

THIS IS THE HOTTEST PPT
TEMPLATE

완친판 18~19 발표

2017010715 허지혜

목차

부제 / 추가내용은 여기

1.18장 – 인터넷 쇼핑몰 정보 크롤링

2.19장 – 인터넷 언론 랭킹 뉴스 크롤링

1. 인터넷 쇼핑몰 정보 크롤링

1. 다양한 카테고리 정보를 주고
원하는 카테고리 정보를 입력 받을 수 있다.
2. 제품별 다양한 정보를 추출하여
txt, csv, xls 형식의 파일로 저장할 수 있다.

아마존닷컴 Best Seller URL : <https://www.amazon.com/bestsellers?ld=NSGoogl>

Amazon Best Sellers

Our most popular products based on sales. Updated hourly.

Any Department

- Amazon Devices & Accessories
- Amazon Launchpad
- Amazon Pantry
- Appliances
- Apps & Games
- Arts, Crafts & Sewing
- Audible Books & Originals
- Automotive
- Baby
- Beauty & Personal Care
- Books
- CDs & Vinyl
- Camera & Photo
- Cell Phones & Accessories
- Clothing, Shoes & Jewelry
- Collectible Currencies
- Computers & Accessories
- Digital Music
- Electronics
- Entertainment Collectibles
- Gift Cards
- Grocery & Gourmet Food
- Handmade Products
- Health & Household
- Home & Kitchen
- Industrial & Scientific
- Kindle Store
- Kitchen & Dining
- Magazine Subscriptions
- Movies & TV
- Musical Instruments
- Office Products
- Patio, Lawn & Garden

Toys & Games

> [See more Best Sellers in Toys & Games](#)

1.



Jenga Classic Game

★★★★★ 13,209

2.



Crayola Washable Kids Paint, 6 Count, Kids At Home Activities, Painting Supplies, Gift

★★★★★ 4,034

3.



Hasbro Connect 4 Game

★★★★★ 9,292

Electronics

> [See more Best Sellers in Electronics](#)

1.



Fire TV Stick streaming media player with Alexa built in, includes Alexa Voice Remote, HD, easy set-up, released 2019

★★★★★ 164,283

2.



Fire TV Stick 4K streaming device with Alexa built in, Ultra HD, Dolby Vision, includes the Alexa Voice Remote

★★★★★ 186,267

3.



Echo Dot (3rd Gen) - Smart speaker with Alexa - Charcoal

★★★★★ 327,590

Camera & Photo

1. 인터넷 쇼핑몰 정보 크롤링

=====




아마존닷컴의 분야별 Best Seller 상품 정보 추출하기

=====

- | | | |
|--------------------------------|------------------------------|-----------------------------|
| 1.Amazon Devices & Accessories | 2.Amazon Launchpad | 3.Appliances |
| 4.Apps & Games | 5.Arts, Crafts & Sewing | 6.Audible Books & Originals |
| 7.Automotive | 8.Baby | 9.Beauty & Personal Care |
| 10.Books | 11.CDs & Vinyl | 12.Camera & Photo |
| 13.Cell Phones & Accessories | 14.Clothing, Shoes & Jewelry | 15.Collectible Currencies |
| 16.Computers & Accessories | 17.Digital Music | 18.Electronics |
| 19.Entertainment Collectibles | 20.Gift Cards | 21.Grocery & Gourmet Food |
| 22.Handmade Products | 23.Health & Household | 24.Home & Kitchen |
| 25.Industrial & Scientific | 26.Kindle Store | 27.Kitchen & Dining |
| 28.Magazine Subscriptions | 29.Movies & TV | 30.Musical Instruments |
| 31.Office Products | 32.Patio, Lawn & Garden | 33.Pet Supplies |
| 34.Prime Pantry | 35.Smart Home | 36.Software |
| 37.Sports & Outdoors | 38.Sports Collectibles | 39.Tools & Home Improvement |
| 40.Toys & Games | 41.Video Games | |

1. 위 분야 중에서 자료를 수집할 분야의 번호를 선택하세요: 3
2. 해당 분야에서 크롤링 할 건수는 몇건입니까?(1-100 건 사이 입력): 3
3. 파일을 저장할 폴더명만 쓰세요(예:c:\temp\):C:\Temp\

요청하신 데이터를 수집하고 있으니 잠시만 기다려 주세요~~

A	B	C	D	E	F	G
	판매순위	제품소개	판매가격	상품평 갯수	상품평점	
0						
1	1	Charmin Ultra Soft Cushiony Touch T	\$11.90	1304	4.7 out of 5 stars	
1						
	2	Snack Pack Chocolate and Vanilla F	\$2.34	3361	4.7 out of 5 stars	
						

저장된 엑셀 파일 예시

2. 실행해보기

부제 / 추가내용은 여기

- Step 1. 필요한 라이브러리와 모듈 코딩 합니다.
- Step 2. 사용자에게 카테고리 메뉴를 보여주고 정보를 입력 받습니다.
- Step 3. 저장될 파일위치와 이름을 지정 한 후
크롬 드라이버를 실행하여 페이지를 엽니다
- Step 4. 화면을 스크롤 해서 아래로 이동한 후 요청된 데이터를 수집합니다.
- Step 5. 검색 결과를 다양한 형태로 저장하기

Step1. 필요한 모듈, 라이브러리 로딩

아마존닷컴 분야별 베스트셀러 상품 크롤러

#Step 1. 필요한 모듈과 라이브러리를 로딩합니다.

```
from bs4 import BeautifulSoup
from selenium import webdriver
```

```
import time
import sys
import re
import math
import numpy
import pandas as pd
import xlwt
import random
import os
```

```
import urllib.request
import urllib
```


Step2.

학습목표 1 : 사용자에게 다양한 메뉴를 보여 준 후 카테고리값을 입력 받아 해당 카테고리 메뉴를 실행한다.

Step 2. 사용자에게 카테고리 메뉴를 보여주고 정보를 입력 받습니다.

```
print("=" *80)
```

```
print("    아마존닷컴의 분야별 Best Seller 상품 정보 추출하기")
```

```
print("=" *80)
```

```
query_txt='아마존닷컴'
```

```
query_url='https://www.amazon.com/bestsellers?ld=NSGoogle'
```

```
sec = input('')
```

1.Amazon Devices & Accessories	2.Amazon Launchpad	3.Appliances
4.Apps & Games	5.Arts, Crafts & Sewing	6.Audible Books & Originals
7.Automotive	8.Baby	9.Beauty & Personal Care
10.Books	11.CDs & Vinyl	12.Camera & Photo
13.Cell Phones & Accessories	14.Clothing, Shoes & Jewelry	15.Collectible Currencies
16.Computers & Accessories	17.Digital Music	18.Electronics
19.Entertainment Collectibles	20.Gift Cards	21.Grocery & Gourmet Food
22.Handmade Products	23.Health & Household	24.Home & Kitchen
25.Industrial & Scientific	26.Kindle Store	27.Kitchen & Dining
28.Magazine Subscriptions	29.Movies & TV	30.Musical Instruments
31.Office Products	32.Patio, Lawn & Garden	33.Pet Supplies
34.Prime Pantry	35.Smart Home	36.Software
37.Sports & Outdoors	38.Sports Collectibles	39.Tools & Home Improvement
40.Toys & Games	41.Video Games	

1.위 분야 중에서 자료를 수집할 분야의 번호를 선택하세요: '')

```
cnt = int(input('    2.해당 분야에서 크롤링 할 건수는 몇건입니까?(1-100 건 사이 입력): '))
```

```
f_dir = input("    3.파일을 저장할 폴더명만 쓰세요(예:c:\###temp###):")
```

```

if sec == '1' :
    sec_name='Amazon Devices and Accessories'
elif sec == '2' :
    sec_name='Amazon Launchpad'
elif sec == '3' :
    sec_name='Appliances'
elif sec == '4' :
    sec_name='Apps and Games'
elif sec == '5' :
    sec_name='Arts and Crafts and Sewing'
elif sec == '6' :
    sec_name='Audible Books and Originals'
elif sec == '7' :
    sec_name='Automotive'
elif sec == '8' :
    sec_name='Baby'
elif sec == '9' :
    sec_name='Beauty and Personal Care'
elif sec == '10' :
    sec_name='Books'
elif sec == '11' :
    sec_name='CDs and Vinyl'
elif sec == '12' :
    sec_name='Camera and Photo'
elif sec == '13' :
    sec_name='Cell Phones and Accessories'
elif sec == '14' :
    sec_name='Clothing and Shoes and Jewelry'
elif sec == '15' :
    sec_name='Collectible Quizzes'

```

```

if cnt > 30 :
    print(" 요청 건수가 많아서 시간이 제법 소요되오니 잠시만 기다려 주세요~~")
else :
    print(" 요청하신 데이터를 수집하고 있으니 잠시만 기다려 주세요~~")

```

```

elif sec == '22' :
    sec_name='Handmade Products'
elif sec == '23' :
    sec_name='Health and Household'
elif sec == '24' :
    sec_name='Home and Kitchen'
elif sec == '25' :
    sec_name='Industrial and Scientific'
elif sec == '26' :
    sec_name='Kindle Store'
elif sec == '27' :
    sec_name='Kitchen and Dining'
elif sec == '28' :
    sec_name='Magazine Subscriptions'
elif sec == '29' :
    sec_name='Movies and TV'
elif sec == '30' :
    sec_name='Musical Instruments'
elif sec == '31' :
    sec_name='Office Products'
elif sec == '32' :
    sec_name='Patio and Lawn and Garden'
elif sec == '33' :
    sec_name='Pet Supplies'
elif sec == '34' :
    sec_name='Prime Pantry'
elif sec == '35' :
    sec_name='Smart Home'
elif sec == '36' :
    sec_name='Software'

```

```

    sec_name='Sports and Outdoors'
    sec_name='Tools and Home Improvement'
    sec_name='Toys and Games'
    sec_name='Video Games'

```

화면 예시

=====

아마존닷컴의 분야별 Best Seller 상품 정보 추출하기

=====

- | | | |
|--------------------------------|------------------------------|-----------------------------|
| 1.Amazon Devices & Accessories | 2.Amazon Launchpad | 3.Appliances |
| 4.Apps & Games | 5.Arts, Crafts & Sewing | 6.Audible Books & Originals |
| 7.Automotive | 8.Baby | 9.Beauty & Personal Care |
| 10.Books | 11.CDs & Vinyl | 12.Camera & Photo |
| 13.Cell Phones & Accessories | 14.Clothing, Shoes & Jewelry | 15.Collectible Currencies |
| 16.Computers & Accessories | 17.Digital Music | 18.Electronics |
| 19.Entertainment Collectibles | 20.Gift Cards | 21.Grocery & Gourmet Food |
| 22.Handmade Products | 23.Health & Household | 24.Home & Kitchen |
| 25.Industrial & Scientific | 26.Kindle Store | 27.Kitchen & Dining |
| 28.Magazine Subscriptions | 29.Movies & TV | 30.Musical Instruments |
| 31.Office Products | 32.Patio, Lawn & Garden | 33.Pet Supplies |
| 34.Prime Pantry | 35.Smart Home | 36.Software |
| 37.Sports & Outdoors | 38.Sports Collectibles | 39.Tools & Home Improvement |
| 40.Toys & Games | 41.Video Games | |

1. 위 분야 중에서 자료를 수집할 분야의 번호를 선택하세요: 3
2. 해당 분야에서 크롤링 할 건수는 몇건입니까?(1-100 건 사이 입력): 3
3. 파일을 저장할 폴더명만 쓰세요(예:c:\temp\): C:\Temp\

요청하신 데이터를 수집하고 있으니 잠시만 기다려 주세요~~

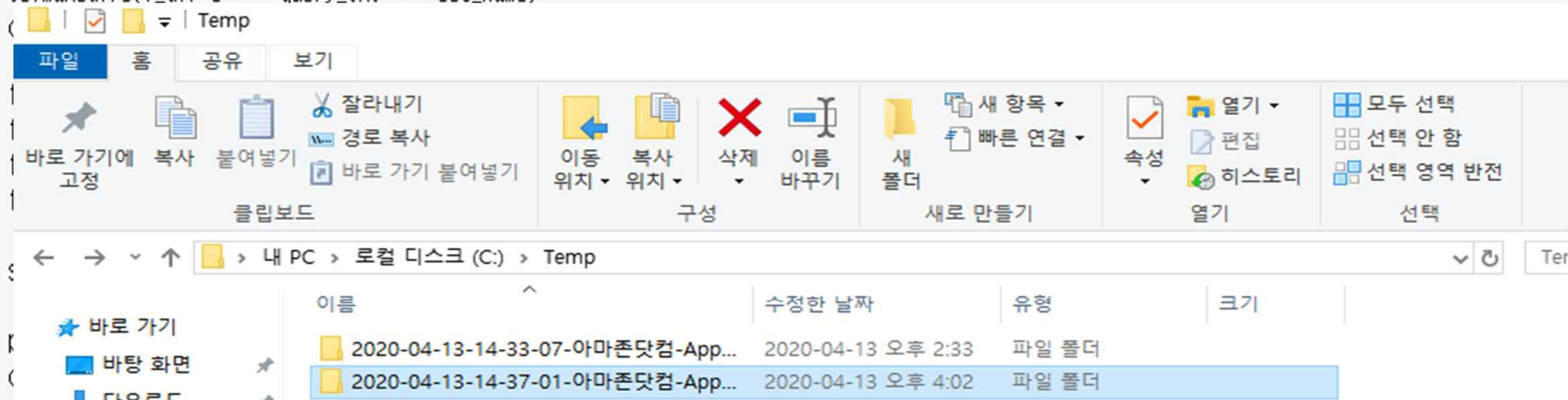
Step3.

Step 3. 저장될 파일위치와 이름을 지정 한 후 크롬 드라이버를 실행하여 페이지를 엽니다

```
now = time.localtime()
```

```
s = '%04d-%02d-%02d-%02d-%02d' % (now.tm_year, now.tm_mon, now.tm_mday, now.tm_hour, now.tm_min, now.tm_sec)
```

```
os.makedirs(f_dir+s+'-'+query_txt+'-'+sec_name)
```



```
driver.get(query_url)
```

```
time.sleep(5)
```

Any Department

Amazon Devices & Accessories
Amazon Launchpad
Appliances
Apps & Games
Arts, Crafts & Sewing
Automotive
Baby
Beauty & Personal Care
Books
CDs & Vinyl
Camera & Photo
Cell Phones & Accessories
Clothing, Shoes & Jewelry
Collectible Coins
Computers & Accessories
Digital Music
Electronics
Entertainment Collectibles
Gift Cards
Grocery & Gourmet Food
Handmade Products
Health & Household
Home & Kitchen
Industrial & Scientific
Kindle Store
Kitchen & Dining
Magazine Subscriptions
Movies & TV
Musical Instruments
Office Products
Patio, Lawn & Garden

1

Toys & Games

› [See more Best Sellers in Toys & Games](#)

1.



Crayola Mini Twistables Crayons, Amazon Exclusive, 50 Count, Great for Coloring Books, Gift

★★★★★ 399

2

2.



Avengers Marvel Legends Series Endgame Power Gauntlet Articulated Electronic Fist

3.



L.O.L. Surprise! Glam Glitter Series Doll with 7 Surprises

★★★★★ 1,638

Electronics

› [See more Best Sellers in Electronics](#)

1.



Fire TV Stick with Alexa Voice Remote, streaming media player

★★★★★ 7,046

2.



Fire TV Stick 4K with Alexa Voice Remote, streaming media player

★★★★★ 19,684

3.



Echo Dot (3rd Gen) - Smart speaker with Alexa - Charcoal

★★★★★ 31,608

분야별 더보기 버튼을 눌러 페이지를 엽니다

```
if sec == '1' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[1]/a").click()
elif sec == '2' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[2]/a").click()
elif sec == '3' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[3]/a").click()
elif sec == '4' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[4]/a").click()
elif sec == '5' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[5]/a").click()
elif sec == '6' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[6]/a").click()
elif sec == '7' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[7]/a").click()
elif sec == '8' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[8]/a").click()
elif sec == '9' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[9]/a").click()
elif sec == '10' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[10]/a").click()
elif sec == '11' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[11]/a").click()
elif sec == '12' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[12]/a").click()
elif sec == '13' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[13]/a").click()
elif sec == '14' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[14]/a").click()
elif sec == '15' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[15]/a").click()
elif sec == '16' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[16]/a").click()
elif sec == '17' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[17]/a").click()
elif sec == '18' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[18]/a").click()
elif sec == '19' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[19]/a").click()
elif sec == '20' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[20]/a").click()
elif sec == '21' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[21]/a").click()
```

```
elif sec == '22' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[22]/a").click()
elif sec == '23' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[23]/a").click()
elif sec == '24' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[24]/a").click()
elif sec == '25' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[25]/a").click()
elif sec == '26' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[26]/a").click()
elif sec == '27' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[27]/a").click()
elif sec == '28' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[28]/a").click()
elif sec == '29' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[29]/a").click()
elif sec == '30' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[30]/a").click()
elif sec == '31' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[31]/a").click()
elif sec == '32' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[32]/a").click()
elif sec == '33' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[33]/a").click()
elif sec == '34' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[34]/a").click()
elif sec == '35' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[35]/a").click()
elif sec == '36' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[36]/a").click()
elif sec == '37' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[37]/a").click()
elif sec == '38' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[38]/a").click()
elif sec == '39' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[39]/a").click()
elif sec == '40' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[40]/a").click()
elif sec == '41' :
    driver.find_element_by_xpath("//*[id='zg_browseRoot']/ul/li[41]/a").click()

time.sleep(1)
```

Step4.

*# 학습목표 2 : 해당 카테고리의 데이터를 수집합니다.
Step 4. 화면을 스크롤해서 아래로 이동한 후 요청된 데이터를 수집합니다.*

```
def scroll_down(driver):
```

```
    driver.execute_script("window.scrollTo(0,9300);")  
    time.sleep(1)
```

```
scroll_down(driver)
```

비트맵 이미지 아이콘을 위한 대체 텍스트를 만듭니다

```
bmp_map = dict.fromkeys(range(0x10000, sys.maxunicode + 1), 0xfffd)
```

이미지 추출 코드 추가

```
img_src2=[]    # 이미지 URL 저장변수  
file_no = 0
```

```
html = driver.page_source
```

```
soup = BeautifulSoup(html, 'html.parser')
```

```
reple_result = soup.select('#zg-center-div > #zg-ordered-list')
```

```
slist = reple_result[0].find_all('li')
```

Step4. 필요한 모듈, 라이브러리 로딩

```
if cnt < 51 :  
    ranking2=[]  
    title3=[]  
    price2=[]  
    score2=[]  
    sat_count2=[]  
    store2=[]  
  
    count = 0  
  
    # 이미지 저장용 폴더 생성하기  
    img_dir = ff_dir+"\\\\images"  
    os.makedirs(img_dir)  
    os.chdir(img_dir)  
  
    for li in slist:  
        # 이미지 저장하기  
        try :  
            photo = li.find('div','a-section a-spacing-small').find('img')['src']  
        except AttributeError :  
            continue  
        file_no += 1  
  
        urllib.request.urlretrieve(photo,str(file_no)+'.jpg')  
        time.sleep(1)  
  
        if cnt == file_no :  
            break  
  
        f = open(ff_name, 'a',encoding='UTF-8')  
        f.write("-----"+"\\n")
```


Step4. 필요한 모듈, 라이브러리 로딩

```
# 가격
try :
    price = li.find('span', 'p13n-sc-price').get_text().replace("₩", "")
except AttributeError :
    price = ''

print("3.가격:", price.replace("₩", ""))
f.write('3.가격:'+ price + "₩")

try :
    sat_count = li.find('a', 'a-size-small a-link-normal').get_text().replace(", ", "")
except (IndexError , AttributeError) :
    sat_count='0'
    print('4.상품평 수: ', sat_count)
    f.write('4.상품평 수:'+ sat_count + "₩")
else :
    print('4.상품평 수:', sat_count)
    f.write('4.상품평 수:'+ sat_count + "₩")

#상품 별점 구하기
try :
    score = li.find('span', 'a-icon-alt').get_text()
except AttributeError :
    score=' '

print('5.평점:', score)
f.write('5.평점:'+ score + "₩")

print("-" *70)
```

1 페이지 정보 추출 후 2 페이지로 넘어가기

```
driver.find_element_by_xpath("//*[@id='zg-center-div']/div[2]/div/ul/li[3]/a").click( )
```

```
print("###")
```

```
print("요청하신 데이터의 수량이 많아 다음 페이지의 데이터를 추출 중이오니 잠시만 기다려 주세요~^^")
```

```
print("###")
```

```
html = driver.page_source
```

```
soup = BeautifulSoup(html, 'html.parser')
```

```
reple_result = soup.select('#zg-center-div > #zg-ordered-list')
```

```
slist = reple_result[0].find_all('li')
```

출력 예시

1. 판매순위: 1

2. 제품소개: Charmin Ultra Soft Cushiony Touch Toilet Paper, Family Mega Rolls, Prime Pantry, 6 Count

3. 가격: \$11.90

4. 상품평 수: 1304

5. 평점: 4.7 out of 5 stars

1. 판매순위: 2

2. 제품소개: Snack Pack Chocolate and Vanilla Pudding Cups Family Pack, 12 Count

3. 가격: \$2.34

4. 상품평 수: 3361

5. 평점: 4.7 out of 5 stars

Step5.

#Step 5. 검색 결과를 다양한 형태로 저장하기

```
amazon_best_seller = pd.DataFrame()  
amazon_best_seller['판매순위'] = ranking2  
amazon_best_seller['제품소개'] = pd.Series(title3)  
amazon_best_seller['판매가격'] = pd.Series(price2)  
amazon_best_seller['상품평 갯수'] = pd.Series(sat_count2)  
amazon_best_seller['상품평점'] = pd.Series(score2)
```

csv 형태로 저장하기

```
amazon_best_seller.to_csv(fc_name, encoding="utf-8-sig", index=True)
```

엑셀 형태로 저장하기

```
amazon_best_seller.to_excel(fx_name, index=True)
```

```
e_time = time.time()  
t_time = e_time - s_time
```

txt 파일에 크롤링 요약 정보 저장하기

```
orig_stdout = sys.stdout  
f = open(ff_name, 'a', encoding='UTF-8')  
sys.stdout = f
```

```

import win32com.client as win32    #pywin32 , pypiwin32 설치후 동작
import win32api    #파이썬 프롤프트를 관리자 권한으로 실행해야 에러없음
                        #파이썬 셸을 관리자 권한으로 실행한 후 불러오기로 이 소스 실행하기
excel = win32.gencache.EnsureDispatch('Excel.Application')
wb = excel.Workbooks.Open(fx_name)
sheet = wb.ActiveSheet
sheet.Columns(3).ColumnWidth = 30    # 이미지 가로 사이즈에 맞게 컬럼 크기 조정
row_cnt = cnt+1
sheet.Rows("2:%s" %row_cnt).RowHeight = 120    # 이미지 세로 사이즈에 맞게 로우 크기 조정

ws = wb.Sheets("Sheet1")
col_name2=[]
file_name2=[]

for a in range(2,cnt+2) :
    col_name='C'+str(a)
    col_name2.append(col_name)

for b in range(1,cnt+1) :
    file_name=img_dir+'###'+str(b)+'.jpg'
    file_name2.append(file_name)

for i in range(0,cnt) :
    rng = ws.Range(col_name2[i])
    image = ws.Shapes.AddPicture(file_name2[i], False, True, rng.Left, rng.Top, 130, 100)
    excel.Visible=True
    excel.ActiveWorkbook.Save()

# Step 6. 요약 정보를 출력하기
print("\n")
print("=" *50)
print("총 소요시간은 %s 초 이며," %t_time)
print("총 저장 건수는 %s 건 입니다 " %count)
print("=" *50)

sys.stdout = orig_stdout
f.close()

print("\n")
print("=" *80)
print("1.요청된 총 %s 건의 리뷰 중에서 실제 크롤링 된 리뷰수는 %s 건입니다" %(cnt,count))
print("2.총 소요시간은 %s 초 입니다 " %round(t_time,1))
print("3.파일 저장 완료: txt 파일명 : %s " %ff_name)
print("4.파일 저장 완료: csv 파일명 : %s " %fc_name)
print("5.파일 저장 완료: xls 파일명 : %s " %fx_name)
print("=" *80)

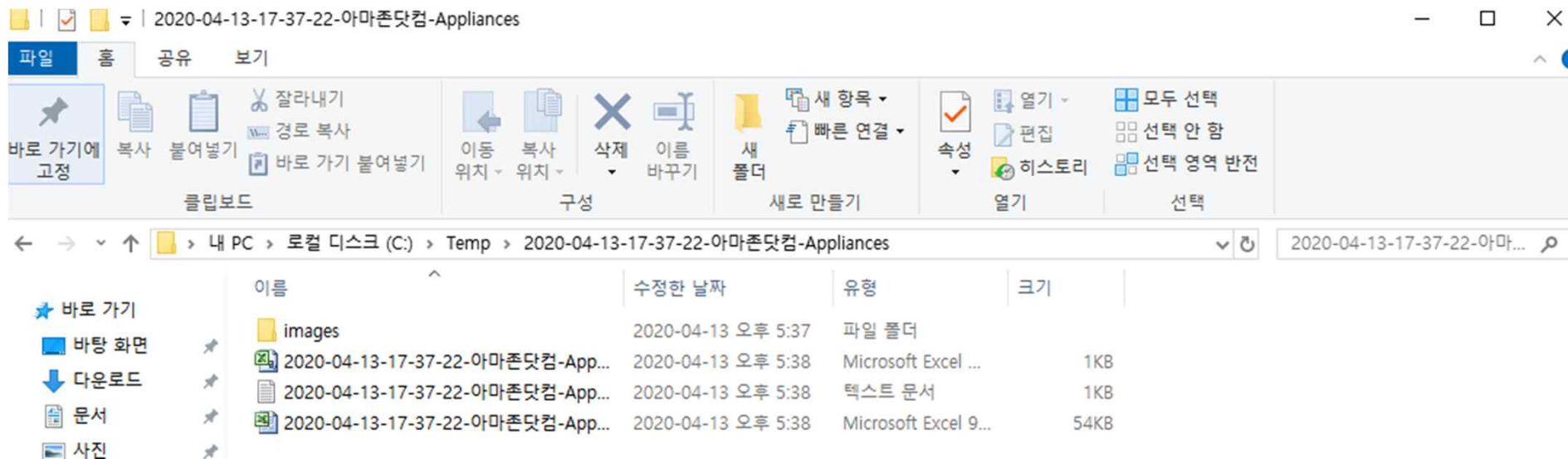
```


Step5.




=====

- 1.요청된 총 3 건의 리뷰 중에서 실제 크롤링 된 리뷰수는 2 건입니다
- 2.총 소요시간은 40.0 초 입니다
- 3.파일 저장 완료: txt 파일명 : c:\Temp\2020-04-13-17-37-22-아마존닷컴-Appliances\2020-04-13-17-37-22-아마존닷컴-Appliances.txt
- 4.파일 저장 완료: csv 파일명 : c:\Temp\2020-04-13-17-37-22-아마존닷컴-Appliances\2020-04-13-17-37-22-아마존닷컴-Appliances.csv
- 5.파일 저장 완료: xls 파일명 : c:\Temp\2020-04-13-17-37-22-아마존닷컴-Appliances\2020-04-13-17-37-22-아마존닷컴-Appliances.xls

=====



Step5.

A	B	C	D	E	F	G
	판매순위	제품소개	판매가격	상품평 갯수	상품평점	
0						
1	1	Charmin Ultra Soft Cushiony Touch T 	\$11.90	1304	4.7 out of 5 stars	
1	2	Snack Pack Chocolate and Vanilla F 	\$2.34	3361	4.7 out of 5 stars	

2. 랭킹 뉴스 크롤링

부제 / 추가내용은 여기

19장 목표

온라인 조선일보 웹 페이지에서 2018년도
경제 분야의 랭킹 뉴스를 추출

URL 주소 : <http://news.chosun.com/ranking/list.html>

=====

조선일보 랭킹 뉴스중 2018년 경제 분야 기사 정보 수집하기

=====

1.파일이 저장될 경로만 쓰세요; (예:c:\₩temp₩): c:\₩data₩

2018년 01월 01일의 뉴스 수집 중 =====

0 : [아파트 별곡]① 자산증식 욕망이 불패 신기루 만들어...주거대안 화두에 진화 거듭
1 : [과학TALK] 전기차도 만드는 3D프린터, 생체·국방 분야로도 확대
2 : [2018년 유통 전망] 최저임금 인상 발등에 불...제품가격 인상 역풍 부나
3 : LG디스플레이, 88인치 8K OLED 디스플레이 세계 최초 개발
4 : 박용만 회장 "낮은 규제 이제 정말 없앨 때...기업들 무력감 크다"
5 : 삼성重 이어 현대重도 어닝쇼크 고백..."부진 장기화" vs "선제적 조치"
6 : "中, 내년 말 D램 양산"...반도체 굴기 현실화
7 : 블랙홀 실체 드러나고...유전자 치료, 사람에 첫 적용
8 : 연초부터 게임 신작 쏟아진다...모바일·PC 동시 출격
9 : LG전자, CES서 인공지능 전시장 'LG 씽큐 존' 전면 배치
10 : "배터리 유상교체?"...애플 보상책 내놔도 국내 집단소송 희망자 18만명
11 : "그때 비트코인 사셔야지!"...세상에서 가장 우울한 사이트
12 : 작년 'CEO 연봉킹' 권오현 회장...200억원으로 추산
13 : 해외여행 간 한국인, 일본보다 800만명 많다
14 : [2018 신년사] 文대통령 "과거 잘못 바로잡는 노력 지속...국민 삶의 질 개선이 최우선 목표"
15 : 부활한 9만9000원 과일세트...유통업계 '김영란법' 개정예 고가 선물 늘린다
16 : 新공정 비용 치솟고 퀄컴 이탈설까지...삼성 파운드리 "속 탄다"
17 : 황금개의 해 'GOLDEN DOG'로 풀어본 식품업계 트렌드
18 : '반도체가 이끈 2017년'...작년 수출액 사상 최대치 기록(종합)
19 : 새해 중국 경제 10대 추세...녹색규제 폭탄·위안화 절하 압력
20 : 올해 부동산 투자 키워드는 '물류창고'...美 경기회복 최대 수혜 예상
21 : 정부, 4조695억원 규모 R&D 종합시행계획 확정
22 : 가상화폐 관련 입법, 국회가 주저주저하는 까닭은
23 : 판교 짝고, 실리콘밸리로 날다
24 : [2018 IT 전망]③ 반도체 초호황 '경착륙'과 중국 ICT 굴기...블록체인 영역 확대
25 : 시총 2조...
26 : 이마트
27 : [2018년
28 : 법원
29 : 신한카
30 건 완료=====

크롤링 실행 화면

2018년 01월 01일의 뉴스 수집 중 =====

- 0: [아파트 별곡]① 자산증식 욕망이 불패 신기루 만들어...주거대안 화두에 진화 거듭
- 1: [과학TALK] 전기차도 만드는 3D프린터, 생체·국방 분야로도 확대
- 2: [2018년 유통 전망] 최저임금 인상 발등에 불...제품가격 인상 역풍 부나
- 3: LG디스플레이, 88인치 8K OLED 디스플레이 세계 최초 개발
- 4: 박용만 회장 "낮은 규제 이제 정말 없앨 때...기업들 무력감 크다"
- 5: 삼성重 이어 현대重도 어닝쇼크 고백..."부진 장기화" vs "선제적 조치"
- 6: "中, 내년 말 D램 양산"... 반도체 굴기 현실화
- 7: 블랙홀 실체 드러나고... 유전자 치료, 사람에 첫 적용
- 8: 연초부터 게임 신작 쏟아진다...모바일·PC 동시 출격
- 9: LG전자, CES서 인공지능 전시장 'LG 씽큐 존' 전면 배치
- 10: "배터리 유상교체?"...애플 보상책 내놔도 국내 집단소송 희망자 18만명
- 11: "그때 비트코인 사셔야지!"...세상에서 가장 우울한 사이트
- 12: 작년 'CEO 연봉킹' 권오현 회장...200억원으로 추산
- 13: 해외여행 간 한국인, 일본보다 800만명 많다
- 14: [2018 신년사] 文대통령 "과거 잘못 바로잡는 노력 지속...국민 삶의 질 개선이 최우선 목표"
- 15: 부활한 9만9000원 과일세트...유통업계 '김영란법' 개정예 고가 선물 늘린다
- 16: 新공정 비용 치솟고 웰컴 이탈설까지...삼성 파운드리 "속 탄다"
- 17: 황금개의 해 'GOLDEN DOG'로 풀어본 식품업계 트렌드
- 18: '반도체가 이끈 2017년'...작년 수출액 사상 최대치 기록(종합)
- 19: 새해 중국 경제 10대 추세...녹색규제 폭탄·위안화 절하 압력
- 20: 올해 부동산 투자 키워드는 '물류창고'...美 경기회복 최대 수혜 예상
- 21: 정부, 4조695억원 규모 R&D 종합시행계획 확정
- 22: 가상화폐 관련 입법, 국회가 주저주저하는 까닭은
- 23: 판교 찍고, 실리콘밸리로 날다
- 24: [2018 IT 전망]③ 반도
- 25: 시총 2위 리플, 3000
- 26: 이마트·신세계 온라인
- 27: [2018년 車 전망] 친
- 28: 법원 "하차요구 무시했다고 죄다 '감금' 아냐"
- 29: 신한카드, 플랫폼 사업그룹 신설... "신한카드 내 14조 규모 디지털 기업"
- 30 건 완료=====

텍스트 형식으로 저장된 파일 예시

뉴스 랭킹 정보 크롤링

- Step 1. 필요한 모듈과 라이브러리를 로딩합니다.
- Step 2. 사용자에게 파일이 저장될 폴더명을 입력 받은 후 파일명을 설정합니다.
- Step 3. 크롬 드라이버를 사용해서 웹 브라우저를 실행합니다.
- Step 4. 날짜를 계산합니다.
- Step 5. 각 날짜별 기사의 Title 을 추출합니다.
- Step 6. 출력 결과를 파일에 저장하기

인터넷 언론 정보 수집하기 - 조선일보 경제 분야 기사 수집

#Step 1. 필요한 모듈과 라이브러리를 로딩합니다.

```
from bs4 import BeautifulSoup
from selenium import webdriver
import time
import sys
import random
import os
```

#Step 2. 사용자에게 파일이 저장될 폴더명을 입력 받은 후 파일명을 설정합니다.

```
print("=" * 80)
print(" 조선일보 랭킹 뉴스중 2018년 경제 분야 기사 정보 수집하기 ")
print("=" * 80)

url = 'http://news.chosun.com/ranking/list.html?site=chosunbiz&score=index&date=' # 경제분야 url

start_date=int(20180101)
end_date=int(20181231)

f_dir=input('1.파일이 저장될 경로만 쓰세요:(예:c:\temp): ')

now = time.localtime()
s = '%04d-%02d-%02d-%02d-%02d' % (now.tm_year, now.tm_mon, now.tm_mday, now.tm_hour, now.tm_min, now.tm_sec)

year='2018'

os.makedirs(f_dir+'조선일보_경제뉴스'+year+'-'+s)
os.chdir(f_dir+'조선일보_경제뉴스'+year+'-'+s)

ff_name=f_dir+'조선일보_경제뉴스'+year+'-'+s+'###'+year+'-'+s+'.txt'
fx_name=f_dir+'조선일보_경제뉴스'+year+'-'+s+'###'+year+'-'+s+'.xls'
```

#Step 3. 크롬 드라이버를 사용해서 웹 브라우저를 실행합니다.

```
s_time = time.time( )

path = "c:/temp/chromedriver_240/chromedriver.exe"
driver = webdriver.Chrome(path)
```

#Step 4. 날짜를 계산합니다.

```
mon=["01","02","03","04","05","06","07","08","09","10","11","12"]
```

```
i=0
```

```
start_date2=[]
```

```
end_date2=[]
```

빈 리스트 생성

월로 나누기

```
for i in range(0,len(mon)) :
```

```
    if mon[i] == "02" :
```

```
        sdate=year+mon[i]+'01'
```

```
        start_date2.append(sdate)
```

```
        edate=year+mon[i]+'28'
```

```
        end_date=edate
```

```
        end_date2.append(end_date)
```

```
    elif mon[i] == "04" or mon[i] == "06" or mon[i] == "09" or mon[i] == "11" :
```

```
        sdate=year+mon[i]+'01'
```

```
        start_date2.append(sdate)
```

```
        edate=year+mon[i]+'30'
```

```
        end_date=edate
```

```
        end_date2.append(end_date)
```

```
    else :
```

```
        sdate=year+mon[i]+'01'
```

```
        start_date2.append(sdate)
```

```
        edate=year+mon[i]+'31'
```

```
        end_date=edate
```

```
        end_date2.append(end_date)
```

#Step 5. 각 날짜별 기사의 Title 을 추출합니다.

```
total_count = 0
```

```
for x in range(0, len(end_date2)) :
```

```
    for i in range(int(start_date2[x]), int(end_date2[x])) :  
        full_url = url+str(i)
```

```
        driver.get(full_url)  
        time.sleep(2)
```

```
        html = driver.page_source  
        soup = BeautifulSoup(html, 'html.parser')
```

```
        count = 0  
        news_no = 1  
        c_year=str(i)[0:4]  
        c_mon=str(i)[4:6]  
        c_day=str(i)[6:]
```

```
        article_result = soup.find('div', class_='list_content rank_numbering')  
        ar_list = article_result.find_all('dl')  
        print("ㄷ")  
        print("%s년 %s월 %s일의 뉴스 수집 중 ===== " % (c_year, c_mon, c_day))
```

```
        f = open(ff_name, 'a', encoding='UTF-8')  
        f.write("%s년 %s월 %s일의 뉴스 수집 중 ===== " % (c_year, c_mon, c_day))  
        f.write("ㄷ")
```

```
        for li in ar_list:  
            title = li.find('div', 'list_inner').find('dt').get_text()  
            print(news_no, ": ", title)
```

```
            f.write(str(news_no) + ": " + title + "ㄷ")  
            time.sleep(0.2)
```

```
            count += 1  
            news_no += 1
```


#Step 2. 사용자에게 파일이 저장될 폴더명을 입력 받은 후 파일명을 설정합니다.

```
print("=" * 80)
```

```
print(" 조선일보 랭킹 뉴스중 2018년 경제 분야 기사 정보 수집하기 ")
```

```
print("=" * 80)
```

```
url = 'http://news.chosun.com/ranking/list.html?site=chosunbiz&scode=index&date=' # 경제분야 url
```

```
start_date=int(20180101)
```

```
end_date=int(20181231)
```

```
f_dir=input('1.파일이 저장될 경로만 쓰세요:(예:c:\temp): ')
```

```
now = time.localtime()
```

```
s = '%04d-%02d-%02d-%02d-%02d-%02d' % (now.tm_year, now.tm_mon, now.tm_mday, now.tm_hour, now.tm_min, now.tm_sec)
```

```
year='2018'
```

```
os.makedirs(f_dir+'조선일보_경제뉴스'+year+'-'+s)
```

```
os.chdir(f_dir+'조선일보_경제뉴스'+year+'-'+s)
```

```
ff_name=f_dir+'조선일보_경제뉴스'+year+'-'+s+'###'+f_dir+'조선일보_경제뉴스'+year+'-'+s+'.txt'
```

```
fx_name=f_dir+'조선일보_경제뉴스'+year+'-'+s+'###'+f_dir+'조선일보_경제뉴스'+year+'-'+s+'.xls'
```

#Step 5. 각 날짜별 기사의 Title 을 추출합니다.

```
total_count = 0
```

```
for x in range(0, len(end_date2)) :
```

```
    for i in range(int(start_date2[x]), int(end_date2[x])) :  
        full_url = url+str(i)
```

```
        driver.get(full_url)  
        time.sleep(2)
```

```
        html = driver.page_source  
        soup = BeautifulSoup(html, 'html.parser')
```

```
        count = 0  
        news_no = 1  
        c_year=str(i)[0:4]  
        c_mon=str(i)[4:6]  
        c_day=str(i)[6:]
```

```
        article_result = soup.find('div', class_='list_content rank_numbering')  
        ar_list = article_result.find_all('dl')  
        print("ㄷ")  
        print("%s년 %s월 %s일의 뉴스 수집 중 ===== " % (c_year, c_mon, c_day))
```

```
        f = open(ff_name, 'a', encoding='UTF-8')  
        f.write("%s년 %s월 %s일의 뉴스 수집 중 ===== " % (c_year, c_mon, c_day))  
        f.write("ㄷ")
```

```
        for li in ar_list:  
            title = li.find('div', 'list_inner').find('dt').get_text()  
            print(news_no, ": ", title)
```

```
            f.write(str(news_no) + ": " + title + "ㄷ")  
            time.sleep(0.2)
```

```
            count += 1  
            news_no += 1
```

=====

조선일보 행킹 뉴스중 2018년 경제 분야 기사 정보 수집하기

=====

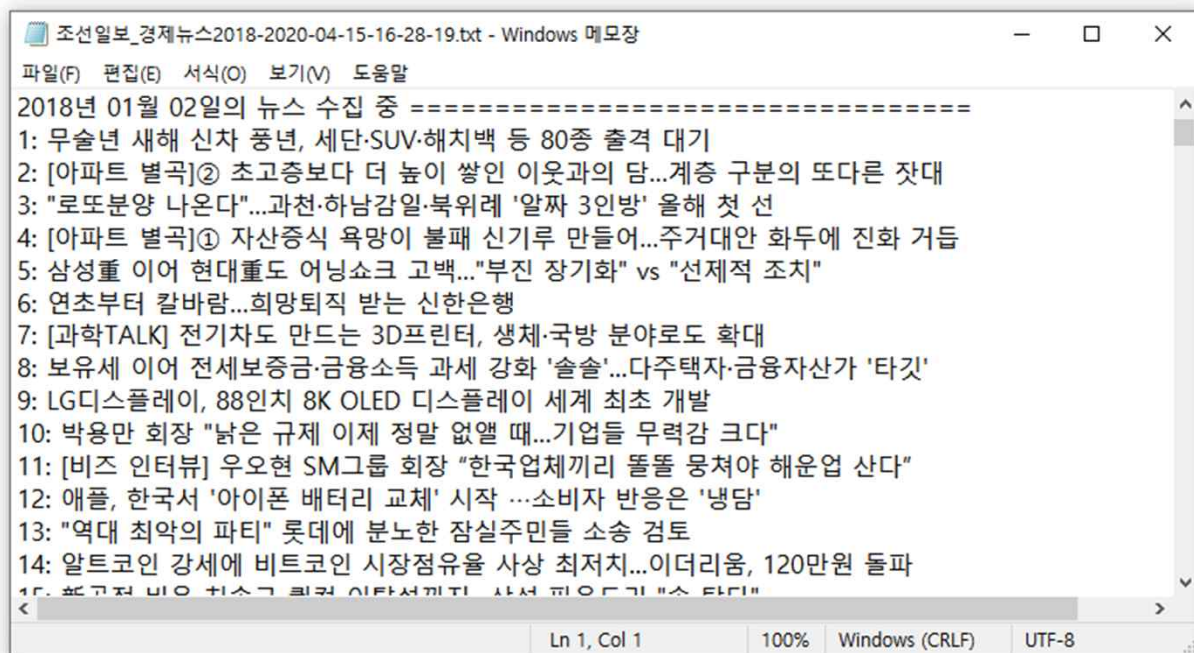
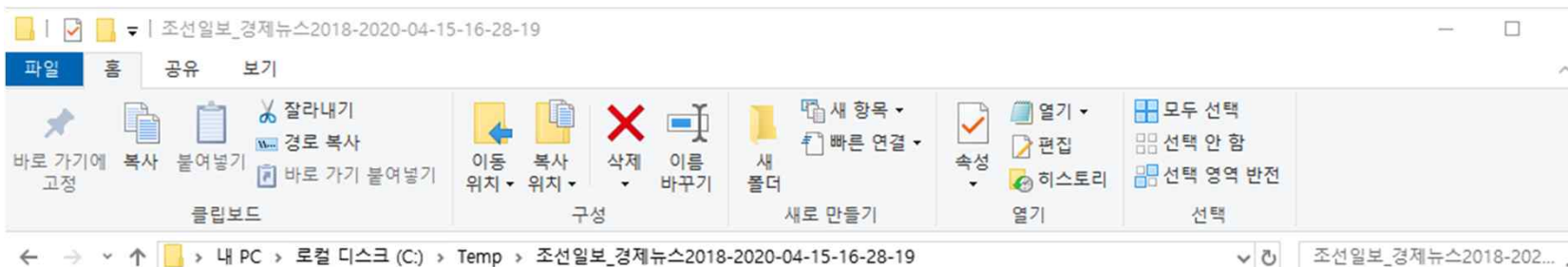
1.파일이 저장될 경로만 쓰세요; (예:c:\temp\): c:\data\

2018년 01월 01일의 뉴스 수집 중 =====

0 : [아파트 별곡]① 자산증식 욕망이 불패 신기루 만들어...주거대안 화두에 진화 거듭
1 : [과학TALK] 전기차도 만드는 3D프린터, 생체·국방 분야로도 확대
2 : [2018년 유통 전망] 최저임금 인상 발등에 불...제품가격 인상 역풍 부나
3 : LG디스플레이, 88인치 8K OLED 디스플레이 세계 최초 개발
4 : 박용만 회장 "낮은 규제 이제 정말 없앨 때...기업들 무력감 크다"
5 : 삼성重 이어 현대重도 어닝쇼크 고백..."부진 장기화" vs "선제적 조치"
6 : "中, 내년 말 D램 양산"...반도체 굴기 현실화
7 : 블랙홀 실체 드러나고...유전자 치료, 사람에 첫 적용
8 : 연초부터 게임 신작 쏟아진다...모바일·PC 동시 출격
9 : LG전자, CES서 인공지능 전시장 'LG 씽큐 존' 전면 배치
10 : "배터리 유상교체?"...애플 보상책 내놔도 국내 집단소송 희망자 18만명
11 : "그때 비트코인 샀어야지!"...세상에서 가장 우울한 사이트
12 : 작년 'CEO 연봉킹' 권오현 회장...200억원으로 추산
13 : 해외여행 간 한국인, 일본보다 800만명 많다
14 : [2018 신년사] 文대통령 "과거 잘못 바로잡는 노력 지속...국민 삶의 질 개선이 최우선 목표"
15 : 부활한 9만9000원 과일세트...유통업계 '김영란법' 개정예 고가 선물 늘린다
16 : 新공정 비용 치솟고 웰컴 이탈설까지...삼성 파운드리 "속 탄다"
17 : 황금개의 해 'GOLDEN DOG'로 풀어본 식품업계 트렌드
18 : '반도체가 이끈 2017년'...작년 수출액 사상 최대치 기록(종합)
19 : 새해 중국 경제 10대 추세...녹색규제 폭탄·위안화 절하 압력
20 : 올해 부동산 투자 키워드는 '물류창고'...美 경기회복 최대 수혜 예상
21 : 정부, 4조695억원 규모 R&D 종합시행계획 확정
22 : 가상화폐 관련 입법, 국회가 주저주저하는 까닭은
23 : 판교 찍고, 실리콘밸리로 날다
24 : [2018 IT 전망]③ 반도체 초호황 '경착륙'과 중국 ICT 굴기...블록체인 영역 확대
25 : 시총 2위 리플, 3000원대 안착...리플은 어떤 가상화폐?
26 : 이마트·신세계 온라인몰, 신년맞이 할인행사 '풍성'
27 : [2018년 車 전망] 친환경차 大戰 본격화...세단·SUV·고성능차 전부문 확산
28 : 법원 "하차요구 무시했다고 죄다 '감금' 아냐"
29 : 신한카드, 플랫폼 사업그룹 신설... "신한카드 내 14조 규모 디지털 기업"

Step6. 저장하기

```
#Step 6. 출력 결과를 파일에 저장하기  
e_time = time.time( )  
t_time = e_time - s_time  
  
print("총 소요시간은 %s 초 입니다 " %round(t_time,1))  
print("총 저장 건수는 %s 건 입니다 " %total_count)  
print("txt 파일 저장 완료: 파일명 : %s " %ff_name)  
  
driver.close( )
```



끝!