웹 크롤링 가이드

서울과학기술대학교 데이터사이언스학과

DDUA 데이터기반 사용자 분석 연구실

INDEX

01. Crawling

02. HTTP & HTML

03. Web Crawling Library

04. Web Crawling Exercise

1

Crawling

Introduction to Crawling

Web Page?

- Web Page는 HTML / CSS / JavaScript 하나로 이루어져 있는걸 말함
- HTML : 하이퍼 텍스트 마크업 언어
- CSS : 종속형 시트 / 캐스케이딩 스타일 시트 (Csscading Style Sheets)로 마크업 언어가 실제 표시되는 방법을 기술하는 언어
- JavaScript : 객체 기반의 스크립트 프로그래밍 언어
- 마크업 언어 : 태그등을 이용하여 문서 데이터의 구조를 명기하는 언어

Introduction to Crawling

API?

- API 서비스란 바로 쓸 수 있는 데이터의 형태 (CSV,Excel 등) 로 제공하는 서비스
- 국내에서는 거의 존재하지가 않음
- 네이버 등에서도 유지가 잘 되고 있지 않음

네이버 오픈 API 목록

네이버 오픈API 목록 및 안내입니다.

API명	설명	호출제한
검색	네이버 블로그, 이미지, 웹, 뉴스, 백과사전, 책, 카페, 지식iN 등 검색	25,000회/일
네이버 로그인	외부 사이트에서 네이버 로그인 기능 구현	없음
네이버 회원 프로필 조회	네이버 회원 이름, 이메일 주소, 휴대전화번호, 별명, 성별, 생일, 연령대, 출생연도, 프로필 조회	없음
Papago 번역	Papago 번역 인공신경망 기반 기계 번역	10,000글자/일
CLOVA Face Recognition	입력된 사진을 입력받아 얼굴윤곽/부위/표정/유명인 닮음도를 리턴	1,000건/일

<List Example of Naver API>

Introduction to Crawling

Crawling?

- 정제되지 않은 **웹 페이지**에서 **필요한 데이터**만 추출해내는 행위
- 크롤링하는 소프트웨어는 크롤러(Crawler)라고 부름
- **API**를 통해 얻어오지 못하는 정보도 크롤링을 통해 얻음

2

HTTP & HTML

Introduction to HTTP & HTML

HTTP (Hyper Text Transfer Protocol)

- HTTP의 기본적 원리는 서버 (Server – 서비스를 제공하는 측) 와 클라이언트(Client – 일반 사용자) 으로 구성 되어있음
- 구글이나 페이스북 등 제공해주는 서비스를 사용하는 대상은 클라이언트이며
- 이를 웹 브라우저를 통해 행동을 하면 클라이언트와 서버간 의 네트워크가 이루어짐

Introduction to HTTP & HTML

HTTP Method

- HTTP 에 따라 하이퍼 텍스트를 주고 받는데 HTTP에서 서버에 요청 할 때 요청의 정보 중에 현재의 요청이 어떤 종류라는 정보를 보내는데 이러한 요청의 종류를 HTTP Method 라고 함

- HTTP Method에는 GET 이라는 서버에서 서버의 자원(텍스트,이미지,압축파일 등) 을 가져올 때 보내는 요청이 존재

- 다운로드, 웹의 글을 보는 것들이 모두 GET을 사용하여 이루어짐

Introduction to HTTP & HTML

HTTP Method

- HTTP 에 따라서 하이퍼 텍스트를 주고 받는데, HTTP에서 서버에 요청 할 때 요청의 여러 정보 중 현재 요청이 어떤 종류의 정보를 보냄

- 이러한 요청을 HTTP Method 라고 정의

- HTTP Method에는 GET 이라는 서버에서 텍스트,이미지,압축파일 등 을 가져올 때 보내는 요청은 모두 GET을 사용하여 이루어짐

Introduction to HTTP & HTML

Request

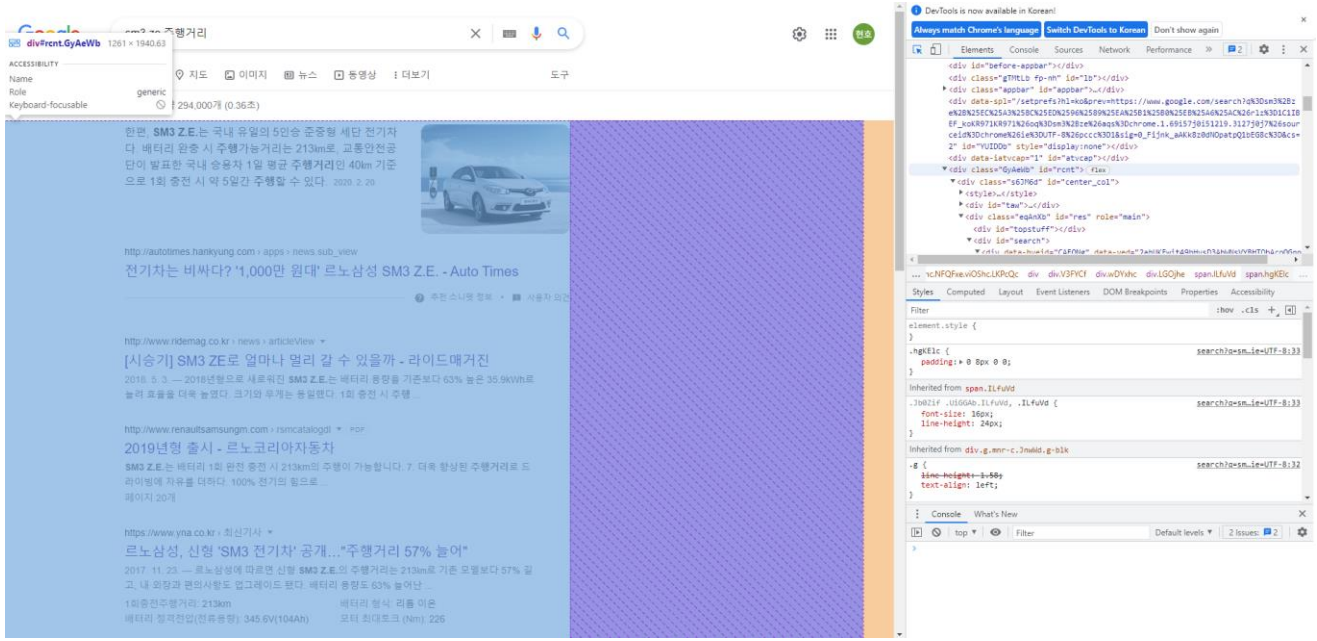
```
import lxm.html
import requests

url = "https://search.naver.com/search.naver?where=nexearch&sm=top_hty&fbm=1&ie=utf8&query=LG+%EC%A0%84%EC%9E%90"

res = requests.get(url)
```

<Example of request>

- Request는 파이썬으로 HTTP 호출하는 프로그램을 작성할 때 종종 사용되는 라이브러리

Introduction to HTTP & HTML

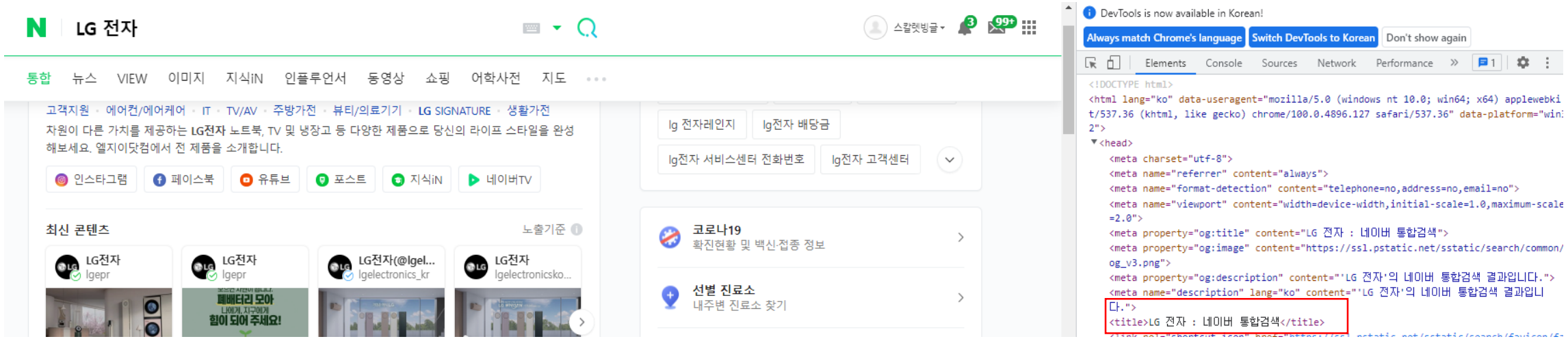
개발자도구



- 웹의 구조 이외에도 다양한 내용이 포함되어 있음
- F12를 통해 간단히 열 수 있다.

Introduction to HTTP & HTML

HTML – 태그(tag)



- HTML 에서 mark up 을 하기 위한 표기법으로 <태그이름> 형식을 사용해서 태그를 달아줌
- 여는 태그와 닫는 태그가 쌍을 이루며 닫는 태그는 태그 이름 앞에 '/'을 붙여 줘야 한다

Introduction to HTTP & HTML

HTML – 요소(element)

```
▼<body class="wrap-new api_animation vsc-initialized">  == $0
    <div id="nxtt_div" style="display:none;position:absolute;border-width:0;z-index:116
    0"></div>
  ▶<div id="u_skip">…</div>
  ▶<div id="wrap">…</div>
  ▶<script type="text/javascript">…</script>
  ▶<script type="text/javascript">…</script>
  </body>
</html>
```

- 여는 태그와 닫는 태그로 감싸인 하나의 영역을 요소 라고 정의
- 요소 안에는 또 다른 요소가 포함 가능

Introduction to HTTP & HTML

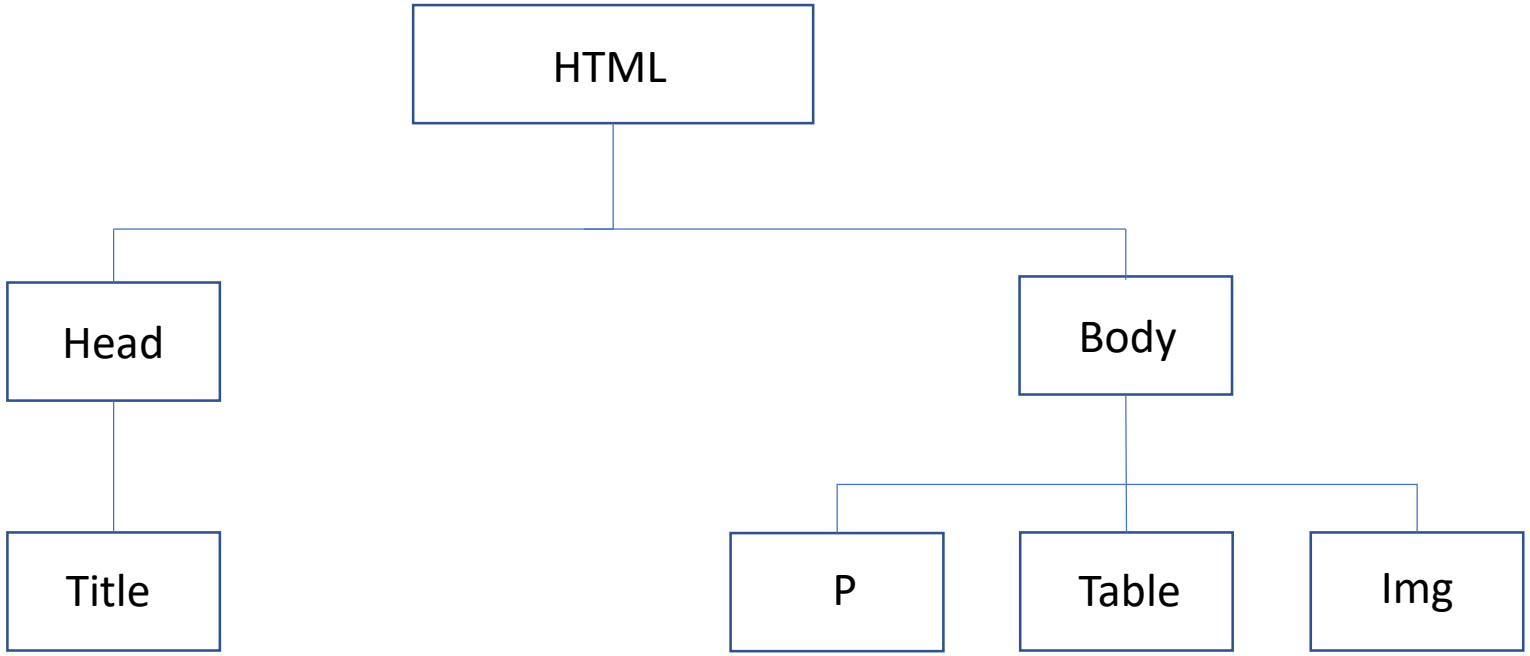
HTML – 속성(attribute)

```
<body class="wrap-new api_animation vsc-initialized">
  <div id="nxtt_div" style="display:none;position:absolute
  0"></div>
  <div id="u_skip">
    <a href="#lnb">…</a>
    <a href="#content">…</a>
  </div>
```

- HTML에서 요소가 추가적인 정보가 필요하다면 태그에 다가 추가적인 속성을 표시해줌
- 속성을 이용해 요소가 가지는 부가적인 정보를 표시

Introduction to HTTP & HTML

HTML Architecture



3

Web Crawling Library

Introduction to Web Crawling Library

Request : 파이썬에서 동작하는 작고 빠른 브라우저

- 웹 서버로 부터 초기 HTML만 받음, 추가 CSS/javascript 처리 불가능함
- 모든 플랫폼에서 구동 가능 함

Selenium : 브라우저를 원격 컨트롤하는 테스팅 라이브러리

- Chrome ,Firefox , IE , Phantom JS 등 에서 사용 가능 함
- 기존 브라우저를 사용하므로 추가 CSS/Java script 처리 지원
- 리소스를 많이 사용하는 단점이 존재 함

Introduction to Web Crawling Library

Selenium

- 셀레니움을 사용하기 이전 준비해야할 단계가 존재함
  1. Pip install 을 통해 셀레니움을 설치 진행
  2. 셀레니움을 통해 제어 하고자 하는 웹 브라우저에 맞는 웹 드라이버 설치

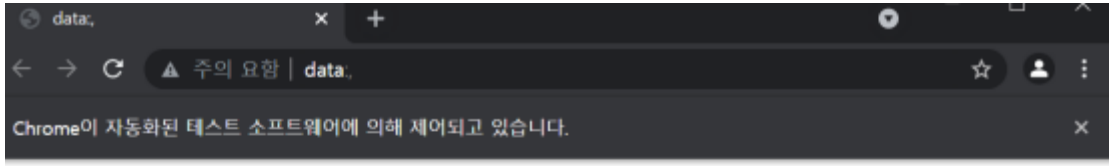
```
(base) C:₩Windows₩system32>pip install selenium
```

Introduction to Web Crawling Library

Selenium

- 웹 드라이버 설치까지 완료했다면 파이썬 환경에서 셀레니움을 불러 올수가 있음

```
from selenium import webdriver
import time


driver = webdriver.Chrome('C:/Users/201/Desktop/chromedriver_win32/chromedriver.exe')
driver.get("https://www.youtube.com/watch?v=lAPAO3qK3_w")
```



Introduction to Web Crawling Library

Selenium Head less

- 셀레니움을 실행 시키면 Web driver에 의해서 크롬 브라우저가 열리는데 크롬 브라우저가 안 열리고 크롤링을 진행하고 싶다면 셀레니움의 Head less 기능을 사용 할 수 있음

```
from selenium import webdriver

options = webdriver.ChromeOptions()
options.add_argument('headless')
driver = webdriver.Chrome('chromedriver.exe')
```

Introduction to Web Crawling Library

BeautifulSoup4 : HTML 파서

- 지정 HTML 로 부터 원하는 위치/형식의 문자열을 획득 가능함
- 주로 Requests에 의해 많이 사용함
- Selenium에서도 사용 가능함

4

Web Crawling Excercise

Introduction to Web Crawling Excercise

HTML PARSER – Daum News

Introduction to Web Crawling Excercise

HTML PARSER – Daum News

1. 필요한 라이브러리 정의

```
import requests
import pandas as pd
from bs4 import BeautifulSoup
```

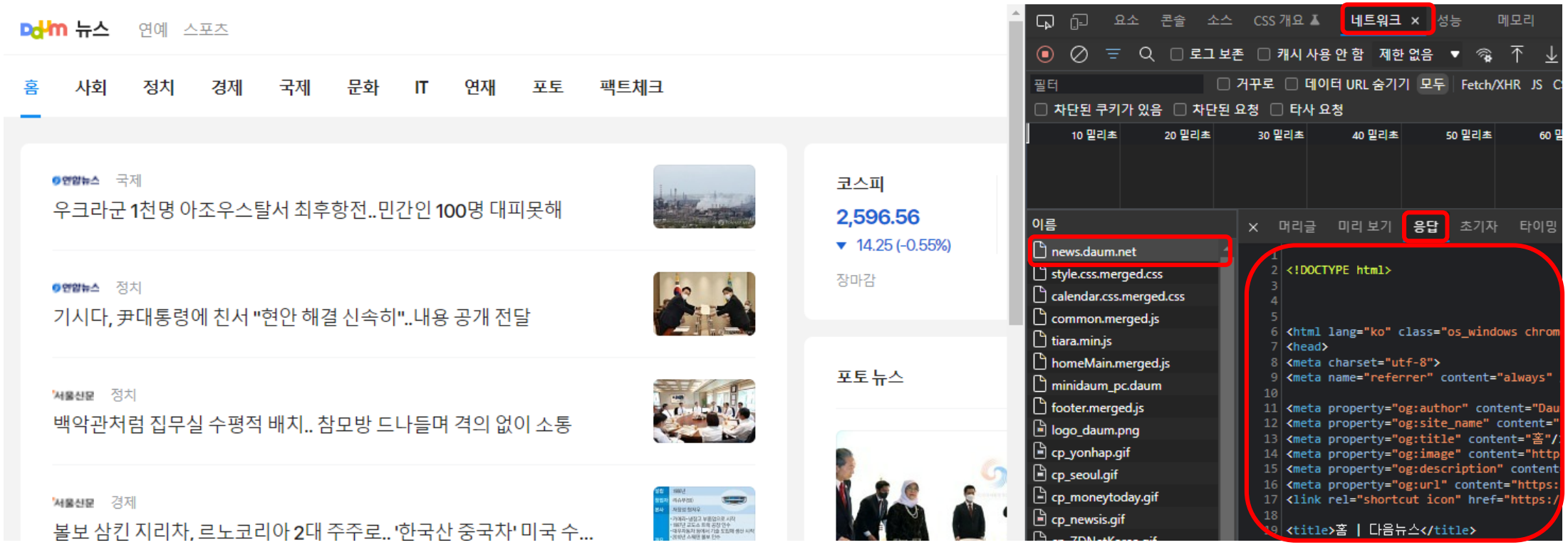
2. requests library를 통해 http 요청을 보내면 요청 값의 소스에 해당하는 응답(response)을 받게 됨

```
res = requests.get("https://news.daum.net/")
```

Introduction to Web Crawling Excercise

HTML PARSER – Daum News

Tip) 웹브라우저의 개발자 도구 –> 네트워크 –> 응답 부분에서 해당하는 주소의 응답값을 확인할 수 있음



Introduction to Web Crawling Excercise

HTML PARSER – Daum News

3. res.content를 통해 불러온 응답값을 확인할 수 있음

```
res.content
```

```
b'\n<!DOCTYPE html>\n\n\n\n<html lang="ko" class="os_unknown none unknown version_0 ">\n<head>\n<meta charset="utf
\xeb\x89\xb4\xec\x8a\xa4" />\n<meta property="og:site_name" content="\xeb\x8b\xa4\xec\x9d\x8c\xeb\x89\xb4\xec\x8a\
ent="https://t1.daumcdn.net/media/img-media/mobile/meta/news.png" />\n<meta property="og:description" content="\xe
s://news.daum.net/" />\n<link rel="shortcut icon" href="https://m2.daumcdn.net/img-media/2010ci/Daum_favicon.ico">
http-equiv="X-UA-Compatible" content="IE=edge">\n\n<link rel="stylesheet" type="text/css" href="//t1.daumcdn.net/m
ref="//t1.daumcdn.net/media/kraken/news/66868dd/calendar.css.merged.css" />\n\n<!--[if lte IE 8]>\n<script src="ht
<![endif]-->\n<!--[if lt IE 9]>\n<link rel="stylesheet" type="text/css" href="//t1.daumcdn.net/media/kraken/news/6
ews/66868dd/common.merged.js"></script>\n\n<script type="text/javascript" src="https://t1.daumcdn.net/tiara/js/v1/
rue\' || false,\n      isSandbox: \'\' || false,\n      isCbt: \'\' || false,\n    }\n  }\n</script>\n        <scr
\n<body>\n<div class="direct-link">\n    <!-- \xec\x9b\xb9\xec\xa0\x91\xea\xb7\xbc\xec\x84\xb1\xec\x9a\xa9 \xeb\xb
-->\n    <a href="#mainContent">\xeb\xb3\xb8\xeb\xac\xb8 \xeb\xb0\x94\xeb\xa1\x9c\xea\xb0\x80\xea\xb8\xb0</a>\n
8\xb0</a>\n</div>\n\n<div class="container-doc">\n<header class="doc-header">\n    <!-- head_top -->\n    <div cla
class="link_daum" data-tiara-layer="gnb default logo">\n                <img src="//t1.daumcdn.net/media/common/ne
</a>\n            <a href="https://news.daum.net/" id="kakaoServiceLogo" data-tiara-layer="GNB service news">\n
</h1>\n        <strong class="screen_out">\xea\xb4\x80\xeb\xa0\xa8 \xec\x84\x9c\xeb\xb9\x84\xec\x8a\xa4</strong>\n
tps://entertain.daum.net" class="link_services" data-tiara-layer="enter">\xec\x97\xb0\xec\x98\x88</a></li>\n
s">\xec\x8a\xa4\xed\x8f\xac\xec\xb8\xa0</a></li>\n        </ul>\n\n<div id="wrapMinidaum"></div>\n<script type="te
c\xa7\x80\xec\xa0\x95 \xea\xb4\x80\xeb\xa0\xa8 \xec\x98\xb5\xec\x85\x98. default: white.\n\t\tbgType: "white", //
a\xb2\xbd \xec\x97\xac\xeb\xb6\x80 flag : default: false - \xeb\x8b\xa4\xec\x9d\x8c\xed\x83\x91\xec\x9c\xbc\xeb\xa
9d\x8c\xed\x83\x91\xec\x9c\xbc\xeb\xa1\x9c \xec\x9d\xb4\xeb\x8f\x99, true - \xed\x98\x84\xec\x9e\xac\xed\x8e\x98\x
```

Introduction to Web Crawling Excercise

HTML PARSER – Daum News

4. BeautifulSoup library를 통해 받아온 응답값(res.content)과 html 구문으로 나누어주는 'html.parser'를 인자값으로 넘겨주면 객체에 응답값이 html 구조로 정리됨

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

```
soup
```

```
<!DOCTYPE html>

<html class="os_unknown none unknown version_0 " lang="ko">
<head>
<meta charset="utf-8"/>
<meta content="always" name="referrer">
<meta content="Daum 뉴스" property="og:author"/>
<meta content="다음뉴스" property="og:site_name"/>
<meta content="홈" property="og:title"/>
<meta content="https://t1.daumcdn.net/media/img-media/mobile/meta/news.png" property="og:image"/>
<meta content="다음뉴스" property="og:description"/>
<meta content="https://news.daum.net/" property="og:url"/>
<link href="https://m2.daumcdn.net/img-media/2010ci/Daum_favicon.ico" rel="shortcut icon"/>
```

Introduction to Web Crawling Excercise

HTML PARSER – Daum News

4. BeautifulSoup library를 통해 받아온 응답값(res.content)과 html 구문으로 나누어주는 'html.parser'를 인자값으로 넘겨주면 객체에 응답값이 html 구조로 정리됨

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

```
soup
```

```
<!DOCTYPE html>

<html class="os_unknown none unknown version_0 " lang="ko">
<head>
<meta charset="utf-8"/>
<meta content="always" name="referrer">
<meta content="Daum 뉴스" property="og:author"/>
<meta content="다음뉴스" property="og:site_name"/>
<meta content="홈" property="og:title"/>
<meta content="https://t1.daumcdn.net/media/img-media/mobile/meta/news.png" property="og:image"/>
<meta content="다음뉴스" property="og:description"/>
<meta content="https://news.daum.net/" property="og:url"/>
<link href="https://m2.daumcdn.net/img-media/2010ci/Daum_favicon.ico" rel="shortcut icon"/>
```

Introduction to Web Crawling Excercise

HTML PARSER – Daum News

5. html 정보가 저장된 soup 객체에서 기사 제목에 해당하는 부분의 요소만 선택
   - 기사 제목은 "html > body > div > main > section > ... > div > strong > a" 부분에 포함되어 있음
   - soup.select('div > strong > a')를 통해 'div > strong > a'에 해당하는 요소들만 불러올 수 있음

```
items = soup.select('div > strong > a')
```

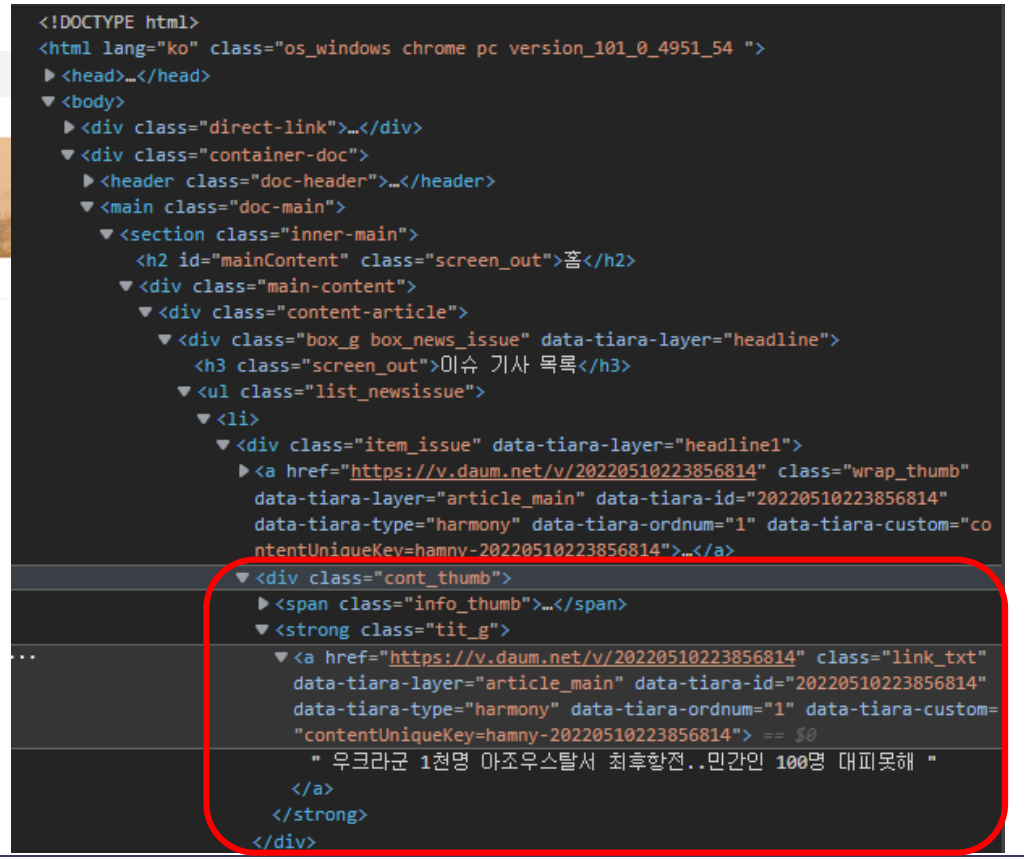
```
items
```

```
[<a class="link_txt" data-tiara-custom="contentUniqueKey=hamny-20220510223454789" data-tiara-id="20220510223454789" data-tiara
ps://v.daum.net/v/20220510223454789">
                                    기시다, 尹대통령에 친서 "현안 해결 신속히"..내용 공개 전달
                                </a>,
 <a class="link_txt" data-tiara-custom="contentUniqueKey=hamny-20220510223111767" data-tiara-id="20220510223111767" data-tiara
ps://v.daum.net/v/20220510223111767">
                                    백악관처럼 집무실 수평적 배치.. 참모방 드나들며 격의 없이 소통
                                </a>,
 <a class="link_txt" data-tiara-custom="contentUniqueKey=hamny-20220510223108765" data-tiara-id="20220510223108765" data-tiara
ps://v.daum.net/v/20220510223108765">
                                    볼보 삼킨 지리차, 르노코리아 2대 주주로.. '한국산 중국차'  미국 수출 야심
```

Introduction to Web Crawling Excercise

HTML PARSER – Daum News

Tip) 원하는 기사 제목은 “div” block 안 “strong” 태그 안 “a” 태그로 이루어져 있음



Introduction to Web Crawling Excercise

HTML PARSER – Daum News

Tip) 리스트의 첫번째 항목에서 text 부분만 추출하고 .strip() 함수를 통해 앞,뒤 공백을 제거하는 방법

```
items[0].text.strip()
```

```
'기시다, 尹대통령에 친서 "현안 해결 신속히"..내용 공개 전달'
```

Tip) 리스트의 첫번째 항목에서 'href' 속성만 불러오는 방법

```
items[0].attrs['href']
```

```
'https://v.daum.net/v/20220510223454789'
```

Introduction to Web Crawling Excercise

HTML PARSER – Daum News

6. text 추출, href 추출을 모든 item에 대하여 적용하여 DataFrame으로 변환하고 excel 파일로 저장함

```
title=[]
url = []

for i in range(len(items)):

    title.append(items[i].text.strip())
    url.append(items[i].attrs['href'])

list_sum = list(zip(title,url))

col = ['title','url']

df = pd.DataFrame(list_sum, columns=col)
df.to_csv('daum_new.xlsx', index=False)
```

Introduction to Web Crawling Excercise

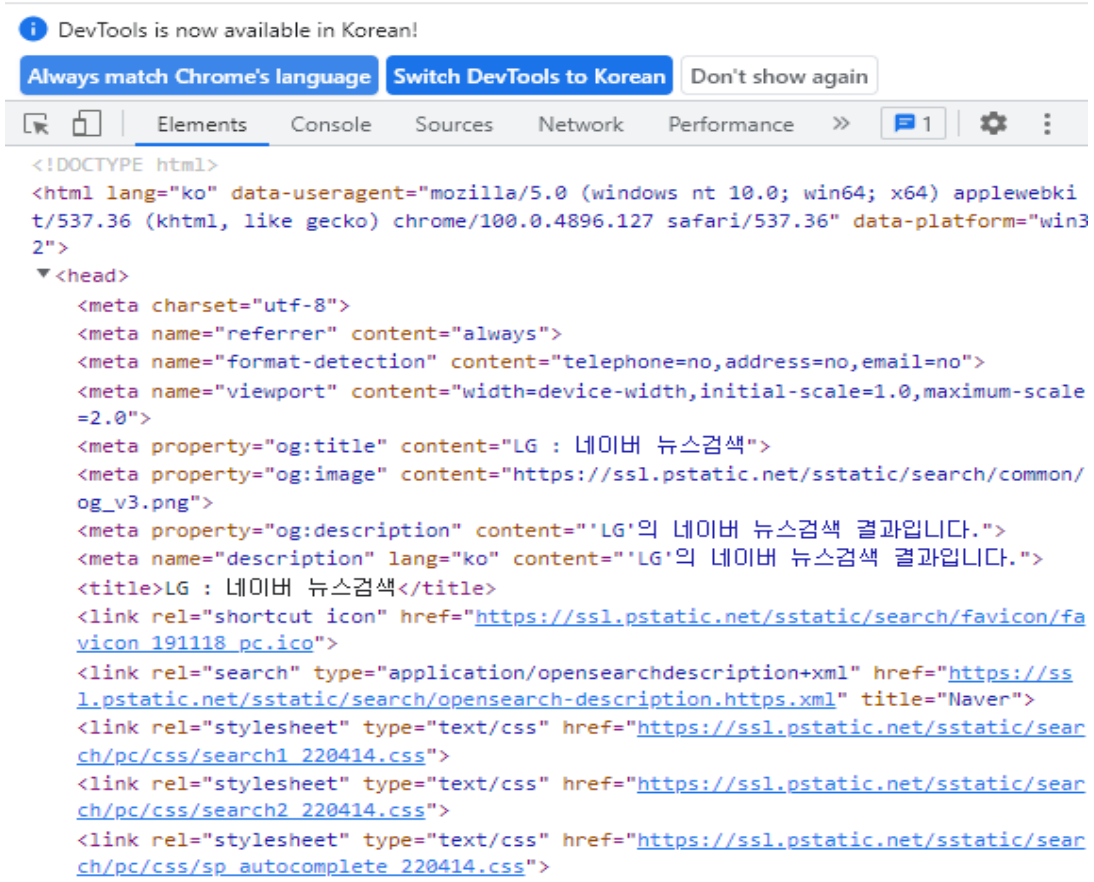
HTML PARSER – Daum News

- 출력 결과

A	B	C	D	E	F
	title	url			
0	'이재명 대항마' 국민의힘 인천시 계양을 후보 누구?..지역정가 '촉각'	https://v.daum.net/v/20220509155232599			
1	英국방장관 "푸틴, 나치급 파시즘 보여줘"..러 전승절 맞아 일침	https://v.daum.net/v/20220509155204585			
2	美 중간선거도 '바이든 vs 트럼프' 구도 될까?.. 바이든, 굳히기 나서	https://v.daum.net/v/20220509155104545			
3	文, 마지막 날까지 '경세제민'..이호승 등 참모진과 열띤 토론	https://v.daum.net/v/20220509154947477			
4	이임 임혜숙 장관 "과기부 역할 더 커질 것..주도적 조정자돼야"	https://v.daum.net/v/20220509154940472			
5	"러 패배도, 우크라 승리도 기준 모호해진 전쟁" WP	https://v.daum.net/v/20220509154855437			
6	민주당, 정호영·원희룡 고발조치..尹 임명강행 수순에 강력 반발	https://v.daum.net/v/20220509154836420			
7	한동훈 "수사종결권 99% 경찰에 주면 기소권까지 주는 셈"	https://v.daum.net/v/20220509154749370			
8	환율 장중 연고점 경신..1.3원 오른 1,274원 마감	https://v.daum.net/v/20220509154741365			
9	김성호 행안부 안전차관..신망 두터운 엘리트 관료[프로필]	https://v.daum.net/v/20220509154732362			
10	아증시 中 제외 일제 급락, 닛케이 2.53%↓ '낙폭최대'	https://v.daum.net/v/20220509154706339			
11	문 대통령 귀향 "지지", "반대" 목소리 고조.."5000명 운집" 文 평산행 D-1	https://v.daum.net/v/20220509154651324			
12	윤석열 취임 D-1, 민주당 용산 집무실 공세..'독단' '불통' 이미지 부각	https://v.daum.net/v/20220509154545271			

Introduction to Web Crawling Excercise

HTML PARSER – Naver News



Introduction to Web Crawling Excercise

HTML PARSER – Naver News

1. 필요한 라이브러리 정의 및 드라이버 세팅

```
from bs4 import BeautifulSoup
import requests
import re
import time
import pandas as pd
from selenium import webdriver
from selenium.webdriver.common.by import By
```

```
options = webdriver.EdgeOptions() # chrome 사용 시 webdriver.ChromeOptions()
options.add_experimental_option("excludeSwitches", ["enable-automation"])
options.add_experimental_option("useAutomationExtension", False)
```

Introduction to Web Crawling Excercise

HTML PARSER – Naver News

2. 입력된 수를 페이지 입력 형식(1, 11, 21, 31 ...)으로 변환해 주는 함수 정의

```
def makePgNum(num):
    if num == 1:
        return num
    elif num == 0:
        return num + 1
    else:
        return num + 9 * (num - 1)
```

Introduction to Web Crawling Excercise

HTML PARSER – Naver News

3. 크롤링할 페이지 url 생성하는 함수 정의(검색어, 크롤링 시작 페이지, 크롤링 종료 페이지)

```
def makeUrl(search, start_pg, end_pg):
    if start_pg == end_pg:
        start_page = makePgNum(start_pg)
        url = "https://search.naver.com/search.naver?where=news&sm=tab_pge&query=" + search + "&start=" + str(
            start_page)
        print("생성url: ", url)
        return url
    else:
        urls = []
        for i in range(start_pg, end_pg + 1):
            page = makePgNum(i)
            url = "https://search.naver.com/search.naver?where=news&sm=tab_pge&query=" + search + "&start=" + str(page)
            urls.append(url)
        print("생성url: ", urls)
        return urls
```

Introduction to Web Crawling Excercise

HTML PARSER – Naver News

4. 검색 키워드 입력 후 시작 페이지와 종료 페이지를 입력

```
search = input("검색할 키워드를 입력해주세요:")
```

```
검색할 키워드를 입력해주세요: LG전자
```

```
page = int(input("\n크롤링할 시작 페이지를 입력해주세요. ex)1(숫자만입력):"))  # ex)1 =1페이지,2=2페이지...
print("\n크롤링할 시작 페이지: ", page, "페이지")
```

```
크롤링할 시작 페이지를 입력해주세요. ex)1(숫자만입력): 1

크롤링할 시작 페이지:  1 페이지
```

```
page2 = int(input("\n크롤링할 종료 페이지를 입력해주세요. ex)1(숫자만입력):"))  # ex)1 =1페이지,2=2페이지...
print("\n크롤링할 종료 페이지: ", page2, "페이지")
```

```
크롤링할 종료 페이지를 입력해주세요. ex)1(숫자만입력): 2

크롤링할 종료 페이지:  2 페이지
```

Introduction to Web Crawling Excercise

HTML PARSER – Naver News

5. 크롤링 할 url 생성

```
search_urls = makeUrl(search, page, page2)
```

```
생성url: ['https://search.naver.com/search.naver?where=news&sm=tab_pge&query=LG전자&start=1', 'https://search.naver.com/search.naver?where=news&sm=tab_pge&query=LG전자&start=11']
```

ex) 크롤링 시작 1 페이지, 종료 2 페이지로 지정 시 브라우저가 1 페이지 크롤링 이후 2 페이지로 변경하여 크롤링

– 각 페이지 당 총 10개의 뉴스 기사가 존재

– 따라서, '~~start=1', '~~start=11' 두 개의 url이 생성된 것을 확인 가능

Introduction to Web Crawling Excercise

HTML PARSER – Naver News

6. Selenium 으로 URL 가져오기

```
driver = webdriver.Edge('msedgedriver.exe') # 크롬 사용 시, webdriver.Chrome('chromedriver.exe')
driver = webdriver.Edge(executable_path='msedgedriver.exe', options=options) # 크롬 사용 시, webdriver.Chrome(executable_path='chromedriver.exe', options=options)
driver.implicitly_wait(3)

naver_urls = []
count = 0

for i in search_urls:
    driver.get(i)
    time.sleep(1)

    a = driver.find_elements(By.CSS_SELECTOR, 'a.info')

    for i in a:
        count += 1
        i.click()

        # 현재탭에 접근
        driver.switch_to.window(driver.window_handles[count - 1])
        time.sleep(3)

        url = driver.current_url
        print(url)

        if "news.naver.com" in url:
            naver_urls.append(url)

        else:
            pass
        driver.switch_to.window(driver.window_handles[0])
```

Introduction to Web Crawling Excercise

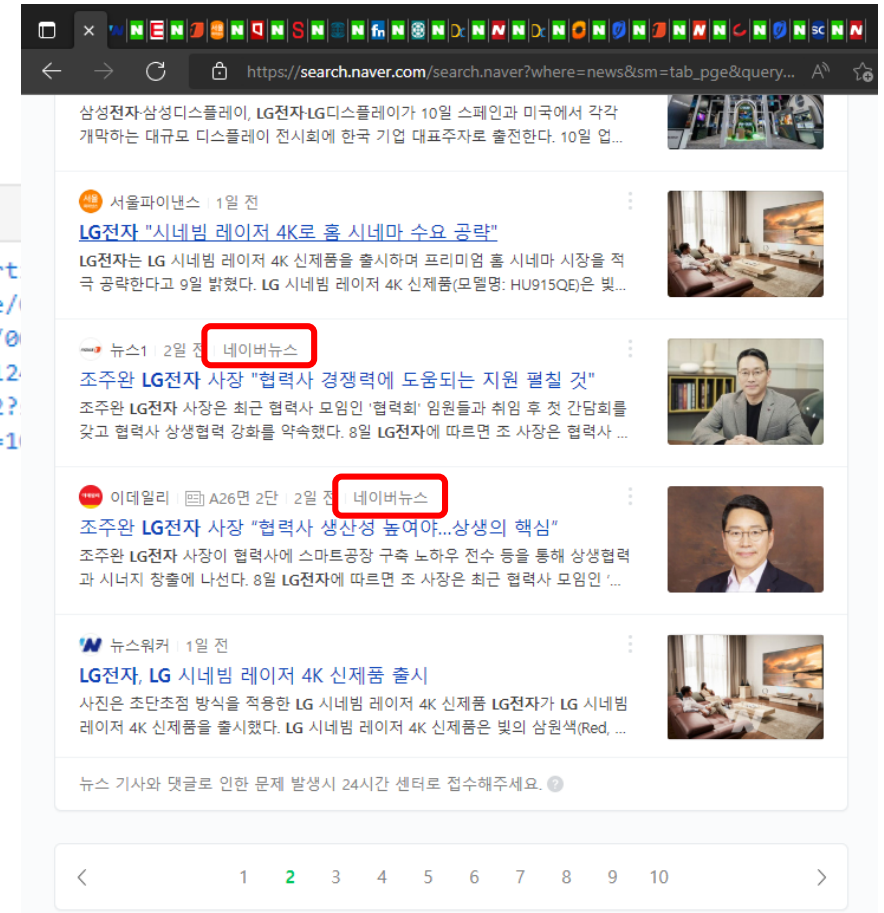
HTML PARSER – Naver News

ex) 개별 신문사만 제공되는 기사를 제외하고 네이버에서 같이 제공되는

총 17개의 뉴스 기사 url이 수집된 것을 확인 가능

```
print(naver_urls)
```

```
['https://n.news.naver.com/mnews/article/003/0011173495?sid=101', 'https://n.news.naver.com/mnews/art
ps://n.news.naver.com/mnews/article/366/0000812297?sid=105', 'https://n.news.naver.com/mnews/article/
n.news.naver.com/mnews/article/001/0013166930?sid=101', 'https://n.news.naver.com/mnews/article/008/0
s.naver.com/mnews/article/003/0011174092?sid=101', 'https://n.news.naver.com/mnews/article/138/000212
er.com/mnews/article/014/0004832474?sid=101', 'https://n.news.naver.com/mnews/article/022/0003693682?
m/mnews/article/277/0005086590?sid=101', 'https://n.news.naver.com/mnews/article/421/0006077135?sid=1
```



Introduction to Web Crawling Excercise

HTML PARSER – Naver News

7. requests와 BeautifulSoup으로 크롤링한 url마다 접근하여 기사 타이틀 및 본문 가져오기

```
headers = {"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) Chrome/98.0.4758.102"}

titles = []
contents = []
for i in naver_urls:
    original_html = requests.get(i, headers=headers)
    html = BeautifulSoup(original_html.text, "html.parser")

    title = html.select("div#ct > div.media_end_head.go_trans > div.media_end_head_title > h2")
    title = ''.join(str(title))
    pattern1 = '<[^>]*>'
    title = re.sub(pattern=pattern1, repl='', string=title)
    titles.append(title)

    content = html.select("div#dic_area")
    content = ''.join(str(content))
    content = re.sub(pattern=pattern1, repl='', string=content)
    pattern2 = """[\n\n\n\n\n// flash 오류를 우회하기 위한 함수 추가\nfunction _flash_removeCallback() {}"""
    content = content.replace(pattern2, '')
    contents.append(content)
```

Introduction to Web Crawling Excercise

HTML PARSER – Naver News

8. 크롤링한 뉴스 제목, 본문을 DataFrame으로 변환 후 CSV로 저장

```
news_df = pd.DataFrame({'title': titles, 'link': naver_urls, 'content': contents})
news_df
news_df.to_csv('NaverNews_%s.csv' % search, index=False, encoding='utf-8-sig')
```

Introduction to Web Crawling Excercise

HTML PARSER – Naver News

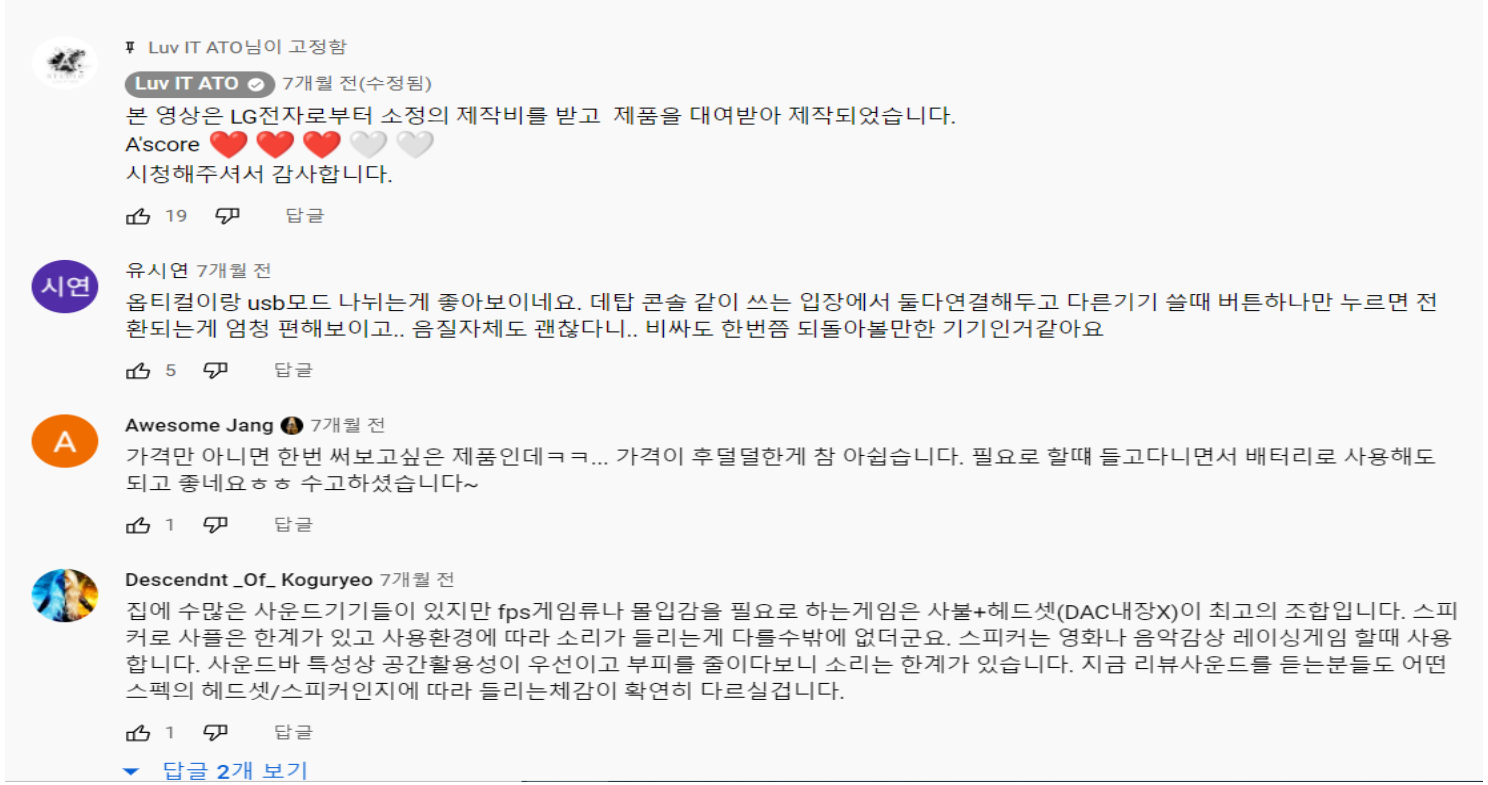
- 출력 결과

[LG전자, 차량용 반도체 내재화 추진...개발 기술 인증받아]	https://n.news.naver.com/mnews/article/001/0013158930?sid=105	[MCU·PMIC 등 차량용 반도체 개발 프로세스 구축..."사업화 검토 단계" LG전자, 차량용 반도체 내재화 추진[LG전자 제공. DB 및 재판매 금지](서울=연합뉴스) 김철선 기자 = LG전자가 차량용 반도체 설계와 구 발 프로세스 전반에 대한 기술을 확보했다. 글로벌 차량용 반도체 부족 사태 장기화에 대응해 자체적인 개발 역량을 확보한 LG전자는 검토해 차량용 반도체 사업 내재화를 추진한다는 방침이다. LG전자는 최근 독일 시험·인증 전문기관 'TUV 라인란드'(TUV Rheinland)로 도체 개발 프로세스에 대한 'ISO 26262' 인증을 받았다고 5일 밝혔다. 이 인증은 차량에 탑재되는 전기·전자 장치의 시스템 오류로 인한 하기 위해 제정된 국제표준화기구(ISO) 자동차 기능 안전 국제표준규격이다. LG전자는 전자제어장치(ECU)와 마이크로컨트롤러유닛(MC 반도체(PMIC) 등 차량용 반도체 개발 프로세스를 구축했다. 특히 이번 인증에서 자동차 기능 안전성 가운데 최고 수준인 '자동차안전무 '(ASIL) D등급의 부품 개발 능력을 인정받았다고 LG전자는 소개했다. ASIL은 사고의 심각도와 발생빈도, 제어 가능성 등에 따라 최저 A D등급까지 4단계로 분류되는데 D등급은 1억 시간 동안 연속 사용했을 때 사고로 이어질 수 있는 고장을 1회 이하로 관리하는 가장 엄격 LG전자의 차량용 반도체 사업은 선행기술 개발을 담당하는 CTO(최고기술책임자) 부문 내 SIC센터 주도로 추진되고 있다. 미래 성장동 레인, 인포테인먼트 등 자동차 부품 사업을 키우고 있는 LG전자는 차량용 반도체 개발 역량까지 확보해 전장 사업 내 시너지 효과를 극대 이다. 회사 관계자는 "차량용 반도체 사업 내재화를 검토 중"이라며 "실제 내재화가 가능할지 선행적으로 기술 개발 역량을 확보하고 사 는 단계"라고 설명했다. 앞서 LG전자는 TUV 라인란드로부터 ADAS(주행보조시스템) 카메라, 인포테인먼트 시스템, 차량용 디스플레이 부품'과 '차량 미디어 부품'의 개발 프로세스에 대해서도 인증받은 바 있다. TUV 라인란드 코리아의 프랭크 주트너(Frank Juettner) 대표 차량용 반도체 개발 프로세스를 성공적으로 구축하고 기능 안전성까지 확보해 앞으로 미래 자동차 산업에서 핵심적인 역할을 하게 될 것 고 말했다. LG전자 SIC센터장 김진경 상무는 "자동차의 핵심 부품인 차량용 반도체의 기능 안전성을 세계 최고 수준으로 개발할 수 있 을 지속해서 확보해 나갈 것"이라고 강조했다. kcs@yna.co.kr]
[스마트폰 철수한 LG전자, 5G 특화망 장비 수주 삼성 제쳤다]	https://n.news.naver.com/mnews/article/366/0000811592?sid=105	[에릭슨엘지, SK네트웍스서비스 5G 장비 공급삼성, 이음5G 1호 기업 네이버에 장비 공급LG CNS, 에릭슨엘지 장비 활용에 무게 에릭슨엘지 CI. /에릭슨엘지 지난해 스마트폰 사업을 철수한 LG전자가 5세대 이동통신(5G) 장비시장에서 삼성전자와 맞대결을 벌이 화망(이음5G) 사업 진출을 선언한 '1호 기업' 네이버클라우드를 삼성전자가 잡은 가운데, LG전자는 SK네트웍스서비스를 공략했다. LG그룹 CNS 역시 에릭슨엘지의 장비 활용 가능성이 큰 만큼 초기 시장에서 LG 측이 삼성을 앞서고 있다는 평가가 나온다.6일 통신장비업계에 따 지는 SK네트웍스 자회사인 SK네트웍스서비스에 이음5G 통신장비를 공급한다. 이로써 LG전자는 삼성전자에 이어 이음5G 통신장비 시장에 진출하게 됐다.앞서 삼성전자는 이음5G 첫 사업자로 선정된 네이버클라우드에 통신장비를 공급하기로 했다. 이음 5G는 통신사 이동통신망 5G 융합서비스를 필요로 하는 기업·기간망 통신사업자가 4.7㎓(기가헤르츠)와 28㎓ 등 전용 주파수를 활용해 망 구축이 가능하도록 돕는 까지 국내서 공식적으로 이음5G 사업 진출 의사를 밝힌 기업은 네이버클라우드, LG CNS, SK네트웍스서비스 등 3곳이다. 네이버클라우드 서비스를 삼성전자와 LG전자가 각각 나눠 가졌고, LG CNS의 경우 LG그룹 계열사인 만큼 에릭슨엘지가 장비를 공급할 것이라는 관측에 두 있다.통신업계에 정통한 한 관계자는 "에릭슨엘지가 합작사라고는 하지만 과거 LG정보통신 출신을 비롯해 LG전자 출신 직원이 상당수 근

Introduction to Web Crawling Excercise

Selenium – Youtube Crawling

- 셀레니움을 통해서 유튜브 댓글을 크롤링이 가능

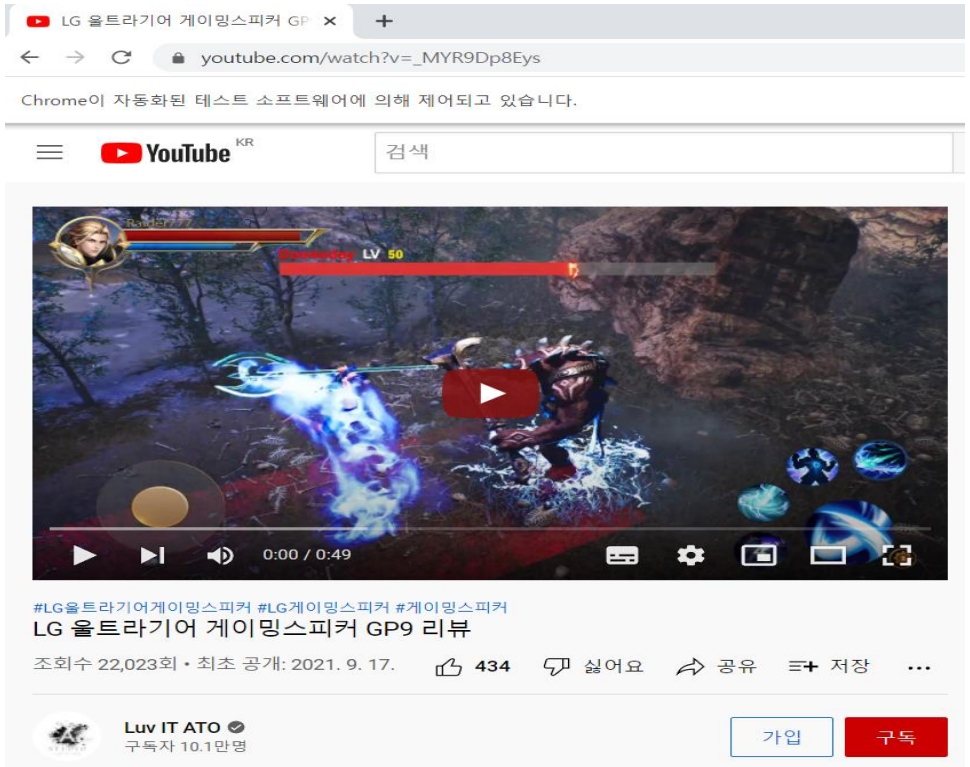


Introduction to Web Crawling Excercise

Selenium – Youtube Crawling

1. 필요한 라이브러리를 불러오고 webdriver 실행

```
from selenium import webdriver
import time
from openpyxl import Workbook
import pandas as pd
from selenium.webdriver.common.keys import Keys
from bs4 import BeautifulSoup

wb = Workbook(write_only=True)
ws = wb.create_sheet()

driver = webdriver.Edge('msedgedriver.exe') # webdriver.Chrome('chromedriver.exe')
driver.get("https://www.youtube.com/watch?v=_MYR9Dp8Eys")
driver.implicitly_wait(3)

time.sleep(1.5)

driver.execute_script("window.scrollTo(0, 800)")
time.sleep(3)
```

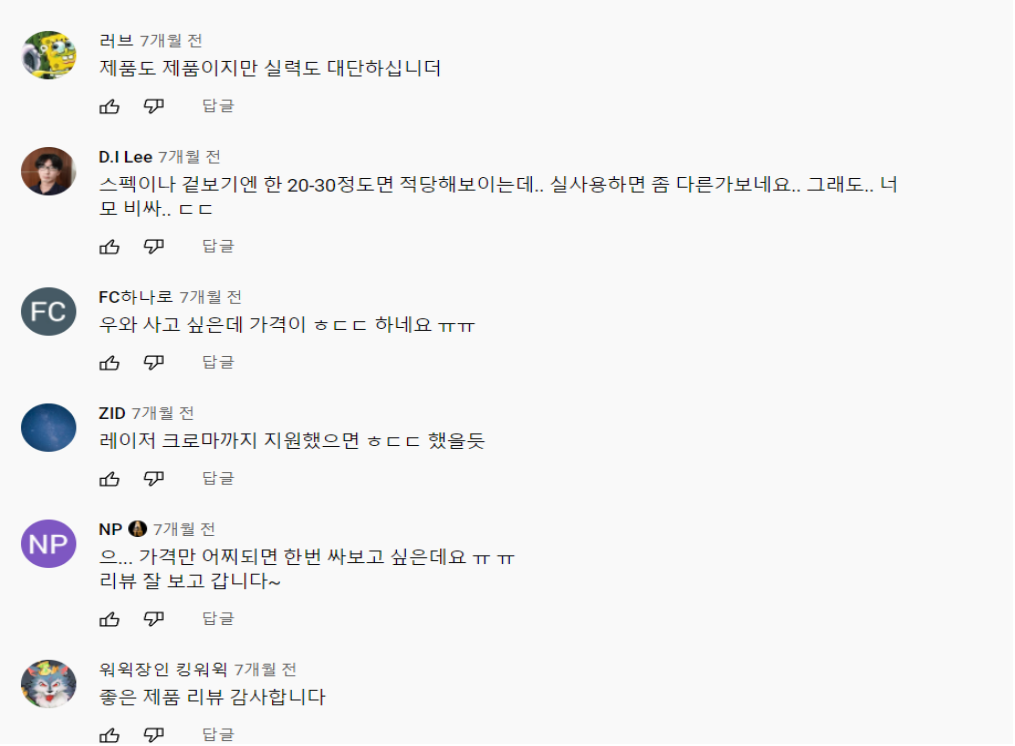


Introduction to Web Crawling Excercise

Selenium – Youtube Crawling

2. 페이지 끝까지 스크롤 및 팝업 닫기 설정 이후 브라우저 스크롤이 이동

```
last_height = driver.execute_script("return document.documentElement.scrollHeight")

while True:
    driver.execute_script("window.scrollTo(0, document.documentElement.scrollHeight);")
    time.sleep(5)

    new_height = driver.execute_script("return document.documentElement.scrollHeight")
    if new_height == last_height:
        break
    last_height = new_height

time.sleep(5)

try:
    driver.find_element_by_css_selector("#dismiss-button > a").click()
except:
    pass
```



Introduction to Web Crawling Excercise

Selenium – Youtube Crawling

3. 정보 추출 및 댓글 정보 Csv 파일로 저장

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

id_list = soup.select("div#header-author > h3 > #author-text > span")
comment_list = soup.select("yt-formatted-string#content-text")

id_final = []
comment_final = []

for i in range(len(comment_list)):
    temp_id = id_list[i].text
    temp_id = temp_id.replace('\n', '')
    temp_id = temp_id.replace('\t', '')
    temp_id = temp_id.replace('    ', '')
    id_final.append(temp_id)

    temp_comment = comment_list[i].text
    temp_comment = temp_comment.replace('\n', '')
    temp_comment = temp_comment.replace('\t', '')
    temp_comment = temp_comment.replace('    ', '')
    comment_final.append(temp_comment)

pd_data = {"아이디" : id_final , "댓글 내용" : comment_final}
youtube_pd = pd.DataFrame(pd_data)
# convert_csv_from_dataframe('title',youtube_pd)
# youtube_pd.to_excel('youtube.xlsx',encoding = 'cp949')
youtube_pd.to_csv(f"youtube_pd_title_5.csv", encoding='utf-8-sig')
driver.close()
```

```
youtube_pd
```

	아이디	댓글 내용
0	Luv IT ATO	본 영상은 LG전자로부터 소정의 제작비를 받고 제품을 대여받아 제작되었습니다. A...
1	유시연	옵티컬이랑 usb모드 나뉘는게 좋아보이네요. 데탑 콘솔 같이 쓰는 입장에서 둘다연결...
2	Awesome Jang	가격만 아니면 한번 써보고싶은 제품인데ㅋㅋ... 가격이 후덜덜한게 참 아쉽습니다. ...
3	Azbel Rex	제가 예전에 AAA급 게임들에 어울리는 스피커를 리뷰해달라고 했는데 해주셨군요!!!...
4	렘푸	내장배터리 모델이라면 usb-c충전으로 변경되었으면 편의성이 더 좋았을텐데
...	...	...
86	원주민	개비싸네
87	가재는게편	6:01 응?
88	m m	안삽니다.
89	김메타	지르...까..?
90	HeiZy	6:00 에 빨간옷 적....

91 rows × 2 columns

Introduction to Web Crawling Excercise

Selenium – Youtube Crawling

- 출력결과

B	C	D	E	F	G	H	I	J	K	L
아이디	댓글 내용									
Luv IT A	본 영상은 LG전자로부터 소정의 제작비를 받고 제품을 대여받아 제작되었습니다. A'score ㅁㅁ시청해주셔서 감사합									
유시연	옵티컬이랑 usb모드 나뉘는게 좋아보이네요. 데탑 콘솔 같이 쓰는 입장에서 둘다연결해두고 다른기기 쓸때 버튼하									
Awesom	가격만 아니면 한번 써보고싶은 제품인데ㅋㅋ... 가격이 후덜덜한게 참 아쉽습니다. 필요로 할때 들고다니면서 배터									
Azbel Re	제가 예전에 AAA급 게임들에 어울리는 스피커를 리뷰해달라고 했는데 해주셨군요!!!!! 사실 스피커를 서라운드급으									
Descend	집에 수많은 사운드기기들이 있지만 fps게임류나 몰입감을 필요로 하는게임은 사불+헤드셋(DAC내장X)이 최고의									
렘푸	내장배터리 모델이라면 usb-c충전으로 변경되었으면 편의성이 더 좋았을텐데									
Corgi Kin	언뜻 보면 스피커를 사용하는 게이밍 데스크탑 환경에 최적화된 올인원 솔루션이라고 볼 수 있지만항상 그렇듯 올									
한형종	커뮤니티에서 실컷 까이던 제품이네요.설명을 들어보니 만능템 입니다.dac와 스피커.마이크 일체형이고 사용하기									
Maior	저는 이 제품 구입한지 3~4일쯤 되었네요.게임 + 영화 + 음악 전부 만족하며 쓰고 있어요.보통 RPG 게임하면 음악									
DMIN	스피커로 마이크가 꽤 괜찮게 들리는게 인상적이네요 ㅎㅎ소리 자체도 나쁘지 않은 것 같은데....역시 가격이 ㅎㅎ									
deko	이정도 가격대라면 후면 전원이랑 유선연결 포트정도는 자석형으로 줘서 블투<->유선 변환 편의성을 극한으로 올									
Lo속기vE	항상 좋은 제품 그리고 한 주간의 뉴스을 잘 전달해주셔서 감사합니다..눈팅과 좋아요는 많이 누르고 가는대 댓글은									
Ghoul_W	Excellent review I would love to pair this speakers with my LG 27GP950-B									
vo nikae	이 영상보고 좋아요 클릭 안 할 수 가없는 완벽한 리뷰 감사합니다 내가 써본거 같은 경험을 했네요!!									
유중훈	헤드셋은 더울 때 가끔 귀가 답답하고 불편했는데 이 스피커 음성채팅 기능만 본다면 게임할 때 나쁘지 않은 선택									
러브	제품도 제품이지만 실력도 대단하십니더									
D.I Lee	스펙이나 겉보기엔 한 20-30정도면 적당해보이는데.. 실사용하면 좀 다른가보네요.. 그래도.. 너모 비싸.. ㄷㄷ									

Introduction to Web Crawling Excercise

Selenium – Naver Blog Crawling

- 셀레늄 을 통해서 네이버 블로그를 키워드 기반으로 크롤링 할 수가 있다.

Introduction to Web Crawling Excercise

Selenium – Naver Blog Crawling

1. 필요한 라이브러리 정의

```
import time
import selenium
import pandas as pd
from selenium import webdriver
```

2. 웹 드라이버와 포스팅 정보를 가져옴

```
def get_post_info(keyword: str, start_date: str, end_date: str, pages: int) -> pd.DataFrame:
    total_header = []
    total_contents = []
    total_date = []
    # chromedriver path
    driver = webdriver.Chrome('chromedriver.exe')

    for j in range(1, pages + 1):
        # first page
        first_url = f"https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=PERIOD&orderBy=sim&startDate=" \
                    f"{start_date}&endDate={end_date}&keyword={keyword}"
        # After first page
        loop_url = f"https://section.blog.naver.com/Search/" \
                   f"Post.nhn?pageNo={j}&rangeType=PERIOD&orderBy=sim&startDate=" \
                   f"{start_date}&endDate={end_date}&keyword={keyword}"
        driver.get(first_url)

        if j != 1:
            driver.get(loop_url)
```

Introduction to Web Crawling Excercise

Selenium – Naver Blog Crawling

3. 페이지별 가져올 정보 정의

```
        for i in range(1, 8):
            time.sleep(1)
            try:
                head = driver.find_element_by_css_selector(
                    f"#content > section > div.area_list_search > div:nth-child({i}) > div > div.info_post > div.desc > a.desc_inner > strong > span").text
            except selenium.common.exceptions.NoSuchElementException:
                head = " "
            try:
                contents = driver.find_element_by_css_selector(
                    f"#content > section > div.area_list_search > div:nth-child({i}) > div > div.info_post > div.desc > a.text").text
            except selenium.common.exceptions.NoSuchElementException:
                contents = " "
            try:
                date = driver.find_element_by_css_selector(
                    f"#content > section > div.area_list_search > div:nth-child({i}) > div > div.info_post > div.writer_info > span.date").text
            except selenium.common.exceptions.NoSuchElementException:
                date = " "

            time.sleep(1)
            print(head)
            print(contents)
            print(date)
            total_header.append(head)
            total_contents.append(contents)
            total_date.append(date)

    return total_header, total_contents, total_date
```

Introduction to Web Crawling Excercise

Selenium – Naver Blog Crawling

4. 검색 형태 지정 및 csv 파일로 저장

```
pages = int(input("몇 페이지 까지 크롤링 할까요? \n => ")) # pages = 12
keyword = input("검색하실 키워드를 입력해주세요 \n =>") # keyword = "LG 전자"
start_date = input("검색 시작일을 입력해주세요 ex) 2000-01-01 \n =>") # start_date = '2019-11-12'
end_date = input("검색 종료일을 입력해주세요 ex) 2000-01-01 \n =>") # end_date = '2019-12-12'

total_header, total_contents, total_date = get_post_info(keyword, start_date, end_date, pages)
data = {"header": total_header, "contents": total_contents, "date": total_date}

df = pd.DataFrame(data, columns=["header", "contents", "date"])
keyword = keyword.replace(" ", "_")
title = keyword+ "_" + str(pages) + "_" + "네이버블로그_크롤링.csv"

df.to_csv('naver.blog.csv', encoding='utf-8-sig')
```

Introduction to Web Crawling Excercise

Selenium – Naver Blog Crawling

5. 검색어 지정 후 webdriver 실행 및 크롤링 진행

```
몇 페이지 까지 크롤링 할까요?
 => 12
검색하실 키워드를 입력해주세요
 => LG 전자
검색 시작일을 입력해주세요 ex) 2000-01-01
 => 2019-11-12
검색 종료일을 입력해주세요 ex) 2000-01-01
 => 2019-12-12
```

Introduction to Web Crawling Excercise

Selenium – Naver Blog Crawling

- 출력 결과

A	B	C	
	header	contents	date
0	LG전자, 2022년형 휘센 듀얼 인버터 제습기 출시	LG전자가 강력하고 빠른 제습 성능은 물론 제품 위생관리까지 대폭 강화한 2022년형 휘센 듀얼 인버터 제	5시간 전
1	LG전자 스타일러 S3WFS 화이트, 지금 구매하면 정말 좋	LG전자 스타일러 S3WFS 화이트, 지금 구매하면 정말 좋은 이유! 여러분! 오늘도 안녕하신가요~ 슈가씨입	13시간 전
2	LG전자 채용 시작! 채용 공고/인재상/자소서 항목/예시	바로 LG전자 채용공고입니다!! 2022 상반기 LG전자 채용이 시작되었는데요. 고객가치혁시부문의 채용연	2시간 전
3	의정부 혼수가전제품/입주이사가전 세일: LG전자 베스	의정부 혼수가전제품/입주이사가전 세일: LG전자 베스트샵 신세계 의정부점 LG전자 베스트샵 신세계백화	2시간 전
4	의정부 결혼/이사가전제품 세일: LG전자 베스트샵 신세	의정부 결혼/이사가전제품 세일: LG전자 베스트샵 신세계백화점 의정부점 - LG전자 베스트샵 신세계백화	2시간 전
5	[LG전자 베스트샵 서안산본점] NEW 프리미엄 랜드마크	가전제품은 오픈매장 혜택이 가장 많다는 건 이미 공공연하게 잘 알려진 사실인데요~ 이번 포스팅에서는	4시간 전
6	아현2구역 이사준비 (LG전자이전서비스)	이사비용에 대해서는 우리가 LG가전이 많아서 LG전자 이전설치서비스를 알아보았다. 1544-7777에 전화	4시간 전