

1. Beautiful Soup

- HTML과 XML 파일로부터 데이터를 추출하기 위한 라이브러리
- \$ pip install beautifulsoup4

```
from bs4 import BeautifulSoup
html_doc = """
<html lang="ko" class="svgless">
<head>
<title>NAVER</title>
</head>
<body>
<p class='main'>Python Crawling Study</p>
<h3 class="ah_ltit">실시간 급상승</h3>
<a href="http://datalab.naver.com/keyword/realtimeList.naver?where=main"
  class="ah_ha" data-clk="lve.rankhistory">
  <span class="ah_ico_datalab">DataLab.</span>급상승 트래킹<span class="ah_ico_hlink"></span>
</a>
<div class="ah_tab">
<a href="#" role="tab" class="ah_tab_btn ah_tab_on" data-tab="1to10" data-clk="lve.tab1">1~10위</a>
<a href="#" role="tab" class="ah_tab_btn" data-tab="11to20" data-clk="lve.tab2">11~20위</a>
</div>
Loum ipsum, Loum ipsum, Loum ipsum, Loum ipsum, Loum ipsum
</body>
</html>
"""

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

print(soup.prettify())
```

1. Beautiful Soup

```
<html class="svgless" lang="ko">
<head>
  <title>
    NAVER
  </title>
</head>
<body>
  <p class="main">
    Python Crawling Study
  </p>
  <h3 class="ah_ltlt">
    실시간 급상승
  </h3>
  <a class="ah_ha" data-clk="lve.rankhistory" href="http://datalab.naver.com/keyword/realtimeList.naver?where=main">
    <span class="ah_ico_datalab">
      DataLab.
    </span>
    급상승 트래킹
    <span class="ah_ico_hlink">
    </span>
  </a>
  <div class="ah_tab">
    <a class="ah_tab_btn ah_tab_on" data-clk="lve.tab1" data-tab="1to10" href="#" role="tab">
      1~10위
    </a>
    <a class="ah_tab_btn" data-clk="lve.tab2" data-tab="11to20" href="#" role="tab">
      11~20위
    </a>
  </div>
  Loum ipsum, Loum ipsum, Loum ipsum, Loum ipsum, Loum ipsum
</body>
</html>
```

1. Beautiful Soup

- `soup.prettify()`
- `soup.title.string`
- `soup.title.parent.name`
- `soup.p`
- `soup.find_all('a')`
- `soup.get_text()`

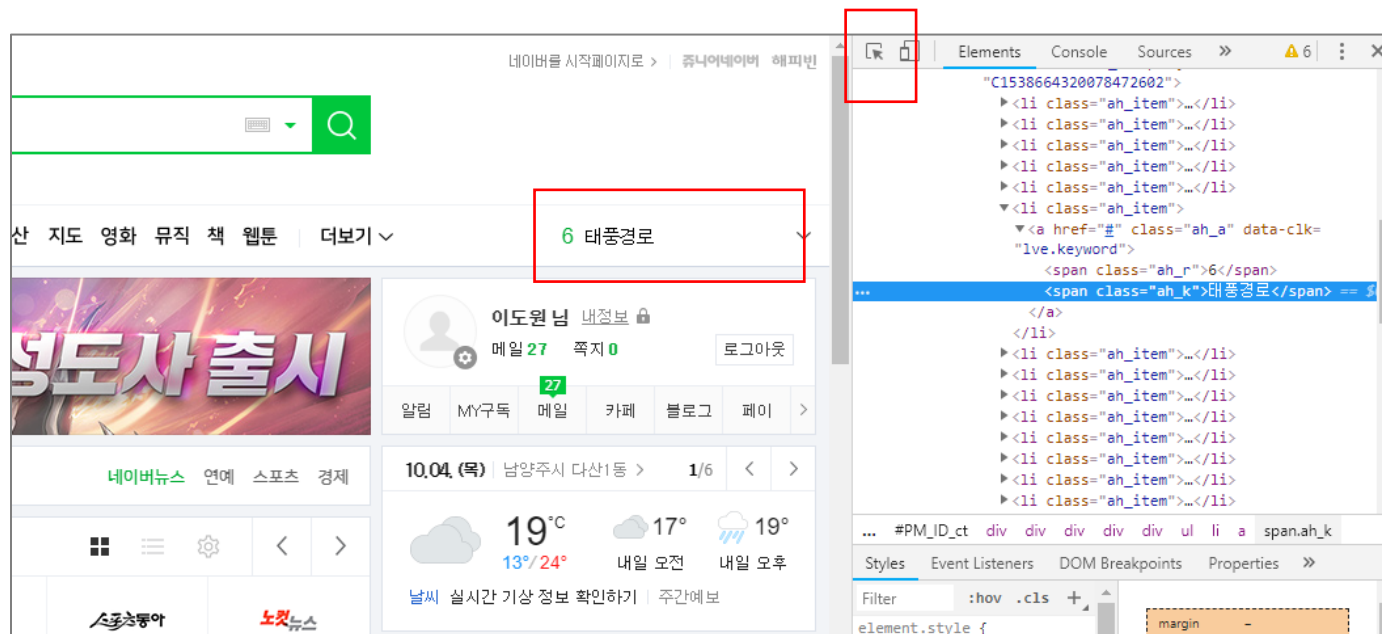
2. Requests

- 인터넷 상의 HTML 파일을 읽어온 다음 Parsing
- \$ pip install requests

```
import requests
URL = 'http://naver.com/'
response = requests.get(URL)
print(response.text)
```

3. Naver 웹 사이트 (실시간 검색어 분석)

- Crawling (or Scraping)
 - 웹 페이지를 그대로 가져온 다음 데이터를 추출하는 것 ➔ Crawler
 - 웹 상의 정보를 자동으로 검색하고 색인 할 때 사용되는 것 ➔ Spider, Bot, Intelligence Agent
- 크롬 웹브라우저 > 개발자모드 > Select 도구 ➔ 크롤링 하길 원하는 부분을 선택



3. Naver 웹 사이트 (실시간 검색어 분석)

```
    </li>
  <li class="ah_item">
    <a href="#" class="ah_a" data-clk=
      "lve.keyword">
      <span class="ah_r">12</span>
      <span class="ah_k">우리은행 채용</span>
    </a>
  </li>
```

- 마우스 우클릭 > Copy > Copy Selector ➔ 복사

```
#PM_ID_ct > div.header > div.section_navbar >
div.area_hotkeyword.PM_CL_realtimeKeyword_base >
div.ah_roll.PM_CL_realtimeKeyword_rolling_base > div > ul > li:nth-child(12) > a > span.ah_k
```

3. Naver 웹 사이트 (실시간 검색어 분석)

```
import requests
from bs4 import BeautifulSoup

req = requests.get('https://www.naver.com/')
source = req.text
soup = BeautifulSoup(source, 'html.parser')

print(soup.select("[Copy Selector]"))
```

```
"ah_k">판토스</span>, <span class="ah_k">내일날씨</span>, <span class="ah_k">서울날씨</span>, <span class="
피니
[<span class="ah_k">구하라</span>, <span class="ah_k">최종범</span>, <span class="ah_k">리벤지 포르노</span>
span
class="ah_k">태풍경로</span>, <span class="ah_k">손 the guest</span>, <span class="ah_k">외모지상주의</spa
<spa
n class="ah_k">판토스</span>, <span class="ah_k">내일날씨</span>, <span class="ah_k">서울날씨</span>, <span
ss="
[<span class="ah_k">구하라</span>, <span class="ah_k">최종범</span>, <span class="ah_k">리벤지 포르노</span>
>, <span class="ah_k">태풍경로</span>, <span class="ah_k">손 the guest</span>, <span class="ah_k">외모지상
주의</span>, <span class="ah_k">판토스</span>, <span class="ah_k">내일날씨</span>, <span class="ah_k">서울
날씨</span>, <span class="ah_k">인피니트</span>, <span class="ah_k">하늘에서 내리는 1억개의 별</span>, <spa
n class="ah_k">우리은행 채용</span>, <span class="ah_k">구하라 남자친구</span>, <span class="ah_k">태풍 콩
레이</span>, <span class="ah_k">태풍</span>, <span class="ah_k">만물상무장아찌</span>, <span class="ah_k">
대구날씨</span>, <span class="ah_k">부산날씨</span>, <span class="ah_k">리벤지 뜻</span>, <span class="ah_k"
">마성의 기쁨</span>]
```

3. Naver 웹 사이트 (실시간 검색어 분석)

```
import requests
from bs4 import BeautifulSoup

req = requests.get('https://www.naver.com/')
source = req.text
soup = BeautifulSoup(source, 'html.parser')

top_list = soup.select("#PM_ID_ct > div.header > div.section_navbar > div.area_hotkeyword.PM_CL_realtimeKeyword_base")

for top in top_list:
    print(top.text)
```

```
구하라
구하라 남자친구
최종범
리벤지 포르노
태풍경로
손 the guest
판토스
외모지상주의
내일날씨
인피니트
서울날씨
하늘에서 내리는 1억개의 별
이상엽
태풍 콩레이
```


4. Selenium

- 웹 애플리케이션 *테스트 프레임워크*
- 웹 사이트에서 버튼 클릭과 같이 이벤트 처리 가능
- JavaScript 실행 가능
- 웹 브라우저 실행을 대신하기 위한 Web Driver 설치 → Selenium이 사용하기 위한 웹 브라우저
 - <http://chromedriver.chromium.org/downloads>
 - 크롬 웹 브라우저와 버전 맞춰서 다운로드

\$ pip install selenium

```
from selenium import webdriver

path = "C:\\Users\\down\\Desktop\\work\\python_수업\\chromedriver_win32\\chromedriver.exe"
driver = webdriver.Chrome(path)
```

```
driver.get("http://google.com/")
```

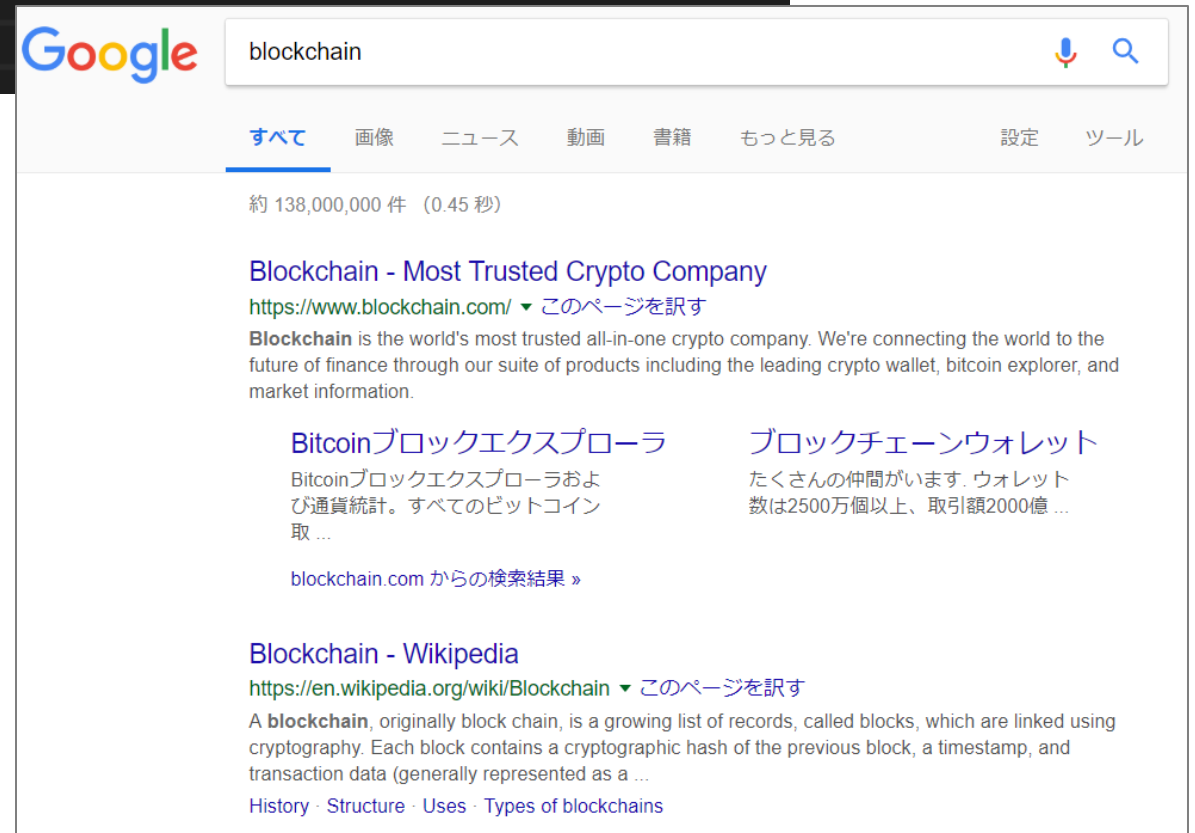
4. Selenium – 구글 검색

```
from selenium import webdriver

path = "C:\\\\Users\\down\\Desktop\\work\\python_수업\\chromedriver_win32\\chromedriver.exe"
driver = webdriver.Chrome(path)

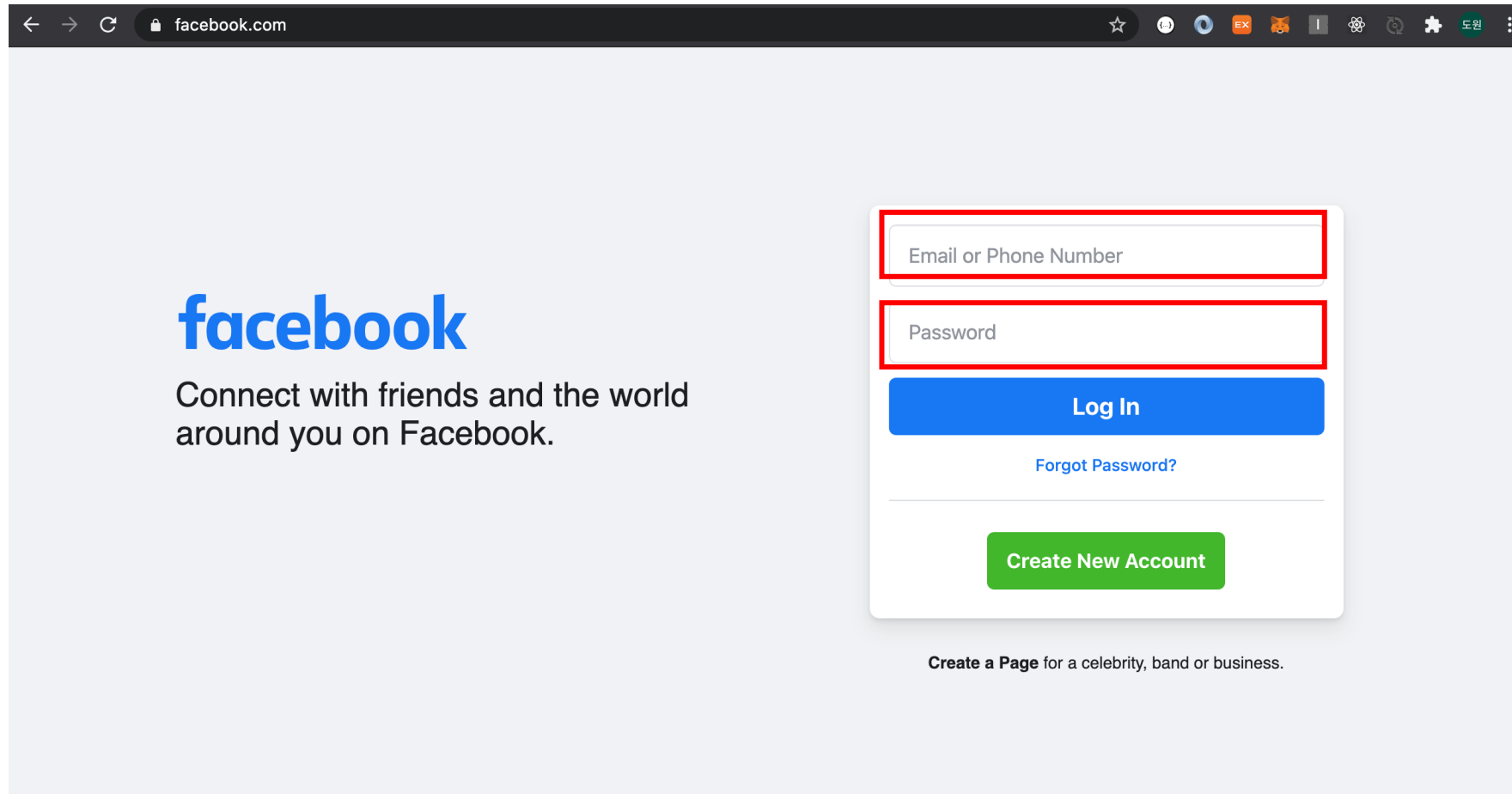
driver.get("http://google.com/")
search_box = driver.find_element_by_name("q")
search_box.send_keys("blockchain")
search_box.submit()
```

실습) 개발자 도구에서 검색어 부분 보기



5. Selenium – 페이스북 로그인

- input 태그에 name이나 id같은 선택자가 있으면 Selenium에서 테스트 가능



실습) 개발자 도구에서 이메일/비밀번호 찾아보기

5. Selenium – 페이스북 로그인

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from bs4 import BeautifulSoup

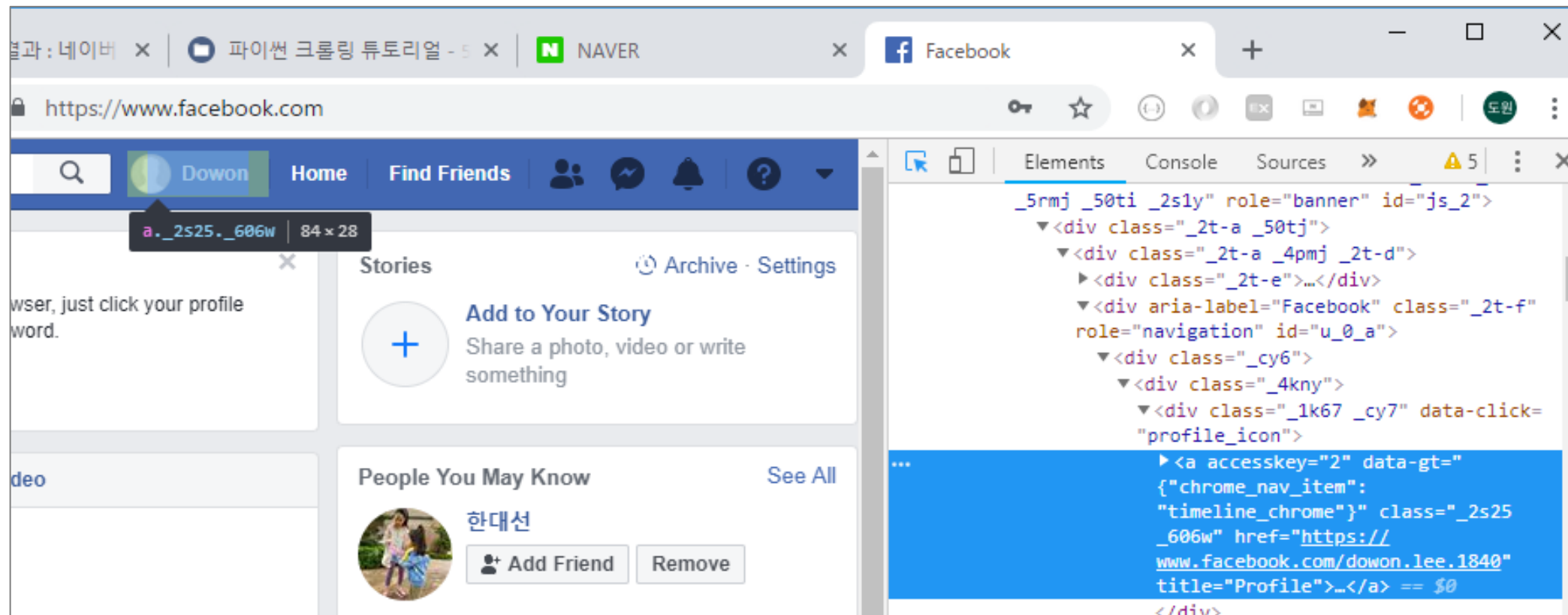
usr = "페이스북 아이디"
pwd = "페이스북 비밀번호"

path = "C:\\Users\\down\\Desktop\\work\\python_수업\\chromedriver_win32\\chromedriver.exe"

driver = webdriver.Chrome(path)
driver.get("https://www.facebook.com/")
assert "Facebook" in driver.title
elem = driver.find_element_by_id("email")
elem.send_keys(usr)
elem = driver.find_element_by_id("pass")
elem.send_keys(pwd)
elem.send_keys(Keys.RETURN)
```

실습) 로그아웃 처리

6. Selenium – 페이스북 프로필



- 프로필 링크 > Copy > Copy XPath

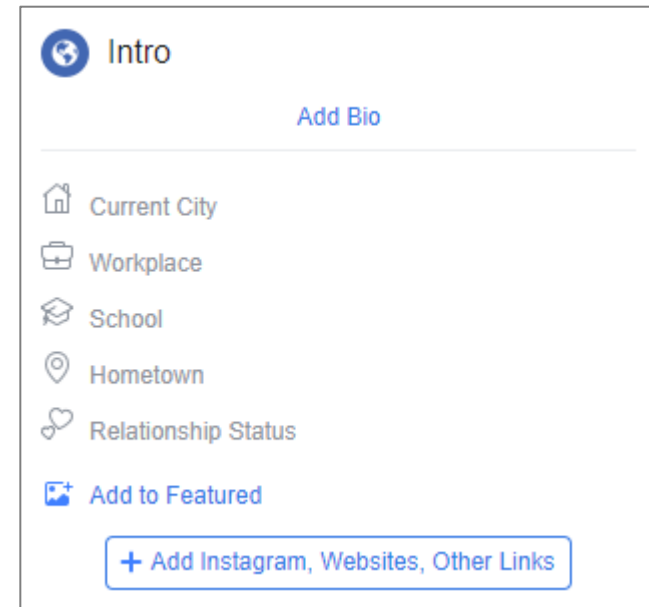
6. Selenium – 페이지스북 프로필

```
a = driver.find_element_by_xpath('//*[@id="mount_0_0"]/div/div[1]/div[1]/div[2]/div[4]/div[1]/div[4]/a')
driver.get(a.get_attribute("href"))
```

- find_elements_by_xpath의 결과는 list ➔ index로 추출하려는 속성에 접근

```
req = driver.page_source
```

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



7. Naver – 뉴스페이지 URL

- 네이버 부동산 뉴스 가져오기

```
from bs4 import BeautifulSoup
import requests

maximum = 0
page = 1

URL = 'http://land.naver.com/news/field.nhn?page=1'
response = requests.get(URL)
source = response.text
soup = BeautifulSoup(source, 'html.parser')

whole_source = ""
URL = 'https://land.naver.com/news/headline.nhn'
response = requests.get(URL)
whole_source = whole_source + response.text
soup = BeautifulSoup(whole_source, 'html.parser')
find_title = soup.select("#content > div.section_headline > ul > li > dl > dt > a")

for title in find_title:
    print(title.text)
```

7. Naver – 뉴스페/이지 URL

[9.13 대책 이후...르포]이사철인데...‘이른 한파’ 시작된 강남 부동산시장

취득세 뛰고 복비도 뛰고...매매 걸림돌 ‘겹겹’

"강북 실수요자만 잡는다"...기준금리 인상 실효성 논란
수도권 1주택자, 교육-근무 목적도 추가 대출 불허
정부, ‘집값 담합’에 칼 빼드나...감정원, 신고센터 운영

규제에도 노원구 부동산 시장 때 아닌 호황...물건 나오는 족족 거래 이뤄져

서울 5억·10억 이상 아파트 나란히 증가

정부 ‘집값 안정’ 그린벨트·금리 카드 만지지만...

국토부, 부실 감정평가 엄벌...징계맨 정보 공개

서울 아파트값 상승률 한달째 둔화... 강남3구 약세

"주택 규제 풍선효과"...경매시장 상가 낙찰가율↑

[2018국감]청년 금수저?... 5년간 20대 임대사업자 9배↑

최다 임대주택 등록자는 ‘부산거주 60대’...혼자서 604채 소유

"보증금 150억 어쩌나"...창원 임대업자 파산에 세입자들 ‘한숨’

7. Naver – 뉴스페이지 URL

- 페이지가 있는 URL 정보 가져오기


```
URL = 'http://land.naver.com/news/field.nhn?page=1'
response = requests.get(URL)
source = response.text
soup = BeautifulSoup(source, 'html.parser')

while 1:
    page_list = soup.findAll("a", {"class": "NP=r:" + str(page)})
    if not page_list:
        maximum = page - 1
        break
    page = page + 1
print("총 " + str(maximum) + " 개의 페이지가 확인 되었습니다.")
```

- 페이지 수 만큼 반복 처리

```
for page_number in range(1, maximum+1):
    URL = 'http://land.naver.com/news/field.nhn?page=' + str(page_number)
    response = requests.get(URL)
    whole_source = whole_source + response.text
```

실습) Github 이슈 가져오기


 Search or jump to... / Pull requests Issues Marketplace Explore

done409 / itman Watch 0 Star 0 Fork 0

[Code](#) **Issues 1** Pull requests 0 Projects 0 Wiki Insights Settings


First issue #1

Open done409 opened this issue 3 hours ago · 1 comment




done409 commented 3 hours ago Owner + 👤 ...

첫번째 이슈입니다.



done409 commented 3 hours ago Owner + 👤 ...

두번째 이슈고요...



Write Preview

AA B i “ > ↺ ☰ ☷ ☹ @ 📌 ↶

Leave a comment

Attach files by dragging & dropping, selecting them, or pasting from the clipboard.

Styling with Markdown is supported

Close issue Comment

Assignees

No one—assign yourself

Labels

None yet

Projects

None yet

Milestone


No milestone

Notifications

Unsubscribe

You're receiving notifications because you authored the thread.

1 participant



8. Scrapy

- 크롤링 vs 스크래핑

크롤링	스크래핑
웹에서 페이지 및 링크 다운로드 (웹을 기반으로 작동)	웹을 포함한 다양한 소스에서 데이터 추출 (반드시 웹과 관련된 것은 아님)
동일한 콘텐츠가 여러 페이지에 업로드 된 것을 인식하지 못하므로 중복 제거는 필수적	특정 데이터를 추출하는 것이므로 중복 제거가 반드시 필요한 것은 아님

8. Scrapy

- 수많은 웹 페이지로 부터 정보를 수집 ➔ 빅데이터로 활용
- Scraping을 위한 라이브러리
 - \$ pip install scrapy
- Scrapy Shell
 - \$ scrapy shell

```
In [1]: fetch('https://news.naver.com/main/list.nhn?mode=LSD&mid=sec&sid1=001')
2018-10-06 00:15:03 [scrapy.core.engine] INFO: Spider opened
2018-10-06 00:15:03 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://news.naver.com/main/list.nhn?mode=LSD&mid=sec&sid1=001> (referer: None)

In [2]: view(response) █
```

8. Scrapy

Books to Scrape We love being scraped!

Home / All products

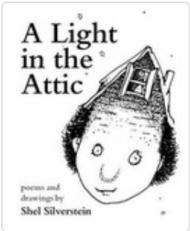
Books

- Travel
- Mystery
- Historical Fiction
- Sequential Art
- Classics
- Philosophy
- Romance
- Womens Fiction
- Fiction
- Childrens
- Religion
- Nonfiction
- Music
- Default
- Science Fiction
- Sports and Games
- Add a comment
- Fantasy
- New Adult
- Young Adult
- Science
- Poetry
- Paranormal

All products

1000 results - showing 1 to 20.

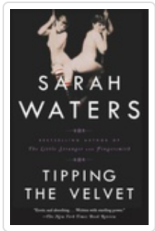
Warning! This is a demo website for web scraping purposes. Prices and ratings here were randomly assigned and have no real meaning.



A Light in the ...

★★★★☆


£51.77



Tipping the Velvet

★★★★☆

£53.74



Soumission

★★★★☆

£50.10

```
<div class="alert alert-warning" role="alert">...</div>
<div>
  <ol class="row">
    ::before
    <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
      <article class="product_pod">
        <div class="image_container">...</div>
        <p class="star-rating Three">...</p>
        <h3>
          <a href="catalogue/a-light-in-the-attic_1000/
            index.html" title="A Light in the Attic">A
            Light in the ...</a> == $0
        </h3>
        <div class="product_price">...</div>
      </article>
    </li>
    <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
      ...</li>
    <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
      ...</li>
    <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
      ...</li>
  </ol>
</div>
```

ol.row li.col-xs-6.col-sm-4.col-md-3.col-lg-3 article.product_pod h3 a

스타일 계산됨 레이아웃 이벤트 리스너 DOM 중단점 속성 접근성

필터 :hov .cls + - < >

```
element.style {
}

a {
  color: #428bca;
  text-decoration: none;
}
```

콘솔 새로운 기능 문제

Highlights from the Chrome 116 update

Improved debugging of missing stylesheets

- \$ print(response.text)

8. Scrapy

- 크롤링 타겟
 - 제목
 - 올린 뉴스 사이트
 - 미리보기 내용

보안/해킹



인색시큐리티, '2018 오프스왓 세미나' 18일 개최

[디지털데일리 홍하나기자] 디지털포렌식 및 네트워크 보안 전문업체인 인색시큐리티(주)가 18일 오프스왓 세미나를 개최한다. 디지털데일리 | 7시간전

//*[@id="main_content"]/div[2]/ul[1]/li[1]/dl/dt[2]/a



사이버공격방어대회·제주사이버보안컨퍼런스 동시 개최

2018년도 사이버공격방어대회와 제주사이버보안컨퍼런스 개최가 한 달 앞으로 다가왔다. 전자신문 | 8시간전

//*[@id="main_content"]/div[2]/ul[1]/li[2]/dl/dt[2]/a



17일 제1회 블록체인 국제 표준 워크숍 열려

분산원장기술표준포럼(의장 류재철)은 17일 섬유센터에서 제1회 블록체인 국제 표준 워크숍을 개최한다. 전자신문 | 10시간전

//*[@id="main_content"]/div[2]/ul[1]/li[3]/dl/dt[2]/a

8. Scrapy

- XPath를 가지고 크롤링

```
//*[@id="main_content"]/div[2]/ul[1]/li[3]/dl/dt[2]/a
```

```
In [10]: response.xpath('//*[@id="main_content"]/div[2]/ul/li/dl/dt[2]/a/text()').extract()
```

[illegible]

8. Scrapy

- CSS를 가지고 크롤링 → 올린 뉴스 사이트 출력

```
response.css('.writing::text').extract()
```

```
·헤럴드POP',  
·파이낸셜뉴스',  
·한국경제',  
·EPA연합뉴스',  
·AP연합뉴스',  
·EPA연합뉴스',  
·EPA연합뉴스',  
·EPA연합뉴스',  
·아시아경제',  
·파이낸셜뉴스',  
·뉴스1',  
·EPA연합뉴스']
```


8. Scrapy

- XPath를 가지고 크롤링 → 미리보기 출력

```
response.css('.lede::text').extract()
```

```
"Egyptian-Saudi Super Cup - press conference Egypt's Zamalek head coach ...",
'Sonia Guajajara In this Sept. 15, 2018 photo, Sonia Guajajara, an indig ...',
"Egyptian-Saudi Super Cup - press conference Egypt's Zamalek head coach ...",
"Egyptian-Saudi Super Cup - press conference Egypt's Zamalek player Haze ...",
"Egyptian-Saudi Super Cup - press conference Egypt's Zamalek head coach ...",
"'나 혼자 산다'에 출연하는 배우 이시언이 관악산 별장을 찾는다. 5일 방송된 MBC '나 혼자 산다'에서는 관악산으
로 물놀이를 ...",
'[인사] 군포시 ㄹ급 승진 △복지문화국장 현승식 △건설교통국장 홍재섭 △군포1동장 안선수 ㄹ급 전보 △기획
재정국장 배재철 △...',
'(평양·서울=뉴스1) 공동취재단, 김다혜 기자 = 10·4선언 11주년 기념 민족통일대회 참석을 위해 평양을 방문한 우
리 방북단이 ...',
"Egyptian-Saudi Super Cup - press conference Egypt's Zamalek head coach ..."]
```

9. Scrapy – Sample

- Scrapy 프로젝트 구조

```
tutorial/
  scrapy.cfg          # deploy configuration file

tutorial/
  __init__.py         # project's Python module, you'll import your code from here

  items.py            # project items definition file

  middlewares.py      # project middlewares file

  pipelines.py        # project pipelines file

  settings.py         # project settings file

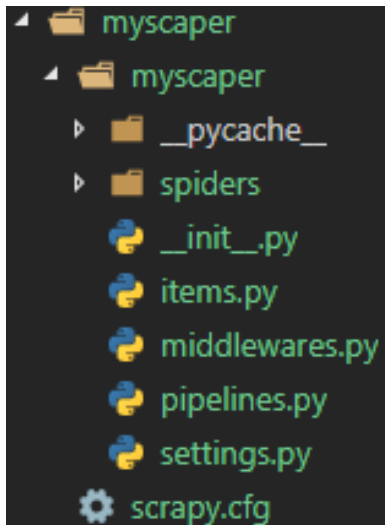
  spiders/
    __init__.py       # a directory where you'll later put your spiders
```

9. Scrapy – Project 작성

- \$ scrapy startproject myscaper

```
PS C:\Users\down\Desktop\work\python_수업\ch17\naverscraper> scrapy startproject myscaper
New Scrapy project 'myscaper', using template directory 'c:\\users\\down\\anaconda3\\lib\\site-pack
ages\\scrapy\\templates\\project', created in:
    C:\Users\down\Desktop\work\python_수업\ch17\naverscraper\myscaper

You can start your first spider with:
    cd myscaper
    scrapy genspider example example.com
PS C:\Users\down\Desktop\work\python_수업\ch17\naverscraper> █
```



9. Scrapy – Spider 작성

- \$ scrapy genspider mybots

“news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=732”

```
PS C:\Users\dowon\Desktop\work\python_수업\ch17\myscaper\myscaper> scrapy genspider mybots "news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=732"
Created spider 'mybots' using template 'basic' in module:
myscaper.spiders.mybots
PS C:\Users\dowon\Desktop\work\python_수업\ch17\myscaper\myscaper> 
```

9. Scrapy – Spider 작성

- spiders/mybots.py 수정

```
# -*- coding: utf-8 -*-
import scrapy

class MybotsSpider(scrapy.Spider):
    name = 'mybots'
    allowed_domains = ['news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=732']
    start_urls = ['http://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=732/']

    def parse(self, response):
        pass
```

9. Scrapy – Spider 작성

- items.py 수정

```
import scrapy

class MyscaperItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    title = scrapy.Field()
    author = scrapy.Field()
    preview = scrapy.Field()
```

9. Scrapy – Spider 작성

- spiders/mybots.py 수정

```
import scrapy
from myscaper.items import MyscaperItem

class MybotsSpider(scrapy.Spider):
    name = 'mybots'
    allowed_domains = ['naver.com']
    start_urls = ['https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=732']

    def parse(self, response):
        titles = response.xpath('//*[@id="main_content"]/div[2]/ul/li/dl/dt[2]/a/text()').extract()
        authors = response.css('.writing::text').extract()
        previews = response.css('.lede::text').extract()

        # for item in zip(titles, authors, previews):
        #     scraped_info = {
        #         'title' : item[0].strip(),
        #         'author' : item[1].strip(),
        #         'preview' : item[2].strip(),
        #     }
        #     yield scraped_info
        items = []
        for idx in range(len(titles)):
            item = MyscaperItem()
            item['title'] = titles[idx]
            item['author'] = authors[idx]
            item['preview'] = previews[idx]
            items.append(item)
        return items
```

9. Scrapy – Spider 작성

- settings.py 수정

```
BOT_NAME = 'myscaper'

SPIDER_MODULES = ['myscaper.spiders']
NEWSPIDER_MODULE = 'myscaper.spiders'

# Crawl responsibly by identifying yourself (and your website) on the user-agent
#USER_AGENT = 'myscaper (+http://www.yourdomain.com)'

# Obey robots.txt rules
ROBOTSTXT_OBEY = False
```

```
FEED_FORMAT = "csv"
FEED_URI = "my_news.csv"
```


9. Scrapy – Spider 실행

- \$ scrapy crawl mybots

```
2018-10-06 00:51:05 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://news.naver.com/main/list.nhn?mode=
{'author': '한국경제',
'preview': '호주 스타트업 트래블바이비트에 250만 달러 투자 [ 오세성 기자 ] 글로벌 가상화폐(암호화폐) 거래소
'세계 주요 공 ...',
'title': '\r\n'
'\t\t\t\t\t\t\t\t\t\t SK인포섹, 사물인터넷(IoT) 보안 가이드북 발간\r\n'
'\t\t\t\t\t\t\t\t\t\t'}
2018-10-06 00:51:05 [scrapy.core.engine] INFO: Closing spider (finished)
2018-10-06 00:51:05 [scrapy.extensions.feedexport] INFO: Stored csv feed (19 items) in: naver_news.csv
2018-10-06 00:51:05 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 262,
'downloader/request_count': 1,
'downloader/request_method_count/GET': 1,
'downloader/response_bytes': 19835,
'downloader/response_count': 1,
'downloader/response_status_count/200': 1,
'finish_reason': 'finished',
'finish_time': datetime.datetime(2018, 10, 5, 15, 51, 5, 626339),
'item_scraped_count': 19,
'log_count/DEBUG': 21,
'log_count/INFO': 8,
```

9. Scrapy – Spider 실행

- \$ CSV 파일 확인

1	author,preview,title
2	
3	디지털데일리, [디지털데일리 홍하나기자] 디지털포렌식 및 네트워크 보안 전문업체인 인섹시큐리티 (대표 김종광)는 악성코드 탐지 전문 업체 오피스왓 ..., "
4	
5	인섹시큐리티, '2018 오피스왓 세미나' 18일 개최
6	
7	"
8	
9	전자신문, 2018년도 사이버공격방어대회와 제주사이버보안컨퍼런스 개최가 한 달 앞으로 다가왔다. 사이버공격방어대회 본선은 29일부터 30일 ..., "
10	
11	사이버공격방어대회·제주사이버보안컨퍼런스 동시 개최
12	
13	"
14	
15	전자신문, "분산원장기술표준포럼 (의장 류재철)은 17일 섬유센터에서 제1회 블록체인 국제 표준 워크숍을 개최한다. 이번 행사는 'ISO, I ...', "
16	
17	17일 제1회 블록체인 국제 표준 워크숍 열려
18	
19	"
20	
21	아이뉴스24, <아이뉴스24> [아이뉴스24 김국배 기자] SK텔레콤이 계열사인 국내 1위 정보보안 업체 SK인포섹 인수 추진을 공식화하고 있 ..., "
22	
23	SK텔레콤, SK인포섹 인수 검토..."융합보안 확장"
24	
25	"
26	
27	한국경제, 10월 14일 전후 테스트넷에 비잔티움 하드포크 적용 전망 메인넷 적용도 연내 이뤄질 듯 [오세성 기자] 이더리움이 테스트넷 ..., "
28	
29	코앞으로 다가온 이더리움 하드포크, PoS 전환 본격화
30	
31	"

9. Scrapy – 실습

- 네이버 영화 평점 가져오기

movie.naver.com/movie/point/af/list.nhn

NAVER 영화

영화홈
상영작 · 예정작
영화랭킹
예매
평점 · 리뷰
네티즌 평점
네티즌 리뷰
다운로드
인디극장

로그인 영화검색 검색

네티즌·관람객 평점

현재 상영작 평점보기
현재 상영작

전체 리스트 총 13,567,582개의 평점이 있습니다.

번호	감상평	글쓴이·날짜
17588413	서터 ★★★★★ 10 오늘 새벽에 해당 영화를 보았습니다. 불쌍한 여인 한명과 나쁜 놈들에 대한 이야기입니다. 천벌을 받은듯 합니다 그놈들 말입니다. 귀신이 무섭지는 안았습니다. 슬픈 이야기였습니다. 고인의 명복을 빕니다. 신고	clwe**** 21.07.10
17588412	더 게임 ★★★★★ 10 미친 개재밌음 마지막 크레딧음악까지 완벽해17여게인을 호러스럽고 기괴하게 담아냄 신고	lo_t**** 21.07.10
17588411	킬러의 보디가드 2 ★★★★★ 4 스토리로 보나 액션으로 보나 전편보다 다운그레이드. 입담만큼은 업그레이드. 신고	hwun**** 21.07.10
17588410	블랙 위도우 ★★★★★ 8 그동안 고생했어 나타샤 신고	yj80**** 21.07.10

영화 인기검색어 더보기

- 블랙 위도우 - 0
- 랑종 ↑ 1
- 발신제한 ↓ 1
- 크루엘라 ↑ 1
- 미드나이트 ↓ 1

2021.07.09

네티즌 최고 평점 더보기

현재 상영영화 모든 영화

- 극장판 귀멸의 칼날. 9.28

이미지 준비중

- 부활: 그 증거 9.27
- 크루엘라 9.25
- 해피 투게더 9.19
- 우드잡 9.12

2021.07.09까지의 관람후 누적 평점

- 제목
- 평점
- 작성자
- 작성일
- 리뷰

9. Scrapy – 실습

- 네이버 영화 평점 가져오기 → 결과 예시

```
1 date,desc,star,title,writer
2 21.07.10,-,2,제8일의 밤,bria****
3 21.07.10,나훈진 감독이 왜 이 영화의 장르를 로맨스라고 했는지 이해된다. 한 남자의 부인을 찾기 위한 혈투 과정. 하정우의 처절한 연기가 돋보인다.,8,황해,mn10****
4 21.07.10,연기 너무 잘하고 스토리도 너무 좋고 그냥 보세요,10,쿨,chfh****
5 21.07.10,정말정말로 재밌네요~,10,이번엔 잘 되겠지,chkw****
6 21.07.10,재평가가 시급히 필요한 영화,10,사바하,thea****
7 21.07.10,스릴러 장르에 충실하면서도 특색있는 연출. 하지만 이 영화의 진정한 가치는 하정우와 김윤석 두 배우의 압도적인 연기력에 있다.,9,추격자,mn10****
8 21.07.10,- 윈터솔저 마블 시리즈중에 토르이후로 젤 많이 봤지싶은;;ㅋ4D까지해서 3~4번정도 본거 같은데수트가 참 멋졌음♡,10,캡틴 아메리카: 윈터 솔저,bd12****
9 21.07.10,-,10,클래식,ybd1****
10 21.07.10,앞부분은 흥미로웠으나 노골적인 자동차 ppl에 윈스러웠고 뻔한 줄거리에 결말이라 지루했다 주인공의 연기력은 굿!,6,발신제한,litt****
11 21.07.10,스토리는 다 예측가능했고 연출이랑 CG는 볼만 했음. 스토리가 너무 부실한것만 빼면 재미있었을텐데 많이 아쉬웠음. 특히 마지막 부분에는 갑자기 급전개에다가 플래그 오지게 세움,6,블랙 위도우,s173****
```

- 리뷰 데이터의 공백 처리 문제

10. Scrapy – 실습 (Kafka 연동)

- settings.py 에 PIPELINE 추가

```
63  # Configure item pipelines
64  # See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
65  ITEM_PIPELINES = {
66      'moviereview.pipelines.MoviereviewPipeline': 300,
67  }
```

10. Scrapy – 실습 (Kafka 연동)

- pipelines.py

```
6 class MoviereviewPipeline:
7     def __init__(self):
8         self.producer = KafkaProducer(acks=0,
9                                       compression_type='gzip',
10                                      bootstrap_servers=['127.0.0.1:9092'],
11                                      value_serializer=lambda x: dumps(x).encode('utf-8'))
12
13
14         self.topic_prefix = "naver_movie2"
15
16     def process_item(self, item, spider):
17         datum = dict(item)
18         prefix = self.topic_prefix
19
20         review_topic = "{prefix}.crawled_review".format(prefix=prefix)
21         self.producer.send(review_topic, datum)
22
23         return item
```

10. Scrappy – 실습 (Kafka 연동)

- KafkaConsumer 예제에서 확인

```
7 | consumer = KafkaConsumer('naver_movie2.crawled_review',
8 |                           bootstrap_servers=['127.0.0.1:9092'],
9 |                           auto_offset_reset='earliest',
10 |                          enable_auto_commit=True,
11 |                          group_id='my-group',
12 |                          value_deserializer=lambda x: loads(x.decode('utf-8')),
13 |                          consumer_timeout_ms=1000
14 |                          )
15
16 | start = time.time()
17 | print("START=", start)
18 | for message in consumer:
19 |     topic = message.topic
20 |     partition = message.partition
21 |     offset = message.offset
22 |     value = message.value
23 |     timestamp = message.timestamp
24 |     datetimeobj = datetime.datetime.fromtimestamp(timestamp/1000)
25
26 |     print("T:{}, P:{}, O:{}, V:{}, D:{}".format(topic, partition, offset, value, datetimeobj))
27
28 | print("Elapsed time=", (time.time() - start))
```


10. Scrapy – 실습 (Kafka 연동)

- KafkaConsumer 예제에서 확인

```
START= 1625924991.798838
```

```
T:naver_movie2.crawled_review, P:0, 0:0, V:{'title': '블랙 위도우 ', 'star': '8', 'desc': '스토리는 진부하긴한데 액션이랑 영상미가 좋아요 크으으 멋져멋져 ', 'writer': 'edh8***', 'date': '21.07.10'}, D:2021-07-10 22:48:56.836000
```

```
T:naver_movie2.crawled_review, P:0, 0:1, V:{'title': '터미네이터: 미래전쟁의 시작 ', 'star': '10', 'desc': '생각할수록 명작 솔직히 5,6보다 재미있는 영화였음 존 코너의 얼굴에 왜 흥터가 생겼는지 떡밥도 풀어주는 훌륭한 영화숨통을 꿰어도 내 손으로 꿰는다 -다크페이트 제임스 카메론-', 'writer': 'babo***', 'date': '21.07.10'}, D:2021-07-10 22:48:56.837000
```

```
T:naver_movie2.crawled_review, P:0, 0:2, V:{'title': '유전 ', 'star': '10', 'desc': '이 영화는 기존 공포영화랑은 다른 공포를 느끼게 해주는데 그 공포감이 기존 공포영화에서 느낄 수 있는 공포감보다 훨씬 높은 공포를 느끼게 해줌 다만 조금 지루할 수 있는 그런 시간대가 있는데 연기력이 너무 좋아서 감탄하고 보게된 공포영화를 좋아하는 사람이라면 또 영화에서 나오는 잔인한장면을 볼 수 있는 사람들 이라면 볼만한 영화임 지금 평점이 7.9인데 평점 준것들을 보면 누구는10점 누구는1점 이런식으로 호불호가 갈림그러니까 공포영화를 좋아하는 사람들은 이 영화의 작품성을 알아보고 고점을 주는거고 반대로 그냥 본 사람들은 불쾌하고 하니까 저점을 주는거임 ', 'writer': 'honr***', 'date': '21.07.10'}, D:2021-07-10 22:48:56.840000
```