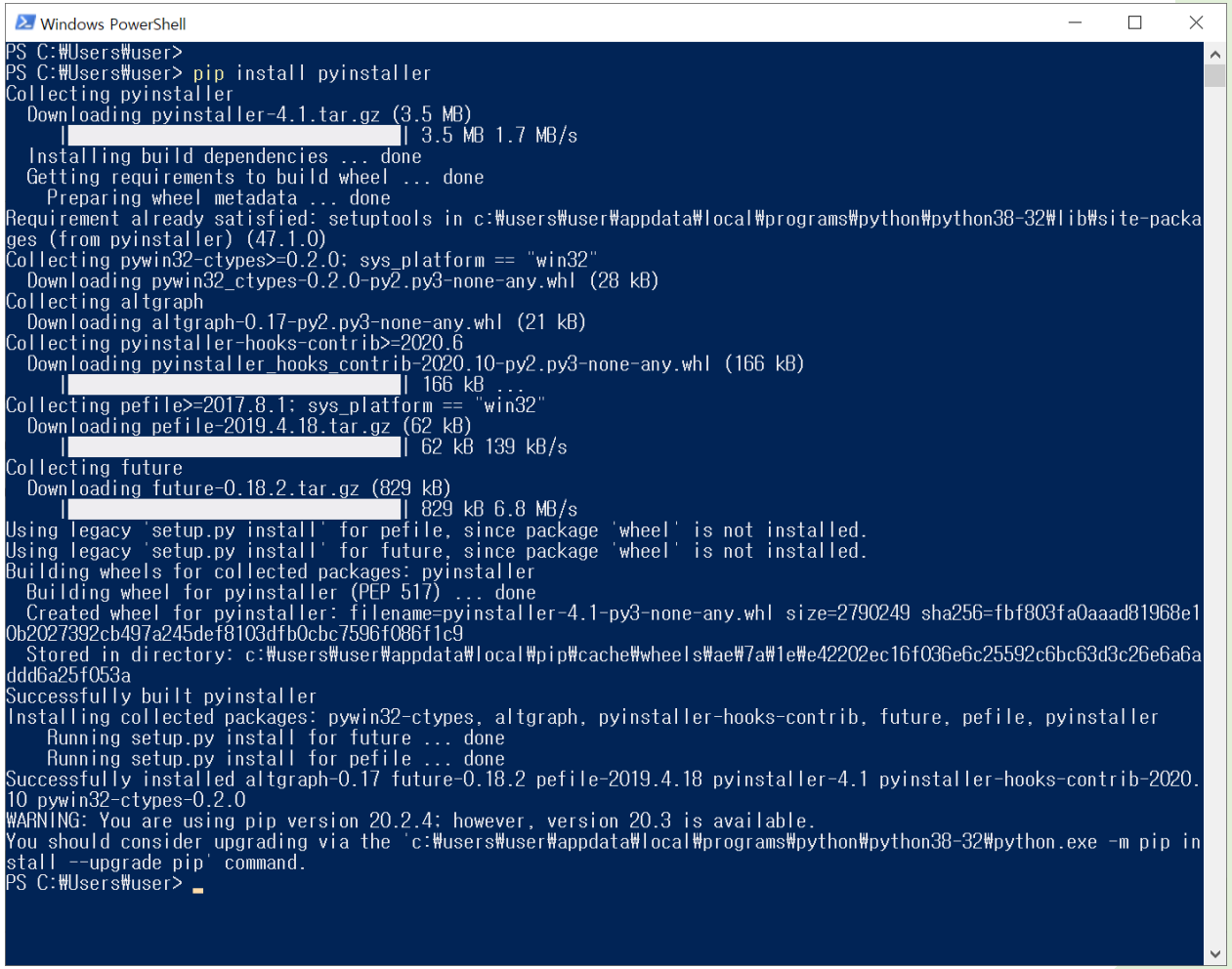
Ln: 2639 Col: 4

06-4 Program 실행 file 만들기

Web Crawling 실행 file 만들기

- pip로 pyinstaller 설치하기

```
C:\Users\user>pip install pyinstaller
```



06-4 Program 실행 file 만들기

Web Crawling 실행 file 만들기

- 작성한 코드 Python(.py) file로 저장

```
import os , codecs, re, datetime, requests
import urllib.request as ur
from bs4 import BeautifulSoup as bs
os.chdir(r'C:\Users\user\python')
url = 'https://news.daum.net/'
f = open(str(datetime.date.today()) + 'articles.txt','w')
soup = bs(ur.urlopen(url).read(),'html.parser')

for i in soup.find_all('div',{"class":"thumb_relate"}):
    try:
        f.write(i.text+'\n')
        f.write(i.find_all('a')[0].get('href') + '\n')
        soup2 = bs(ur.urlopen(i.find_all('a')[0].get('href')).read(),'html.
parser')
        for j in soup2.find_all('p'):
            f.write(j.text)
    except:
        pass

f.close()
```

06-4 Program 실행 file 만들기

Web Crawling 실행 file 만들기

- Python file 실행 file로 만들기

```
pyinstaller --onefile [파이썬 파일].py
```

```
pyinstaller --onefile article_collector.py
```

>> pyinstaller --onefile web_crawling_3.py

Windows PowerShell

```
PS C:₩과소사₩과소사-강의예제₩web_crawling> ls


    디렉터리: C:₩과소사₩과소사-강의예제₩web_crawling


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
d-----        2020-12-02  오후 12:46                __pycache__
-a----        2020-12-02  오후 10:59           4320 article_New.txt
-a----        2020-12-02  오후 10:59         138501 article_total.txt
-a----        2020-12-02  오후 10:58            180 links.txt
-a----        2020-12-02  오후 12:45            621 usecsv.py
-a----        2020-12-02   오후 9:39           1786 web_crawling_1.py
-a----        2020-12-02   오후 9:51           3047 web_crawling_2.py
-a----        2020-12-02  오후 10:58           3477 web_crawling_3.py


PS C:₩과소사₩과소사-강의예제₩web_crawling> pyinstaller --onefile web_crawling_3.py
62 INFO: PyInstaller: 4.1
62 INFO: Python: 3.8.5
62 INFO: Platform: Windows-10-10.0.18362-SP0
62 INFO: wrote C:₩과소사₩과소사-강의예제₩web_crawling₩web_crawling_3.spec
62 INFO: UPX is not available.
62 INFO: Extending PYTHONPATH with paths
['C:₩₩과소사₩₩과소사-강의예제₩₩web_crawling', 'C:₩₩과소사₩₩과소사-강의예제₩₩web_crawling']
78 INFO: checking Analysis
78 INFO: Building Analysis because Analysis-00.toc is non existent
78 INFO: Initializing module dependency graph...
78 INFO: Caching module graph hooks...
93 INFO: Analyzing base_library.zip ...
2294 INFO: Processing pre-find module path hook distutils from 'c:₩₩users₩₩user₩₩appdata₩₩local₩₩programs₩₩python₩₩
python38-32₩₩lib₩₩site-packages₩₩PyInstaller₩₩hooks₩₩pre_find_module_path₩₩hook-distutils.py'.
2295 INFO: distutils: retargeting to non-venv dir 'c:₩₩users₩₩user₩₩appdata₩₩local₩₩programs₩₩python₩₩python38-32₩₩
lib'
5412 INFO: Caching module dependency graph...
5537 INFO: running Analysis Analysis-00.toc
5537 INFO: Adding Microsoft.Windows.Common-Controls to dependent assemblies of final executable
  required by c:₩users₩user₩appdata₩local₩programs₩python₩python38-32₩python.exe
5647 INFO: Analyzing C:₩과소사₩과소사-강의예제₩web_crawling₩web_crawling_3.py
5771 INFO: Processing pre-safe import module hook urllib3.packages.six.moves from 'c:₩₩users₩₩user₩₩appdata₩₩local₩
₩programs₩₩python₩₩python38-32₩₩lib₩₩site-packages₩₩PyInstaller₩₩hooks₩₩pre_safe_import_module₩₩hook-urllib3.packag
es.six.moves.py'.
7219 INFO: Processing module hooks...
```

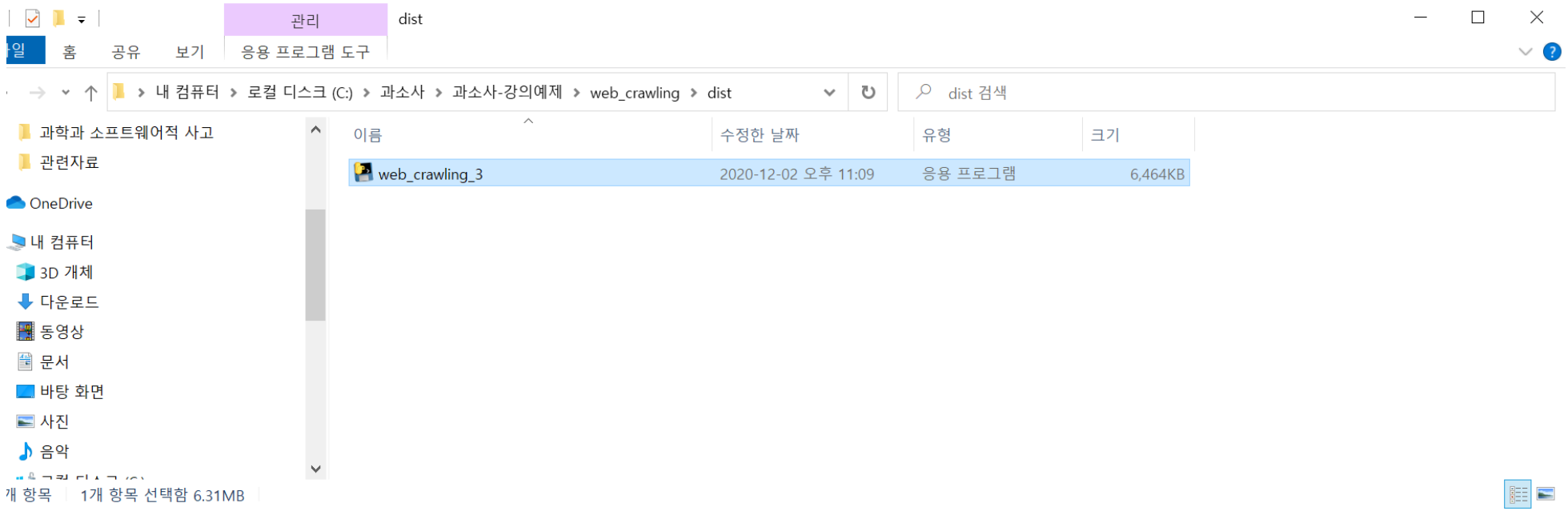
Windows PowerShell

```
PS C:₩과소사₩과소사-강의예제₩web_crawling>
PS C:₩과소사₩과소사-강의예제₩web_crawling>
PS C:₩과소사₩과소사-강의예제₩web_crawling>
PS C:₩과소사₩과소사-강의예제₩web_crawling> ls


    디렉터리: C:₩과소사₩과소사-강의예제₩web_crawling


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
d-----        2020-12-02  오후 11:08                build
d-----        2020-12-02  오후 11:08                dist
d-----        2020-12-02  오후 11:09                __pycache__
-a----        2020-12-02  오후 10:59           4320 article_New.txt
-a----        2020-12-02  오후 10:59         138501 article_total.txt
-a----        2020-12-02  오후 10:58            180 links.txt
-a----        2020-12-02  오후 12:45            621 usecsv.py
-a----        2020-12-02   오후 9:39           1786 web_crawling_1.py
-a----        2020-12-02   오후 9:51           3047 web_crawling_2.py
-a----        2020-12-02  오후 10:58           3477 web_crawling_3.py
-a----        2020-12-02  오후 11:08            927 web_crawling_3.spec


PS C:₩과소사₩과소사-강의예제₩web_crawling> cd dist
PS C:₩과소사₩과소사-강의예제₩web_crawling₩dist> ls


    디렉터리: C:₩과소사₩과소사-강의예제₩web_crawling₩dist


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        2020-12-02  오후 11:09        6618826 web_crawling_3.exe


PS C:₩과소사₩과소사-강의예제₩web_crawling₩dist>
```

06-4 Program 실행 file 만들기

Web Crawling 실행 file 만들기

- ‘web_crawling_3.exe’ 가 생성

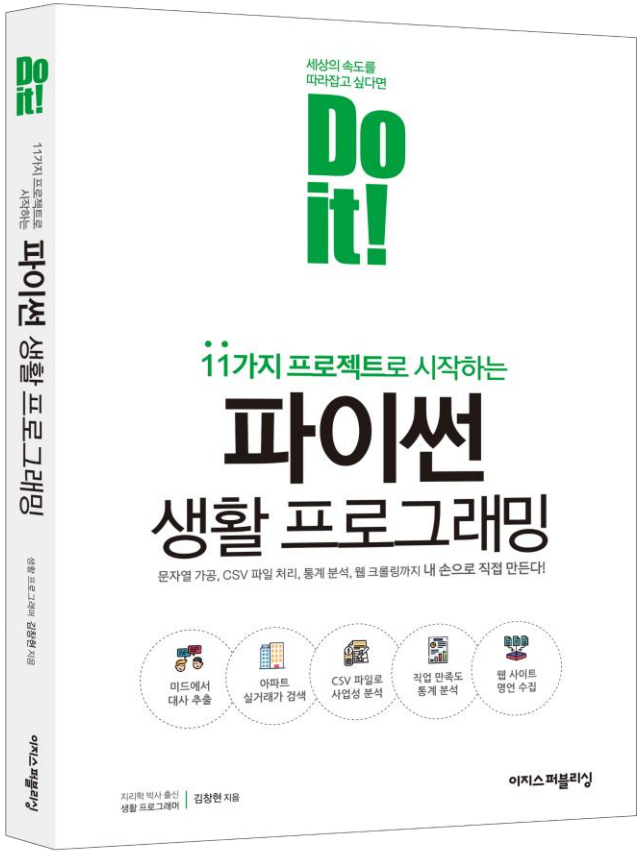


06-4 Program 실행 file 만들기

Web Crawling 실행 file 만들기

- web_crawling_3.exe 실행
- Program이 crawling 후 앞에서 실행시 생성하였던 file 생성

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
이지스퍼블리싱
지리학 박사 출신 생활 프로그래머 | 김창현 지음