

14기 정규세션

ToBig's 13기 조상연

Crawling

크롤링

Content Lists

Unit 01 | 크롤링을 위한 개념

Unit 01 | 크롤링 개념 (1) HTTP

Unit 01 | 크롤링 개념 (2) HTML

Unit 02 | 크롤링 활용 및 실습

Unit 03 | 과제

Unit 00 | 크롤링 전체 개념

기본 웹 통신 이해
(HTTP)

Postman 활용법

XmlHttpRequest 이해

GraphQL 심

웹 문서 구조 이해
(HTML)

BeautifulSoup 활용법

Session, Cookie 인증

스마트폰 앱 데

Python 라이브러리 활용
(Requests)

Selenium 활용법

Javascript 이해 및 활용

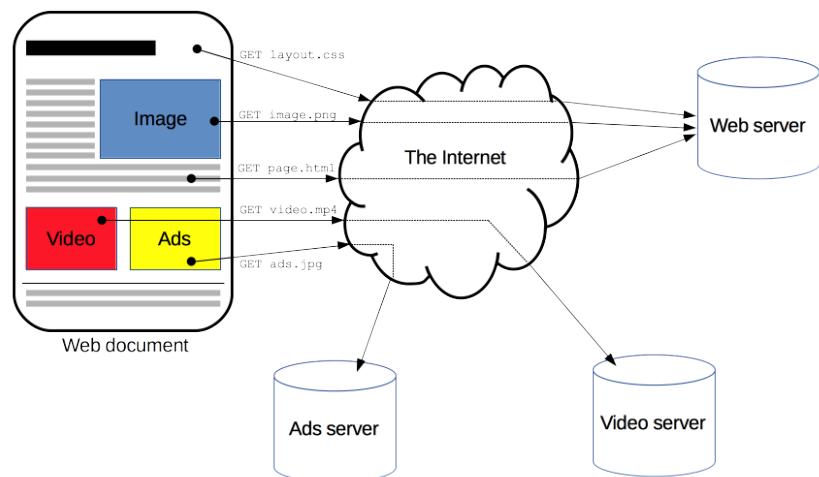
서버사이드

Unit 01 | 크롤링을 위한 기초 개념

1.1. 기본 웹 통신 이해 (HTTP)

HTTP는 HTML 문서와 같은 리소스들을 가져올 수 있도록 해주는 [프로토콜](#)입니다.
HTTP는 웹에서 이루어지는 모든 데이터 교환의 기초이며, 클라이언트-서버 프로토콜이기도 합니다.

- MDN



HTTP 개요도 - MDN

Unit 01 | 크롤링을 위한 기초 개념

1.1. 기본 웹 통신 이해 (HTTP)

1. Hyper Text

말 그대로 **초월한 (Hyper) 문서 (Text)**, **Hyper Link**를 통해 즉시 접근이 가능한 문서를 의미

2. Hyper Text Markup Language (HTML)

이러한 **Hyper Text**를 **구조적으로 적기 위한 언어**, **태그**를 활용

3. Hyper Text Transfer Protocol (HTTP)

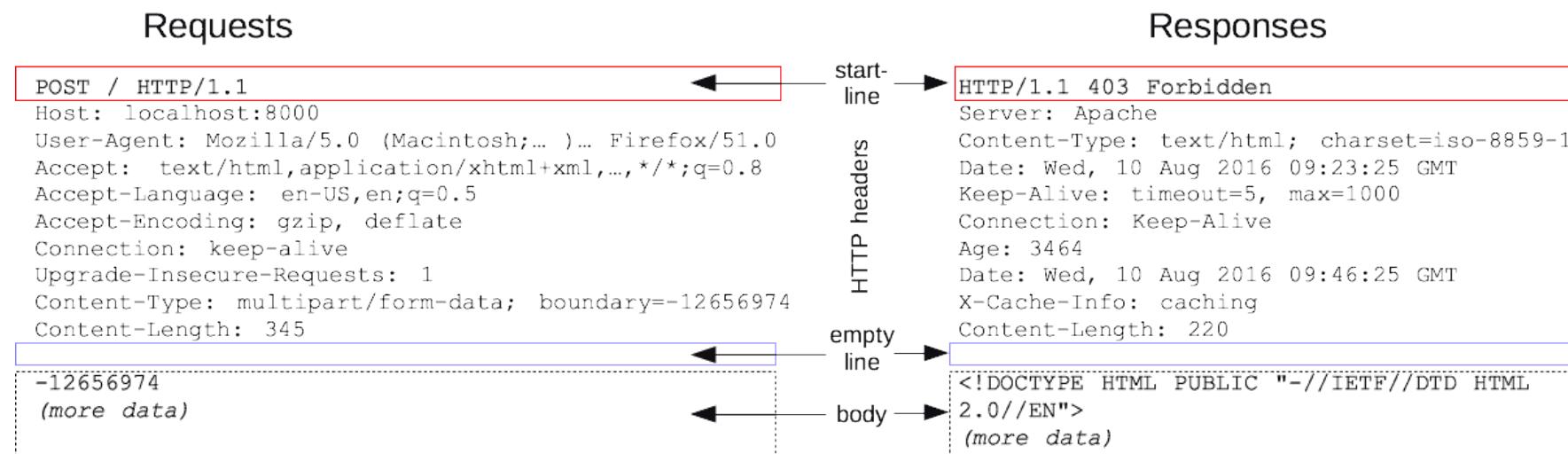
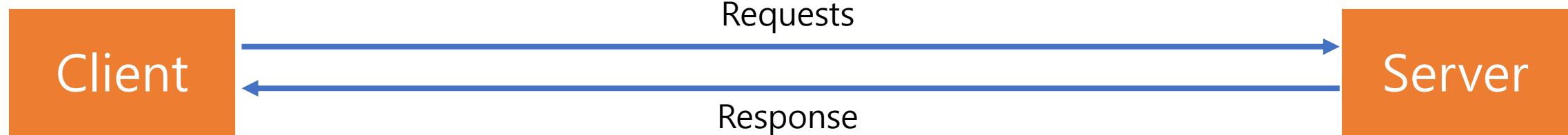
마찬가지로 **Hyper Text**를 **전송하기 위한 프로토콜**

4. Protocol

통신을 위한 약속, 통신의 유형, 내용 등을 미리 정의

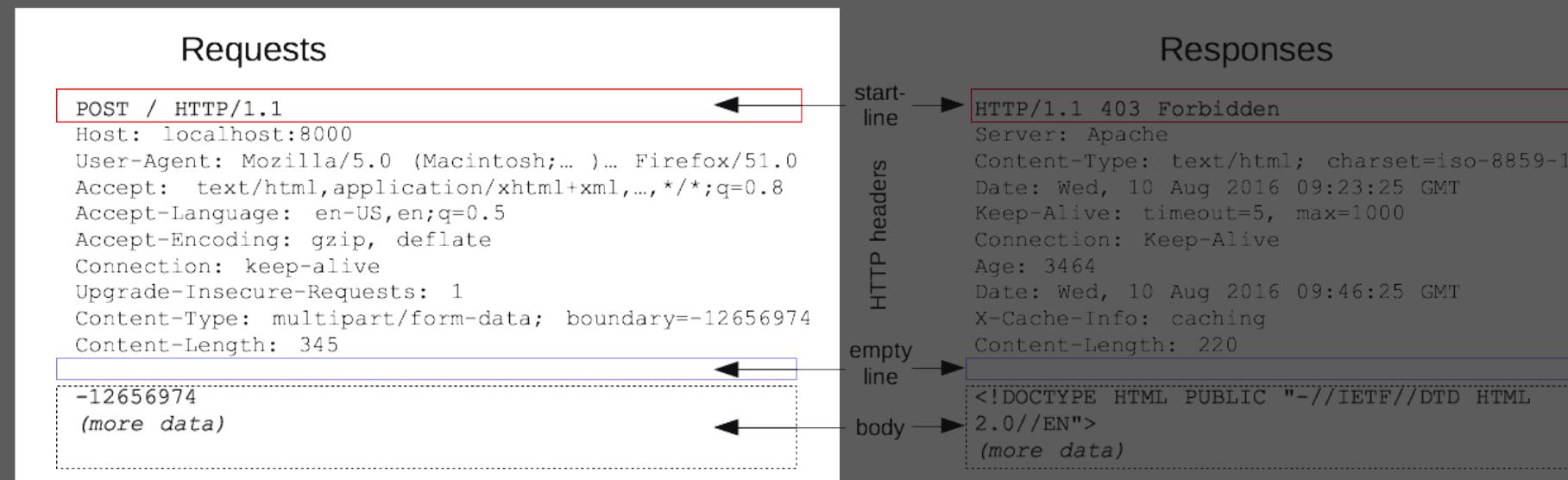
Unit 01 | 크롤링을 위한 기초 개념

1.1. 그 약속



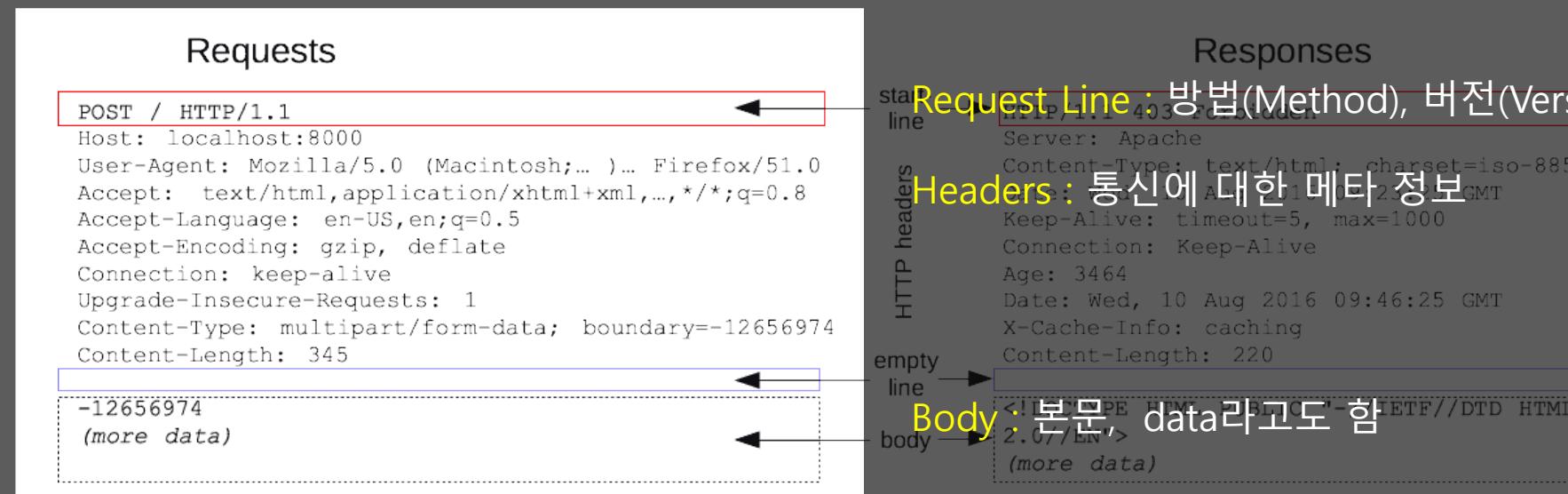
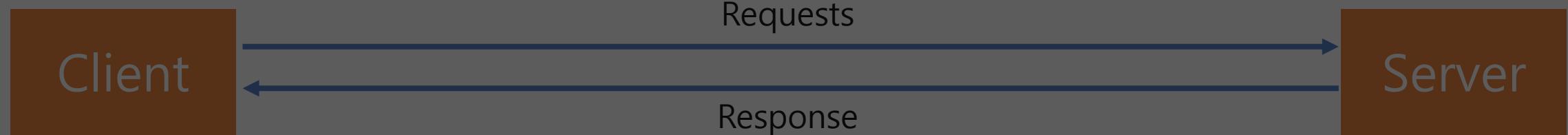
Unit 01 | 크롤링을 위한 기초 개념

1.1. 그 약속



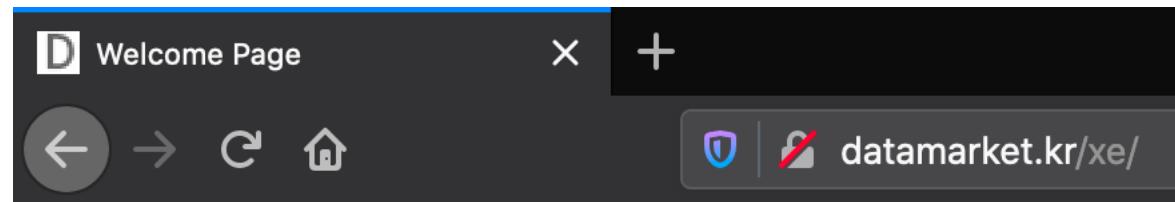
Unit 01 | 크롤링을 위한 기초 개념

1.1. 그 약속



Unit 01 | 크롤링을 위한 기초 개념

1.1. 기본 웹 통신 이해 (HTTP) : 데이터마켓 접속



200	GET	datamarket.kr	/xe/	browsing-c...	html	8.58 KB	50.21 KB
200	GET	pagead2.googlesyndic...	activevie...	osd_listene...	gif	1.06 KB	42 B
200	GET	pagead2.googlesyndic...	activevie...	osd_listene...	gif	1.06 KB	42 B
200	GET	datamarket.kr	autolink.js...	script	js	2.62 KB (raced)	2.37 KB
200	GET	datamarket.kr	jquery-ui....	script	js	202.07 KB (raced)	201.82 KB

Unit 01 | 크롤링을 위한 기초 개념

1.1. 기본 웹 통신 이해 (HTTP)

Requests

Headers Cookies Request Response Timings Stack Trace

Filter Headers

▶ GET http://datamarket.kr/xe/

Status	200 OK
Version	HTTP/1.1
Transferred	8.58 KB (50.21 KB size)

▶ Response Headers (389 B) Raw

▶ Request Headers (650 B) Raw

- Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
- Accept-Encoding: gzip, deflate
- Accept-Language: en-US,en;q=0.5
- Cache-Control: max-age=0
- Connection: keep-alive
- Cookie: mobile=false; user-agent=c6c7a9a1c2c713d151327e9b155be92d; PHPSESSID=s5ir376vtt342ed82es4rocdp5; wcs_bt=1275f81a1c37904:1597686136; _ga=GA1.2.267574067.1597686131; _gid=GA1.2.462394718.1597686131; __gads=ID=d0d113b761773ecb:T=1597686131:S=ALNI_MbF2PnsmuxNPFY1ecxhZQiz9ij38A
- Host: datamarket.kr
- Upgrade-Insecure-Requests: 1
- User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15; rv:79.0) Gecko/20100101 Firefox/79.0

Response

Raw

▼ Response Headers (389 B)

- Cache-Control: no-store, no-cache, must-revalidate
- Cache-Control: post-check=0, pre-check=0
- Connection: close
- Content-Encoding: gzip
- Content-Type: text/html; charset=UTF-8
- Date: Mon, 17 Aug 2020 17:54:13 GMT
- Expires: Mon, 26 Jul 1997 05:00:00 GMT
- Last-Modified: Mon, 17 Aug 2020 17:54:14 GMT

Headers Cookies Request **Response** Timings Stack Trace

▼ Preview

close_btn

사이트 로그인

아이디를 입력해주세요 id_img 패스워드를 입력해주세요 pw_img

로그인유지

[회원가입 아이디/비밀번호를 잊어 버리셨나요?](#)

- [Array틀박스](#)
 - 소개
 - 공지사항
 - 멤버
 - 프로젝트
 - 포토 게시판
- [Array빅데이터강의](#)
 - 빅데이터?
 - R
 - SAS
 - SQL
 - EXCEL
 - 대시보드

Unit 01 | 크롤링을 위한 기초 개념

1.1. 기본 웹 통신 이해 (HTTP): Method

GET

POST

정보를 얻기 위함 (GET)

정보를 쓰기 위함 (POST)

Query String 활용

*Query String: 주소 뒤에 "?"를 붙인 뒤 정보 전송

Request Body 활용

보안 우수

Unit 01 | 크롤링을 위한 기초 개념

1.1. 기본 웹 통신 이해 (HTTP): Headers

▼ Request Headers (573 B)		Raw <input type="checkbox"/>
②	Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8	
②	Accept-Encoding: gzip, deflate, br	
②	Accept-Language: en-US,en;q=0.5	
②	Connection: keep-alive	
초중요	Cookie: EME_ED	
중요	Host: www.naver.com	
중요	Referer: https://www.google.com/	
②	TE: Trailers	
②	Upgrade-Insecure-Requests: 1	
중요	User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15; rv:79.0) Gecko/20100101 Firefox/79.0	
중요	Content-Type (Post에서 중요)	

Unit 01 | 크롤링을 위한 기초 개념

1.1. 기본 웹 통신 이해 (HTTP): Cookie 초중요

- HTTP는 기본적으로 **Stateless Protocol** 이다.
- 이러한 한계를 극복하기 위해 서버에서 브라우저에 심는 것이 바로 **쿠키**
 - ✓ 쿠키를 통해 서버는 유저가 누구인지 알 수 있다. (즉 유저 상태를 보존할 수 있다.)
 - ✓ 쿠키는 유저의 정보를 담는 것이 일반적으로 사이트에 접속하거나 로그인을 할 때 발급된다.
- 이때 서버는 쿠키의 유효기간과 여러 설정을 할 수 있다.
 - ✓ 유효기간을 늘릴 수록 오래 브라우저에 남아 계속 활용할 수도 있다.
 - ✓ 혹은 Session, 즉 브라우저를 종료하면 삭제되게 하여 보안을 높일 수도 있다.



Unit 01 | 크롤링을 위한 기초 개념

NAVER 보안 로그인 접속 시

Name	Value	Domain	Path	Expires / Max-Age
NID_SAUTO		.nid.naver.com	/	2020-09-16T18:45:38.225Z
nid_enctp		.nid.naver.com	/	2020-11-15T18:45:27.762Z
NID_JKL		.naver.com	/	Session
NID_SES		.naver.com	/	Session
NID_AUT		.naver.com	/	Session
JSESSIONID		www.naver.com	/	Session
PM_CK_loc		www.naver.com	/	2020-08-18T10:46:58.330Z
NRTK		.naver.com	/	2021-08-17T18:45:39.000Z

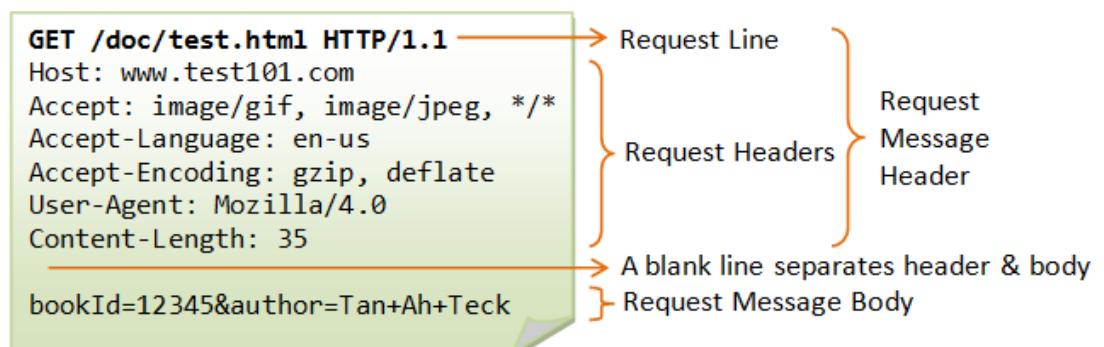
NAVER 자동 로그인 접속 시

Name	Value	Domain	Path	Expires / Max-Age
JSESSIONID		www.naver.com	/	Session
NID_AUT		.naver.com	/	2022-08-17T18:47:03.702Z
NID_JKL		.naver.com	/	2022-08-17T18:47:03.702Z
NID_SAUTO		.nid.naver.com	/	2020-09-16T18:45:38.225Z
NID_SES		.naver.com	/	2020-09-16T18:47:05.409Z
NM_RTK_VIEW_GUIDE		www.naver.com	/	2021-08-07T12:50:07.000Z
NM_THEME_EDIT		www.naver.com	/	2030-08-15T18:47:04.000Z
NNB		.naver.com	/	2050-01-01T09:00:00.022Z

Unit 01 | 크롤링을 위한 기초 개념

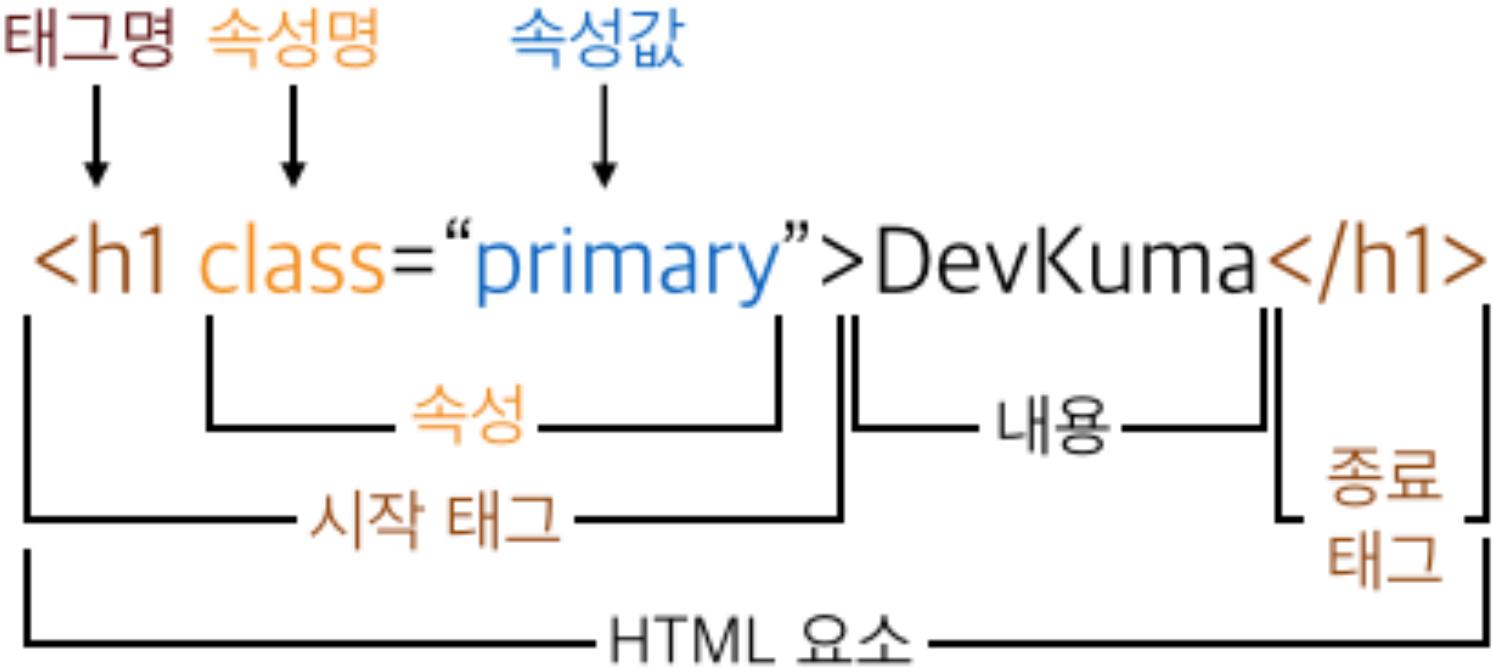
1.1. 기본 웹 통신 이해 (HTTP) 정리

1. HTTP는 웹 문서 전송을 위한 약속이다.
2. 그 약속의 방법(Method) 중 대표적으로 **GET과 POST**가 있다.
3. **Header**는 통신에 대한 기본적인 정보이다.
4. 그 중 **Cookie는 유저의 인증**과 연관이 있다.
5. **GET은 Query를 통해, POST는 Body(data)를 통해 정보를 전달한다.**



Unit 01 | 크롤링을 위한 기초 개념

1.2. HTML과 Selector



Unit 01 | 크롤링을 위한 기초 개념

1.2. HTML과 Selector

1. Selector 를 이용해 HTML 문서 중 원하는 정보를 가져올 수 있다
2. 특히 class 와 id 속성이 중요하다.
3. Class 는 . 을 통해 선택할 수 있고
Id는 # 을 통해 선택할 수 있다.
4. 너무 깊게 알기 보단 활용만 하면 좋음 (크롬 툴 활용)
5. 유용한 selector도 많아서 아래 게임을 통해 한번 익혀보면 좋음

<https://flukeout.github.io/>

```
<div id="header">
  <!-- 상단배너 -->

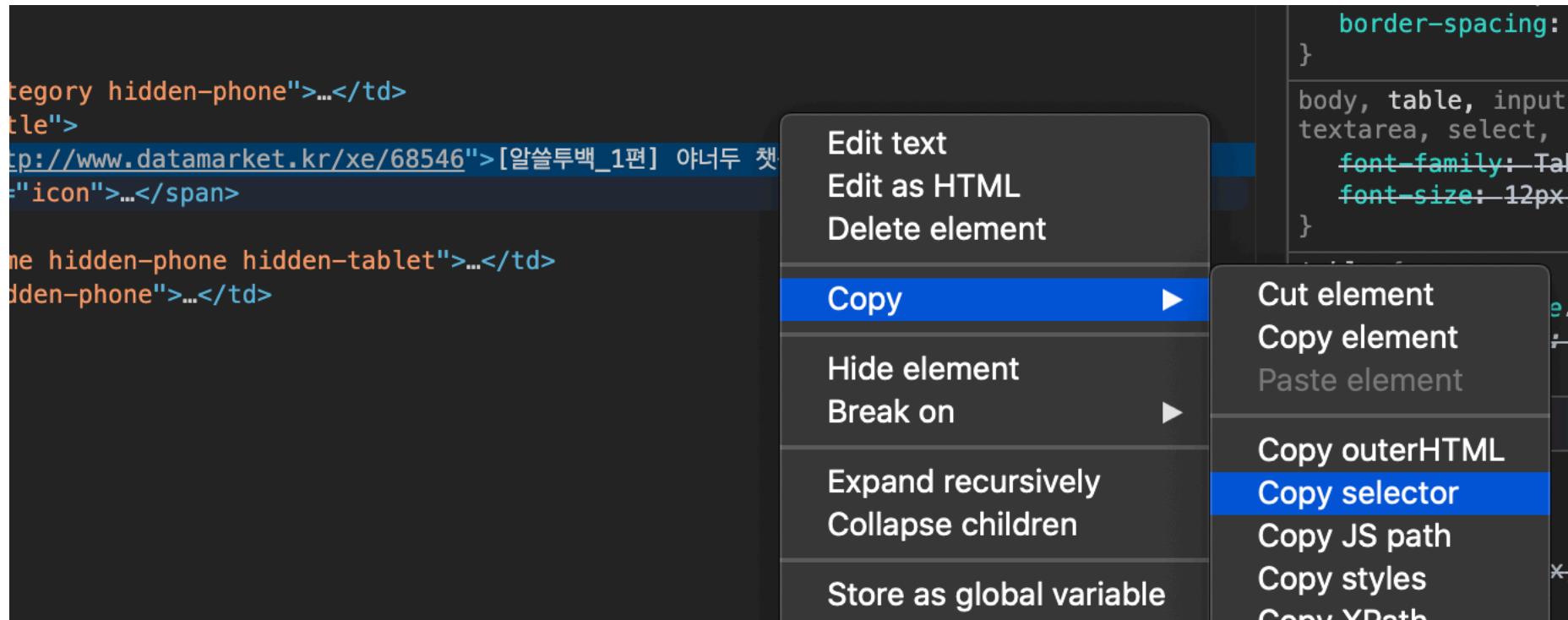
  <!-- 메뉴바 -->
  <div class="menu_bg"></div>
  <div class="menu">
    <div class="gnb fix_width">

      <div class = "logo" onclick="window.location.href='http://www.datamarket.kr';">
        <a ></a>           <a class="text_strong"></a>           <a ></a>
      </div>

      <div class="fix_width">
        <ul>
          <li class="main_hover main_active">
            <div class="pipe_wrap"><!--파이프 사용여부-->
              <span class="pipe1"></span>
              <span class="pipe2"></span>
            </div>
            <a href="/xe/page_QEhq64" class="hover_1 active_1">
              <!-- 리스트 미니 이미지 -->
              
              <li>
                <!-- 3차 메뉴 있을시 표식-->
                <a href="/xe/page_SKdp53" class="active_a">소개</a>
                </li><li>
                <!-- 3차 메뉴 있을시 표식-->
                <a href="/xe/board_lh0x96" class="active_a">공지사항</a>
                </li><li>
                <!-- 3차 메뉴 있을시 표식-->
                <a href="/xe/page_ryhi16" class="active_a">멤버</a>
                </li><li>
                <!-- 3차 메뉴 있을시 표식-->
                <a href="/xe/board_pdzw77" class="active_a">프로젝트</a>
                </li><li>
                <!-- 3차 메뉴 있을시 표식-->
                <a href="/xe/board_FdiN37" class="active_a">포토 게시판</a>
              </li>
            </ul>
          </li>
        </ul>
      </div>
    </div>
  </div>
</div>
```

Unit 01 | 크롤링을 위한 기초 개념

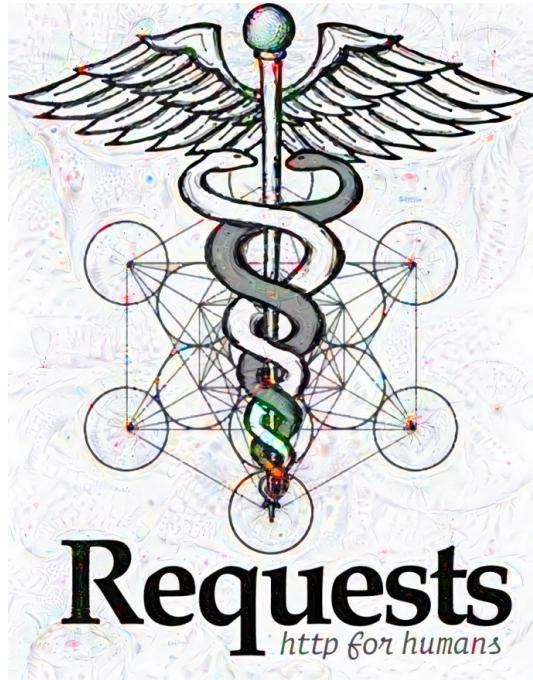
1.2. HTML과 Selector



```
#body > div.content_wrap > div.contents > div:nth-child(2)  
> div > div > table > tbody > tr:nth-child(1) > td.title > a
```

Unit 01 | 크롤링을 위한 기초 개념

1.3. Python Requests



Python HTTP Client

pip install requests

Unit 01 | 크롤링을 위한 기초 개념

1.3. Python Requests: GET

```
# URL
url = "https://search.naver.com/search.naver"

# Parameter -> Query가 됨 https://search.naver.com/search.naver?where=news&sm=tab_jum&query=테넷
my_params = {'where': 'news',
              'sm': 'tab_jum',
              'query': '테넷'}

# Headers 설정
my_headers = {
    "referer": "https://www.tobigs.com"
}

# 전송
res = requests.get(url, params=my_params, headers=my_headers)
res # 상태코드 200으로 성공

<Response [200]>

res.text[:500] # 결과

'<!doctype html> <html lang="ko"> <head> <meta charset="utf-8"> <meta name="referrer" content="al
ways"> <meta name="format-detection" content="telephone=no,address=no,email=no"> <meta name="vie
wport" content="width=device-width,initial-scale=1.0,maximum-scale=2.0"> <meta property="og:title"
content="테넷 : 네이버 뉴스검색"/> <meta property="og:image" content="https://ssl.pstatic.net/sstat
ic/search/common/og_v3.png"> <meta property="og:description" content="\'테넷\'의 네이버 뉴스검색 결과입니
다."> <meta name="description"
```

URL

headers

params

Requests.get

Response.text || Response.json()

Unit 01 | 크롤링을 위한 기초 개념

1.3. Python Requests: POST

```
import requests

url = "https://api.everytime.kr/find/lecture/article/list"

payload = {
    'school_id': [REDACTED],
    'limit_num': 2,
    'lecture_id': [REDACTED]

headers = {
    'Cookie': '_ga=GA1.2.168813388.1596099998; _gid=GA1.2.1403299065.1597760084; _gat_gtag_UA_220221',
    'Content-Type': 'application/x-www-form-urlencoded'
}

response = requests.post(url, headers=headers, data = payload)

print(response.text[:500])
<?xml version="1.0" encoding="UTF-8"?>
<response lectureId="1688764">
    <lecture name="컴퓨터네트워크론" professor="김유성" campus="" />
    <rate>4.55</rate>
    <details assessment_grade="학점느님" assessment_homework="보통" assessment_team="없음" assessment_attendance="전자출결" exam_times="두 번"/>
    <semesters>
        <semester year="2020" semester="2"/>
        <semester year="2020" semester="1"/>
        <semester year="2019" semester="2"/>
        <semester year="2018" semester="2"/>
        <semester year="2018" semester="1"/>
        <se
```

URL

headers

data

Requests.**post**

Response.text || Response.json()

Unit 01 | 크롤링을 위한 기초 개념

1.4. 크롤링 불법성 윤리

대법원 "웹사이트 무단 크롤링은 불법"

김동훈 기자 99re@bizwatch.co.kr
2017.09.27(수) 17:37

잡코리아, 사람인 상대 9년 소송전서 승소
크롤링(데이터 수집) 법적 기준 정립



웹사이트 콘텐츠를 긁어오는 '크롤링'을 이용해 확보한 콘텐츠를 자신의 영업에 무단 사용하는 것은 데이터베이스(DB)권 침해 행위라는 대법원 판단이 나왔다. 이는 온라인 웹사이트를 운영하는 사업자 사이에서 광범위하게 이용되는 크롤링에 대한 법적 판단의 기준을 세웠다는 점에서 주목된다.

<http://news.bizwatch.co.kr/article/mobile/2017/09/27/0023>

法 “여기어때, 야놀자 정보 무단수집 맞다”...前 대표 ‘유죄’

징역 1년6개월에 집행유예 2년 선고...회사도 벌금형

백봉삼 기자 | 입력 : 2020/02/11 15:44 -- 수정: 2020/02/11 16:31 | 중기/벤처

<https://zdnet.co.kr/view/?no=20200211153634>

공대생이 '실시간 주문'한 마스크...할머니는 5시간 기다렸다

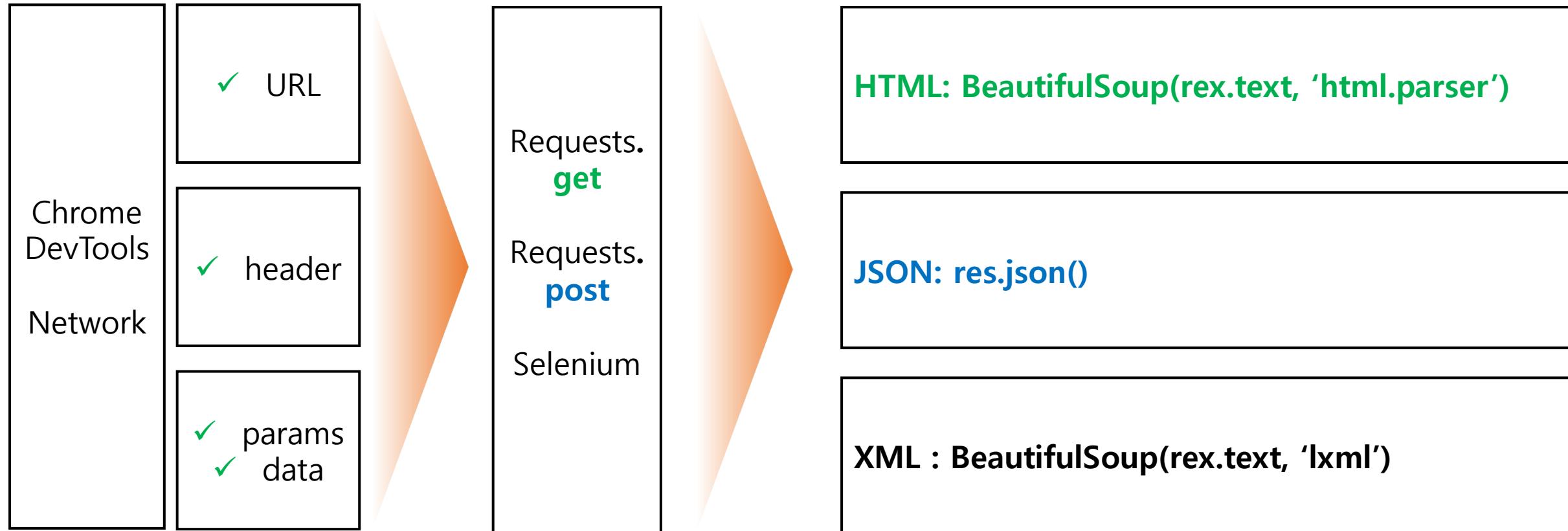
등록 : 2020-03-04 20:43 수정 : 2020-03-05 09:46

만약 조사 도중 매크로 프로그램으로 마스크를 산 사실이 확인되면, 경찰은 그 사람을 '업무방해 혐의'로 입건하고 본격 수사에 나서는데요. 실제로 지난달 29일 경기도 수원에 사는 20대 이 모 씨가 입건돼 조사를 받고 있습니다. 이 씨는 지난달 초부터, 지인의 쿠팡 아이디 8개를 빌려 자신의 컴퓨터 1대로 마스크 9천여장을 사들인 뒤 2배 가격에 되판 혐의를 받고 있습니다. 쿠팡에선 컴퓨터 1대로 여러 아이디 접속이 불가능하고 아이디당 마스크 구매 수량도 제한돼 있었지만, 매크로 공격 앞에선 무용지물이었습니다.

<http://news.bizwatch.co.kr/article/mobile/2017/09/27/0023>

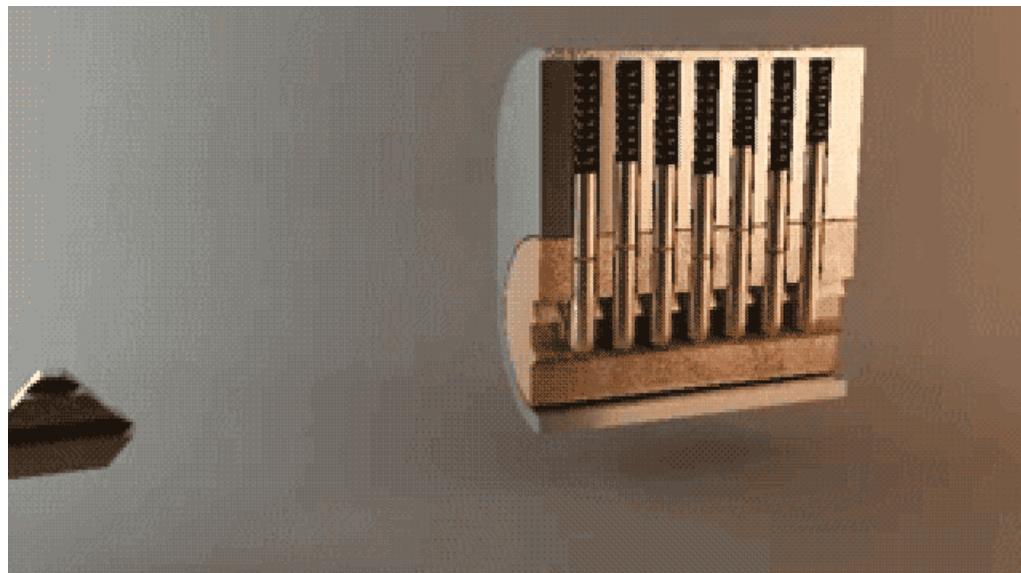
Unit 02 | 크롤링 활용 및 실습

2.0. 크롤링 전체 흐름



Unit 02 | 크롤링 활용 및 실습

2.0. 크롤링 전체 흐름



Unit 02 | 크롤링 활용 및 실습

2.0. 크롤링 전체 흐름

실습으로~

Unit 03 | 과제

1) 네이버 랭킹 뉴스 크롤링 https://news.naver.com/main/ranking/popularDay.nhn?rankingType=popular_day&date=20200818

1) 수집 내용

- 1) 많이 본 뉴스 – 섹션별 (정치 ~ IT/과학) Top5 기사 제목, 신문사
- 2) 해당 기사별 기사 내용, 리액션 (좋아요 ~ 후속기사 원해요)

많이 본 뉴스 | 일간 누적 집계한 조회수입니다. 총 누적수와는 다를 수 있습니다.



2) 수집 방법 (택 1)

- 1) [기본] Requests , BeautifulSoup, **Selenium**
- 2) [심화] Requests, BeautifulSoup (+ 멀티프로세싱)

3) 수집 범위 및 저장

- 1) 2019년 7월 21일 ~ 2020년 8월 20일 (동작 가능, 실제 구동 x)
- 2) 하나의 파일로 저장 (방식 자유)
- 3) Ex) 총 6 섹션 * Top 5 * 365일 = 10950 rows

정치		더보기 >
1	이나연, 검사 후 자가격리...확진자 쓴 방송사 마이크 사용 더불어민주당 이나연 당 대표 후보가 18일 자가 격리에 들어갔다. 이 후보 측은 이날 공지를 통... 한국경제 235,748	더보기
2	"다망한 정권, 정은경이 지탱" "폭망 지름길" 여야 때린 진... 중앙일보 227,911	
3	"녹음기 켜놔야 한다"...文대통령-김종인 만남 결렬의 이유 머니투데이 220,786	
4	수도권 교회 예배·클럽부페·노래방 금지 조선일보 199,690	
5	정의당 원내대표 "박정희는 반민족행위자...무덤 파내자" 조선일보 192,133	



Unit 03 | 과제

주의 사항 및 발전 사항

1. ! 코드로 동작 가능하되 실제 파일로 저장은 하지 않아도 됨!

너무 큰 파일의 경우 git에 올릴 때 오류가 발생할 수 있음 (15MB를 넘었다면 .gitignore에 추가 혹은 삭제)

2. 심화 방법 사용 시 가점 (만약 기본과 속도 비교까지 한다면 무조건 우수과제)

3. Top 5 초과로 수집 시 가점

4. 댓글 (글쓴이, 추천, 비추, 내용, 과거 이력) 까지 수집 시 가점

댓글까지 수집한다면 댓글은 다른 파일로 저장하는게 좋습니다! 물론 어느 기사에 대한 것인지 정보를 같이 포함해야 합니다.

Q & A

들어주셔서 감사합니다.