



Web Crawling

Syllabus - Sequential(Time series data)

4월 11일	시계열 데이터 수집 (웹 크롤링)
4월 14일	5.02.시계열데이터분석
4월 15일	기온 데이터 실습 (EDA, 데이터 전처리 등)
4월 16일	MLP
4월 17일	MLP 5.03.MLP.BackProp.example
4월 18일	회사 견학
4월 21일	DL 모델 이해 및 적용 (CNN)
4월 22일	DL 모델 이해 및 적용 (RNN)
4월 23일	DL 모델 이해 및 적용 (LSTM)
4월 24일	DL 모델 이해 및 적용 (AE)
4월 25일	5.06.EXAM.y24 - DL 모델 종합 실습 (모델링, 결과 시각화, 모델 해석 등)
4월 28일	프로젝트

Web Crawling

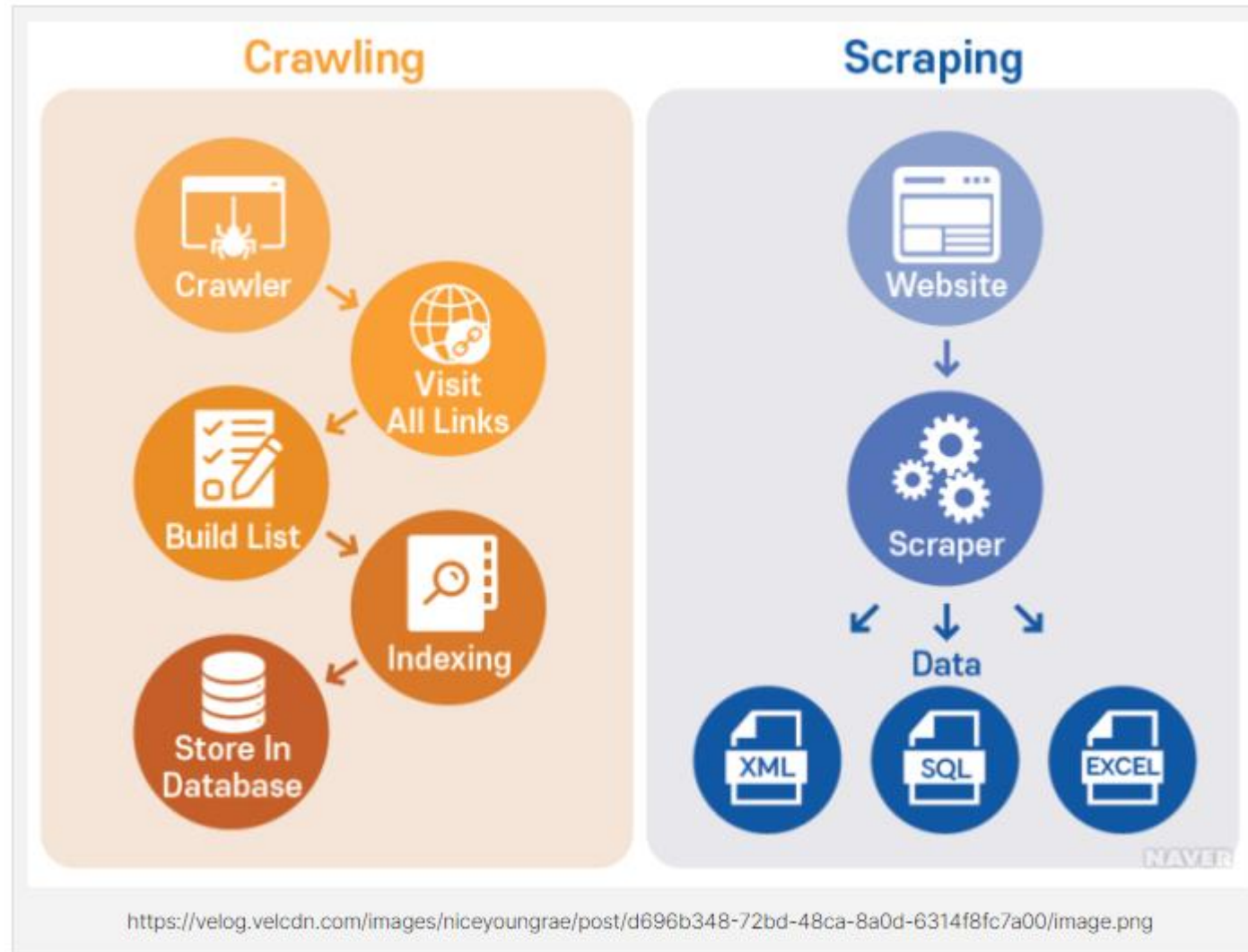
웹 크롤링(Web Crawling)

- 웹상의 정보들을 탐색하고 수집(인덱싱)하는 작업
- 웹 크롤러 (Crawler) 라고 불리는 자동화된 프로그램이 웹사이트를 순회하며 데이터 수집 링크를 따라가며 다양한 웹 페이지로 이동

웹 스크래핑(Web Scraping)

- 웹 페이지에서 필요한 데이터를 추출하는 작업 ('스크래퍼' 가 작업)
 - 1) (목표로 하는) 특정 웹 사이트(웹 페이지) HTTP GET 요청 전송
 - 2) 사이트 응답 수신(웹페이지 수신)
 - 3) 수신된 HTML 문서를 분석하여 데이터 추출

Web Crawling



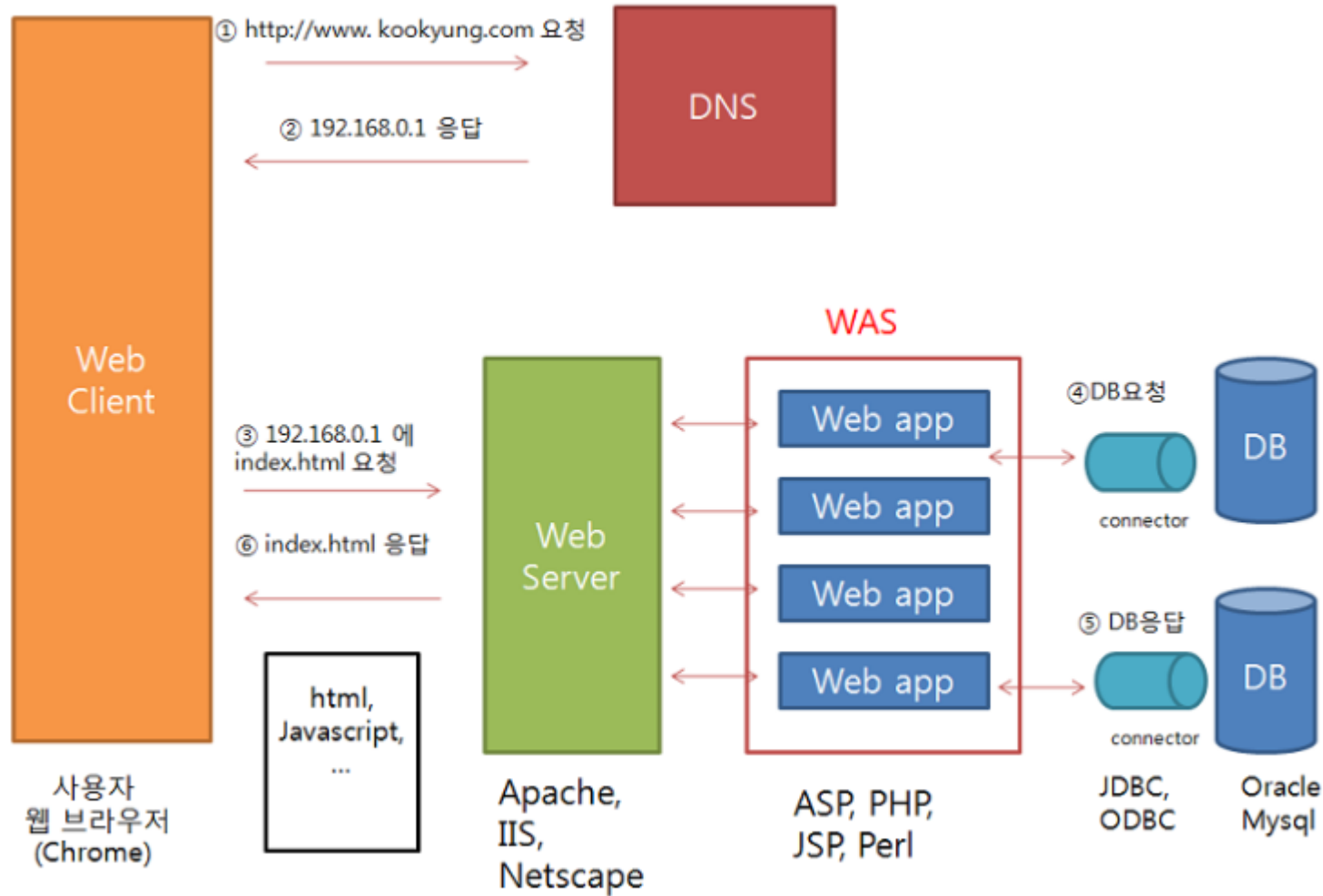
대규모의 웹 페이지 인덱싱

특정 정보, 데이터 추출

Web Crawling

Web

Web - 구조



Web Crawling

Web

Web - 구조



```
Microsoft Windows [Version 10.0.19045.5011]
(c) Microsoft Corporation. All rights reserved.

C:\Users\01>ping www.naver.com

Ping e6030.a.akamaiedge.net [104.109.240.195] 32바이트 데이터 사용 :
104.109.240.195의 응답 : 바이트=32 시간=3ms TTL=56
104.109.240.195의 응답 : 바이트=32 시간=4ms TTL=56
104.109.240.195의 응답 : 바이트=32 시간=5ms TTL=56
104.109.240.195의 응답 : 바이트=32 시간=3ms TTL=56

104.109.240.195에 대한 Ping 통계 :
    패킷 : 보냄 = 4, 받음 = 4, 손실 = 0 (0% 손실),
왕복 시간(밀리초):
    최소 = 3ms, 최대 = 5ms, 평균 = 3ms

C:\Users\01>
```

Web Crawling

Web

Web - 구조

```
C:\Users\01 >ipconfig
```

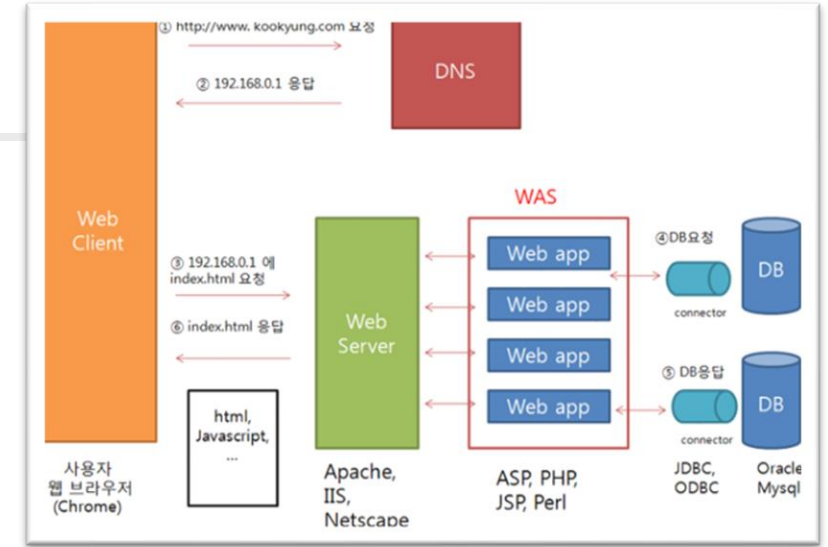
```
연결별 DNS 접미사 . . . . . :  
링크-로컬 IPv6 주소 . . . . . : fe80::a43:af6e:2fdb:d2ed%16  
IPv4 주소 . . . . . : 10.230.101.84  
서브넷 마스크 . . . . . : 255.255.252.0  
기본 게이트웨이 . . . . . : 10.230.100.1
```

Web Crawling

Web

Web - 주요 구성 요소

- **웹 페이지(Web Page):** 웹 상에서 정보를 제공하는 문서
주로 HTML(HyperText Markup Language)로 작성
+ CSS(Cascading Style Sheets)와 JavaScript 등
- **웹 사이트(Web Site):** 공통된 목적 또는 주제를 가진 관련 웹 페이지들의 모음 = 서비스
보통 메인 페이지(홈페이지)를 통해 여러 하위 페이지로 구성
- **웹 서버(Web Server):** 웹 페이지, 사이트, (or 웹 애플리케이션)을 호스팅, 사용자의 요청에 따라 이들을
웹 브라우저로 전송하는 프로그램(서버) cf) 물리적 서버 = 장비
- **웹 브라우저(Web Browser):** 사용자가 웹 서핑(검색 등)을 위해 사용하는 소프트웨어
HTML 문서를 해석(파싱)하여 사용자에게 시각적으로 표시
- **URL(Uniform Resource Locator):** 인터넷 상의 자원(웹 페이지, 이미지, 동영상 등)의 위치를 지정하는
표준화된 주소 체계 ex) <http://www.naver.com>
- **HTTP(Hypertext Transfer Protocol):** 클라이언트와 서버 간에 웹 페이지 및 기타 콘텐츠를 전송하기 위해
사용되는 프로토콜 cf) https, ftp



Web Crawling

Web

Web – 기본 동작 원리

1. 요청(Request)

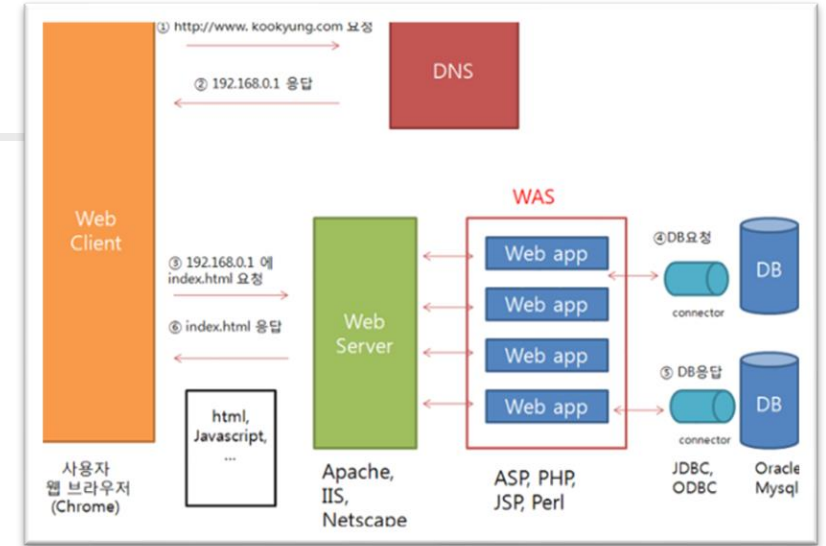
사용자가 웹 브라우저를 통해 특정 URL을 입력 or 링크 클릭을 통해 요청,
웹 브라우저가 해당 웹 서버에 요청 전달

2. 응답(Response)

웹 서버가 요청 받은 웹 페이지를 찾아, 해당 페이지의 HTML, CSS, JavaScript 파일 등을 웹 브라우저로 전송

3. 렌더링(Rendering):

웹 브라우저는 파일들을 수신, 해석하여 사용자에게 시각적으로 표현



Web Crawling

Web

Web - URL

host path fragment
http://www.google.com:5883/search/food.html?topic=pizza#top
protocol port query

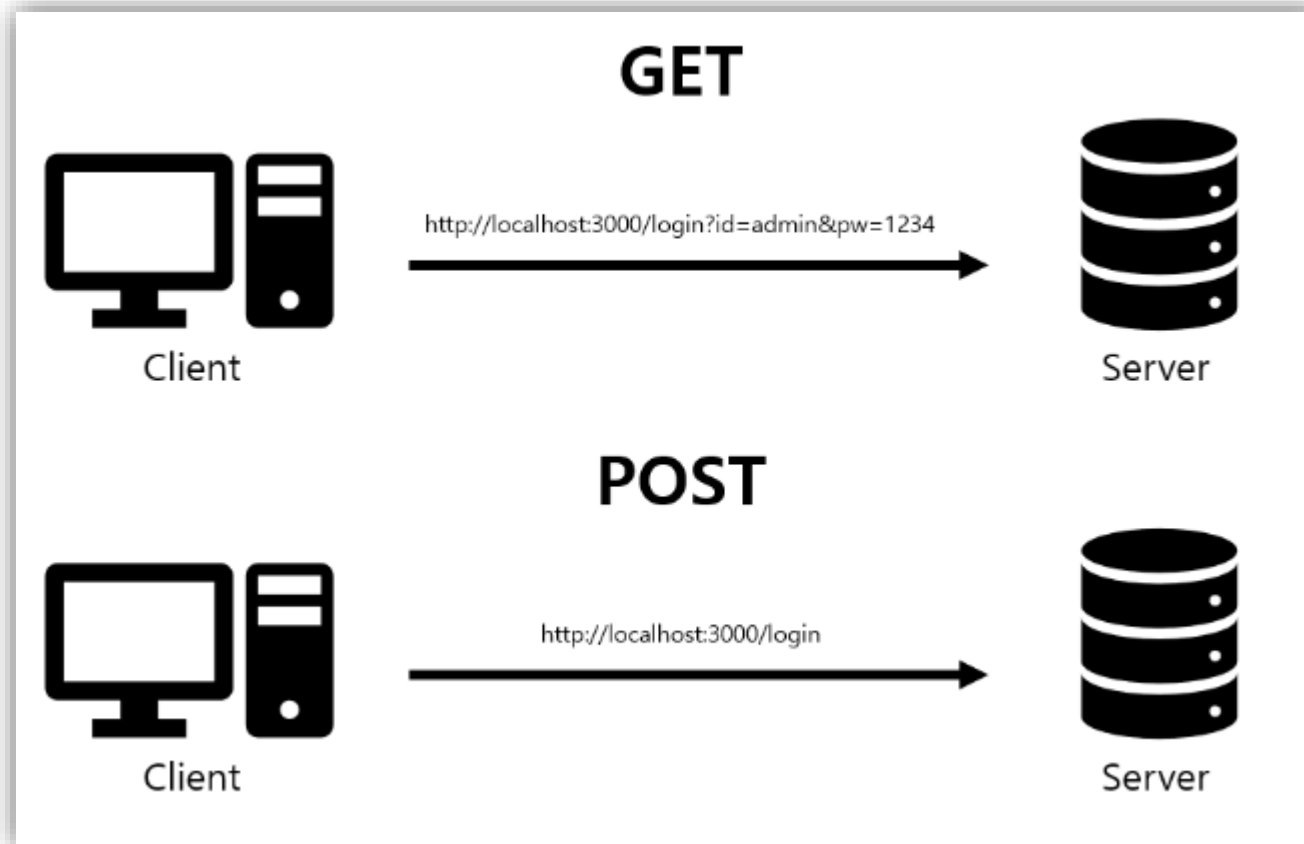
Web Crawling

Web

Request – GET, POST

1. 요청(Request)

사용자가 웹 브라우저를 통해 특정 URL을 입력 or 링크 클릭을 통해 요청,
웹 브라우저가 해당 웹 서버에 요청 전달



Web Crawling

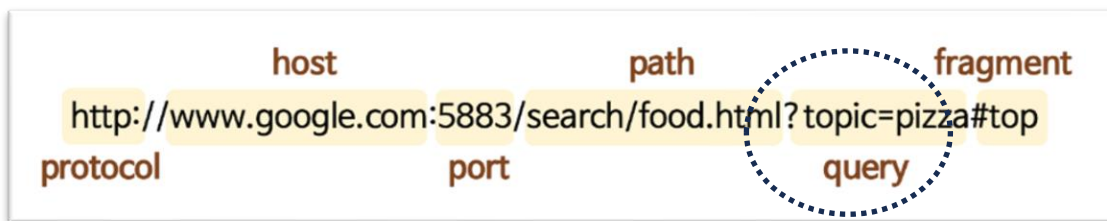
Web

Request - GET

1. 요청(Request)

사용자가 웹 브라우저를 통해 특정 URL을 입력 or 링크 클릭을 통해 요청,
웹 브라우저가 해당 웹 서버에 요청 전달

GET 이란?



GET 은 클라이언트에서 서버로 어떠한 리소스로 부터 정보를 요청하기 위해 사용되는 메서드이다.

예를들면 게시판의 게시물을 조회할 때 쓸 수 있다.

GET을 통한 요청은 URL 주소 끝에 파라미터로 포함되어 전송되며, 이 부분을 쿼리 스트링 (query string) 이라고 부른다.

방식은 URL 끝에 "?" 를 붙이고 그다음 변수명1=값1&변수명2=값2... 형식으로 이어 붙이면 된다.

예를들어 다음과 같은 방식이다.

`www.example.com/show?name1=value1&name2=value2`

서버에서는 name1 과 name2 라는 파라미터 명으로 각각 value1 과 value2 의 파라미터 값을 전달 받을 수 있다.

Web Crawling

Web

Request - POST

1. 요청(Request)

사용자가 웹 브라우저를 통해 특정 URL을 입력 or 링크 클릭을 통해 요청,
웹 브라우저가 해당 웹 서버에 요청 전달

POST 란?

POST는 클라이언트에서 서버로 리소스를 생성하거나 업데이트하기 위해 데이터를 보낼 때 사용 되는 메서드다. 예를들면 게시판에 게시글을 작성하는 작업 등을 할 때 사용할 된다.

POST는 전송할 데이터를 HTTP 메시지 body 부분에 담아서 서버로 보낸다. (body의 타입은 Content-Type 헤더에 따라 결정 된다.)

GET에서 URL의 파라미터로 보냈던 name1=value1&name2=value2가 body에 담겨 보내진다고 생각하면 된다.

POST로 데이터를 전송할 때 길이 제한이 따로 없어 용량이 큰 데이터를 보낼 때 사용하거나 GET처럼 데이터가 외부적으로 드러나는건 아니어서 보안이 필요한 부분에 많이 사용된다.

(하지만 데이터를 암호화하지 않으면 body의 데이터도 결국 볼 수 있는건 똑같다.)

POST를 통한 데이터 전송은 보통 HTML form을 통해 서버로 전송된다.

Web Crawling

Web

Request - POST

1. 요청(Request)

사용자가 웹 브라우저를 통해 특정 URL을 입력 or 링크 클릭을 통해 요청,
웹 브라우저가 해당 웹 서버에 요청 전달

소식 게시물 제목30

게시글 삭제

게시글 편집 확인

제 목 *

소식 게시물 제목30

분 류

소식

작 성 일

2023-12-26 15:00:27

작 성 자 *

개발자

단락

B

I

A

소식 게시물 내용30

컬럼명	데이터 타입	설명
ID	INTEGER	게시글의 고유 번호 (기본 키)
Category	TEXT	게시글의 카테고리
Title	TEXT	게시글의 제목
Content	TEXT	게시글의 내용
Author	TEXT	게시글 작성자의 이름
DatePosted	TEXT	게시글이 작성된 날짜와 시간

Web Crawling

Web

HTML

```
<!doctype html> <html lang= ko class= fzoom > <head> <meta charset= utf-8 > <meta name= Heferrer content= origin > <meta http-equiv= X-UA-Compatible content= ie=edge > <meta name= viewport content= width=1190> <title>NAVER</title> <meta name= apple-mobile-web-app-title content= NAVER/> <meta name= robots content= index,nofollow/> <meta name= description content= 네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요/> <meta property= og:title content= 네이버> <meta property= og:url content= https://www.naver.com/> <meta property= og:image content= https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png> <meta property= og:description content= 네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요/> <meta name= twitter:card content= summary> <meta name= twitter:title content= > <meta name= twitter:url content= https://www.naver.com/> <meta name= twitter:image content= https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png> <meta name= twitter:description content= 네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요/> <link rel= shortcut icon type= image/x-icon href= /favicon.ico?1> <link rel= apple-touch-icon-precomposed href= https://s.pstatic.net/static/www/nFavicon96.png/> <link rel= apple-touch-icon sizes= 114x114 href= https://s.pstatic.net/static/www/u/2014/0328/mma_204243574.png/> <link rel= apple-touch-icon href= https://s.pstatic.net/static/www/u/2014/0328/mma_20432863.png/> <link rel= stylesheet href= https://ssl.pstatic.net/sstatic/search/pc/css/sp_autocomplete_231130.css> <script>window.gladsdk=window.gladsdk||{};window.gladsdk.cmd=window.gladsdk.cmd||[];window.ndpsdk=window.ndpsdk||{};window.ndpsdk.cmd=window.ndpsdk.cmd||[];window.ndpsdk.polyfill=window.ndpsdk.polyfill||{};var g_ssc=navertop.v5;window.nsc=g_ssc;window.nmain=window.nmain||{};window.nmain.jsOrigin=www</script> <script async src= https://ssl.pstatic.net/tveta/libs/ndpsdk/prod/ndp-loader.js> </script> <script async src= https://ssl.pstatic.net/tveta/libs/glad/prod/gfp-core.js> </script> <script src= https://ssl.pstatic.net/tveta/libs/assets/is/pc/main/min/pc.veta.core.m defer= defer> </script> <script>window["EAGER-DATA"] = window["EAGER-DATA"] || {};  
window["EAGER-DATA"]["PC-FEED-WRAPPER"] = { "@type": "BLOCK", "blocks": [{ "@type": "BLOCK", "blocks": [{ "@type": "PC-CAS-EDIT-CONTENTS-BLOCK", "blocks": null, "materi [{ "@type": "MATERIAL-PC-FEED-EDIT", "title": "돌고 도는 정비 주기를", "url": "https://mycar.naver.com/?topCardType=maintenance", "image": { "url": "https://s.pstatic.net/static/www/mobile/edit/20230829_1095/upload_1693273391781UBW8.jpg", "highlight": "네이버 마이카", "adMark": false, "desc": "이제 외출 필요 없어요!", "desc2": 가능한 모델을 버튼으로 확인하세요!", "clickCode": "mycar", "_id": "6548800558a9ce6eb3e8a64e", "@type": "MATERIAL-PC-FEED-EDIT", "title": "푸조 시승신청", "url": "https://naver.me/x10YIRAV", "image": { "url": "https://s.pstatic.net/static/www/mobile/edit/20240118_1095/upload_1705546451159U0ZFU.jpg", "highlight": "네이버 자동차", "adMark": true, "desc": "시승 가능한 모델을 버튼으로 확인하세요!", "clickCode": "epeugeot", "_id": "6548800558a9ce6eb3e8a64f", "@type": "MATERIAL-PC-FEED-EDIT", "title": "지프 시승신청", "url": "https://naver.me/FfAydygVz", "image": { "url": "https://s.pstatic.net/static/www/mobile/edit/20240102_1095/upload_17041847906004W9hN.jpg", "highlight": "네이버 자동차", "adMark": true, "desc": "시승 가능한 모델을 버튼으로 확인하세요!", "clickCode": "JEEP", "_id": "6548800558a9ce6eb3e8a650", "@type": "MATERIAL-PC-FEED-EDIT", "title": "BYD 시승신청", "url": "https://naver.me/xME4yDOE", "image": { "url": "https://s.pstatic.net/static/www/mobile/edit/20240119_1095/upload_1705655134064hPNzp.jpg", "highlight": "네이버 자동차", "adMark": true, "desc": "시승 가능한 모델을 버튼으로 확인하세요!", "clickCode": "BYD", "_id": "65918205372986ca693489f2", "@type": "MATERIAL-PC-FEED-EDIT", "title": "마세라티 시승신청", "url": "https://naver.me/5hJfkFTo", "image": { "url": "https://s.pstatic.net/static/www/mobile/edit/20240220_1095/upload_17084131287790VFgR.jpg", "highlight": "네이버 마이카", "adMark": true, "desc": "시승 가능한 모델을 버튼으로 확인하세요!", "clickCode": "MA", "_id": "65ba6085660c4e2ad3043c5c", "excludeInPaging": false, "positionForPaging": 0, "realtime": false, "orderRandom": true, "_id": null, "@type": "PC-CAS-EDIT-CONTENTS-BLOCK", "@code": null, "@template": "NONE", "@flowId": null, "@flowExecutionId": null, "@provider": null, "@lastModifiedAt": null, "materials": null, "excludeInPaging": false, "positionForPaging": 0, "ltime": false, "_id": null, "@type": "BLOCK", "@code": "PC-FEED-CARGAME-CAS-EDIT", "@template": "PC-FEED-CAS-EDIT", "@flowId": null, "@flowExecutionId": null, "@provider": null, "@lastModifiedAt": null, "@type": "BLOCK", "blocks": null, "materials": [{ "@type": "MATERIAL-PC-FEED", "title": "현대자동차 아이오닉 5 N, 2024 올해의 차에 선정", "url": "https://post.naver.com/viewer/postView.naver?volumeNo=37365459&memberNo=32414000", "image": { "url": "https://s.pstatic.net/dthumb.phinf/?src=%22https%3A%2F%2Fs.pstatic.net%2Fstatic%2Fwww%2Fmobile%2Fedit%2F20240221_1095%2Fupload_1708495663873o7zq6.jpg%22&type=ff364_236&service=navermain", "source": { "name": "글로벌오토뉴스", "service": "POST", "image": { "url": "https://s.pstatic.net/post.phinf/20160721_160/globalautonews_14690765574890FTG0_JPEG/globalautonews_6042517971570971942.jpg?type=f120_120", "_id": "6441ca92c499cc07f334c614", "@type": "MATERIAL-PC-FEED", "title": "토요타 신형 랜드크루저, 미국 가격은 57,345달러부터", "url": "https://post.naver.com/viewer/postView.naver?volumeNo=37372479&memberNo=8910941", "image": { "url": "https://s.pstatic.net/dthumb.phinf/?src=%22https%3A%2F%2Fs.pstatic.net%2Fstatic%2Fwww%2Fmobile%2Fedit%2F20240222_1095%2Fupload_1708558842824PT52L.jpg%22&type=ff364_236&service=navermain", "source": { "name": "오토얼라인먼트", "service": "POST", "image": { "url": "https://s.pstatic.net/post.phinf/MjAyMDExMTY2fMTcxMDAxNjcwNzI0NjYyMTAx.OxcZh151m-R7saHkrkKGrzOL5H76No9.XxDcU6oEmSsg.ed713d-XxnmLycuMIU1vgjs1u0Mws0Ttwq1kG_penLOg.PNG/post_8767636244181691749.png?type=f120_120", "_id": "6441ca92c499cc07f334c615", "@type": "MATERIAL-PC-FEED", "title": "₩6300억에 660대 한정싱아우디, '더 뉴 RS 6 아반트 GT' 2분기 출시", "url": "https://post.naver.com/viewer/postView.naver?volumeNo=37367831&memberNo=5894665", "image": { "url": "https://s.pstatic.net/dthumb.phinf/?src=%22https%3A%2F%2Fs.pstatic.net%2Fstatic%2Fwww%2Fmobile%2Fedit%2F20240221_1095%2Fupload_1708495668626soNvl.jpg%22&type=ff364_236&service=navermain", "source": { "name": "오토모닝", "service": "POST", "image": { "url": "https://s.pstatic.net/post.phinf/MjAyMDExMTY2fMTY2fMDAxNjA2MjE1OTM3OTg0.2xeJNrGJC1j_zJoE2p4SnsqfdYvW8mCJrgJFLeJ5og.qjeLcGv2-pmY1NLj-RJLJL07STCnBMsJ7xUkVbMrz0g.JPEG/post_2734046224559142214.jpeg?type=f120_120", "_id": "6441ca92c499cc07f334c616", "@type": "MATERIAL-PC-FEED", "title": "V6 얇은 마세라티 그레칼레 트로페오 V8 트로페오와 비교하니?!", "url": "https://tv.naver.com/v/47229691", "image": { "url": "https://s.pstatic.net/dthumb.phinf/?src=%22https%3A%2F%2Fs.pstatic.net%2Fstatic%2Fwww%2Fmobile%2Fedit%2F20240220_1095%2Fupload_1708402933329sd986.jpg%22&type=ff364_236&service=navermain", "source": { "name": "모터피디
```

Web Crawling

Web

HTML

① {
 <?xml version="1.0" encoding="UTF-8" ?>
 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
 "http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
 <html>
 ③ {
 <head>
 ④ {
 <title>Pride and Prejudice</title>
 <link rel="stylesheet" href="css/style.css" type="text/css" />
 </head>
 <body>
 ⑤ {
 ⑥ {
 <p>...</p>
 </body>
 </html>
 }

HTML 요소

태그 이름 속성명 속성값 내용

↓ { { {

<p class="para">TCPschool.com</p>

{ }

시작 태그 종료 태그

Web Crawling

Web

HTML - 태그

텍스트 관련 태그

종류	설명
<h1>	제목을 나타냄
<p>	텍스트 단락. 내용이 길면 웹브라우저 창의 너비에 맞게 줄바꿈
 	줄바꿈.
<blockquote>	내용 인용. 다른 내용보다 들여쓰고 단락으로 표시됨.
	굵게 표시. 화면 낭독기에서 강조해서 읽음.
	굵게 표시. 중요하지 않음
	중요한 텍스트를 기울여 표시(문장). 화면 낭독기에서 강조해서 읽음.
<i>	기울여 표시. 중요하지 않음.
<ins>	추가한 내용을 표시
	삭제한 내용을 표시
<sup>	위 첨자
<sub>	아래 첨자

목록 관련 태그

종류	설명
	순서 있는 목록의 시작과 끝을 나타냄.
	순서 없는 목록의 시작과 끝을 나타냄.
	목록의 각 항목을 나타냄.
<dl>	설명 목록의 시작과 끝을 나타냄.
<dt>	설명 목록의 이름(제목)
<dd>	설명 목록의 값(설명)

Web Crawling

Web

HTML - 태그

표 관련 태그

종류	설명
<table>	표의 시작과 끝을 나타냄.
<caption>	표의 제목
<tr>	표의 행
<td>	표의 셀
<th>	제목 셀
<thead>	표 구조에서 제목 부분
<tbody>	표 구조에서 본문 부분
<tfoot>	표 구조에서 요약이나 정리 부분
<col>	표에서 열의 스타일 지정
<colgroup>	표에서 여러열을 한꺼번에 묶어 스타일 지정

멀티미디어 관련 태그

종류	설명
<object>	멀티미디어, PDF 파일 등 다양한 형식의 파일 삽입
<embed>	audio, video, object 태그를 지원하지 않을 경우 멀티미디어 파일 삽입
	이미지 파일 삽입
<audio>	오디오 파일 삽입
<video>	비디오 파일 삽입)

Additional Material

xml

Web Crawling

Web

Web browser

웹 브라우저 시장 점유율

- 구글 크롬 (62.9%)
- 사파리 (15.97%)
- 파이어폭스 (4.33%)
- UC 브라우저 (2.94%)
- 오페라 (2.34%)



Web Crawling

웹 스크래핑

Web browser

- 웹브라우저의 역할
 - 특정 페이지의 raw data를 해당 raw data가 저장되어 있는 컴퓨터 (서버) 에서 다운로드 받아서 user-friendly 형태로 display
- 특정 웹 페이지의 정보를 추출(= 웹 스크래핑) 하기 위해서는 프로그램(파이썬 등)을 통해 웹 페이지를 로컬로 다운로드 받아야 함

Web Crawling

웹 스크래핑

웹 페이지에서 필요한 데이터를 추출하는 작업

ex)

- Movie information

https://www.imdb.com/showtimes/location?ref=shlc_sh

- 책 제목

<https://www.yes24.com/Product/Goods/118579613>

- 신문기사

<https://n.news.naver.com/mnews/article/001/0014176648?sid=104>

Web Crawling

Source code

- 예제 page
 - https://www.imdb.com/showtimes/location?ref=shlc_sh
 - 페이지의 소스코드를 보기 위해서는
마우스 오른쪽 버튼 클릭 -> '소스 코드 보기' 클릭

Web Crawling

Source code

▪ 소스 코드 구성요소

- Tags
 - `<tag_name> content </tag_name>`
 - Each tag has a different role
 - ` content ` for making the content bold
 - You don't have to understand what each tag does!!!
- Tags contain the information (usually text) that we want to collect

Web Crawling

Web scraping

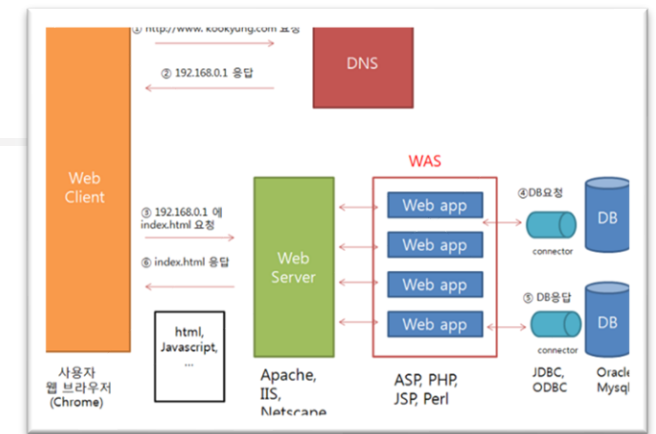
- **웹스크래핑 process**
 - Decide what you need and where you can obtain it
 - Download source code of the page
 - Extract the information
 - **Find the tag** including the information
 - Extract the information
 - Save the information

Web Crawling

Web scraping

- 웹 스크래핑 프로그램의 두가지 역할

역할	관련 모듈
<u>Web communication</u> => download the (raw) data (i.e., source codes)	requests / urllib
<u>Information extraction</u> => extract the information from the source codes and saves it	BeautifulSoup



Web Crawling

Web Data 수집

▪ 웹페이지 소스 코드 다운로드

웹페이지로부터 정보를 다운로드(추출)하기 위해서는 먼저 소스 코드를 다운로드 필요

- 웹페이지의 소스 코드는 서버 즉, 웹 서비스용 컴퓨터에 저장, 서비스
- 웹 브라우저의 주소 창에서 찾고자 하는 URL 주소와 웹 브라우저를 사용하여 웹페이지 소스 코드 다운로드
- requests 모듈 사용

Web Crawling

requests

▪ requests - 역할

- Send HTTP/1.1 requests
- Communicate with a server in Python without using a browser

▪ 관련 모듈

- urllib, urllib2
- Requests is much simpler, fast, and provides more features
- Can add headers, form data, multipart files, and parameters with simple Python dictionaries, and access the response data in the same way

Web Crawling

requests

▪ Requests 모듈

- import requests
- url = 'http://www.daum.net'
- r=requests.get(url)
- r.text
- more information at

<https://docs.python-requests.org/en/latest/>

Web Crawling

Urllib.requests

▪ “requests” 이외의 방법들

- ```
from urllib.request import urlopen
html = urlopen(url)
print(html.read()) # source codes
print(html.read().decode(html.headers.get_content_charset()))
```
- ```
soup = BeautifulSoup(html, 'lxml')
```

▪ BeautifulSoup

- 'bs4' 모듈에 저장된 Python 클래스
- 태그를 기반으로 특정 정보를 찾아내거나 탐색

```
import requests
from bs4 import BeautifulSoup

# 웹 페이지 가져오기
URL = 'https://www.google.com/'
page = requests.get(URL)

# BeautifulSoup 객체 생성
soup = BeautifulSoup(page.content, 'html.parser')

# h1 태그의 텍스트 추출
h1_tag_text = soup.find('h1').text
print(h1_tag_text)
```

Web Crawling

BeautifulSoup

- <http://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- 설치
 - pip install beautifulsoup4
 - BS was originally developed as a HTML parser
 - XML parser: lxml parser, but current BS also supports xml parser
- Importing
 - from bs4 import BeautifulSoup

Web Crawling

BeautifulSoup

▪ Html을 BS object 로 변환

- `r = requests.get (url)`
- `html=r.text`
- `soup=BeautifulSoup(html, 'lxml')`
- BeautifulSoup의 역할
 - HTML 페이지를 중첩된 데이터 구조로 표현, 하고 페이지 파싱
 - 페이지 파싱을 위해 'BeautifulSoup()' 생성자 사용, 해당 페이지를 BS 객체로 변경

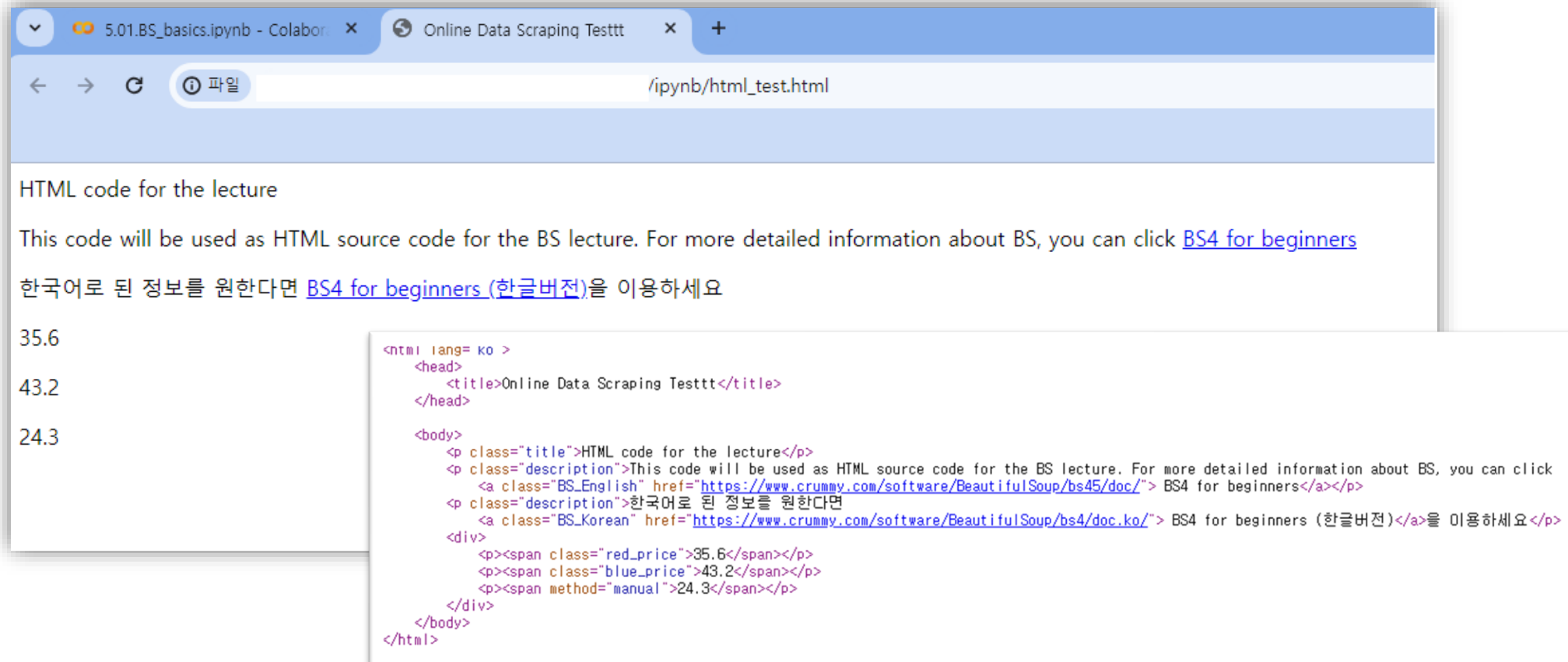
```
soup = BeautifulSoup("<html> <body>Hello World! </body> </html>")  
print(soup.pretify())
```

Web Crawling

실습 - BeautifulSoup

• 5.01. BS_basics.ipynb

HTML 파일에서 정보 추출하기



The screenshot shows a Jupyter Notebook interface with two tabs: "5.01.BS_basics.ipynb - Colaboratory" and "Online Data Scraping Testtt". The active tab displays the following content:

HTML code for the lecture

This code will be used as HTML source code for the BS lecture. For more detailed information about BS, you can click [BS4 for beginners](https://www.crummy.com/software/BeautifulSoup/bs4/doc/)

한국어로 된 정보를 원한다면 [BS4 for beginners \(한글버전\)](https://www.crummy.com/software/BeautifulSoup/bs4/doc.ko/)을 이용하세요

35.6

43.2

24.3

```
<html lang= ko >
<head>
  <title>Online Data Scraping Testtt</title>
</head>
<body>
  <p class="title">HTML code for the lecture</p>
  <p class="description">This code will be used as HTML source code for the BS lecture. For more detailed information about BS, you can click
    <a class="BS_English" href="https://www.crummy.com/software/BeautifulSoup/bs4/doc/"> BS4 for beginners</a></p>
  <p class="description">한국어로 된 정보를 원한다면
    <a class="BS_Korean" href="https://www.crummy.com/software/BeautifulSoup/bs4/doc.ko/"> BS4 for beginners (한글버전)</a>을 이용하세요</p>
  <div>
    <p><span class="red_price">35.6</span></p>
    <p><span class="blue_price">43.2</span></p>
    <p><span method="manual">24.3</span></p>
  </div>
</body>
</html>
```

Web Crawling

실습 - BeautifulSoup

- Open an example html file
 - `html_test.html`
- Accessing a tag
 - `f = read('html_test.html', 'r')`
 - `source = f.read()`
 - `soup = BeautifulSoup(source, 'lxml')`
 - `soup.tag_name`
 - `soup.title`
 - `soup.title.text`

Web Crawling

실습 - BeautifulSoup

- **find(), find_all()**
 - 탐색하고자하는 tag의 조건 (이름, 그외 다른 속성 정보)을 인자로 입력
 - find() : 해당 조건을 만족하는 첫번째 tag 반환
 - find_all() : 해당 조건을 만족하는 모든 tag들 반환
- Examples
 - soup.**find**('a')
 - soup.**find_all**('a')
 - 리스트 데이터가 결과로 반환
 - 찾고자 하는 tag를 indexing을 통해서 접근 가능

Web Crawling

실습 - BeautifulSoup

- Tag의 속성 (attribute) 정보 이용
 - Tag 이름 or 속성 정보를 이용해서 특정 tag를 탐색 가능
 - 속성 정보는 시작 tag에 저장
 - attrs 파라미터 사용

soup.find(tag_name, attrs={'attr_name':'value'})

```
).eq(1).attr("class", "lnk_bn");
```

```
).attr("width", "180");  
)attr("height", "300");
```

Web Crawling

실습 - BeautifulSoup

- If) 원하는 정보가 시작태그와 종료태그 사이에 저장되어 있는 텍스트 정보가 아니라, 특정 속성의 값인 경우
 - BeautifulSoup에서 제공하는 get() 함수 이용
.get('attr_name')
 - 두번째 링크의 url 정보 가져오기
.get('herf')

```
<html lang= ko >
<head>
  <title>Online Data Scrapping Testtt</title>
</head>
<body>
  <p class="title">HTML code for the lecture</p>
  <p class="description">This code will be used as HTML source code for the BS lecture. For more detailed information about BS, you can click
    <a class="BS_English" href="https://www.crummy.com/software/BeautifulSoup/bs4/doc/"> BS4 for beginners</a></p>
  <p class="description">한국어로 된 정보를 원한다면
    <a class="BS_Korean" href="https://www.crummy.com/software/BeautifulSoup/bs4/doc.ko/"> BS4 for beginners (한글버전)</a>을 이용하세요</p>
  <div>
    <p><span class="red_price">35.6</span></p>
    <p><span class="blue_price">43.2</span></p>
    <p><span method="manual">24.3</span></p>
  </div>
</body>
</html>
```

Web Crawling

실습 - BeautifulSoup

- BeautifulSoup 기타 기능
 - `.find_all([' h1',' h2',' h3',' h4',' h5',' h6'])`
찾고자 하는 태그 목록 제공
 - `.find_all('span', attrs={'class':{'green', 'red'}})`
다양한 속성 값 서치

Web Crawling

실습문제 - BeautifulSoup

• 5.01. BS_basics.p.ipynb

연습 문제 : 아래 URL에서 정보 추출하기

- URL

<http://www.yes24.com/Product/Goods/119293186>

1. 위 URL의 HTML 파일을 로컬 폴더에 저장

119293186.html

2. 아래 정보를 추출

- 책제목
- 저자
- 출판사
- 출간일
- 평점
- 판매가
- 할인율



```
<span class="gd_pubArea">
  <span class="gd_auth">
    <a href="https://www.yes24.com/Product/Search?domain=ALL&query=북">
      </span>
    <em class="divi"></em>
    <span class="gd_pub"><a href="javascript:void(0);">
      <em class="divi"></em>
      <span class="gd_date">2023년 06월 15일</span>
    </span>

  <span class="gd_ratingArea">
    <span id="spanGdRating" class="gd_rating">
      <a href="javascript:void(0);">
        <span class="moreRatingArea">
          <span class="moreRatingBtn">
            <a href="javascript:void(0);">
              </span>
            <span class="moreRatingLi">
              <span class="moreRatingIDiv">
```



Web Crawling

실습문제 - BeautifulSoup

<http://www.yes24.com/24/Category/BestSeller>

- 베스트셀러 페이지에 존재하는 각 책들에 대한 책 정보를 추출하기

3 .



[도서] 나는 메트로폴리탄 미술관의 경비원입니다 / 파츠키 브랑리 지/김희영, 조한주 역 | 용인지식하우스 | 2023년 11월

(H) (8월 말까지) 책값이 50% 할인된 가격에 판매 중입니다 (포인트 차감)

15,750원 (10% 할인) | 870원

한해까지 420,632 | 회원리뷰(7947) | ★★★★★ 9.5


22시까지 주문하면 당일 아침 7시 전(14%, 화) 도착예정

#탈라리아 #나폴리시계 #독자선정작 #코레아출판협회대표 #미용한의원세정제

미리보기 | [이벤트](#) | [세종](#) | *책이 아직 없습니다. 나중에 다시 확인해주세요. (2023.12.15)

[카드에 추가](#)
[바구니에](#)
[리스토어에 추가](#)

4 .



[도서] 세이노의 가르침: 피노다 전 세계 삶에 대한 70만 부 기념 에디션 / 세이노(Seyno) 지 | 해리슨 | 2023년 03월

6,480원 (10% 할인) | 360원

한해까지 2,044,512 | 회원리뷰(2,0747) | ★★★★★ 9.0


22시까지 주문하면 당일 아침 7시 전(14%, 화) 도착예정

#올해의책 #공공선화 #공공선화 #살아있는책 #한글서체

미리보기

[카드에 추가](#)
[바구니에](#)
[리스토어에 추가](#)

5 .



[역사] 분리 독립: 최후의 독립을 위한 최후의 투쟁 / 정주희 지/이수진 역 | 시사 | 2024년 03월

(H) (한복) 한복이슈 노트 (포인트 차감)

26,820원 (10% 할인) | 1,400원

한해까지 64,620

최저가 10% 할인(예약구매)

#국립중앙도서관 #조선시대 #한글서체

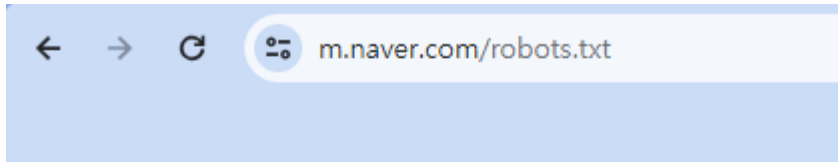
미리보기 | [이벤트](#) | [세종](#) | *책이 아직 없습니다. 나중에 다시 확인해주세요. (2023.12.15)

[카드에 추가](#)
[바구니에](#)
[리스토어에 추가](#)

Web Crawling

웹 크롤링 - API

- 크롤링 허용 여부 확인
 - 크롤링 허용 여부를 확인하기 위해 주소 창에 '크롤링할 주소/ robots.txt'를 입력
 - If) robots.txt 파일이 없다면 수집에 대한 정책이 없으니 크롤링을 해도 된다는 의미



```
User-agent: *  
Disallow: /  
Allow: /$
```

표시	허용 여부
User-agent: * Allow: / 또는 User-agent: * Disallow:	모든 접근 허용
User-agent: * Disallow: /	모든 접근 금지
User-agent: * Disallow: /user/	특정 디렉토리만 접근 금지

Web Crawling

웹 크롤링 - API

- 웹 크롤링 : 웹 페이지에서 필요한 데이터를 추출하는 작업
- 웹 크롤링의 한계
 - 웹 크롤링 수행 시 해당 웹사이트의 이용 약관 및 정책에 따라 제한되는 문제 발생 할 수 있음
 - 많은 웹사이트와 기관들은 자체 데이터를 보호하기 위해 이용 약관에서 자동화된 방법으로는 데이터 접근 제한
 - ex) 공공 기관



API

- 실습
법원 데이터 API로 가져오기

<https://open.law.go.kr/LSO/main.do> 회원가입 / 로그인



Web Crawling

광주 기온 데이터

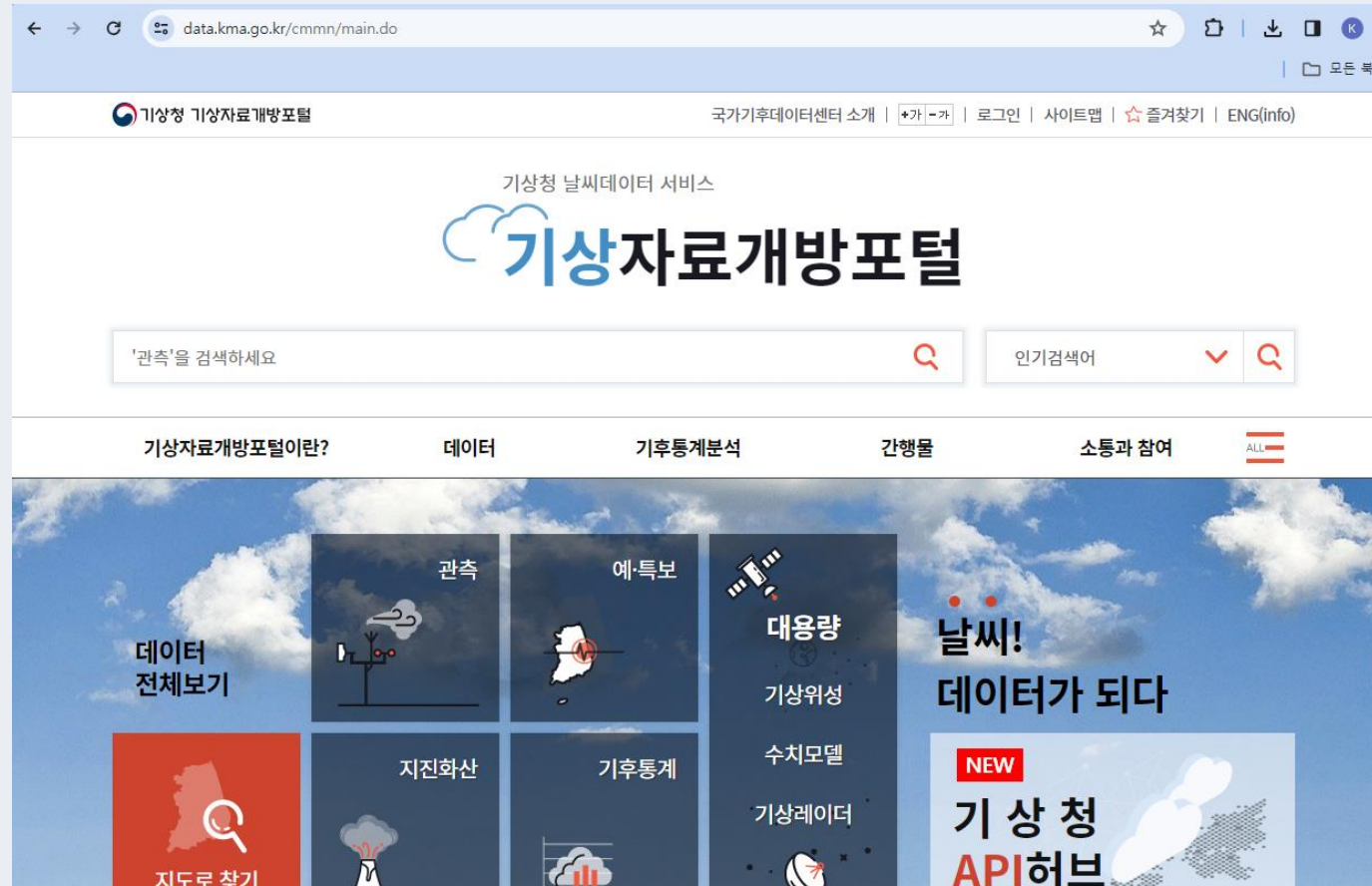
- TASK

1960년 1월 부터 2021년 8월 까지의 광주지역 일일 기온 데이터 수집하기

Web Crawling

광주 기온 데이터

https://data.kma.go.kr/cmmn/main.do



Web Crawling

광주 기온 데이터

<https://data.kma.go.kr/climate/RankState/selectRankStatisticsDivisionList.do?pgmNo=179>

 기상청 기상자료개방포털

국가기후데이터센터 소개 | [*가|-가](#) | 로그인 | 사이트맵 | [☆ 즐겨찾기](#) | [ENG\(info\)](#)



기상자료개방포털이란?

데이터

기후통계분석

간행물

소통과 참여

ALL


기후통계분석

평년값 

통계분석 

조건별통계

기온분석

강수량분석

다중지점통계

24절기

순위값

장마

Home > 기후통계분석 > 통계분석 > 조건별통계

조건별통계

▪ 자료설명 [> 사용법](#)

기온, 강수량, 바람 자료 대상으로 원하는 조건의 자료를 검색할 수 있습니다.

* '지역/지점'의 '지역'은 전국 및 광역 단위의 평균 제공(1973년~)
- 전국 및 광역별 평균 산출에 사용되는 지점은 62개 지점이며 제주도 4개 지점은 제주지역 산출에만 사용됩니다.
[전국/지역별 통계 산출 지점 정보(더보기)]

▪ 검색조건

▪ 분류

지상 

▪ 지역/지점

광주 

[선택](#)

▪ 요소

기온 

▪ 기간

월 

 2014  년 ~ 2024  년

▪ 조건

☐ 요소

평균기온 

 < 



[선택](#) 

☒ 월

01 

 월 ~ 12  월

Web Crawling

광주 기온 데이터

temp_mean_gwangju.csv

	1960_1_m	1960_2_m	1960_3_m	1960_4_m	1960_5_m	1960_6_m	1960_7_m	1960_8_m	1960_9_m	1960_10_n	1960_11_n	1960_12_n	1961_1_m	1961_2_m	1961_3_m
0	2.7	1.7	7.9	3.6	13.5	17.3	24	27.1	24.9	18.8	14.2	1.8	-8.3	-6.8	9.4
1	2.6	0.9	8.5	4.2	17	19	24.1	27.9	24.8	17	10.1	2.1	-3.1	-2.4	13.1
2	6.4	2.5	6.4	7.4	21.5	19.2	24.5	28.6	25.3	15.2	9.4	4.6	0	-2.9	10.9
3	10.9	0.8	8.4	8.2	17.6	18	25.4	28.8	22.2	16.6	10.9	5.4	-3	-0.2	6.2
4	-0.2	1.2	8.7	6	15.6	19.9	25.2	26.2	21.6	16.3	10.2	3.5	-4.3	3.5	4.1
5	-2.2	6.5	12.7	9.2	13.4	20.2	24.3	28	19.5	17.9	9.8	0.6	-3	5	4.8
6	-0.3	9.7	13.7	6.1	14.8	21.4	23.3	28.4	21.3	18.4	10.7	1.6	-2.3	3.2	2.3
7	2.2	8.6	12.1	9.3	14.5	23.5	24.8	28.1	19.2	17	11.4	4.3	1.2	1.1	1.1
8	4.1	6.1	11.8	12.8	12.1	23.2	23.7	28.1	19.7	15.6	10.4	6.7	3.1	1.6	1
9	5.3	0.3	9.8	10.7	12.3	21.9	26.1	27.5	18.7	16.7	10.5	8.3	2.1	0.3	1.1
10	2	0	9.8	10.8	14.2	20.6	26.9	27.8	20.2	16.8	9.3	10	-6.2	1.4	2.9
11	0.4	-1.3	5	11.2	18.2	17.7	26.7	26.3	23.4	17.1	10	5.5	-5.4	-0.7	7.4
12	1.8	0.9	0.7	12.3	18.1	21.9	26.9	26.4	23.8	18.2	6.9	4.2	-2.3	1	9.6
13	3	1.9	1.8	9.1	14.6	22.8	26.6	24.9	19.7	18.1	7.2	6.7	-1.6	-2.7	5.6
14	1.5	0.3	3.6	10.6	17.1	21.5	26.6	26.4	19.3	19	7.5	4.9	-1.6	-2.3	6
15	-2.1	2.2	6.7	7.7	18.2	21.5	26.8	26.1	18.8	15.6	8.4	2.6	-3.7	1	8.3
16	-3.4	4	7.1	6.5	20	20.5	27.3	27.5	17.6	13.2	10.7	1.3	-4.7	1.3	8.5
17	-0.5	0.4	8.2	8.6	17.5	21.5	27.6	27.2		13	10	-3.6	-4	0.3	13.6
18	2.6	3.2	10	11.1	17.9	21.9	27.3	27.6		14.2	10.6	-1.3	-4.3	3	11.2
19	2.7	5	12	10.2	14.9	24.3	27.3	28.1		14.9	11.2	0.4	-1.3	5.4	9.5
20	-0.3	1.1	6.4	10.8	14.4	20.6	26.8	28.6		16.6	16.5	1.1	2.7	3.6	7.4

Web Crawling

광주 기온 데이터

- 실습

`temp_mean_gwangju.csv`

`5.02TempAnal.ipynb` - 5.02.시계열데이터분석

Web Crawling

- 실습
시계열 데이터 준비

Yahoo Finance: 주식 시장 데이터

<https://help.yahoo.com/kb/SLN2311.html>

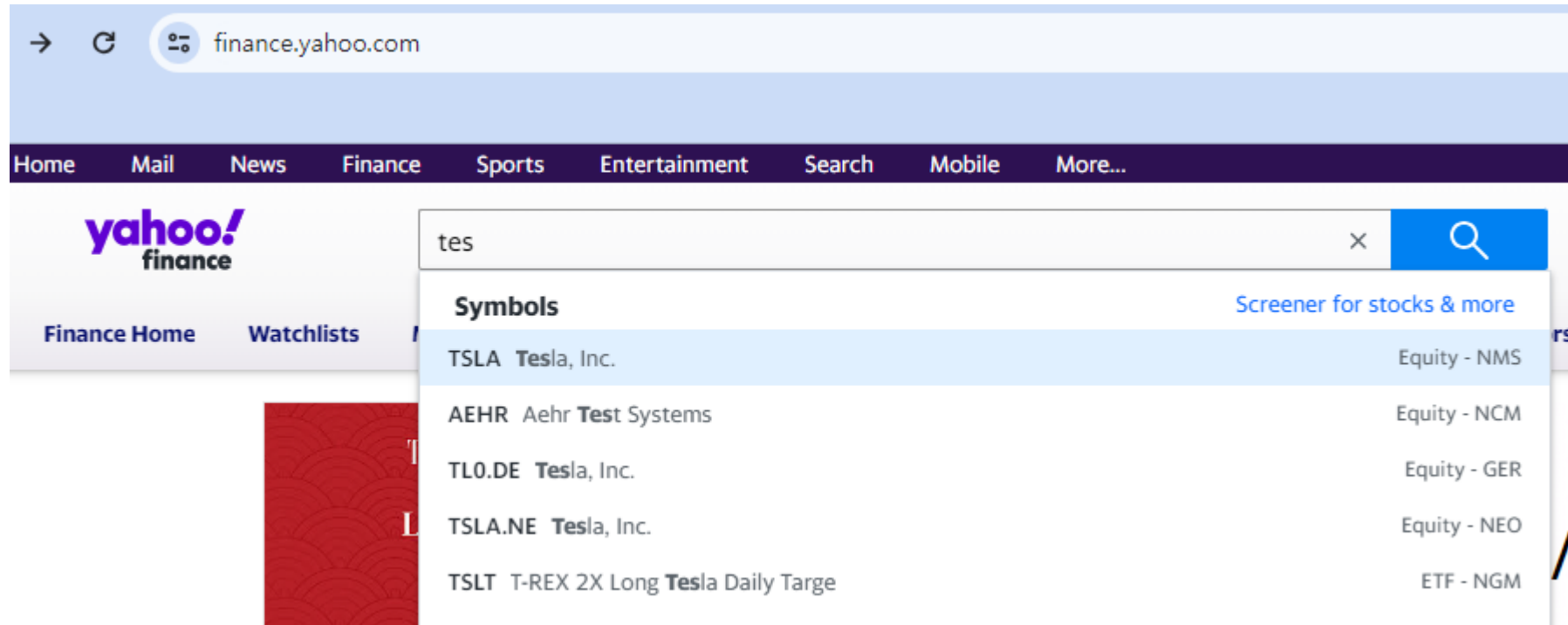
— Save historical data from a desktop browser

- 1 Go to Yahoo Finance.
- 2 Enter a quote into the search field.
- 3 Select a quote in the search results to view it.
- 4 Click **Historical Data**.
- 5 Select a Time Period, data to Show, and Frequency.
- 6 Quote data will refresh automatically.
- 7 To use the data offline, click **Download**.

+ Save historical data from a mobile browser

Web Crawling

- 실습
시계열 데이터 준비



Web Crawling


→ ↻ 🔍 finance.yahoo.com/quote/TSLA?.tsrc=fin-srch

Home Mail News Finance Sports Entertainment Search Mobile More...


yahoo!
finance


Search for news, symbols or companies 🔍


Finance Home Watchlists My Portfolio Markets News Videos Screeners Personal Finance Crypto Sectors


 참여 하얏트 호텔에서
최대 15% 할인 혜택을 받으세요
하얏트 월드 멤버십의 리워드 혜택을 누려 보세요


지금 예약하기
약관이 적용됩니다

S&P Futures
5,099.75
+2.00 (+0.04%) 

Dow Futures
39,147.00
+24.00 (+0.06%) 

Nasdaq Futures
18,039.75
-7.75 (-0.04%) 

Russell 2000 Futures
2,016.30
+0.30 (+0.01%) 

Crude Oil
78.33
-0.28 (-0.36%) 


Tesla, Inc. (TSLA)
NasdaqGS - NasdaqGS Real Time Price. Currency in USD

☆ Follow

197.41 +2.64 (+1.36%)
At close: 04:00PM EST

196.00 -1.41 (-0.71%)
After hours: 05:00PM EST

Summary Chart Conversations Statistics **Historical Data** Profile Financials Analysis Options Holders Sustainability

Previous Close 194.77 Market Cap 628.709B 1D 5D 1M 6M YTD 1Y 5Y Max  Full screen

Open 194.00 Beta (5Y Monthly) 2.43

NEW: Experience our best charts yet. ✕

Web Crawling

197.41

+2.64 (+1.36%)

At close: 04:00PM EST

196.00

-1.41 (-0.71%)

After hours: 08:00PM EST

Summary

Chart

Conversations

Statistics

Historical Data

Profile

Financials

Analysis

Options

Holders

Sustainabili

St

Adobe Stock

Adobe Stock에서 10개의 레이아웃
템플릿 및 목업을 무료로 이용하세요.

무료 체험하기

Time Period: Feb 22, 2023 - Feb 22, 2024

Show: Historical Prices

Frequency: Daily

Apply

Currency in USD

Date	Open	High	Low	Close*	Adj Close**	Volume
Feb 22, 2024	194.00	198.32	191.36	197.41	197.41	92,081,183
Feb 21, 2024	193.36	199.44	191.95	194.77	194.77	103,844,000
Feb 20, 2024	196.13	198.60	189.13	193.76	193.76	104,545,800
Feb 16, 2024	202.06	203.17	197.40	199.95	199.95	111,173,600

Download

THANK YOU