

15기 정규세션

ToBig's 14기 강의정

Crawling

즐거운 크롤링

Contents

Unit 01 | WEB 동작 원리

Unit 02 | HTML

Unit 03 | Crawling 실습

Unit 04 | 과제

Unit 01 | WEB 동작 원리

1. 주소창에 news.daum.net 입력

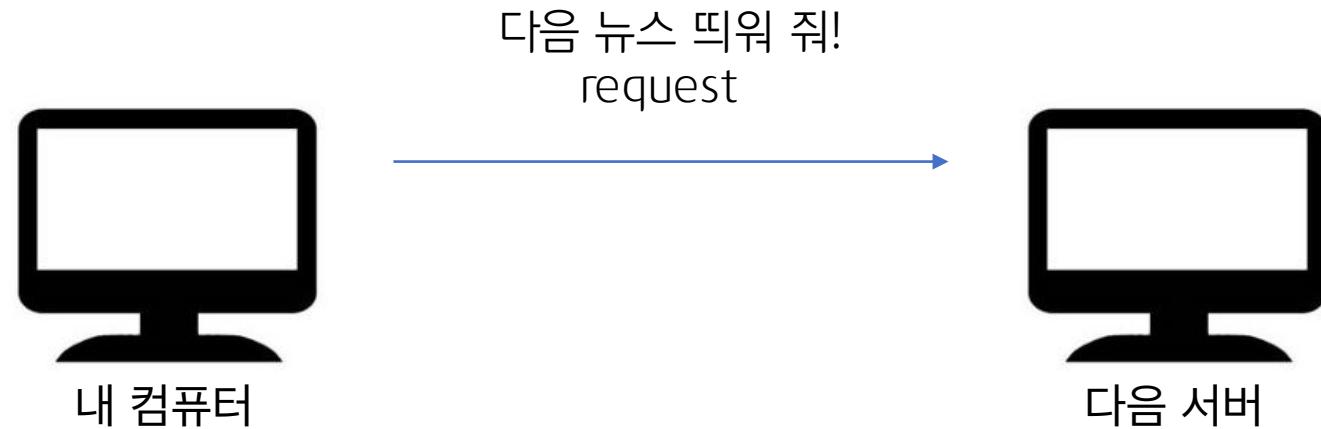
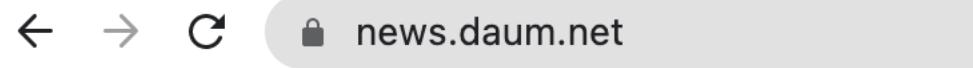


2. '경제' 버튼 클릭



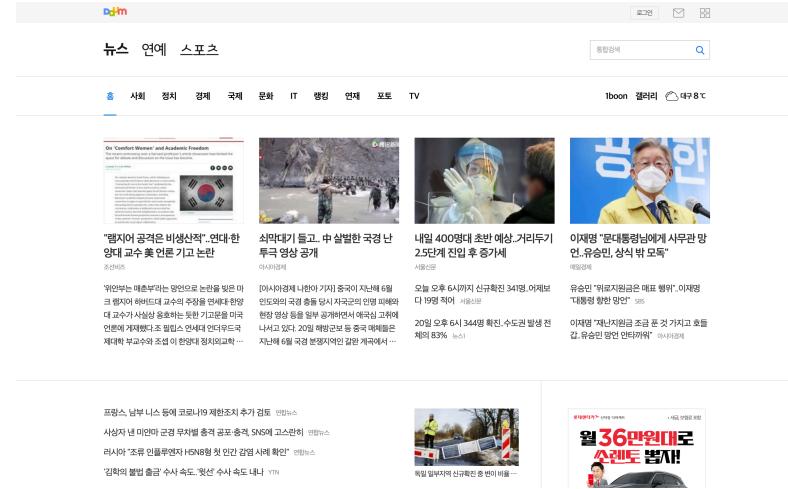
Unit 01 | WEB 동작 원리

1. 주소창에 news.daum.net 입력
2. '경제' 버튼 클릭



Unit 01 | WEB 동작 원리

1. 주소창에 news.daum.net 입력
2. '경제' 버튼 클릭

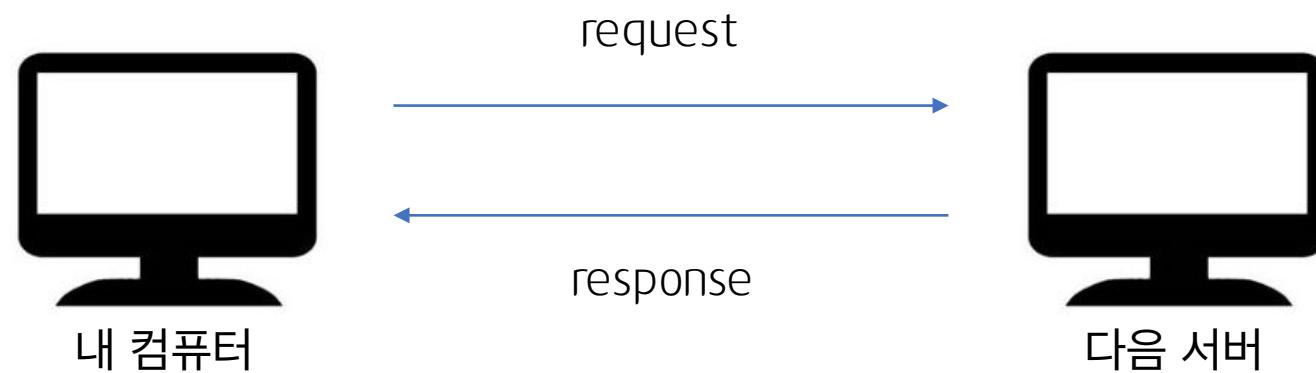


Unit 01 | WEB 동작 원리

1. 주소창에 news.daum.net 입력

← → ⌂ news.daum.net/economic#1

2. '경제' 버튼 클릭



Unit 01 | WEB 동작 원리

손수 경제 뉴스 기사 5000천건의 내용을 수집하려면?

<https://news.daum.net/economic#1> 접속 →

0번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

1번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

2번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

...

4998번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

4999번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

Unit 01 | WEB 동작 원리

손수 경제 뉴스 기사 5000천건의 내용을 수집하려면?

파이썬으로?

1. 원하는 웹 페이지에 접속하기

<https://news.daum.net/econ/idx>

2. 웹 페이지에서 원하는 내용 찾기

0번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

1번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

2번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

...

4998번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

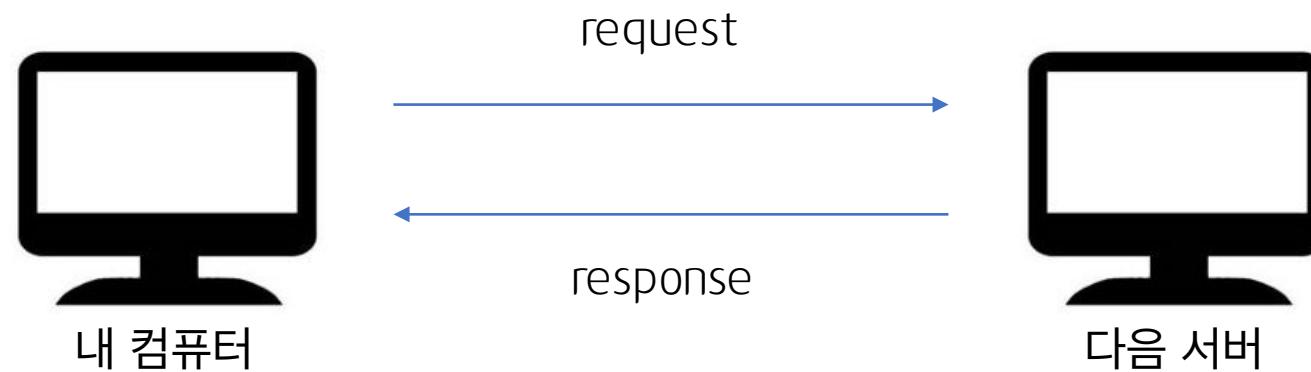
4999번째 뉴스기사 클릭 -> 기사 내용 ctrl C ctrl V -> 뒤로가기

Unit 01 | WEB 동작 원리

HTTP(HyperText Transfer Protocol)

웹에서 이루어지는 모든 데이터 교환의 기초이며, 클라이언트-서버 프로토콜.

하나의 완전한 문서는 텍스트, 레이아웃 설명, 이미지, 비디오, 스크립트 등 불러온(fetched) 하위 문서들로 재구성됩니다. -MDN



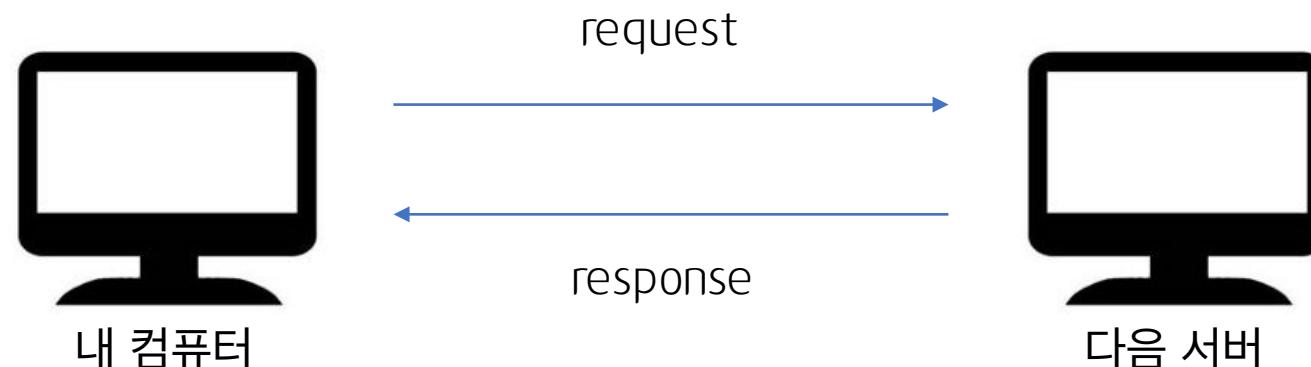
Unit 01 | WEB 동작 원리

HTTP(HyperText Transfer Protocol)

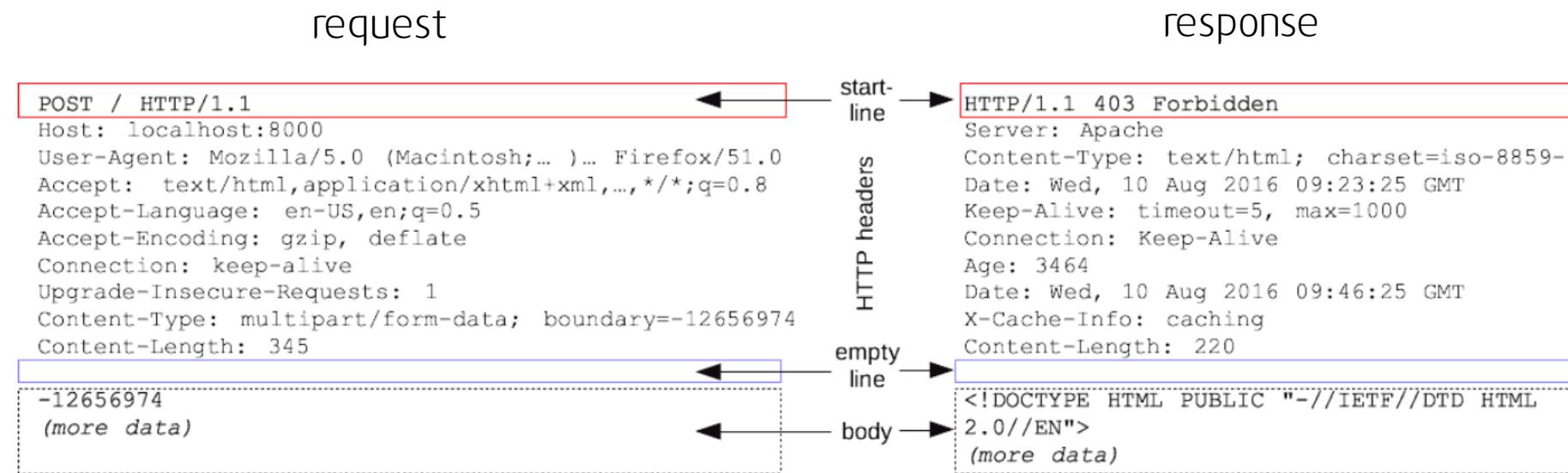
웹에서 이루어지는 모든 데이터 교환의 기초이며, 클라이언트-서버 프로토콜.

하나의 완전한 문서는 텍스트, 레이아웃 설명, 이미지, 비디오, 스크립트 등 불러온(fetched) 하위 문서들로 재구성됩니다. -MDN

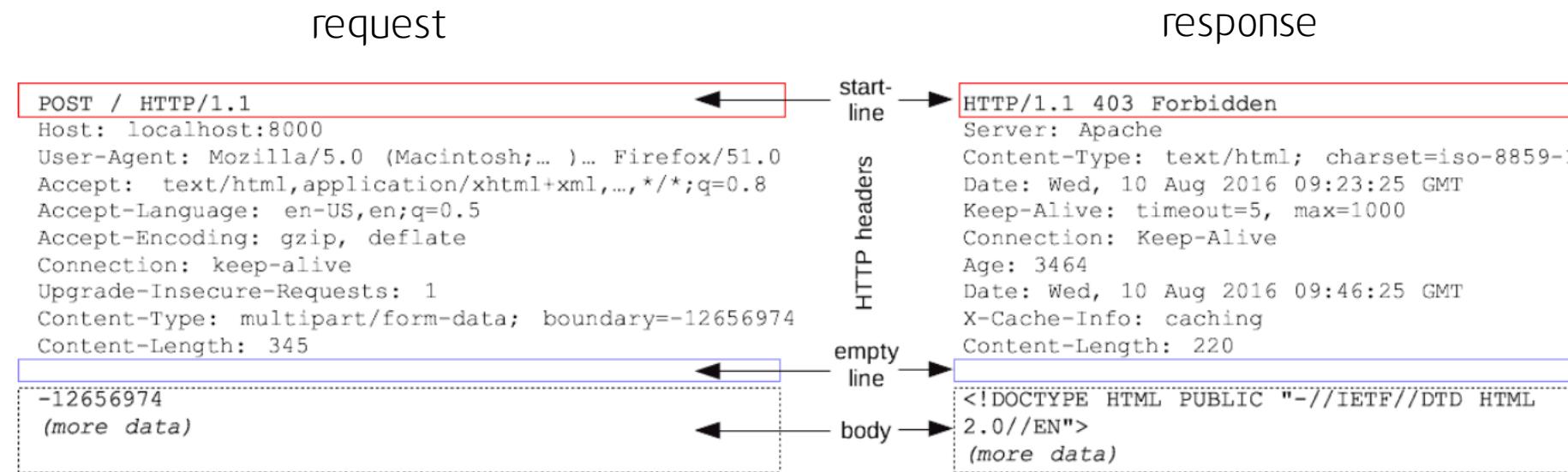
Protocol은 Language와 같다!
ex) TCP, UDP, ICMP, IP



Unit 01 | WEB 동작 원리

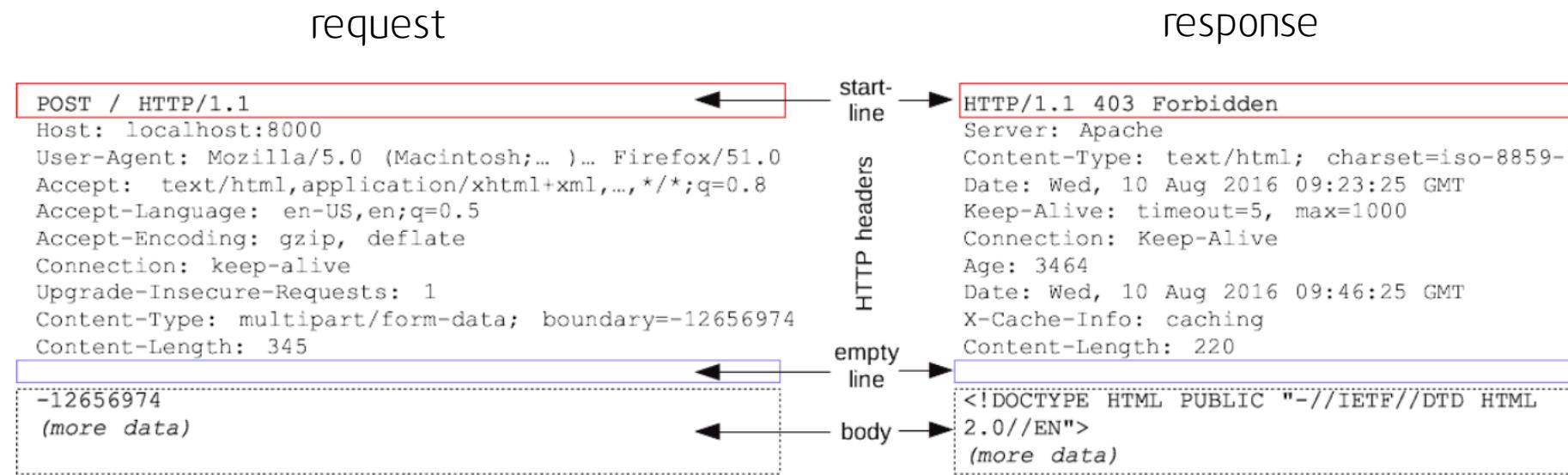


Unit 01 | WEB 동작 원리



HTTP Method
GET : 서버에 resource를 요청할 때
POST : 서버에 Input Data를 전송할 때
PUT : 서버에 resource를 전송할 때
DELETE : 서버 resource를 삭제할 때

Unit 01 | WEB 동작 원리



Status Code
2xx : 성공
3xx : 리다이렉션
4xx : 클라이언트 에러
5xx : 서버 에러

Unit 01 | WEB 동작 원리

news.daum.net 페이지를 요청하는 GET Request를 잘 만들면
다음 뉴스의 웹 소스코드를 얻을 수 있겠다..!

```
<!DOCTYPE html>
...<html lang="ko" class="os_mac chrome pc version_88_0_4324_150" =
  ▶ <head>...</head>
  ▼ <body class="bg_news">
    ▼ <div id="kakaoIndex">
      <a href="#kakaoBody">본문 바로가기</a>
      <a href="#kakaoGnb">메뉴 바로가기</a>
    </div>
    ▼ <div id="kakaoWrap" class="news_type1">
      ▼ <div id="wrapMinidaum">
        ▶ <div id="minidaum">...</div>
      </div>
      ▼ <script type="text/javascript">
        var minidaum_options = {
          bgType : "white",
          enableLogoutRetun : true,
          disableTracker : true, // false일 경우 사용,
          true일 경우 미사용
          enableShield : false
        };
      </script>
      <script src="https://go.daum.net/minidaum_pc.daum"></script>
    ▼ <div id="kakaoHead" role="banner">
      ▶ <div class="head_media">...</div>
      ▼ <div id="kakaoGnb" role="navigation">
        ▶ <div class="inner_gnb">...</div>
      </div>
      <script src="/t1.daumcdn.net/media/kraken/news/be1fb6d/gn
      b.merged.js"></script>
    ▼ <style>
```

Unit 01 | WEB 동작 원리

URL

<https://www.naver.com>

-> https://search.naver.com/search.naver?where=nexearch&sm=top_hty&fbm=1&ie=utf8&query=아이폰

← → C search.naver.com/search.naver?where=nexearch&sm=top_hty&fbm=1&ie=utf8&query=아이폰

The screenshot shows the Naver search results page for the query "아이폰". The URL in the address bar is https://search.naver.com/search.naver?where=nexearch&sm=top_hty&fbm=1&ie=utf8&query=아이폰. The search bar contains the same query. The main content area displays a news snippet from Apple's official website for the iPhone 12 mini, followed by a list of news articles under the heading "뉴스토Pic".

N | 아이폰

통합 이미지 쇼핑 VIEW 뉴스 지식IN 지도 실시간검색 지식백과 ... 검색옵션 ^

정렬 ▾ 기간 ▾ 영역 ▾ 옵션유지 상세검색 ▾

Apple 공식 홈페이지

Apple 한국 공식 사이트
iPhone 12 mini
반가워요 5G. 스피드 그 이상의 스피드.
[보상판매](#) · [비교하기](#)

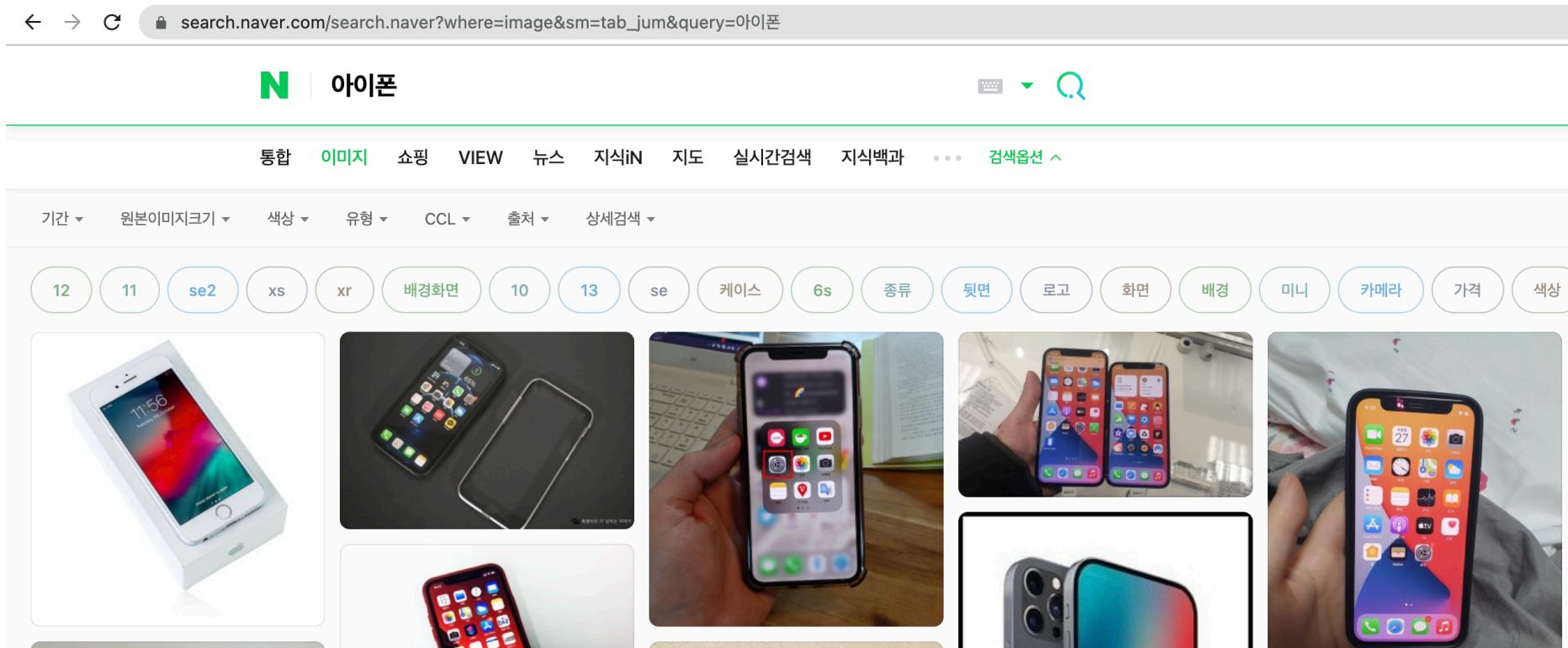
뉴스토Pic

뉴스 연예·스포츠

- 1 KB국민은행 한국판 뉴딜 ...
- 2 경주 전복 어선 선원 2명 ...
- 3 금융위 여전사 유동성관리 ...
- 4 이재용 삼성전자 부회장 ...
- 5 금감원 올해 종합검사 2배 ...
- 6 한전 콘크리트관 담합 ...
- 7 1월 주택 거래량 전달 대비...
- 8 지난해 재무제표 고의위반 ...
- 9 금감원 사업보고서 종점 점...
- 10 허리 지병 최정우 포스코 ...

Unit 01 | WEB 동작 원리

search.naver.com/search.naver?where=image&sm=tab_jum&query=%EC%9D%98%EB%8A%84



Unit 01 | WEB 동작 원리

URL

N https://search.naver.com/search.naver?where=image&sm=tab_jum&query=아이폰



Unit 02 | HTML

HTML(HyperText Markup Language)

WEB 페이지에 개발에 쓰이는 언어

```
<!DOCTYPE html>
...<html lang="ko" class="os_mac chrome pc version_88_0_4324_150 "> =
  ▶ <head>...</head>
  ▼ <body class="bg_news">
    ▼ <div id="kakaoIndex">
      <a href="#kakaoBody">본문 바로가기</a>
      <a href="#kakaoGnb">메뉴 바로가기</a>
    </div>
    ▼ <div id="kakaoWrap" class="news_type1">
      ▼ <div id="wrapMinidaum">
        ▶ <div id="minidaum">...</div>
      </div>
      ▼ <script type="text/javascript">
        var minidaum_options = {
          bgType : "white",
          enableLogoutRetun : true,
          disableTracker : true, // false일 경우 사용,
          true일 경우 미사용
          enableShield : false
        };
      </script>
      <script src="https://go.daum.net/minidaum_pc.daum"></script>
    ▼ <div id="kakaoHead" role="banner">
      ▶ <div class="head_media">...</div>
      ▼ <div id="kakaoGnb" role="navigation">
        ▶ <div class="inner_gnb">...</div>
      </div>
      <script src="//t1.daumcdn.net/media/kraken/news/be1fb6d/gn
      b.merged.js"></script>
    ▼ <style>
```

Unit 02 | HTML

HTML(HyperText Markup Language)

WEB 페이지에 개발에 쓰이는 언어



Unit 02 | HTML

HTML(HyperText Markup Language)
WEB 페이지에 개발에 쓰이는 언어

- html 프로그래밍
- html 프로그래밍 언어
- html 프로그래밍 언어인가

개발자를 찾는 방법



Unit 02 | HTML

HTML(HyperText Markup Language)



Unit 02 | HTML

id

하나의 HTML 문서 안에서 id값은 unique하다.

```
93      </div>
94  </form>
95</div>
96<script type="text/javascript" src="https://search1.daumcdn.net/search/suggest_pc/suggest-1.2.16.min.js"></script>
97<script src="//t1.daumcdn.net/media/kraken/news/belfb6d/suggest.merged.js"></script>
98
99</div>
100
101<div id="kakaoGnb" role="navigation">
102  <div class="inner_gnb">
103    <h2 class="screen_out">뉴스 메인메뉴</h2>
104    <ul class="gnb_comm" data-tiara-layer="GNB default">
105      <li ><a href="/" class="link_gnb link_gnb1" data-tiara-layer="media_home"><span class="screen_out">선택됨</span><span class="inner_bar"></span></span></a></li>
106      <li ><a href="/society" class="link_gnb link_gnb2" data-tiara-layer="media_society"><span class="ir_wa">사회</a></li>
107      <li ><a href="/politics" class="link_gnb link_gnb3" data-tiara-layer="media_politics"><span class="ir_wa">정치</span></a></li>
108      <li class="on"><a href="/economic" class="link_gnb link_gnb4" data-tiara-layer="media_economic"><span class="ir_wa">경제</span></a></li>
109      <li ><a href="/foreign" class="link_gnb link_gnb5" data-tiara-layer="media_foreign"><span class="ir_wa">국제</a></li>
110      <li ><a href="/culture" class="link_gnb link_gnb6" data-tiara-layer="media_culture"><span class="ir_wa">문화</a></li>
111      <li ><a href="/digital" class="link_gnb link_gnb7" data-tiara-layer="media_digital"><span class="ir_wa">IT</span></a></li>
```

Unit 02 | HTML

class

하나의 HTML 문서 안에서 동일한 class가 존재할 수 있다.

```
<a href="http://search.daum.net/search?w=tot&DA=23W&rtmaxcoll=Z8T&q=울릉군날씨" class="list-item"
    <span class="ico weather ico_weather4">흐림</span>
    <span class="txt_weather">
        울릉/독도
        <span class="num_heat">12</span> °C
    </span>
</a>
</li>
<li>
    <a href="http://search.daum.net/search?w=tot&DA=23W&rtmaxcoll=Z8T&q=춘천시날씨" class="list-item"
        <span class="ico weather ico_weather4">흐림</span>
        <span class="txt_weather">
            춘천
            <span class="num_heat">-1</span> °C
        </span>
    </a>
</li>
<li>
```

Unit 03 | Crawling 실습

Requests



BeautifulSoup

The BeautifulSoup logo features the word "Beautiful" in a bold, black, sans-serif font, followed by "Soup" in a stylized, italicized font where the 'f' has a long, sweeping loop extending downwards and to the right.

Selenium



Unit 03 | Crawling 실습

Requests



Requests
http for humans

BeautifulSoup

BeautifulSoup

Selenium

Selenium

실습으로! week5_crawling.ipynb

Unit 03 | Crawling 실습

```

<!-- 종목 -->
▼<div class="box_type_l">
  ▶<div class="tab_style_1">...</div>
  <!-- [D] 활성화된 탭메뉴에 따라 blind text 변경해주세요 -->
  <h4 class="blind">코스피</h4>
  ▼<table summary="코스피 시세정보를 선택한 항목에 따라 정보를 제공합니다." cellpadding="0" cellspacing="0" class="type_2">
    <caption>코스피</caption>
    ▶<colgroup>...</colgroup>
    ▶<thead>...</thead>
    ▼<tbody>
      ▶<tr>...</tr>
      ▼<tr onmouseover="mouseOver(this)" onmouseout="mouseOut(this)" style="background-color: #f0f0f0; border-bottom: 1px solid black; padding: 5px; position: relative; width: 100%; height: 100%;">
        <td class="no">1</td>
        ▼<td>
          <a href="/item/main.nhn?code=005930" class="title">삼성전자</a> == $0
        </td>
        <td class="number">53,700</td>
      ▶<td class="number">...</td>
      ▶<td class="number">...</td>
        <td class="number">100</td>
        <td class="number">3,205,773</td>
        <td class="number">5,969,783</td>
        <td class="number">57.57</td>
        <td class="number">9,684,758</td>
        <td class="number">8.91</td>
        <td class="number">19.63</td>
      ▶<td class="center">...</td>
      </tr>
      ▼<tr onmouseover="mouseOver(this)" onmouseout="mouseOut(this)" style="background-color: #f0f0f0; border-bottom: 1px solid black; padding: 5px; position: relative; width: 100%; height: 100%;">
        <td onmouseover="mouseOver(this)" onmouseout="mouseOut(this)" style="background-color: #f0f0f0; border-bottom: 1px solid black; padding: 5px; position: relative; width: 100%; height: 100%;">
      
```

코스피		코스닥						
N	종목명	현재가	전일비	등락률	액면가	시가총액	상장주식수	
1	삼성전자	53,700	▲ 900	+1.70%	100	3,205,773	5	
2	SK하이닉스	85,200	▲ 1,800	+2.16%	5,000	620,258		
3	삼성전자우	43,100	▲ 750	+1.77%	100	354,664		
4	NAVER	173,500	▼ 6,500	-3.61%	100	285,951		
5	현대차	125,500	▲ 1,000	+0.80%	5,000	268,154		
6	삼성바이오로직스	398,000	▲ 2,500	+0.63%	2,500	263,337		
7	현대모비스	257,500	▲ 11,500	+4.67%	5,000	245,415		
8	셀트리온	186,000	▼ 500	-0.27%	1,000	238,708		

Unit 03 | Crawling 실습

```

<!-- 종목 -->


<!-- [D] 활성화된 탭메뉴에 따라 blind text 변경해주세요 -->
    <h4 class="blind">코스피</h4>
    <table summary="코스피 시세정보를 선택한 항목에 따라 정보를 제공합니다." cellpadding="0" cellspacing="0" class="type_2">
        <caption>코스피</caption>
        <colgroup>...</colgroup>
        <thead>...</thead>
        <tbody>
            <tr>
                <td colspan="10" class="blank_08"></td>
            </tr>
            <tr onmouseover="mouseOver(this)" onmouseout="mouseOut(this)" style="background-color: #f0f0f0;">
                <td class="no">1</td>
                <td>
                    <a href="/item/main.nhn?code=005930" class="tltle">삼성전자</a>
                </td>
                <td class="number">53,700</td>
                <td class="number">...</td>
                <td class="number">...</td>
                <td class="number">100</td>
                <td class="number">3,205,773</td>
                <td class="number">5,969,783</td>
                <td class="number">57.57</td>
                <td class="number">9,684,758</td>
            </tr>
            <tr onmouseover="mouseOver(this)" onmouseout="mouseOut(this)" style="background-color: #f0f0f0;">
                <td class="center">...</td>
            </tr>
            <tr onmouseover="mouseOver(this)" onmouseout="mouseOut(this)" style="background-color: #f0f0f0;">...</tr>
            <tr onmouseover="mouseOver(this)" onmouseout="mouseOut(this)" style="background-color: #f0f0f0;">...</tr>
            <tr onmouseover="mouseOver(this)" onmouseout="mouseOut(this)" style="background-color: #f0f0f0;">...</tr>
        </tbody>
    </table>


```

div.box_type_l > table > tbody > tr > a

코스피		코스닥					
N	종목명	현재가	전일비	등락률	액면가	시가총액	상장주
1	삼성전자	53,700	▲ 900	+1.70%	100	3,205,773	5
2	SK하이닉스	85,200	▲ 1,800	+2.16%	5,000	620,258	
3	삼성전자우	43,100	▲ 750	+1.77%	100	354,664	
4	NAVER	173,500	▼ 6,500	-3.61%	100	285,951	
5	현대차	125,500	▲ 1,000	+0.80%	5,000	268,154	
6	삼성바이오로직스	398,000	▲ 2,500	+0.63%	2,500	263,337	
7	현대모비스	257,500	▲ 11,500	+4.67%	5,000	245,415	
8	셀트리온	186,000	▼ 500	-0.27%	1,000	238,708	
9	LG화학	217,500	▼ 2,000	-0.89%	5,000	224,121	

Unit 03 | Crawling 실습

```
html = """
<div class="rate_info">
    <div class="no_today">
        <div class="blind">
            53,700
        </div>
    </div>
    <table class="no_info">
        <tr>
            <td class="first">
                <span class="sptxt sp_txt2">전일</span>
                <em>
                    <span class="blind">52,800</span>
                    <span class="no5">5</span>
                    <span class="no2">2</span>
                    <span class="shim">, </span>
                    <span class="no8">8</span>
                    <span class="no0">0</span>
                    <span class="no0">0</span>
                </em>
            </td>
        </tr>
    </table>
</div>
<div class="tab_con1">
</div>
"""
```

삼성전자 전일 주가

Unit 03 | Crawling 실습

```
html = """
<div class="rate_info">
    <div class="no_today">
        <div class="blind">
            53,700
        </div>
    </div>
    <table class="no_info">
        <tr>
            <td class="first">
                <span class="sptxt sp_txt2">전일</span>
                <em>
                    <span class="blind">52,800</span>
                    <span class="no5">5</span>
                    <span class="no2">2</span>
                    <span class="shim">,</span>
                    <span class="no8">8</span>
                    <span class="no0">0</span>
                    <span class="no0">0</span>
                </em>
            </td>
        </tr>
    </table>
</div>
<div class="tab_con1">
</div>
"""
"""
```

```
bs.select_one('.rate_info')
```

Unit 03 | Crawling 실습

```
html = """
<div class="rate_info">
    <div class="no_today">
        <div class="blind">
            53,700
        </div>
    </div>
    <table class="no_info">
        <tr>
            <td class="first">
                <span class="sptxt sp txt2">전일</span>
                <em>
                    <span class="blind">52,800</span>
                    <span class="no5">5</span>
                    <span class="no2">2</span>
                    <span class="shim">,</span>
                    <span class="no8">8</span>
                    <span class="no0">0</span>
                    <span class="no0">0</span>
                </em>
            </td>
        </tr>
    </table>
</div>
<div class="tab_con1">
</div>
"""
```

```
bs.select_one('.rate_info').find(text =“전일”).parent
```

Unit 03 | Crawling 실습

```
html = """
<div class="rate_info">
    <div class="no_today">
        <div class="blind">
            53,700
        </div>
    </div>
    <table class="no_info">
        <tr>
            <td class="first">
                <span class="sptxt sp_txt2">전일</span>
                <em>
                    <span class="blind">52,800</span>
                    <span class="no5">5</span>
                    <span class="no2">2</span>
                    <span class="shim">,</span>
                    <span class="no8">8</span>
                    <span class="no0">0</span>
                    <span class="no0">0</span>
                </em>
            </td>
        </tr>
    </table>
</div>
<div class="tab_con1">
</div>
"""
```

bs.select_one('.rate_info').find(text =“전일”).parent.parent

Unit 03 | Crawling 실습

```
html = """
<div class="rate_info">
    <div class="no_today">
        <div class="blind">
            53,700
        </div>
    </div>
    <table class="no_info">
        <tr>
            <td class="first">
                <span class="sptxt sp_txt2">전일</span>
                <em>
                    <span class="blind">52,800</span>
                    <span class="no5">5</span>
                    <span class="no2">2</span>
                    <span class="shim">,</span>
                    <span class="no8">8</span>
                    <span class="no0">0</span>
                    <span class="no0">0</span>
                </em>
            </td>
        </tr>
    </table>
</div>
<div class="tab_con1">
</div>
"""
```

```
bs.select_one('.rate_info').find(text = "전일").parent
    .parent
    .select_one('.blind')
```

Unit 04 | 과제

국내증시 시각총액 페이지 크롤링
(https://finance.naver.com/sise/sise_market_sum.nhn)

수집 내용

1. 기업 리스트 수집 (기업명, 기업 상세페이지 링크)
2. 기업 상세 정보 수집(주가, 시가, 시가총액, 시가총액순위)

주의사항 및 발전사항

- 모든 Unit Test를 통과시켜야 합니다.
- 과제 제출은 week5_crawling_assignment.ipynb만 제출합니다.(csv 결과파일 제외)
- 추가 정보 수집시 **가산점** 드립니다!

Q & A

들어주셔서 감사합니다.