웹이란?

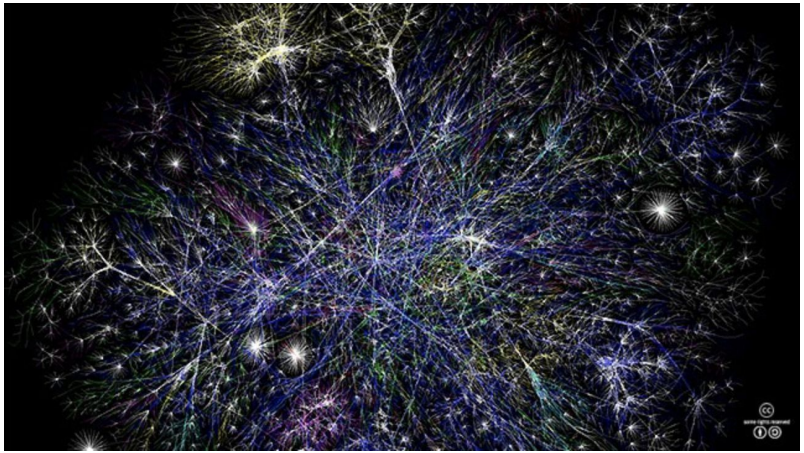
- 크롤링 -

인터넷

1969년 미국 국방부 산하의 고등연구계획국에서 아르파넷(ARPANET) 구축

1983년 다양한 목적으로 사용하고자 목적때문에 군사용 네트워크 기능을 분리

컴퓨터가 다양해지며 프로토콜(통신규약) 의 제정비가 필요해 TCP/IP 도입



주요 IP주소를 연결해구현한 인터넷지도

이미지출처
wikipedia.org

WorldWideWeb 의 도입

인터넷을 프로그램끼리 통신이나 메일전송 정도로 제한적으로 사용

팀 버너스리 박사는 문서속에 연결된 특정항목이 또 다른 문서로 연결되는 정보검색 시스템을 제시

WWW이라는 세계적인 정보공유공간 및 하이퍼텍스트(서로 연결된 문서), 웹사이트 제작언어인 HTML(HyperText Markup Language) 개념등장

WWW는 1991년 8월 6일 처음으로 서비스 시작

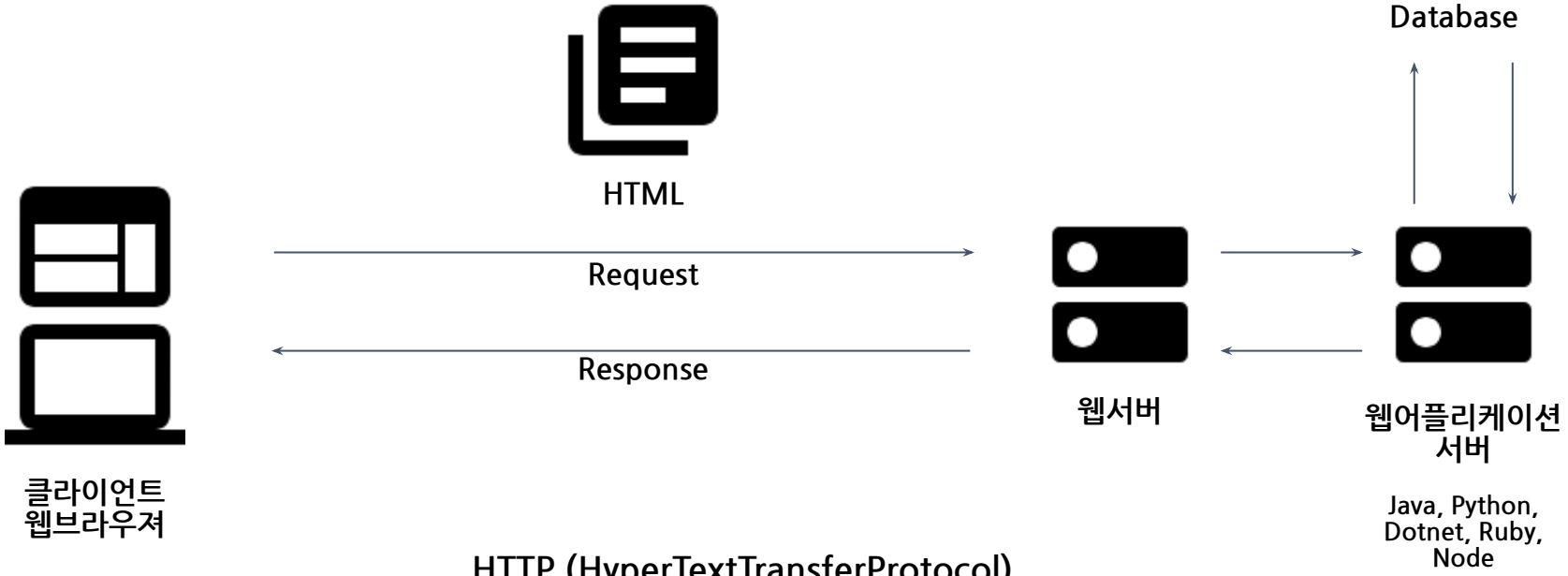
1993년 그래픽기반 웹 브라우저 모자이크가 등장

이미지출처
SDF 2013 사무국

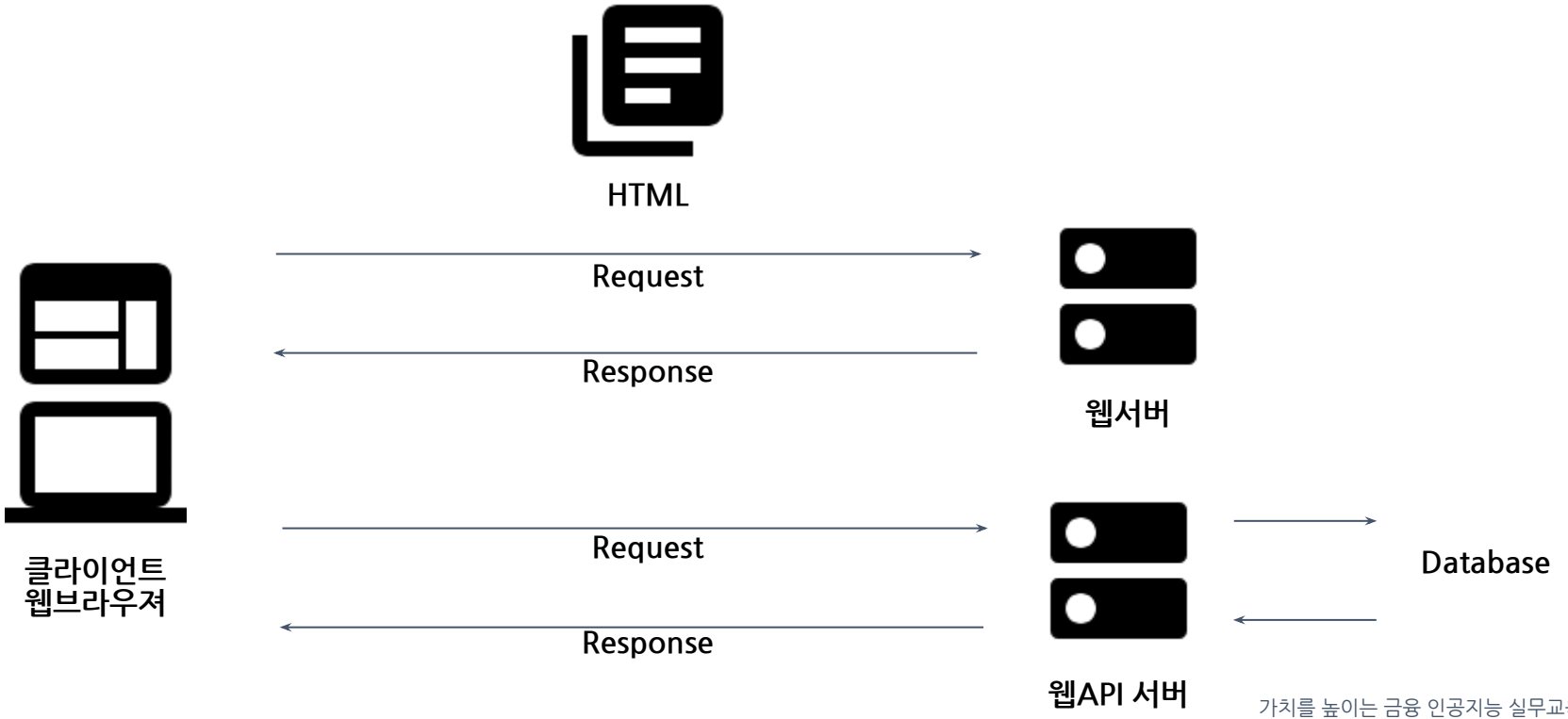
웹의 기본구조

HTTP (HyperTextTransferProtocol)
웹브라우저와 웹서버 사이의 전송규약

동적웹 구조 (Dynamic Web)



비동기방식



웹페이지 구조

웹페이지는 주로 HTML, CSS, Javascript 로 이루어져있습니다.

HTML - 주로 문서의 구조를 담당
CSS - 문서를 꾸며주는 역활 (폰트크기, 요소의 위치 등등)
Javascript - 문서내의 동작, 기능

https://www.w3schools.com/code/tryit.asp?filename=G3AL9ZB9RMAR

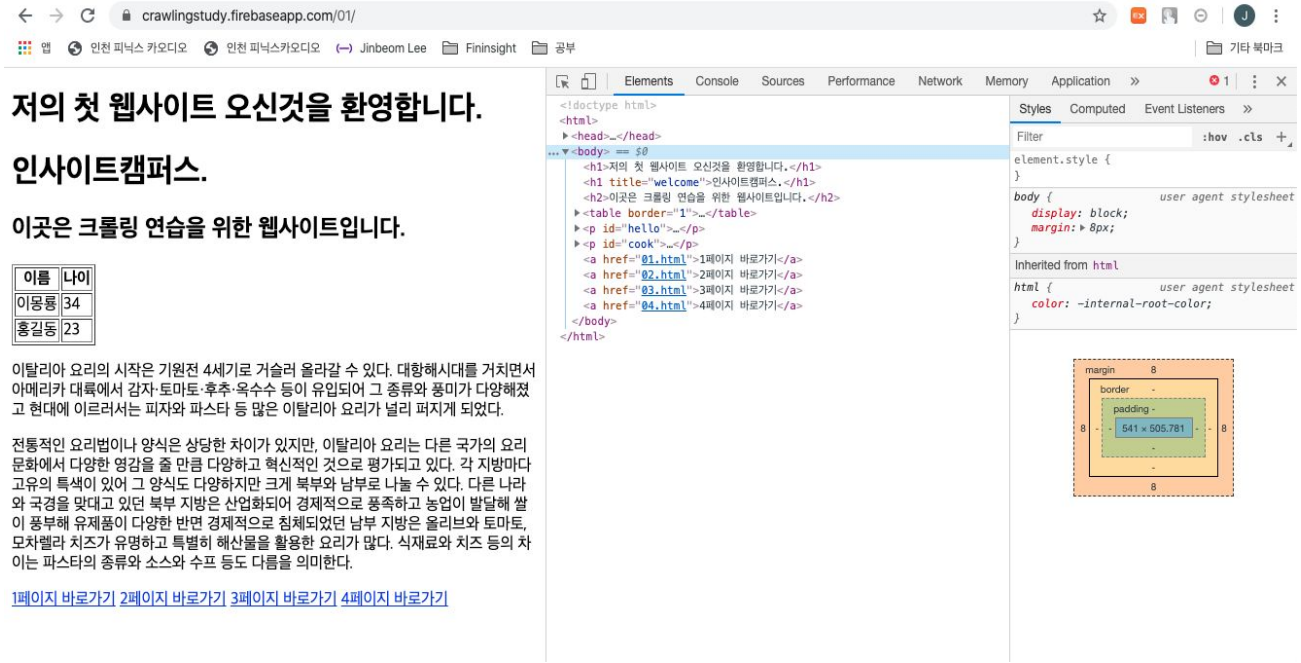
HTML 연습

w3schools.com

크롬 개발자도구

- 크롤링 -

크롬 개발자도구



F12 버튼을 눌러 활성화

연습사이트 URL

https://scrapying-study.firebaseapp.com/01

크롬 개발자도구

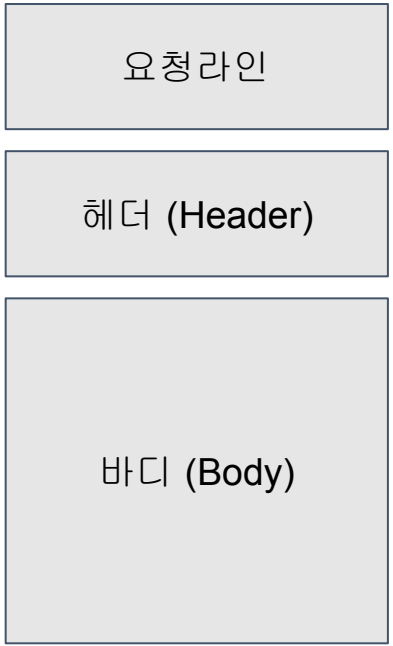
Elements - 화면에 보이는 HTML 문서의 구조
Console - 자바스크립트 결과 출력
Sources - 웹페이지에 사용된 소스코드들
Performance - 웹페이지 성능체크
Network - 웹페이지를 표시하는데 다운로드된 파일 및 요청한 서비스
Memory - 웹페이지 메모리사용률
Application - 웹페이지에서 사용하는 브라우져 스토리지정보 (storage, web DB, cookies 등)

연습사이트 URL
https://scrapying-study.firebaseapp.com/01

HTTP 구조

- 크롤링 -

HTTP 구조



General
Request URL - 요청 URL
Request Method - 요청방식 (Post, Get, Put, Delete)
Status Code - 응답코드

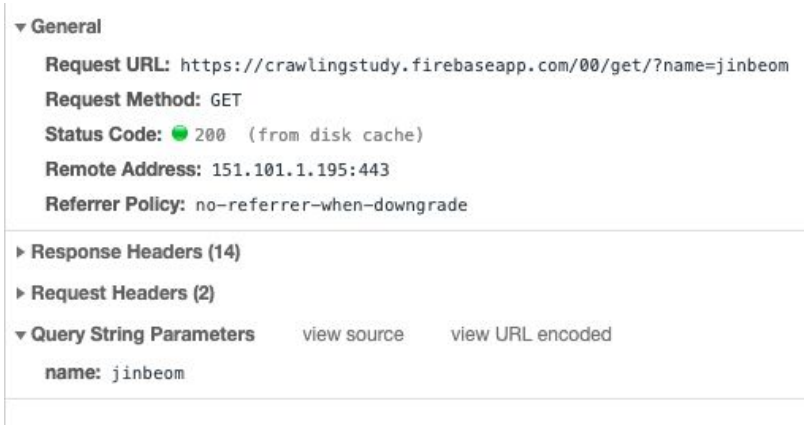
헤더
accept: application/json
user-agent: Mozilla/5.0
content-type: application/x-www-form-urlencoded; charset=UTF-8

바디
전송하고싶은 데이터

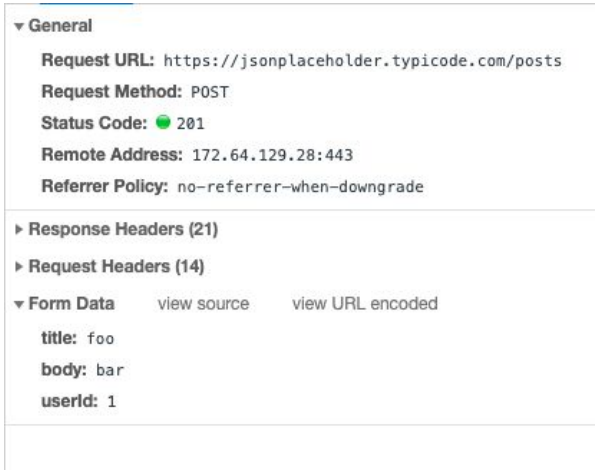
연습사이트 URL
https://scrapying-study.firebaseapp.com/00/get/?name=ronen
https://scrapying-study.firebaseapp.com/00/post/

Get, Post 구조차이

Get 은 요청시 Body 대신 URL 활용



Post, Put, Delete 는 Body 사용



연습사이트 URL

https://scrapying-study.firebaseapp.com/00/get/?name=ronen

https://scrapying-study.firebaseapp.com/00/post/

HTML

- 크롤링 -

나의 첫 웹사이트 (문서의 기본구조)

```
<!DOCTYPE html>
<html>
<head>
        <title>타이틀</title>
</head>
<body>
        <h1>저의 첫 웹사이트 오신것을 환영합니다.</h1>
</body>
</html>
```

https://www.w3schools.com/code/tryit.asp?filename=G3AL9ZB9RMAR
https://scrapying-study.firebaseapp.com/00/index.html

태그의 구조

<태그>내용</태그>

<태그 속성="속성값">내용</태그>

하이퍼텍스트

다음페이지
다음페이지

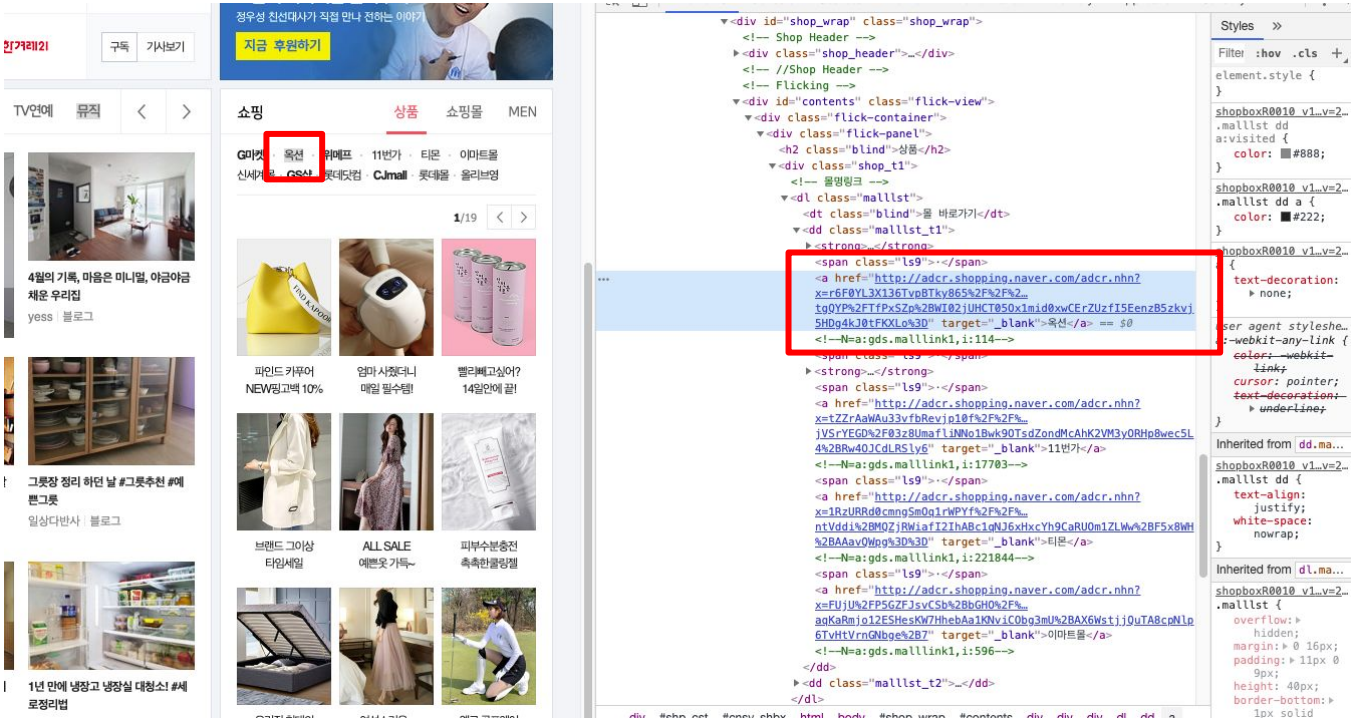
네이버로이동
네이버로이동

href : 이동할 주소나 페이지

target : 페이지를 어떻게 열것인가
_blank - 새창으로
_self - 현재페이지
_parent - 부모프레임
_top - 가장 상위프레임

https://www.w3schools.com/code/tryit.asp?filename=G3ALNE6T9O8J

하이퍼텍스트 사용예



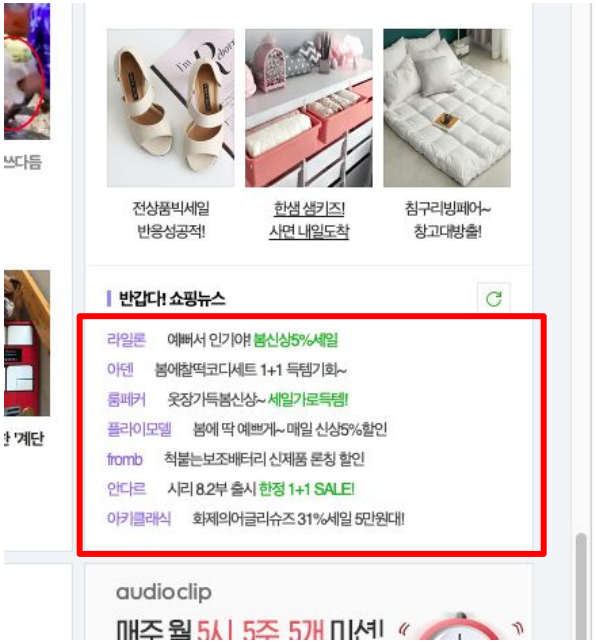
목록태그

```
<ul>
    <li>첫번째</li>
    <li>두번째</li>
    <li>세번째</li>
</ul>
```

```
<ol>
    <li>첫번째</li>
    <li>두번째</li>
    <li>세번째</li>
</ol>
```

https://www.w3schools.com/code/tryit.asp?filename=G3ALZC29GOUL

목록태그 사용예




```
▼<div class="lst_gift">
  <h4 class="blind">상품 리스트</h4>
 ▶<ul class="giftlst_u">…</ul>
  <!-- 배너 영역 시작 -->
  <!-- v1 -->
 ▼<div id="contents_productAdBanner">
  ▼<div class="gift_news">
   ▶<div class="gift_news_header">…</div>
   ▼<ul class="gift_news_list">
     ▶<li class="gift_news_item">…</li>
     ▶<li class="gift_news_item">…</li>
     ▶<li class="gift_news_item">…</li>
     ▶<li class="gift_news_item">…</li>
     ▶<li class="gift_news_item">…</li>
     ▶<li class="gift_news_item">…</li>
     ▶<li class="gift_news_item">…</li>
    </ul>
   </div>
   <!-- 배너 정보 -->
   <input class="bannerExposePage" type="hidden" value="1"
   name="bannerExposePage">
   <input class="bannerOrder" type="hidden" value="5"
   name="bannerOrder">
   <input class="bannerTotalCount" type="hidden" value="5"
   name="bannerTotalCount">
  </div>
  <!-- 배너 영역 종료 -->
 ▶<div class="page">…</div>
</div>
<input class="exposePage" type="hidden" value="1" name=
```

제목태그

<h1>제목1</h1>
<h2>제목2</h2>
<h3>제목3</h3>
<h4>제목4</h4>
<h5>제목5</h5>
<h6>제목6</h6>

https://www.w3schools.com/code/tryit.asp?filename=G3AMATLD8I2W

이미지태그

```
<img src="img.png" alt="이미지" title="이미지"/>
<a href="http://naver.com"><img src="img.png" alt="이미지" title="이미지"/></a>
```

https://www.w3schools.com/code/tryit.asp?filename=G3AMHDIVQK1V

이미지태그 사용예

단락태그

```
<p>
    이탈리아 요리의 시작은 기원전 4세기로 거슬러 올라갈 수 있다. 대항해시대를 거치면서 아메리카 대륙에서 감자·토마토·후추·옥수수 등이 유입되어 그 종류와 풍미가 다양해졌고 현대에 이르러서는 피자와 파스타 등 많은 이탈리아 요리가 널리 퍼지게 되었다.
</p>
<p>
    전통적인 요리법이나 양식은 상당한 차이가 있지만, 이탈리아 요리는 다른 국가의 요리 문화에서 다양한 영감을 줄 만큼 다양하고 혁신적인 것으로 평가되고 있다. 각 지방마다 고유의 특색이 있어 그 양식도 다양하지만 크게 북부와 남부로 나눌 수 있다. 다른 나라와 국경을 맞대고 있던 북부 지방은 산업화되어 경제적으로 풍족하고 농업이 발달해 쌀이 풍부해 유제품이 다양한 반면 경제적으로 침체되었던 남부 지방은 올리브와 토마토, 모차렐라 치즈가 유명하고 특별히 해산물을 활용한 요리가 많다. 식재료와 치즈 등의 차이는 파스타의 종류와 소스와 수프 등도 다름을 의미한다.
</p>
<p>
    카스틸리오네가 남긴 다량의 유화 작품과 서양식 투시법을 중국 전통 안료와 융합한 선법화 등은 당시 궁정을 중심으로 이루어진 중국과 서양의 회화 교류에 큰 영향을 끼쳤다. 중국의 전통화풍과 혼합된 서양화법은 궁정 화단을 중심으로 마진(馬晉)과 황족 출신 화가 부설재(溥雪齋) 등에게 계승되었고, 베이징을 중심으로 중국 근대 화단에까지 영향을 주었다.
</p>
```

https://www.w3schools.com/code/tryit.asp?filename=G3AMN2G62B5Q

텍스트를 꾸미는 인라인 태그들

```
<p>
    전통적인 요리법이나 <span>양식은 상당한 차이가 있지만</span>, 이탈리아 요리는 다른 국가의 요리 문화에서 다양한 영감을 줄 만큼 다양하고 혁신적인 것으로 평가되고 있다. 각 지방마다 고유의 특색이 있어 그 양식도 다양하지만 크게 북부와 남부로 나눌 수 있다. 다른 나라와 국경을 맞대고 있던 북부 지방은 산업화되어 경제적으로 풍족하고 농업이 발달해 쌀이 풍부해 유제품이 다양한 반면 경제적으로 침체되었던 남부 지방은 올리브와 <b>토마토</b>, <em>모차렐라 치즈</em>가 유명하고 특별히 해산물을 활용한 요리가 많다. 식재료와 치즈 등의 차이는 파스타의 종류와 소스와 수프 등도 다름을 의미한다.
</p>
```

https://www.w3schools.com/code/tryit.asp?filename=G3ANI5JYTQHU

그룹태그

```
<div>
        <a href="http://naver.com"><img src="img.png" alt="이미지" title="이미지"/></a>
        <p>
                이탈리아 요리의 시작은 기원전 4세기로 거슬러 올라갈 수 있다. 대항해시대를 거치면서 아메리카 대륙에서
                감자·토마토·후추·옥수수 등이 유입되어 그 종류와 풍미가 다양해졌고 현대에 이르러서는 피자와 파스타 등 많은 이탈리아
                요리가 널리 퍼지게 되었다.
        </p>
</div>
```

https://www.w3schools.com/code/tryit.asp?filename=G3AN7A9ENEG2

테이블 #1

```
<table>
  <tr>
    <td>표1</td>
    <td>표2</td>
  </tr>
  <tr>
    <td>표3</td>
    <td>표4</td>
  </tr>
</table>
```

```
<table>
  <thead>
    <tr>
      <th>이름</th>
      <th>나이</th>
    </tr>
  </thead>
  <tbody>
    <tr>
      <td>이몽룡</td>
      <td>34</td>
    </tr>
    <tr>
      <td>홍길동</td>
      <td>23</td>
    </tr>
  </tbody>
</table>
```

https://www.w3schools.com/code/tryit.asp?filename=G3AMZEOA81YH

테이블 #2

```
<table border="1">
  <tr>
    <td colspan="2">표1</td>
  </tr>
  <tr>
    <td>표3</td>
    <td>표4</td>
  </tr>
</table>
```

```
<table border="1">
  <tr>
    <td rowspan="2">표1</td>
    <td>표3</td>
  </tr>
  <tr>
    <td>표4</td>
  </tr>
</table>
```

https://www.w3schools.com/code/tryit.asp?filename=G3AMZEOA81YH

폼태그

```
<input type="text" value="값을 입력해주세요"/><br/>
<textarea cols="50" rows="5">값을 입력해주세요</textarea>
<select>
  <option value="1">첫번째</option>
  <option value="2">두번째</option>
  <option value="3">세번째</option>
</select><br/>
<input type="submit" value="전송"><br/>
<input type="button" value="버튼">
```

https://www.w3schools.com/code/tryit.asp?filename=G3ANCCUNSVOE

태그에 이름붙이기

ID (#)

태그에 고유한 이름
해당 이름은 한번만 사용해야함
CSS 에서 검색시 # 사용

CLASS (.)

태그에 이름을 붙여준다
여러번 중복해서 사용가능
CSS 에서 검색시 . 사용

```
<div id="menu_container">
        <div class="menu">메뉴1</div>
        <div class="menu">메뉴2</div>
        <div class="menu">메뉴3</div>
</div>
```

https://www.w3schools.com/code/tryit.asp?filename=G3ANV2EVYOKH

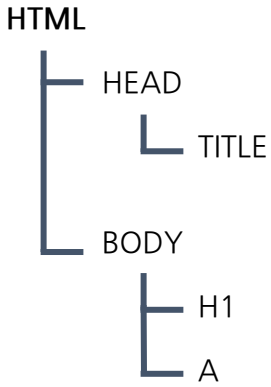
HTML 기본구조

TREE 구조

```
<!DOCTYPE html>

<html>
        <head>
                <title>타이틀</title>
        </head>

        <body>
                <h1>저의 첫 웹사이트 오신것을 환영합니다.</h1>
                <a href="http://naver.com">네이버</a>
        </body>
</html>
```



HTML 태그는 **HEAD, BODY** 자식태그를 가지고있다.
HTML 태그는 **HEAD, BODY, TITLE, H1, A** 자손태그를 가지고있다.

HTML 기본구조

태그란?

```
<!DOCTYPE html>

<html>
        <head>
                <title>타이틀</title>
        </head>

        <body>
                <h1>저의 첫 웹사이트 오신것을 환영합니다.</h1>
                <a href=”http://naver.com”>네이버</a>
        </body>
</html>
```

<태그명 옵션="옵션값"> 내용 </태그명>

네이버

https://developer.mozilla.org/ko/docs/Web/HTML/Element

HTML 기본 크롤링

- 크롤링 -

request 객체

HTTP 요청을 보내는 모듈 (웹사이트 접속)

```
import requests

URL =
'https://scrapying-study.firebaseapp.com/01/'

response = requests.get(URL)

print(response.status_code)
print(response.text)
```

연습사이트 URL
https://scrapying-study.firebaseapp.com/01/

소스코드
HTML기본크롤링01.ipynb

request 객체

Header
추가

```
import requests

URL = 'https://scrapying-study.firebaseapp.com/01/'

headers = {'Content-Type': 'application/json; charset=utf-8'}
response = requests.get(URL, headers=headers)


print(response.status_code)
print(response.text)
```

연습사이트 URL
https://scrapying-study.firebaseapp.com/01/

소스코드
HTML기본크롤링01.ipynb

request 객체

Get 방식
데이터전송

```
import requests

URL = 'https://scrapying-study.firebaseapp.com/01/'

data = {'name': 'ronen'}
response = requests.get(URL, data=data)

print(response.status_code)
print(response.text)
```

연습사이트 URL
https://scrapying-study.firebaseapp.com/01/

소스코드
HTML기본크롤링01.ipynb

BeautifulSoup

HTML 에서 손쉽게 원하는 데이터를 가져올수있도록 지원

pip install beautifulsoup4

```
import requests
from bs4 import BeautifulSoup

URL = 'https://scrapying-study.firebaseapp.com/01/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.find("title")
print(result)
print(result.text)
```

연습사이트 URL

https://scrapying-study.firebaseapp.com/01/

소스코드

HTML기본크롤링01.ipynb

BeautifulSoup

```
URL = 'https://scrapying-study.firebaseapp.com/01/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.find("title")
print(result)
print(result.text)
```

soup.find(“p”) - 가장 상단에 있는태그 하나만
soup.find_all(“p”, limit=2) - 일치하는 모든태그 (limit : 가져올개수제한, 생략가능)

soup.find(“th”,“tablehead”) - 옵션값이 class가 tablehead 인것
soup.find(“th”, class_=“tablehead”) - 옵션값이 class가 tablehead 인것
soup.find(“th”, attrs={“class”:“tablehead”}) - 옵션값 class가 tablehead 인것 (옵션명 변경가능)
soup.find(“h1”, attrs={“title”:“welcome”}) - 옵션값 title이 welcome 인것 (옵션명 변경가능)
soup.find(id=”hello”) - 옵션값이 id가 hello 인것

연습사이트 URL
https://scrapying-study.firebaseapp.com/01/

소스코드
HTML기본크롤링01.ipynb

BeautifulSoup

```
URL = 'https://scrapying-study.firebaseapp.com/01/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
results = soup.find("a")
print(results.text)
print(results.attrs["href"])
```

```
result = soup.find(“table”)
result2 = result.find(“tbody”)
```

결과값에서 다시 검색가능

result.**text** - 태그내에 내용만 추출
result.**attrs**[“옵션명"] - 태그내에 옵션을 추출

연습사이트 URL
https://scrapying-study.firebaseapp.com/01/

소스코드
HTML기본크롤링01.ipynb

실습01

HTML 문서내에 ID가 cook 인 태그내의 내용을 출력해주세요

결과 :

전통적인 요리법이나 양식은 상당한 차이가 있지만, 이탈리아 요리는 다른 국가의 요리 문화에서 다양한 영감을 줄 만큼 다양하고 혁신적인 것으로 평가되고 있다. 각 지방마다 고유의 특색이 있어 그 양식도 다양하지만 크게 북부와 남부로 나눌 수 있다. 다른 나라와 국경을 맞대고 있던 북부 지방은 산업화되어 경제적으로 풍족하고 농업이 발달해 쌀이 풍부해 유제품이 다양한 반면 경제적으로 침체되었던 남부 지방은 올리브와 토마토, 모차렐라 치즈가 유명하고 특별히 해산물을 활용한 요리가 많다. 식재료와 치즈 등의 차이는 파스타의 종류와 소스와 수프 등도 다름을 의미한다.

연습사이트 URL
https://scrapying-study.firebaseapp.com/01/

소스코드
HTML기본크롤링01.ipynb

실습02

HTML 문서내에 TABLE 내에 th 와 td에 있는 값들을 크롤링해 아래와같은 딕셔너리 형태를 만들어보세요

결과 :

[{'이름': '이몽룡', '나이': '34'}, {'이름': '홍길동', '나이': '23'}]

연습사이트 URL
https://scrapying-study.firebaseapp.com/01/

소스코드
HTML기본크롤링01.ipynb

실습03

HTML 문서내에 모든 A 태그에 링크된 페이지에 있는 내용을 읽어 출력해주세요

결과 :

크롤링 연습사이트 01-1 페이지입니다.
크롤링 연습사이트 01-2 페이지입니다.
크롤링 연습사이트 01-3 페이지입니다.
크롤링 연습사이트 01-4 페이지입니다.

연습사이트 URL
https://scrapying-study.firebaseapp.com/01/

소스코드
HTML기본크롤링01.ipynb

CSS 셀렉터를 활용한 크롤링

- 크롤링 -

CSS란?

Cascading Style Sheets

HTML 요소들이 어떻게 보이는가를 정의하는 언어
1994년 10월 10일, 하콤 비움 리가 처음 제안하였다.
최신버전은 CSS3

내용출처
https://ko.wikipedia.org/

CSS 사용방법

1. 태그내에 지정

```
<div style="border:1px solid red;font-size:20px">
안녕하세요 DIV 박스입니다.
</div>
```

CSS 사용방법

2. 내부 스타일시트

```
<head>
<style type="text/css">

body { font-size:9pt; }
.content {
    border:1px solid red;
    font-size:20px;
}

</style>
</head>
<body>
        <div class="content">
                안녕하세요 DIV 박스입니다.
        </div>
</body>
```

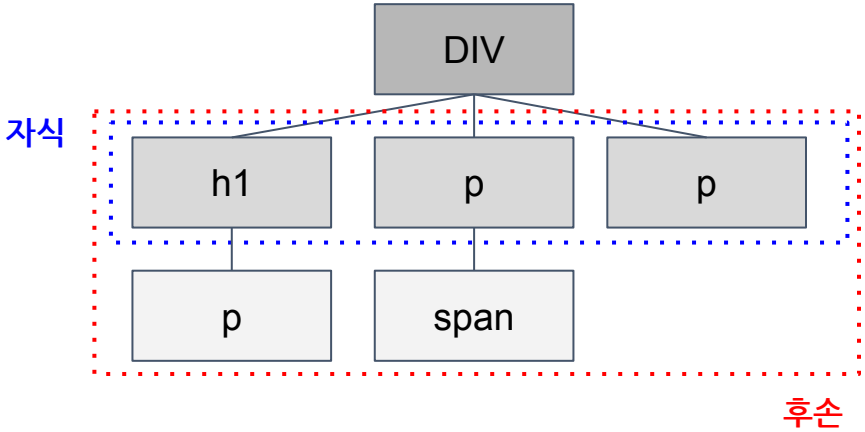
CSS 사용방법

3. 외부 스타일시트

```
<head>
<link rel="stylesheet" type="text/css" href="mystyle.css">
</head>
<body>
        <div class=”content”>
                안녕하세요 DIV 박스입니다.
        </div>
</body>
```

CSS 셀렉터

1. **태그 셀렉터 (HTML 태그를 활용)** - p, h1, h2
2. **ID 셀렉터 (ID 속성을 활용)** - #title
3. **class 셀릭터 (class 속성을 활용)** - .content
4. **속성 셀렉터 (태그내 속성을 활용)** - a[href], a[target="_blank"]
   셀렉터[어트리뷰트~="값"] - 해당 단어를 포함
   셀렉터[어트리뷰트^="값"] - 해당 값으로 시작
   셀렉터[어트리뷰트$="값"] - 해당 값으로 끝나는
   셀렉터[어트리뷰트*="값"] - 해당 값을 포함하는
5. **후손셀렉터 (해당 태그 내에 포함되는 태그)** - div p
6. **자식셀렉터 (해당 태그 바로 안에 포함되는 태그)** - div > p



BeautifulSoup

```
import requests
from bs4 import BeautifulSoup

URL = 'https://scrapying-study.firebaseapp.com/02/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select("#title")
print(result[0])
print(result[0].text)
```

soup.select_one(셀렉터) - 셀렉터에 일치하는 하나의 태그만
soup.select(셀렉터) - 셀렉터에 일치하는 모든태그

연습사이트 URL
https://scrapying-study.firebaseapp.com/02/

소스코드
CSS셀렉터01.ipynb

BeautifulSoup

태그 셀렉터 (HTML 태그를 활용) - p, h1, h2

```
URL = 'https://scrapying-study.firebaseapp.com/02/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select("title")
print(result)
```

결과 :

[<title>크롤링 연습사이트 02</title>]

연습사이트 URL

https://scrapying-study.firebaseapp.com/02/

소스코드

CSS셀렉터01.ipynb

BeautifulSoup

ID 셀렉터 (ID 속성을 활용) - #title

```
URL = 'https://scrapying-study.firebaseapp.com/02/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select("#title")
print(result)
```

결과 :

```
[<div class="bold" id="title">
    안녕하세요
</div>]
```

연습사이트 URL

https://scrapying-study.firebaseapp.com/02/

소스코드

CSS셀렉터01.ipynb

BeautifulSoup

class 셀릭터 (class 속성을 활용) - .content

```
URL = 'https://scrapying-study.firebaseapp.com/02/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select(".bold")
print(result)
```

결과 :

```
[<li class="blue">두번째리스트</li>,
<li class="blue">세번째리스트</li>,
<p class="blue">두근거리고 <span>익숙한 듯 편안해</span>
</p>,
<p class="blue">오래된 친구같아</p>]
```

연습사이트 URL
https://scrapying-study.firebaseapp.com/02/

소스코드
CSS셀렉터01.ipynb

BeautifulSoup

속성 셀렉터 (태그내 속성을 활용) - a[href], a[target="_blank"]

셀렉터[어트리뷰트~=”값”] - 해당 단어를 포함
셀렉터[어트리뷰트^=”값”] - 해당 값으로 시작
셀렉터[어트리뷰트$=”값”] - 해당 값으로 끝나는
셀렉터[어트리뷰트*=”값”] - 해당 값을 포함하는

```
URL = 'https://scrapying-study.firebaseapp.com/02/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select('a[target="_blank"]')
print(result)
```

결과 :

[네이버]

연습사이트 URL

https://scrapying-study.firebaseapp.com/02/

소스코드

CSS셀렉터01.ipynb

BeautifulSoup

속성 셀렉터 (태그내 속성을 활용) - a[href], a[target="_blank"]

셀렉터[어트리뷰트~=”값”] - 해당 단어를 포함
셀렉터[어트리뷰트^=”값”] - 해당 값으로 시작
셀렉터[어트리뷰트$=”값”] - 해당 값으로 끝나는
셀렉터[어트리뷰트*=”값”] - 해당 값을 포함하는

```
URL = 'https://scrapying-study.firebaseapp.com/02/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select('a[href$=".com"]')
print(result)
```

결과 :

[네이버,
구글]

연습사이트 URL

https://scrapying-study.firebaseapp.com/02/

소스코드

CSS셀렉터01.ipynb

BeautifulSoup

후손셀렉터 (해당 태그 내에 포함되는 태그) - div p

```
URL = 'https://scrapying-study.firebaseapp.com/02/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select('div#winter p')
print(result)
```

결과 :

```
[<p>온세상이 떨릴듯</p>,
 <p class="blue">두근거리고 <span>익숙한 듯 편안해</span></p>,
 <p>네가 느껴져</p>,
 <p class="blue">오래된 친구같아</p>]
```

연습사이트 URL
https://scrapying-study.firebaseapp.com/02/

소스코드
CSS셀렉터01.ipynb

BeautifulSoup

자식셀렉터 (해당 태그 바로 안에 포함되는 태그) - div > p

```
URL = 'https://scrapying-study.firebaseapp.com/02/'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select('div#winter > p')
print(result)
```

결과 :

[온세상이 떨릴듯 ,
두근거리고 익숙한 듯 편안해]

연습사이트 URL
https://scrapying-study.firebaseapp.com/02/

소스코드
CSS셀렉터01.ipynb

실습01

사이트내에 인기검색종목과 주요해외지수를 각각 크롤링하여 종목명과 주가지수를 아래와같이 리스트로 정리해주세요

결과 :

[['써니전자', '5,000'], ['삼성전자', '55,200'], ['안랩', '81,000'], ['케이엠더블..', '57,300'], ['피피아이', '12,600'], ['KT&G', '92,500'], ['삼성전자우', '45,600'], ['대양금속', '10,550'], ['SK하이닉스', '94,700'], ['SK텔레콤', '234,000']]

[['다우산업', '28,647.43'], ['나스닥', '9,015.03'], ['홍콩H', '11,320.56'], ['상해종합', '3,085.20'], ['니케이225', '23,656.62']]

연습사이트 URL
https://scrapying-study.firebaseapp.com/03/

소스코드
CSS셀렉터01.ipynb

실습02

사이트내에 인기검색종목과 주요해외지수를 각각 크롤링하여 종목명과 상한, 하한여부를 아래와같이 리스트로 정리해주세요

결과 :

[['써니전자', '상한'], ['삼성전자', '하한'], ['안랩', '상한'], ['케이엠더블..', '상한'], ['피피아이', '상한'], ['KT&G', '하한'], ['삼성전자우', '상한'], ['대양금속', '하한'], ['SK하이닉스', '상한'], ['SK텔레콤', '하한']]

[['다우산업', '상한'], ['나스닥', '상한'], ['홍콩H', '상한'], ['상해종합', '상한'], ['니케이225', '하한']]

연습사이트 URL
https://scrapying-study.firebaseapp.com/03/

소스코드
CSS셀렉터01.ipynb

실습03

사이트내에 인기검색종목과 주요해외지수를 각각 상한가인 종목만 크롤링하여 종목명과 주가지수를 아래와같이 리스트로 정리해주세요

결과 :

[['써니전자', '5,000'], ['안랩', '81,000'], ['케이엠더블..', '57,300'], ['피피아이', '12,600'], ['삼성전자우', '45,600'], ['SK하이닉스', '94,700']]

[['다우산업', '28,647.43'], ['나스닥', '9,015.03'], ['홍콩H', '11,320.56'], ['상해종합', '3,085.20']]

연습사이트 URL
https://scrapying-study.firebaseapp.com/03/

소스코드
CSS셀렉터01.ipynb

실습04

분양중인 아파트 정보를 크롤링하여 아래와 같이 딕셔너리 형태로 정리해주세요

결과 :

[{'이름': 'H하우스장위', '분양가': '16000', '유형': '아파트', '분양유형': '일반민간임대', '세대수': '분양 134세대', '평형': '45㎡~65㎡'},
{'이름': '고덕리엔파크2단지 장기전세', '분양가': '38400', '유형': '아파트', '분양유형': '장기전세주택', '세대수': '분양 1세대', '평형': '149㎡'},
{'이름': '신정이펜하우스3단지 장기전세', '분양가': '39040', '유형': '아파트', '분양유형': '장기전세주택', '세대수': '분양 1세대', '평형': '148㎡'},
{'이름': '천왕이펜하우스2단지 장기전세', '분양가': '38240', '유형': '아파트', '분양유형': '장기전세주택', '세대수': '분양 1세대', '평형': '142㎡'},
{'이름': '송파파크데일2단지 장기전세', '분양가': '45600', '유형': '아파트', '분양유형': '장기전세주택', '세대수': '분양 1세대', '평형': '150㎡'}]

연습사이트 URL
https://scrapying-study.firebaseapp.com/03/

소스코드
CSS셀렉터01.ipynb

실전01
- 크롤링 -

실전01

네이버 영화랭킹

```
import requests
from bs4 import BeautifulSoup

URL = 'https://movie.naver.com/movie/sdb/rank/rmovie.nhn'
response = requests.get(URL)
soup = BeautifulSoup(response.text, "html.parser")
result = soup.select(".list_ranking tr > td.title a")

for i, v in enumerate(result):
    print(i+1, v.text)
```

소스코드

실전01.ipynb

실습01

네이버 주가 크롤링

https://finance.naver.com/sise/sise_quant.nhn

코스피 코스닥

N.	종목명	현재가	전일비	등락률	거래량	거래대금	매수호가	매도호가	시가총액	PER	ROE
1	KODEX 200선물인버스2X	5,270	▼ 400	-7.05%	228,349,079	1,217,092	5,270	5,275	19,293	N/A	N/A
2	KODEX 레버리지	12,680	▲ 845	+7.14%	120,363,769	1,511,682	12,675	12,680	33,830	N/A	N/A
3	삼성중공업	6,970	▲ 1,080	+18.34%	110,900,770	728,045	6,970	6,980	43,911	-3.06	-21.88
4	KODEX 인버스	6,095	▼ 225	-3.56%	55,716,039	341,589	6,095	6,100	10,020	N/A	N/A
5	KODEX 코스닥150선물인버스	6,220	▲ 20	+0.32%	50,528,871	311,701	6,220	6,225	4,715	N/A	N/A
6	삼성전자	54,500	▲ 3,100	+6.03%	48,787,603	2,629,934	54,500	54,600	3,253,531	17.39	8.69
7	두산인프라코어	6,250	▲ 710	+12.82%	43,248,149	272,059	6,250	6,260	13,010	6.37	11.59
8	KODEX 코스닥150 레버리지	10,010	▼ 80	-0.79%	37,575,643	384,722	10,005	10,010	12,172	N/A	N/A
9	미래산업	83	▲ 2	+2.47%	32,447,838	2,668	82	83	700	-16.60	-12.97
10	문배철강	3,060	▲ 140	+4.79%	29,730,396	96,750	3,060	3,065	627	37.32	2.14
11	삼성 레버리지 WTI원유 선물 ETN	455	▲ 45	+10.98%	29,475,904	13,511	455	460	933	N/A	N/A
12	태평양물산	2,220	▲ 90	+4.23%	26,576,427	64,174	2,220	2,230	1,108	2,220.00	6.99
13	쌍방울	1,015	▼ 5	-0.49%	22,618,490	24,258	1,015	1,020	1,414	-2.08	-19.68
14	이아이디	235	▼ 7	-2.89%	20,921,183	4,979	234	235	1,289	19.58	0.44
15	신한 레버리지 WTI원유 선물 ETN(H)	360	▲ 40	+12.50%	20,820,339	7,456	360	365	1,368	N/A	N/A
16	KODEX WTI원유선물(H)	6,095	▲ 325	+5.63%	18,983,196	114,853	6,090	6,095	14,765	N/A	N/A
17	파미셀	22,000	▼ 650	-2.87%	15,429,756	348,822	22,000	22,050	13,191	309.86	8.78
18	한화생명	1,620	▲ 60	+3.85%	15,158,508	24,384	1,620	1,625	14,070	13.97	0.51
19	화인베스틸	2,560	▲ 245	+10.58%	14,705,935	38,214	2,555	2,560	756	-10.08	-7.77
20	팬오션	3,860	▲ 235	+6.48%	14,606,724	56,803	3,855	3,860	20,634	14.46	5.49
21	신성통상	1,450	▼ 50	-3.33%	13,713,658	20,196	1,445	1,450	2,084	-55.77	2.26
22	에이프로젠 KIC	2,960	▼ 240	-7.50%	13,583,125	40,907	2,960	2,965	4,648	105.71	4.70
23	모나미	4,645	▼ 115	-2.42%	12,891,194	61,134	4,640	4,645	878	-30.36	-2.39
24	마니커	978	▼ 16	-1.61%	11,730,931	11,556	977	978	1,550	-4.51	-18.14
25	SK하이닉스	88,700	▲ 5,400	+6.48%	11,589,791	1,007,398	88,600	88,700	645,738	41.43	4.25

1. 품목명과 현재가를 크롤링해주세요.

결과
1 KODEX 200선물인버스2X 5,270
2 KODEX 레버리지 12,680
3 삼성중공업 6,970
4 KODEX 인버스 6,095
5 KODEX 코스닥150선물인버스 6,220
6 삼성전자 54,500
7 두산인프라코어 6,250
8 KODEX 코스닥150 레버리지 10,010
9 미래산업 83
10 문배철강 3,060
.
.
.

2. 전일대비 상승한 항목만 품목명, 현재가, 전일비를 크롤링해주세요.

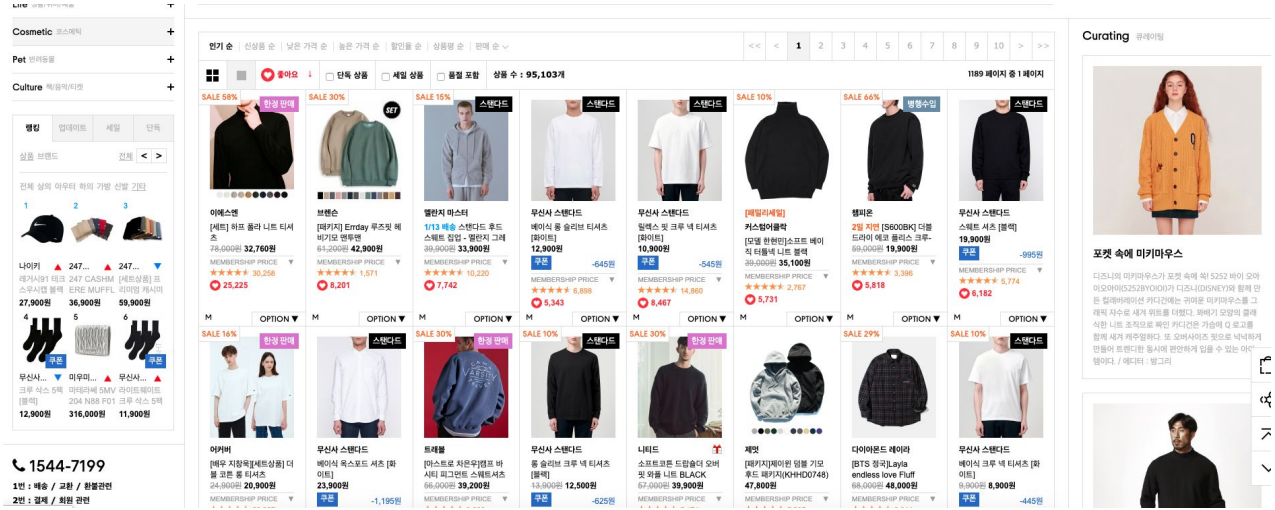
결과
1 KODEX 200선물인버스2X 5,270 400
4 KODEX 인버스 6,095 225
8 KODEX 코스닥150 레버리지 10,010 80
13 쌍방울 1,015 5
14 이아이디 235 7
17 파미셀 22,000 650
21 신성통상 1,450 50
22 에이프로젠 KIC 2,960 240
23 모나미 4,645 115
24 마니커 978 16
30 삼성 인버스 2X WTI원유 선물 ETN 2,765 365
33 엔케이 1,165 30
.
.
.

소스코드
실전01.ipynb

실습02

무신사 쇼핑몰의 상의(TOP) 카테고리 첫페이지의 제품들의 브랜드명, 제품명, 가격을 아래와같이 크롤링해주세요.
https://store.musinsa.com/app/items/lists/001

성공하신분들은 총 10페이지를 크롤링해주세요.



결과:

브랜드 : 86로드
제품명 : FUTURE RACING T-SHIRTS BLUE
가격 : 15,750원

브랜드 : 86로드
제품명 : FUTURE RACING T-SHIRTS GRAY
가격 : 15,750원

브랜드 : 리스펙트
제품명 : [리스펙트 X 도리] 바나나 반팔티 (블랙)
가격 : 19,000원

브랜드 : 리스펙트
제품명 : [리스펙트 X 도리] 바나나 반팔티 (화이트)
가격 : 19,000원

브랜드 : 왕관
제품명 : 피그먼트 오버핏 티셔츠 (Charcoal)
가격 : 48,000원

브랜드 : 왕관
제품명 : 피그먼트 오버핏 티셔츠 (Pink)
가격 : 48,000원

브랜드 : 집시
제품명 : 오버사이즈 쿨링베이스 티셔츠 -다크 올리브-
가격 : 27,900원

.
.
.

소스코드
실전01.ipynb

실습03

네이버뉴스 크롤링

아래 URL 의 뉴시스 매체의 어제일자 모든 기사를 아래와같이 제목과 본문을 분류하여 크롤링해주세요

https://news.naver.com/main/list.nhn?mode=LPOD&mid=sec&oid=003

티웨이항공 탑승권 제시하면 베어스타운 할인 혜택

티웨이항공, 겨울 시즌 맞이 스키장 제휴 프로모션 실시[서울=뉴시스] 오동현 기자 = 티웨이항공과 베어스타운 리조트가 겨울 시즌 스키장 이용객들을 대상으로 제휴 프로모션을 실시한다고 5일 밝혔다.이번 '스키타러 티웨이랑 갈곰' 프로모션은 티웨이항공 탑승권을 소지한 고객 대상 현장 할인 혜택 제공과 SNS 참여 경품 이벤트로 진행된다.국내선과 국제선 상관없이 6개월 이내 사용한 티웨이항공 탑승권을 제시하면 베어스타운 리조트에서 리프트 40%, 장비 렌탈 40%, 의류 렌탈 20%, 눈썰매장 40%의 할인이 제공된다. 객실도 주중 55%, 주말 45% 할인 제공되며 유선으로 예약 가능하다. 현장 할인 프로모션은 베어스타운 스키장 시즌이 운영되는 오는 3월 8일까지 진행된다.베어스타운 리조트 내 SNS 인증 이벤트도 진행된다. 리조트 내에 설치된 티웨이항공 유니폼을 입은 곰 조형물과 사진을 찍은 후 개인 SNS(인스타그램) 계정에 필수 해시태그(스키타고대만갈곰, 티웨이랑갈곰, 베어스타운, 티웨이항공)와 함께 업로드하면 참여된다.추첨을 통해 1등 3명에게 티웨이항공 대만 노선 왕복 항공권 2매씩, 2등 5명에게 베어스타운 콘도 숙박권 1매씩, 3등 10명에게는 티웨이항공 모형 비행기 1대씩 푸짐한 경품을 증정한다. SNS 인증 이벤트 기간은 1월 7일부터 2월 29일까지다. 티웨이항공 관계자는 "고객들을 위해 겨울 시즌 즐길 수 있는 스키장과 함께하는 프로모션을 마련했다"라며 "스키장에서 할인도 받고 SNS 인증 이벤트 참여로 대만 여행의 행운도 가져가시기 바란다"고 전했다.☞공감언론 뉴시스 odong85@newsis.com▶ K-Artprice 모바일 오픈! 미술작품 가격을 공개합니다▶ 뉴시스 채널 구독하고 에어팟 프로 받아가세요▶ 뉴시스 빅데이터 MSI 주가시세표 바로가기

울산 마트 주차장 차량서 불...인명피해 없어

[울산=뉴시스] 박수지 기자 = 5일 오후 8시 5분께 울산시 북구 한 마트 주차장에 주차돼 있던 차량에서 불이나 소방당국이 진화작업을 벌이고 있다.2019.01.05.(사진=울산소방본부 제공)[울산=뉴시스]박수지 기자 = 5일 오후 8시 5분께 울산시 북구 한 마트 주차장에 주차돼 있던 차량에서 화재가 발생했다. 차량 엔진룸에서 시작된 불은 약 200만원의 재산피해(소방서 추산)를 내고 10여분 만에 진화됐다. 다행히 인명피해는 없었다. 경찰과 소방당국은 정확한 화재 원인을 조사 중이다.☞공감언론 뉴시스 parksj@newsis.com▶ K-Artprice 모바일 오픈! 미술작품 가격을 .

.
.
.
.
.

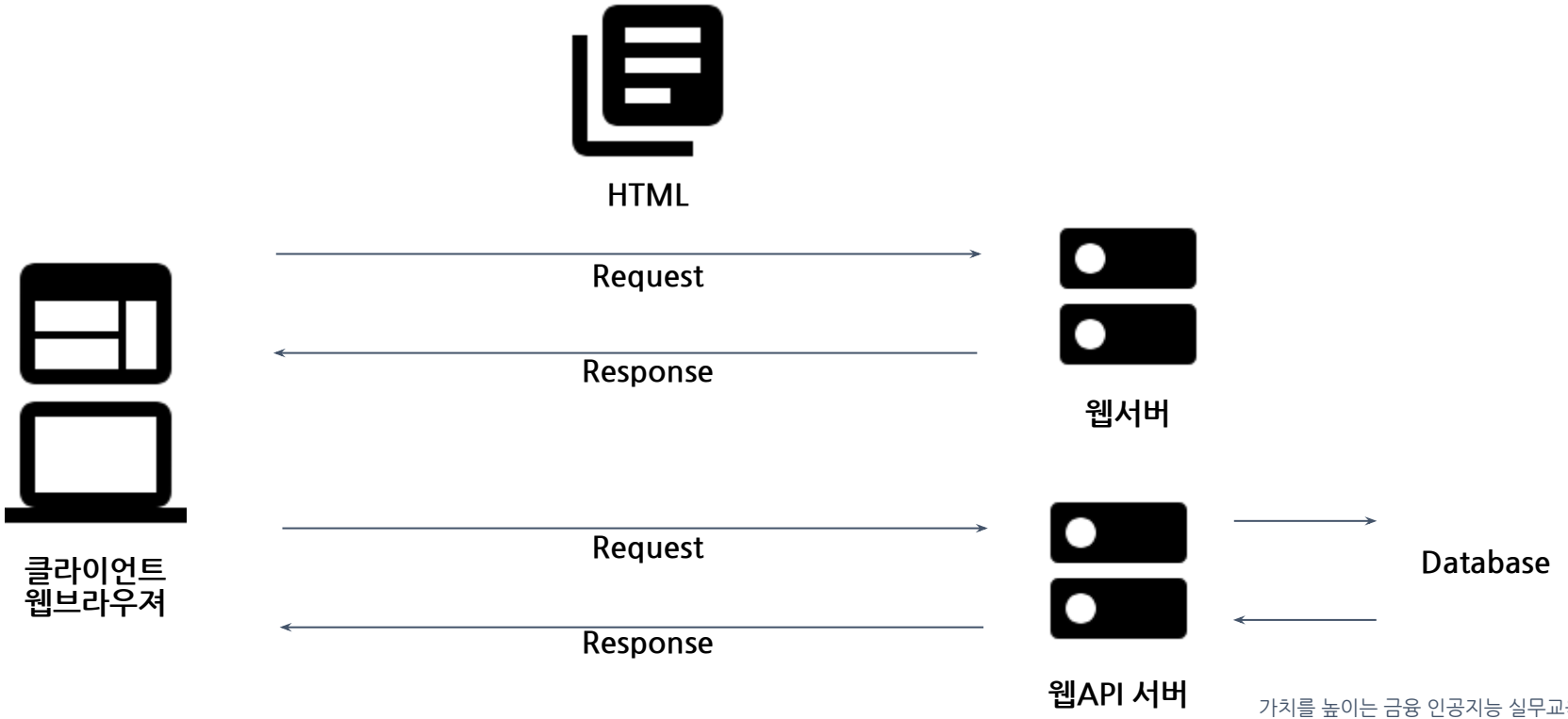
소스코드

실전01.ipynb

비동기방식 데이터 크롤링

- 크롤링 -

비동기방식



비동기방식

```
import requests
from bs4 import BeautifulSoup
response = requests.get("https://scrapying-study.firebaseapp.com/04/")
print(response.text)
```

결과 :

```
<!doctype html>
<html>
<head>
    <meta charset="utf-8">
    <script src="https://code.jquery.com/jquery-3.4.1.js"
integrity="sha256-WpOohJOqMqqyKL9FccASB9O0KwACQJpFTUBLTYOVvVU="
crossorigin="anonymous"></script>
    <script>
        ...
    </script>
    <title>Demo</title>
</head>
<body>
    <div id="post" style="width:500px;">
    </div>
</body>
</html>
```

? 데이터가 없음

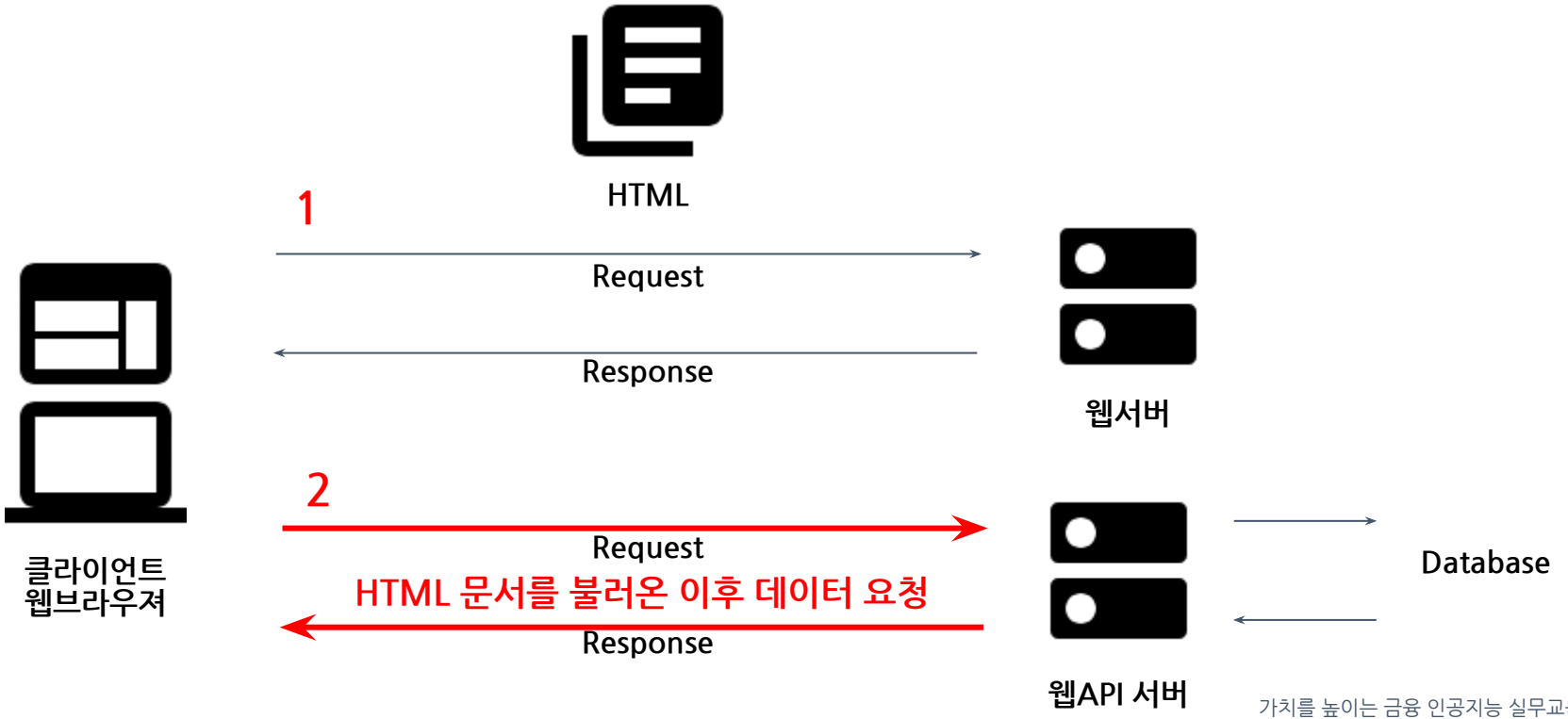
연습사이트 URL

https://scrapying-study.firebaseapp.com/04/

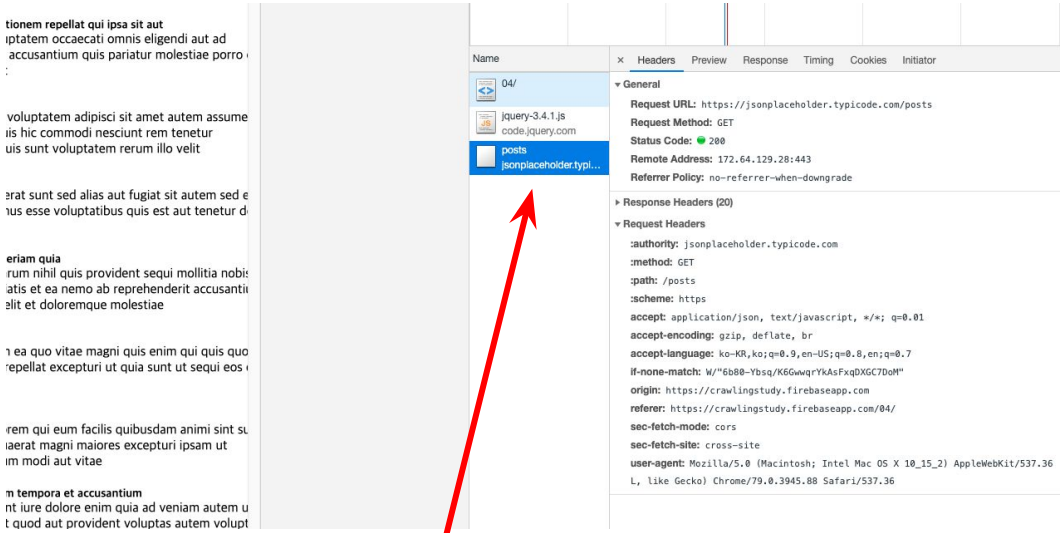
소스코드

비동기방식데이터크롤링ipynb

비동기방식



비동기방식



개발자도구의 네트워크탭을 확인하여 서비스요청 확인

연습사이트 URL

https://scrapying-study.firebaseapp.com/04/

소스코드

비동기방식데이터크롤링ipynb

가치를 높이는 금융 인공지능 실무교육

Insight campus

비동기방식

```
import requests
from bs4 import BeautifulSoup
response = requests.get("https://jsonplaceholder.typicode.com/posts")
print(response.text)
```

결과가 JSON? 형식으로

결과


```
[
  {
    "userId": 1,
    "id": 1,
    "title": "sunt aut facere repellat provident occaecati excepturi optio reprehenderit",
    "body": "quia et suscipit₩nsuscipit recusandae consequuntur expedita et cum₩nreprehenderit
molestiae ut ut quas totam₩nnostrum rerum est autem sunt rem eveniet architecto"
  },
  {
    "userId": 1,
    "id": 2,
    "title": "qui est esse",
    "body": "est rerum tempore vitae₩nsequi sint nihil reprehenderit dolor beatae ea dolores
neque₩nfugiat blanditiis voluptate porro vel nihil molestiae ut reiciendis₩nqui aperiam non
debitis possimus qui neque nisi nulla"
  },
  {
    "userId": 1,
    "id": 3,
    "title": "ea molestias quasi exercitationem repellat qui ipsa sit aut",
    "body": "et iusto sed quo iure₩nvoluptatem occaecati omnis eligendi aut ad₩nvoluptatem
doloribus vel accusantium quis pariatur₩nmolestiae porro eius odio et labore et velit aut"
  },
```


연습사이트 URL

https://scrapying-study.firebaseapp.com/04/

소스코드

비동기방식데이터크롤링ipynb

가치를 높이는 금융 인공지능 실무교육

Insight campus

JSON

JSON은 자바스크립트 객체 문법을 따르는 문자 기반의 데이터 포맷이다.
파이썬의 딕셔너리(사전) 형태와 비슷


```
[
  {
    "userId": 1,
    "id": 1,
    "title": "sunt aut facere repellat provident occaecati excepturi optio reprehenderit",
    "body": "quia et suscipit₩nsuscipit recusandae consequuntur expedita et cum₩nreprehenderit molestiae ut ut quas totam₩nnostrum rerum est autem sunt rem eveniet architecto"
  },
  {
    "userId": 1,
    "id": 2,
    "title": "qui est esse",
    "body": "est rerum tempore vitae₩nsequi sint nihil reprehenderit dolor beatae ea dolores neque₩nfugiat blanditiis voluptate porro vel nihil molestiae ut reiciendis₩nqui aperiam non debitis possimus qui neque nisi nulla"
  }
]
```


JSON 형태 데이터를 파이썬 딕셔너리로

import json
json.loads(JSON형식의 텍스트)

연습사이트 URL
https://scrapying-study.firebaseapp.com/04/

소스코드
비동기방식데이터크롤링ipynb

비동기방식

```
import json

import requests
from bs4 import BeautifulSoup
response =
requests.get("https://jsonplaceholder.typicode.com/posts")
result_dic = json.loads(response.text)
print(result_dic)
```

연습사이트 URL

https://scrapying-study.firebaseapp.com/04/

소스코드

비동기방식데이터크롤링ipynb

크롤링 데이터 JSON 형태로 저장

json.dump() - json 파일로저장
json.load() - json 파일읽기

```
import json

import requests
from bs4 import BeautifulSoup
response = requests.get("https://jsonplaceholder.typicode.com/posts")
result_dic = json.loads(response.text)

with open("data.json", "w") as json_file:
    json.dump(result_dic, json_file)
```

```
import json

with open("data.json", "r") as json_file:
    result = json.load(json_file)

print(result)
```

연습사이트 URL
https://scrapying-study.firebaseapp.com/04/

소스코드
비동기방식데이터크롤링ipynb

실습01

아래 사이트를 크롤링하여 아래와 같이 각각 을에 id와 title 그리고 url을 리스트형식으로 담고 최종 json 파일 형태로 저장해보세요.

결과 :


```
[
  {
    "id": 1,
    "title": "accusamus beatae ad facilis cum similique qui sunt",
    "url": "https://via.placeholder.com/600/92c952",
  },
  {
    "id": 2,
    "title": "reprehenderit est deserunt velit ipsam",
    "url": "https://via.placeholder.com/600/771796",
  },
...
```


연습사이트 URL

https://scrapying-study.firebaseapp.com/05/

소스코드

비동기방식데이터크롤링ipynb

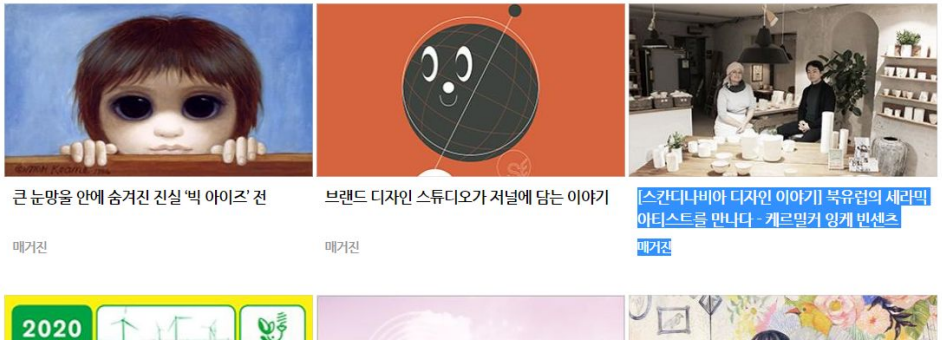
가치를 높이는 금융 인공지능 실무교육

Insight campus

실전02
- 크롤링 -

실전02-1

디자인정글 메인페이지 데이터 크롤링 (제목, 카테고리)
더 보기를 눌러서 나오는 추가데이터도 모두 크롤링
https://www.jungle.co.kr/



결과 :

큰 눈망울 안에 숨겨진 진실 '빅 아이즈' 전
매거진

브랜드 디자인 스튜디오가 저널에 담는 이야기
매거진

[스칸디나비아 디자인 이야기] 북유럽의 세라믹 아티스트를 만나다 - 케르밀커 잉케 빈센츠
매거진

소스코드
실전02.ipynb

실전02-2

로켓펀치 채용페이지 총 10페이지 크롤링 (회사명, 회사설명, 채용정보(회사별 여러개))

https://www.rocketpunch.com/jobs

결과 :

[{'name': '에이블리코퍼레이션',
'description': '700만 다운로드 패션 커머스 플랫폼 에이블리 입니다.\xa0',
'detail': ['iOS 개발자 (병역특례가능)', '안드로이드 개발자
(병역특례가능)']},
{'name': 'InHandPlus, Inc.',
'description': '기술 기반 헬스케어 솔루션 개발\xa0',
'detail': ['Python / Django 백엔드 개발자(경력)']},
….

소스코드

실전02.ipynb

Selenium

- 크롤링 -

설치

pip install selenium

https://sites.google.com/a/chromium.org/chromedriver/

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

chromedriver = 'c:/webdriver/chromedriver'
driver = webdriver.Chrome(chromedriver)
```

소스코드

Selenium01.ipynb

접속 및 객체셀렉터

driver.get('https://www.jungle.co.kr/')

find_element_by_name()
find_elements_by_name()

find_elements_by_tag_name()
find_elements_by_tag_class_name()
find_elements_by_css_selector()
find_elements_by_xpath()

element.text - 태그 내 텍스트
element.get_attribute - 옵션(속성)

소스코드
Selenium01.ipynb

클릭 및 키보드입력

클릭
element.click()

키보드입력
element2.send_keys("python")
element2.send_keys(Keys.RETURN)

ARROW_DOWN / ARROW_LEFT / ARROW_RIGHT / ARROW_UP BACKSPACE / DELETE / HOME / END /INSERT / ALT / COMMAND / CONTROL / SHIFT ENTER / ESCAPE /SPACE / TAB F1 / F2 / F3 ...../ F12

소스코드
Selenium01.ipynb

driver.implicitly_wait(), sleep()

처음 접속시 대기
(페이지 읽기가 끝나면 진행)

```
driver.implicitly_wait(5)
```

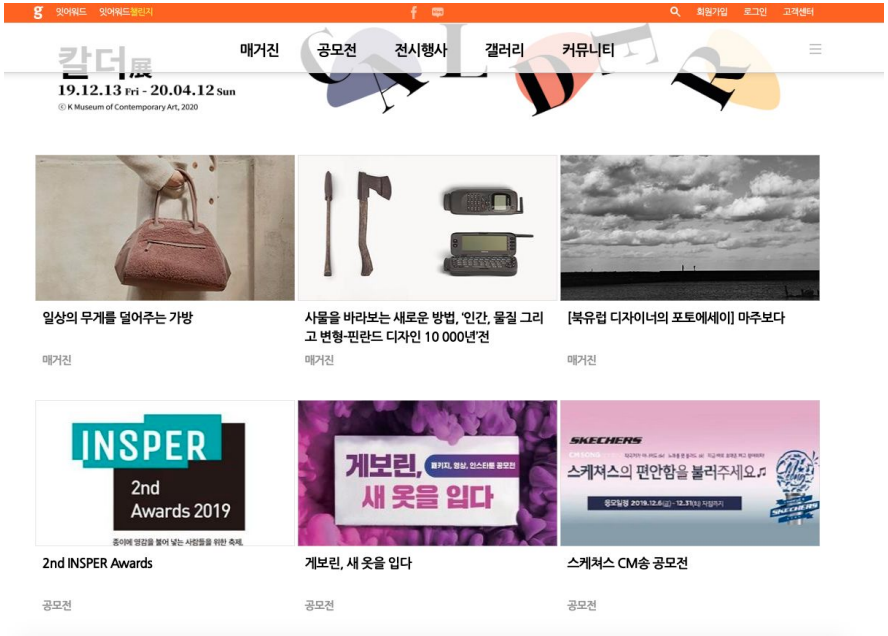
잠시 멈추는 함수

```
import time
time.sleep()
```

소스코드
Selenium01.ipynb

실습01

디자인정글 메인페이지 데이터 셀레니움으로 크롤링
https://www.jungle.co.kr/



소스코드
Selenium01.ipynb

실전03

로켓펀치 채용페이지 총 10페이지 셀레니움으로 크롤링 (회사명, 회사설명, 채용정보(회사별 여러개))

https://www.rocketpunch.com/jobs

결과 :

[{'name': '에이블리코퍼레이션',
'description': '700만 다운로드 패션 커머스 플랫폼 에이블리 입니다.\xa0',
'detail': ['iOS 개발자 (병역특례가능)', '안드로이드 개발자
(병역특례가능)']},
{'name': 'InHandPlus, Inc.',
'description': '기술 기반 헬스케어 솔루션 개발\xa0',
'detail': ['Python / Django 백엔드 개발자(경력)']},
....

```
driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
```

소스코드
실전03.ipynb

공공데이터 API 활용

- 크롤링 -

공공데이터API활용

공공데이터 포털 - https://www.data.go.kr/

도로명주소조회