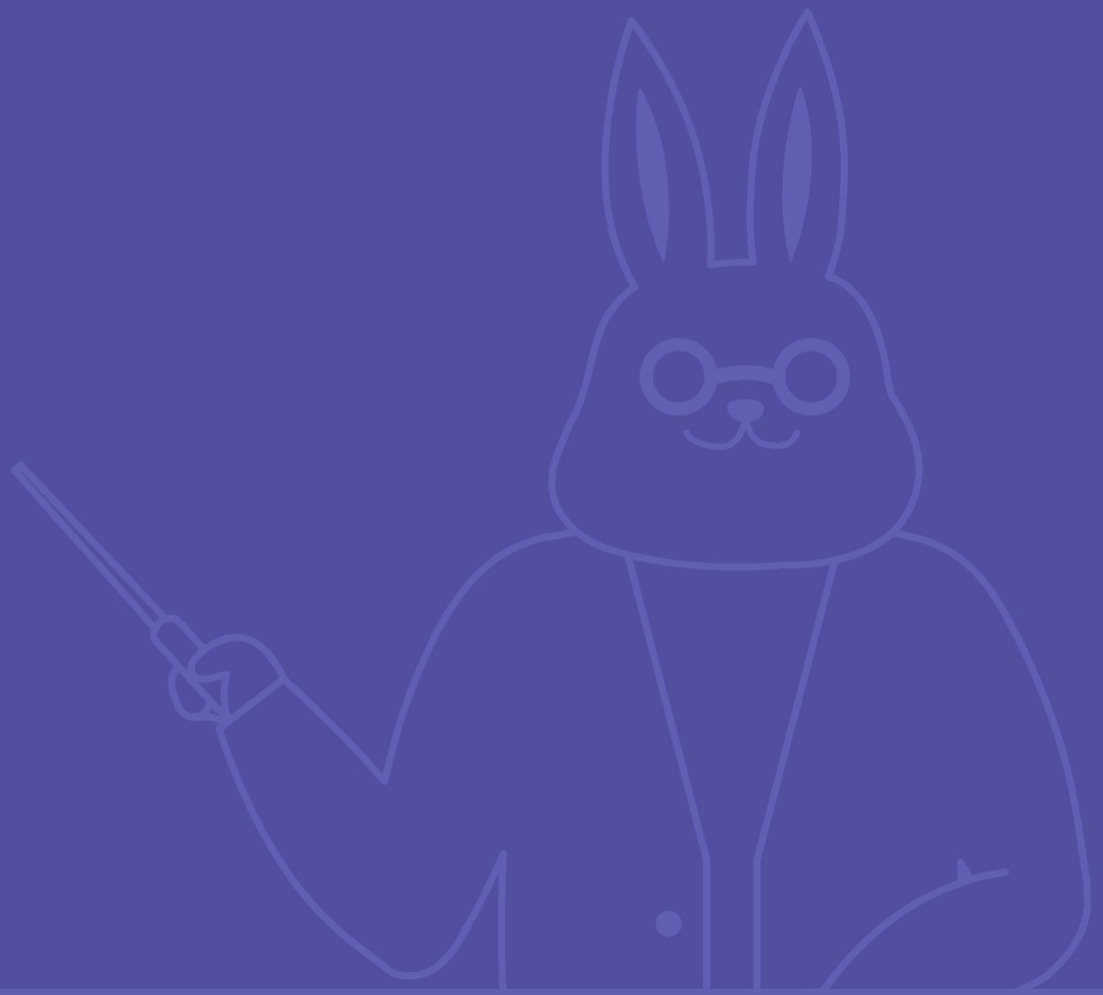
파이썬 크롤링 입문

01 HTML 훑어보기



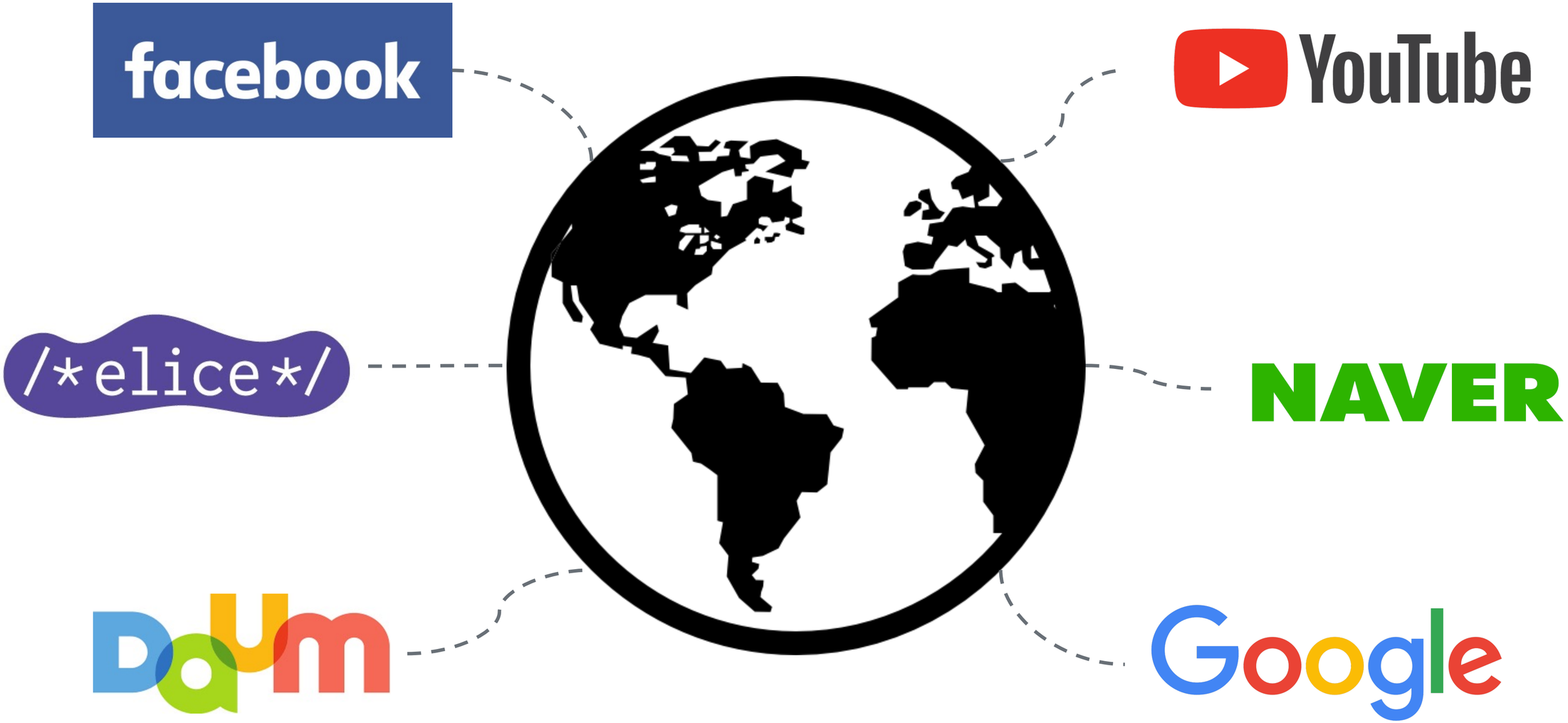
/* elice */

목차

01

크롤링에 관하여

World Wide Web



Web Crawler (Web Spider)

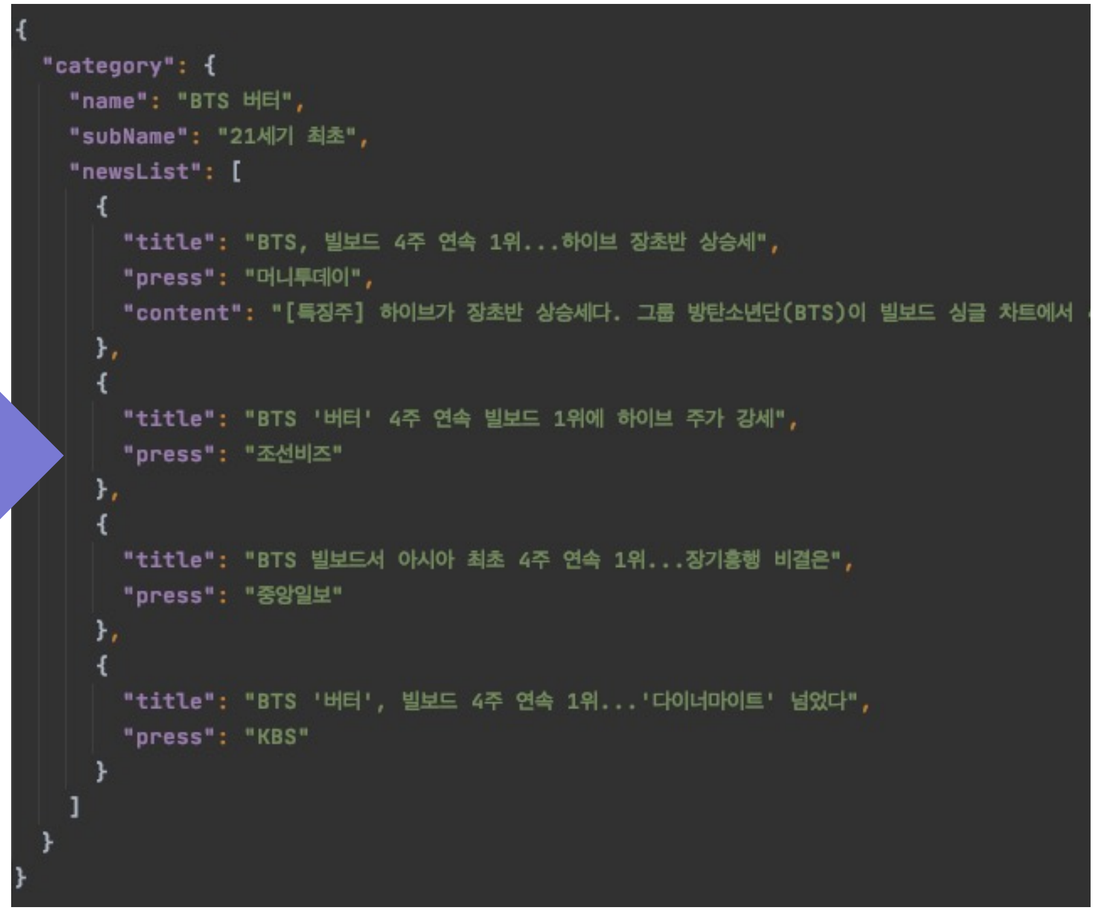
수많은 웹 사이트를 탐험하며 페이지를 수집하는 시스템을

Web Crawler 또는 **Web Spider**라고 한다.

신문/잡지 스크랩

신문이나 잡지에서 **필요한** 부분만 잘라 **내** 노트에 붙였던 것

Web Scraping

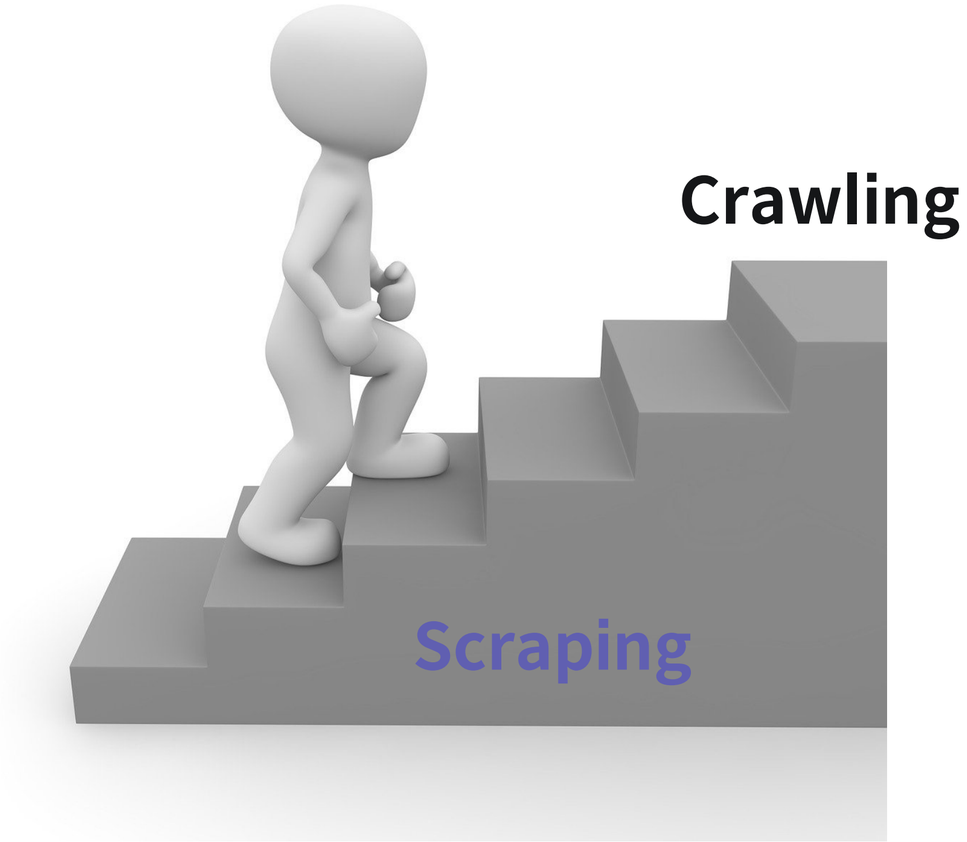


```
{
  "category": {
    "name": "BTS 버터",
    "subName": "21세기 최초",
    "newsList": [
      {
        "title": "BTS, 빌보드 4주 연속 1위...하이브 장초반 상승세",
        "press": "머니투데이",
        "content": "[특징주] 하이브가 장초반 상승세다. 그룹 방탄소년단(BTS)이 빌보드 싱글 차트에서
      },
      {
        "title": "BTS '버터' 4주 연속 빌보드 1위에 하이브 주가 강세",
        "press": "조선비즈"
      },
      {
        "title": "BTS 빌보드서 아시아 최초 4주 연속 1위...장기흥행 비결은",
        "press": "중앙일보"
      },
      {
        "title": "BTS '버터', 빌보드 4주 연속 1위...'다이너마이트' 넘었다",
        "press": "KBS"
      }
    ]
  }
}
```

특정 웹 페이지 내용 중 **원하는 부분**을

내가 **원하는 형식**으로 만드는 것

Crawling vs Scraping



크롤링 기술을 배우기 위한 발판, **스크래핑**

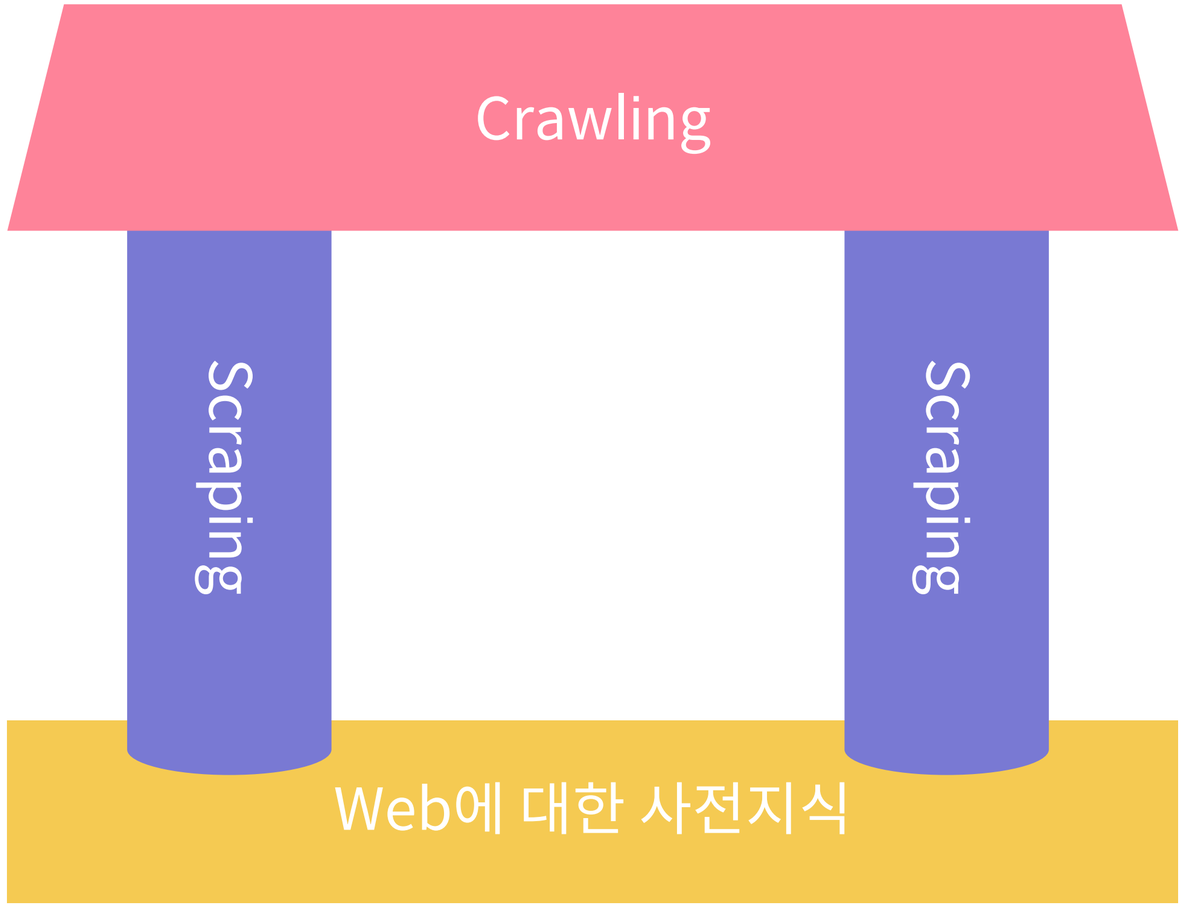
How to scrap?


```
{
  "category": {
    "name": "BTS 버터",
    "subName": "21세기 최초",
    "newsList": [
      {
        "title": "BTS, 빌보드 4주 연속 1위...하이브 장초반 상승세",
        "press": "머니투데이",
        "content": "[특징주] 하이브가 장초반 상승세다. 그룹 방탄소년단(BTS)이 빌보드 싱글 차트에서
      },
      {
        "title": "BTS '버터' 4주 연속 빌보드 1위에 하이브 주가 강세",
        "press": "조선비즈"
      },
      {
        "title": "BTS 빌보드서 아시아 최초 4주 연속 1위...장기흥행 비결은",
        "press": "중앙일보"
      },
      {
        "title": "BTS '버터', 빌보드 4주 연속 1위...'다이너마이트' 넘었다",
        "press": "KBS"
      }
    ]
  }
}
```


웹 페이지 작성 규칙에 따름

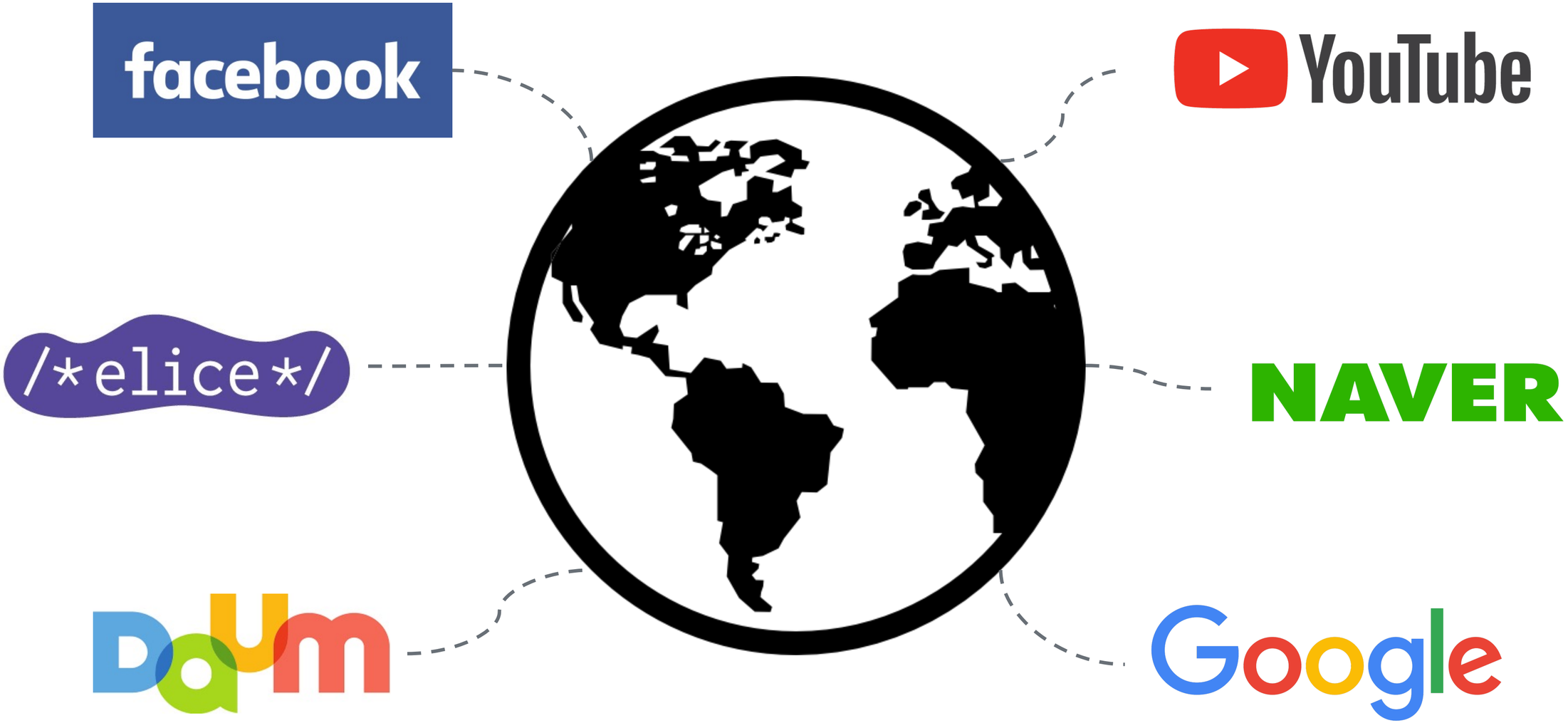
규칙을 이용해 정리 가능

스크래핑을 배우기 전에

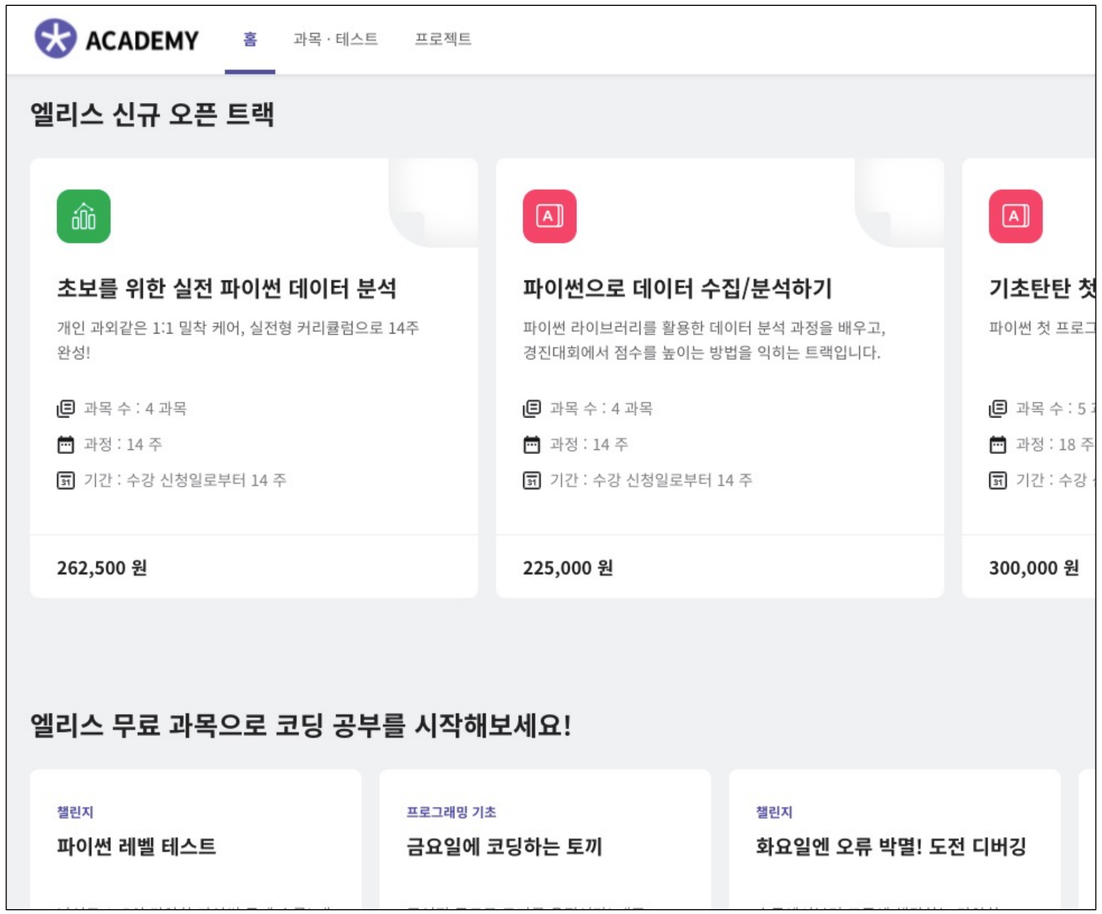


Web에 대한 지식이 선행되어야 함

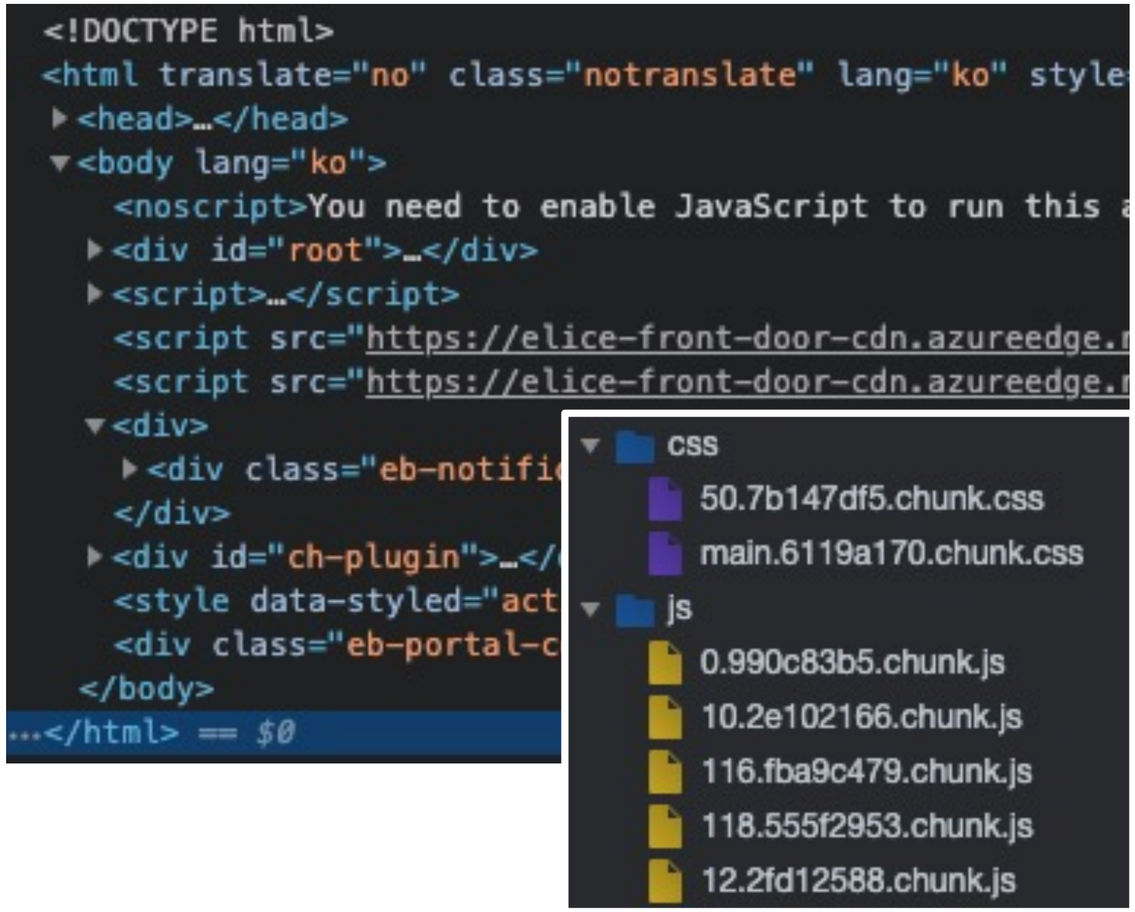
World Wide Web



웹 페이지의 구조



엘리스 아카데미 웹 페이지



웹 페이지를 구성하는 html과 여러 파일

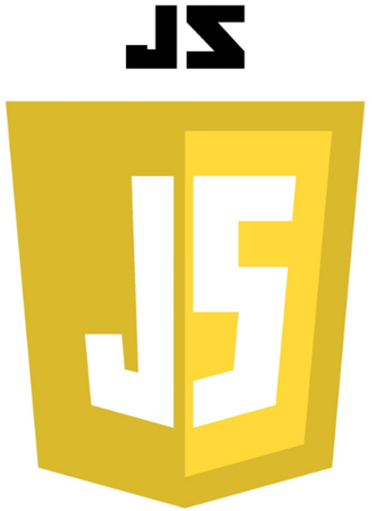
HTML, CSS, JavaScript

- HTML

: 정보 및 설계도

- CSS

: 디자인 및 스타일링



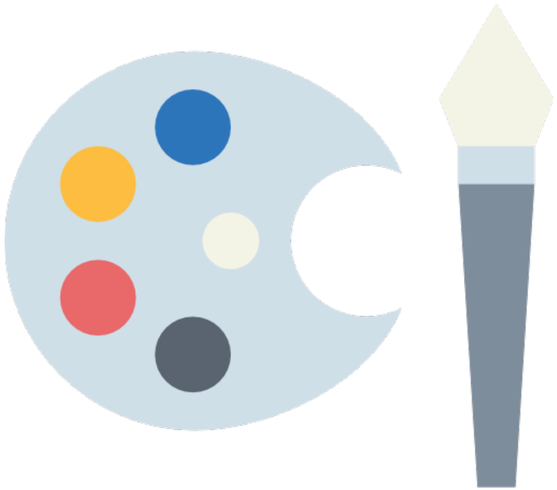
- JavaScript

: 기능과 효과

웹 사이트 제작 = 건물 짓기

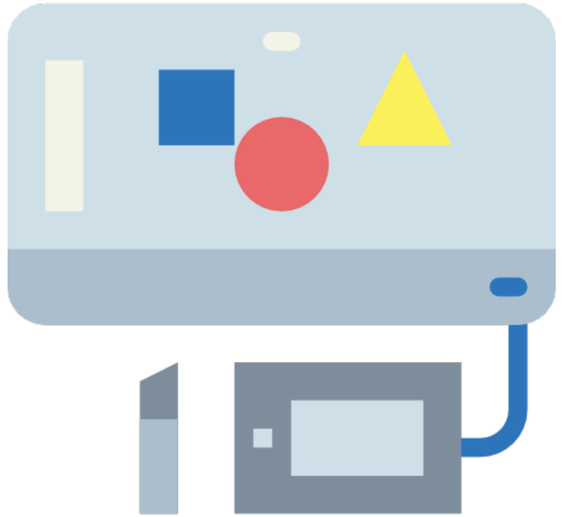
- HTML

: 건물 설계도



- CSS

: 인테리어 디자인



- JavaScript

: 기능과 효과

02

HTML - 태그 (1)

HTML이란?

Hyper Text Markup Language

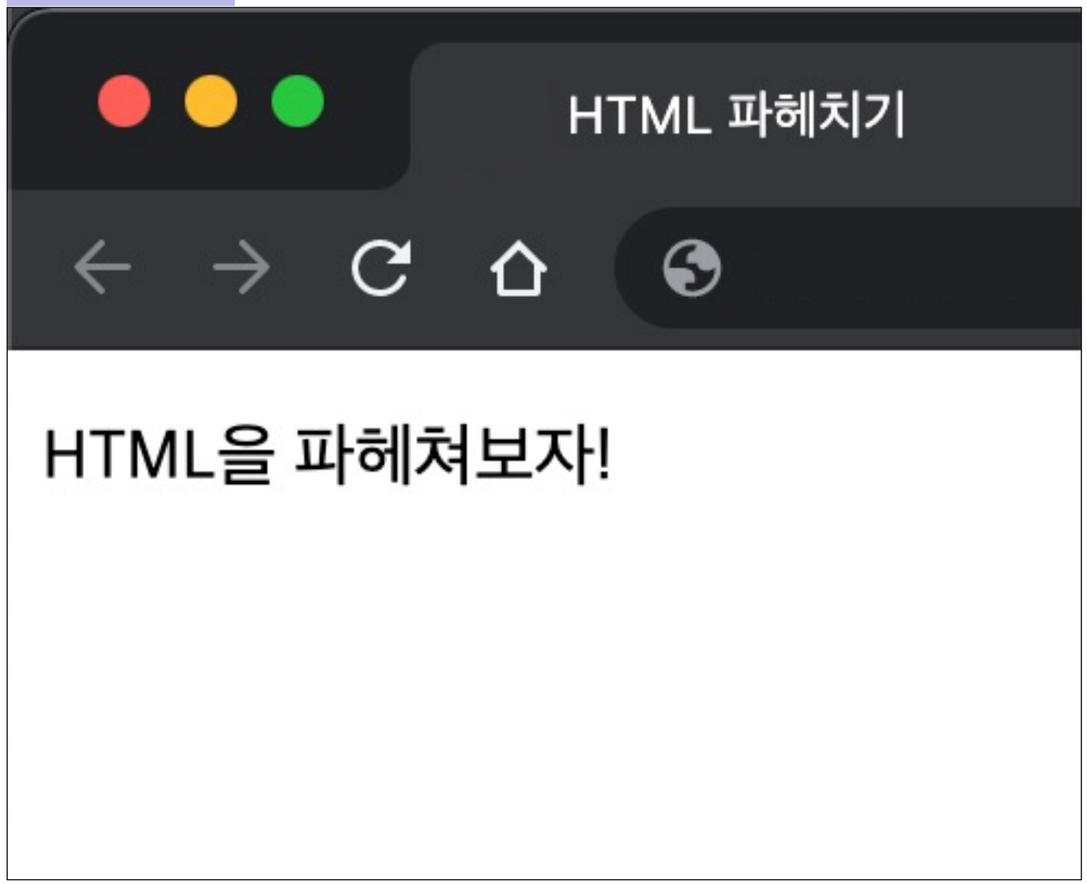
웹사이트에서 **눈에 보이는 정보**나 **특정 구역**을 설정할 때 사용하는 언어

간단한 웹 페이지 예시

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <title>HTML 파헤치기</title>
  </head>
  <body>
    <p>HTML을 파헤쳐보자!</p>
  </body>
</html>
```

실행결과



HTML과 요소

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <title>HTML 파헤치기</title>
  </head>
  <body>
    <p>HTML을 파헤쳐보자!</p>
  </body>
</html>
```

head 요소

title 요소

HTML

- 일련의 **요소**들로 이루어짐

요소 (Element)

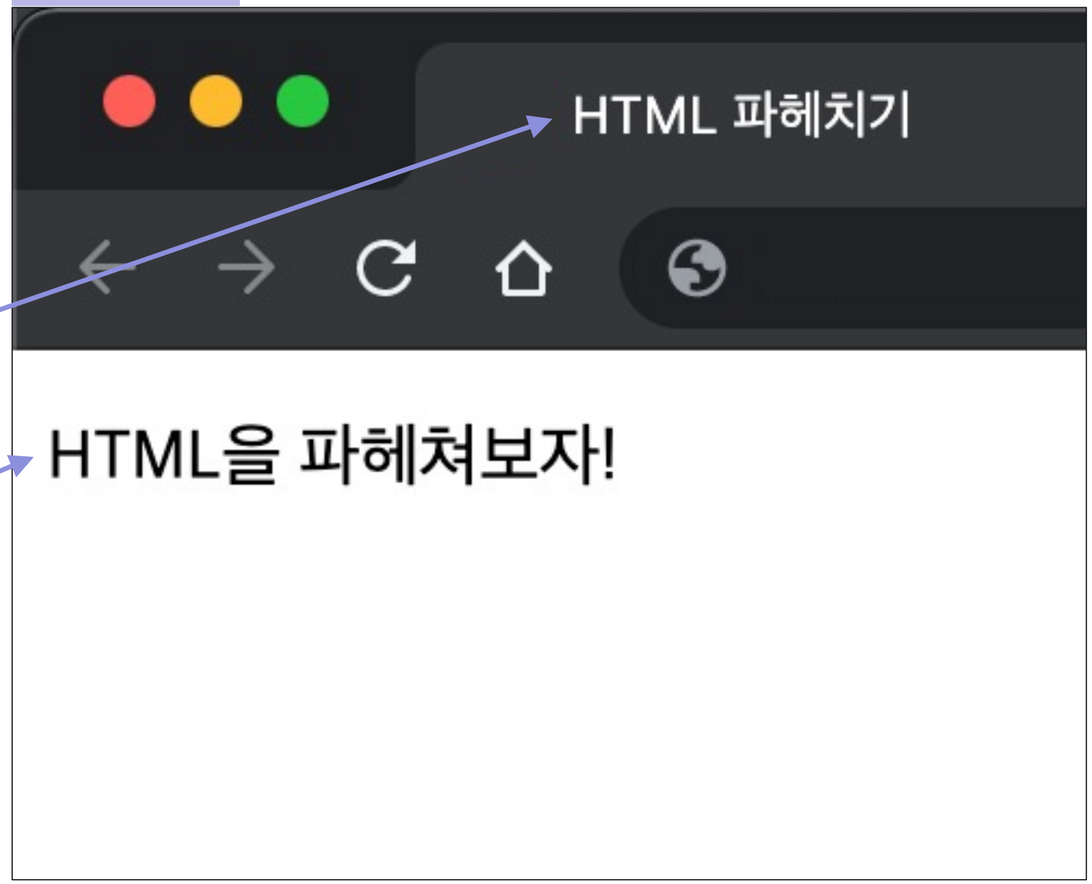
- 웹 페이지를 구성하는 것
- 콘텐츠(텍스트, 이미지 등)를 **다른 형식**으로 보이게 하거나, **특정한 방식**으로 동작하게 함

요소에 따라 달라지는 위치

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <title>HTML 파헤치기</title>
  </head>
  <body>
    <p>HTML을 파헤쳐보자!</p>
  </body>
</html>
```

실행결과



요소 (Element)

요소 작성법

```
<여는 태그 속성 = "속성값"> 텍스트 컨텐츠 </닫는 태그>
```

태그

- 요소의 특징을 나타냄
- “<”, **태그 이름**, “>”으로 작성

텍스트 컨텐츠

- 요소가 나타낼 텍스트

요소 (Element)

요소 작성법

```
<여는 태그 속성 = "속성값"> 텍스트 컨텐츠 </닫는 태그>
```

속성

- 태그가 가진 **기본 특징**이나 **동작**을 변경할 수 있게 해줌

속성값

- 해당 속성의 값
- 속성에 따라 **생략**될 수 있음

빈 요소

빈 요소 작성법

```
<여는 태그 속성 = "속성값">
```

특징

- 컨텐츠 및 닫는 태그가 없음

태그 파헤치기: <!DOCTYPE>, html

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <title>HTML 파헤치기</title>
  </head>
  <body>
    <p>HTML을 파헤쳐보자!</p>
  </body>
</html>
```

<!DOCTYPE html>

- HTML 문서임을 선언하는 요소
- 문서 맨 윗 줄에 작성하며, html 태그 안에 쓰지 않는다.

<html> … </html>

- HTML 문서의 시작과 끝을 의미
- 모든 다른 요소들은 **<html> 요소** 안에 입력

태그 파헤치기: head, body

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <title>HTML 파헤치기</title>
  </head>
  <body>
    <p>HTML을 파헤쳐보자!</p>
  </body>
</html>
```

<head> … </head>

- 웹사이트에 대한 설명(=메타데이터)을 적어 넣는 요소
- 정확히 한 개의 **<title> 요소**를 포함해야 함

<body> … </body>

- 웹사이트에 표시될 내용을 적어 넣는 요소
- 앞으로 배우는 대부분의 요소는 이곳에 작성

태그 파헤치기: meta, title

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <title>HTML 파헤치기</title>
  </head>
  <body>
    <p>HTML을 파헤쳐보자!</p>
  </body>
</html>
```

<meta>

- 빈 요소 (닫는 태그 없음)
- 해당 HTML 문서의 특징 및 설명을 적는 요소
- charset 속성으로 문자 인코딩 선언

<title> … </title>

- 웹 사이트의 제목을 적는 요소

태그 종합

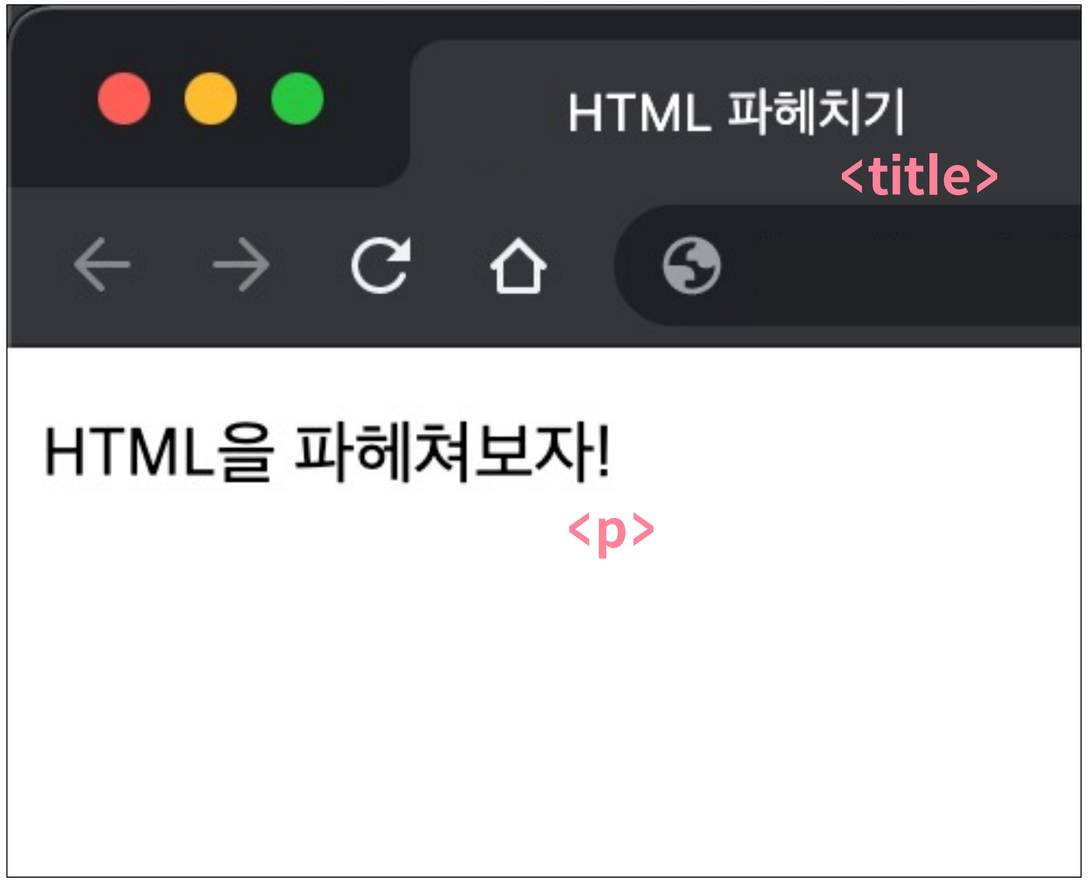
HTML

```
<!DOCTYPE html>          <!-- HTML5 문서 선언 -->
<html>                   <!-- HTML 문서의 시작과 끝 -->
  <head>                 <!-- 문서와 관련된 요약 정보 정리 -->
    <meta charset="utf-8">          <!-- 정보 및 설명 -->
    <title>HTML 파헤치기</title>    <!-- 웹사이트 제목 -->
  </head>
  <body>                 <!-- 웹사이트 내용 -->
    <p>HTML을 파헤쳐보자!</p>
  </body>
</html>
```

03

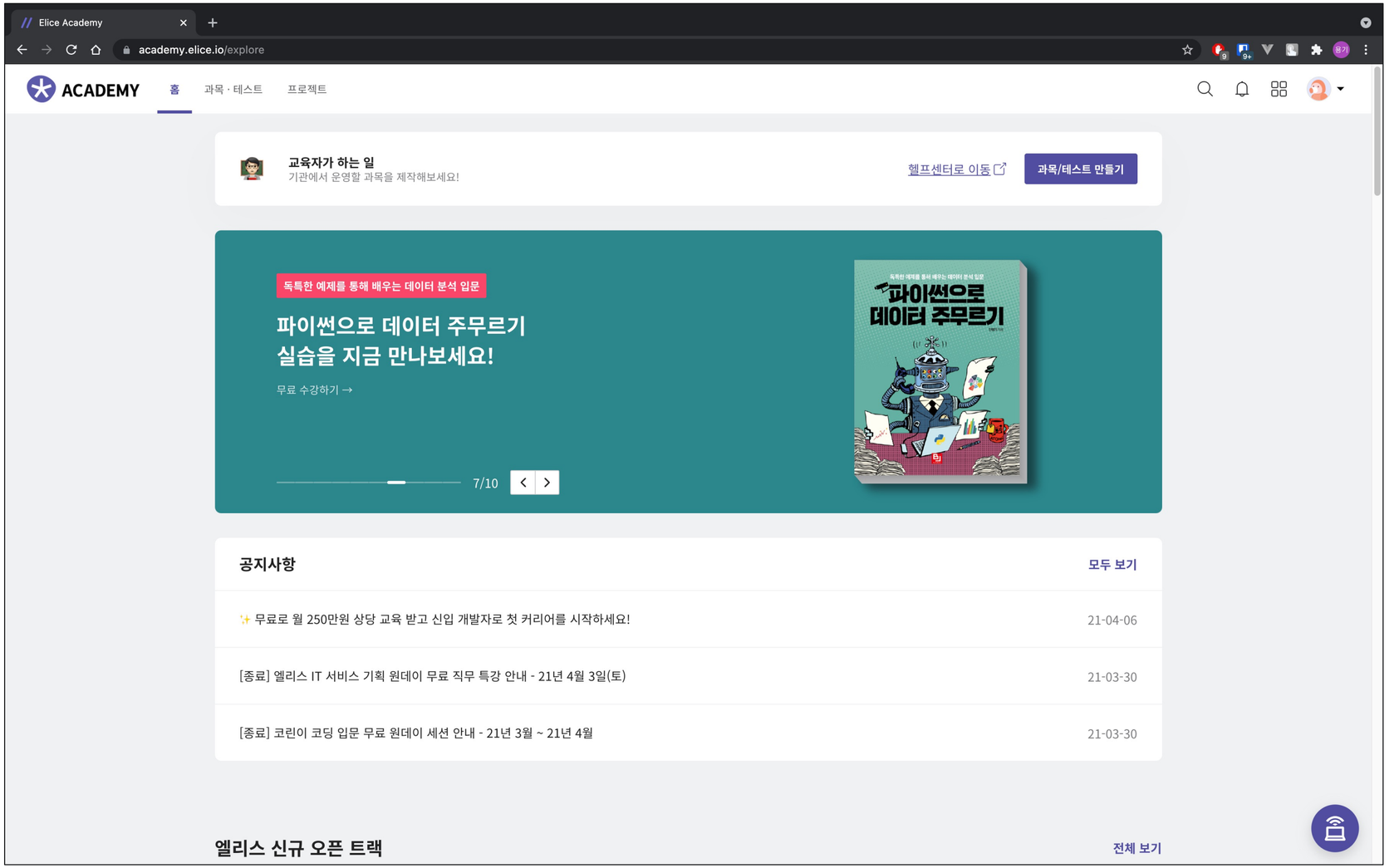
HTML - 태그 (2)

간단한 웹 페이지 예시



<body>의 영역

<body>



<body>를 꾸며주는 수많은 요소들

address article aside footer header h1 h2 h3 h4 h5 h6 main nav section blockquote dd div dl dt figcaption figure hr li ol p pre ul a abbr b bdo br cite code data dfn em i kbd mark q rb rp rt rtc ruby s samp small span strong sub sup time u var wbr area audio img map track video embed iframe object param canvas noscript script caption col colgroup table tbody td th thead tr button datalist fieldset form input label legend meter optgroup option output progress select textarea details dialog menu summary slot template

<body>를 꾸며주는 수많은 요소들

address article aside footer header h1 h2 h3 h4 h5 h6 main nav section
blockquote dd div dl dt figcaption figure hr li ol p pre ul a abbr b bdo br cite
code data dfn em i kbd mark q rb rp rt rtc ruby s samp small span strong sub
sup time u var wbr area audio img map track video embed iframe object param
canvas noscript script caption col colgroup table tbody td th thead tr button
datalist fieldset form input label legend meter optgroup option output
progress select textarea details dialog menu summary slot template

여러 웹 페이지에서 **일반적**으로 사용하고, **직관적**이어서
이해하기 쉬운 요소 위주로 학습

태그 파헤치기: p

HTML

```
<p>Paragraph 1
   Paragraph 1</p>
<p>Paragraph 2
   Paragraph 2</p>
```

실행결과

Paragraph 1 Paragraph 1

Paragraph 2 Paragraph 2

…

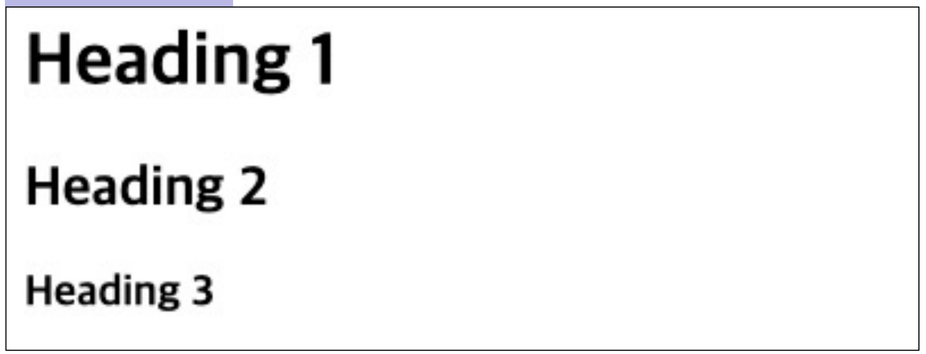
- paragraph의 약자로, **문단**을 나타냄
- 문단 간에는 한 줄의 간격 생성

태그 파헤치기: h

HTML

```
<h1>Heading 1</h1>
<h2>Heading 2</h2>
<h3>Heading 3</h3>
```

실행결과



<h?> … </h?>

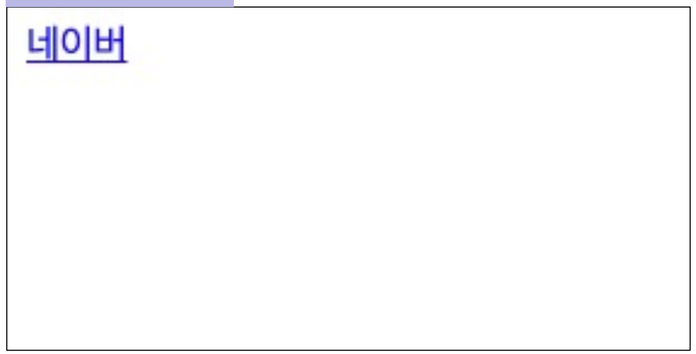
- heading의 약자로 **제목** 또는 **부제목**을 표현
- 숫자 값이 클수록 폰트 사이즈가 작음
- <h1> 태그는 가장 중요한 제목이므로, 하나의 html 문서에서 **한 번만** 사용되는 것이 옳다.

태그 파헤치기: a

HTML

```
<a href="naver.com" target="_blank">네이버</a>
```

실행결과



<a> …

- anchor의 줄임말로 다른 URL로 연결하는 **하이퍼링크**를 만드는 태그
- 콘텐츠에는 링크의 목적지에 대한 설명이 들어감

href

- 하이퍼링크의 목적지 **URL**을 나타내는 속성

target

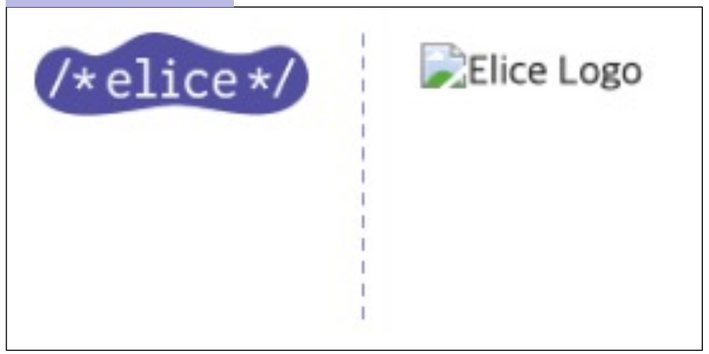
- 해당 링크를 보여줄 **위치**를 정하는 속성

태그 파헤치기: img

HTML

```
<img src="logo.png" alt="Elice Logo">
```

실행결과



- 빈 요소
- 문서에 **이미지**를 넣고자 할 때 사용하는 태그

src

- 보여주고자 하는 이미지의 **경로**
- **필수** 속성

alt

- 이미지의 **대체 텍스트**

04

HTML - 태그 (3)

<body>를 꾸며주는 수많은 요소들

address article aside footer header h1 h2 h3 h4 h5 h6 main nav section
blockquote dd div dl dt figcaption figure hr li ol p pre ul a abbr b bdo br cite
code data dfn em i kbd mark q rb rp rt rtc ruby s samp small span strong sub
sup time u var wbr area audio img map track video embed iframe object param
canvas noscript script caption col colgroup table tbody td th thead tr button
datalist fieldset form input label legend meter optgroup option output
progress select textarea details dialog menu summary slot template

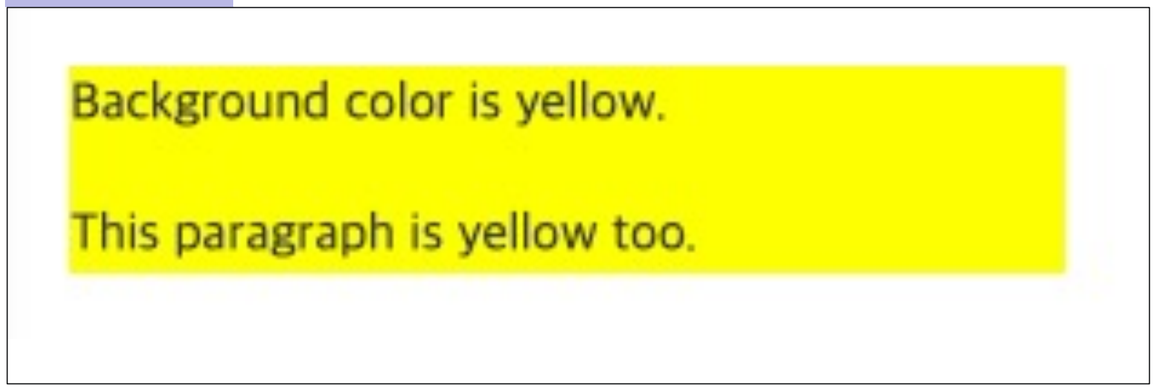
여러 웹 페이지에서 **일반적**으로 사용하고, **직관적**이어서
이해하기 쉬운 요소 위주로 학습

태그 파헤치기: div

HTML

```
<div style="background-color: yellow;">
  <p>Background color is yellow.</p>
  <p>This paragraph is yellow too.</p>
</div>
```

실행결과



…

- 요소 자체는 아무것도 표현하지 않는 **컨테이너**
- 여러 요소를 하나의 **구역**으로 묶어 꾸미기 쉽게 해줌

태그 파헤치기: div

HTML

```
<div style="background-color: yellow;">
  <p>Background color is yellow.</p>
  <p>This paragraph is yellow too.</p>
</div>
```

실행결과

Background color is yellow.

This paragraph is yellow too.

```
<p style="background-color: yellow;">
  Background color is yellow.</p>
<p style="background-color: yellow;">
  This paragraph is yellow too.</p>
```

Background color is yellow.

This paragraph is yellow too.

태그 파헤치기: span

HTML

```
<p>
  Elice is the
  <span style="color: red;">best</span>
  platform of the world.
</p>
```

실행결과

Elice is the best platform of the world.

…

- 요소도 자체만으론 아무 것도 표현하지 않음
- 주로 **문장 중간**에서 일부분을 꾸며줄 때 사용

태그 파헤치기: span

HTML

```
<p>
  Elice is the
  <span style="color: red;">best</span>
  platform of the world.
</p>
```

실행결과

Elice is the best platform of the world.

```
<p>
  Elice is the
  <div style="color: red;">best</div>
  platform of the world.
</p>    <!-- 잘못된 문법 -->
```

Elice is the

best

platform of the world.

태그 파헤치기: ul

HTML

```
<ul>
    <li>HTML</li>
    <li>CSS</li>
    <li>JavaScript</li>
</ul>
```

실행결과

- HTML
- CSS
- JavaScript

 …

- 비정렬 목록
- **순서가 없는** 리스트를 표현할 때 사용
- **불릿**으로 표현됨

 …

- 목록의 **항목**
- 정렬 태그인 **<ol>, ** 안에서 사용해야 함

태그 파헤치기: ol

HTML

```
<ol>
    <li>HTML</li>
    <li>CSS</li>
    <li>JavaScript</li>
</ol>
```

실행결과

1. HTML
2. CSS
3. JavaScript

 …

- 정렬 목록
- **순서가 있는** 리스트를 표현할 때 사용
- **숫자 또는 알파벳**의 오름차순으로 표현됨

05

HTML - 전역 속성

태그와 속성

예시

```
<a href="naver.com" target="_blank">네이버</a>
<img src="logo.png" alt="Elice Logo">
```

href, target 속성은 **<a> 태그,**
src, alt 속성은 **<img> 태그**와 짝을 이룸

전역 속성 (global attributes)

모든 요소가 **공통**으로 사용할 수 있는 속성

속성 파헤치기: style

HTML

```
<div style="background-color: yellow;">
  <p style="color: red;">Red</p>
  <p style="color: red;
    font-size: 24px;">Big Red</p>
</div>
```

실행결과

style

- 요소에 **CSS 스타일**을 적용
- **테스트 용도**로 적합

속성 파헤치기: class

CSS

```css
.warning { background-color: yellow; }
.red { color: red; }
.big { font-size: 24px; }
```

HTML

```html
<div class="warning">
  <p class="red">Red</p>
  <p class="big red">Big Red</p>
</div>
```

실행결과

class

- 특정 요소에 스타일 또는 스크립트를 적용하기 위해 사용
- 각 요소는 여러 **class**를 가질 수 있음

속성 파헤치기: id

CSS

```css
.warning { background-color: yellow; }
.red { color: red; }
.big { font-size: 24px; }
#black-p { color: black; }
```

HTML

```html
<div class="warning">
  <p id="black-p">Red</p>
  <p class="big red">Big Red</p>
</div>
```

실행결과

id

- 특정 요소에 스타일 또는 스크립트를 적용하기 위해 사용
- 각 요소는 하나의 id만 가질 수 있음

속성 파헤치기: hidden

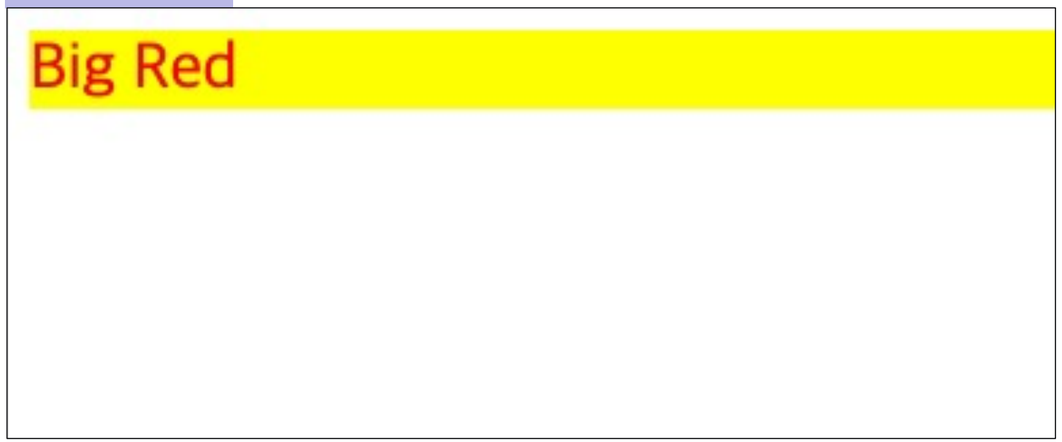
CSS

```css
.warning { background-color: yellow; }
.red { color: red; }
.big { font-size: 24px; }
#black-p { color: black; }
```

HTML

```html
<div class="warning">
  <p hidden">Red</p>
  <p class="big red">Big Red</p>
</div>
```

실행결과



hidden

- 해당 요소를 보이지 않게 함
- `hidden="hidden"` 대신 간략히 hidden만 적을 수 있음

속성 파헤치기: title

HTML

```
<img src="html5.png" title="HTML5">
<p>html 로고</p>
```

실행결과

html 로고

title

- 요소와 관련된 툴팁 제공

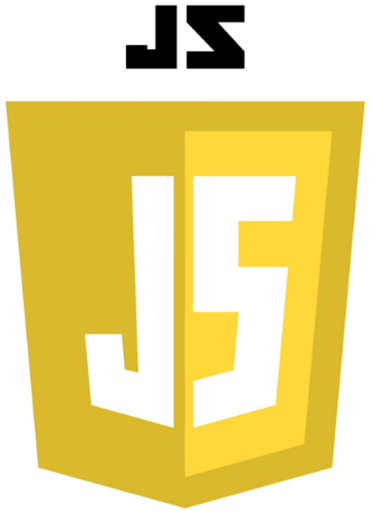
06

맺으며

웹 페이지 구성

- HTML
: 정보 및 설계도

- CSS
: 디자인 및 스타일링



- JavaScript
: 기능과 효과

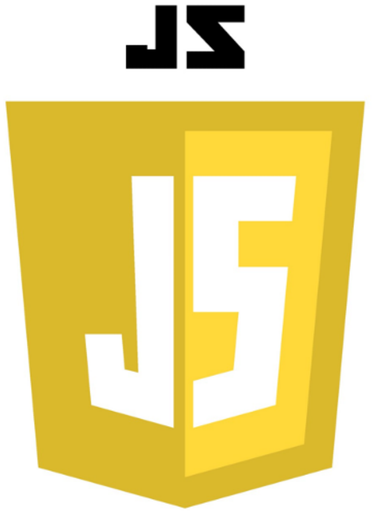
웹 페이지 구성

- HTML

: **정보** 및 설계도

- CSS

: 디자인 및 스타일링



- JavaScript

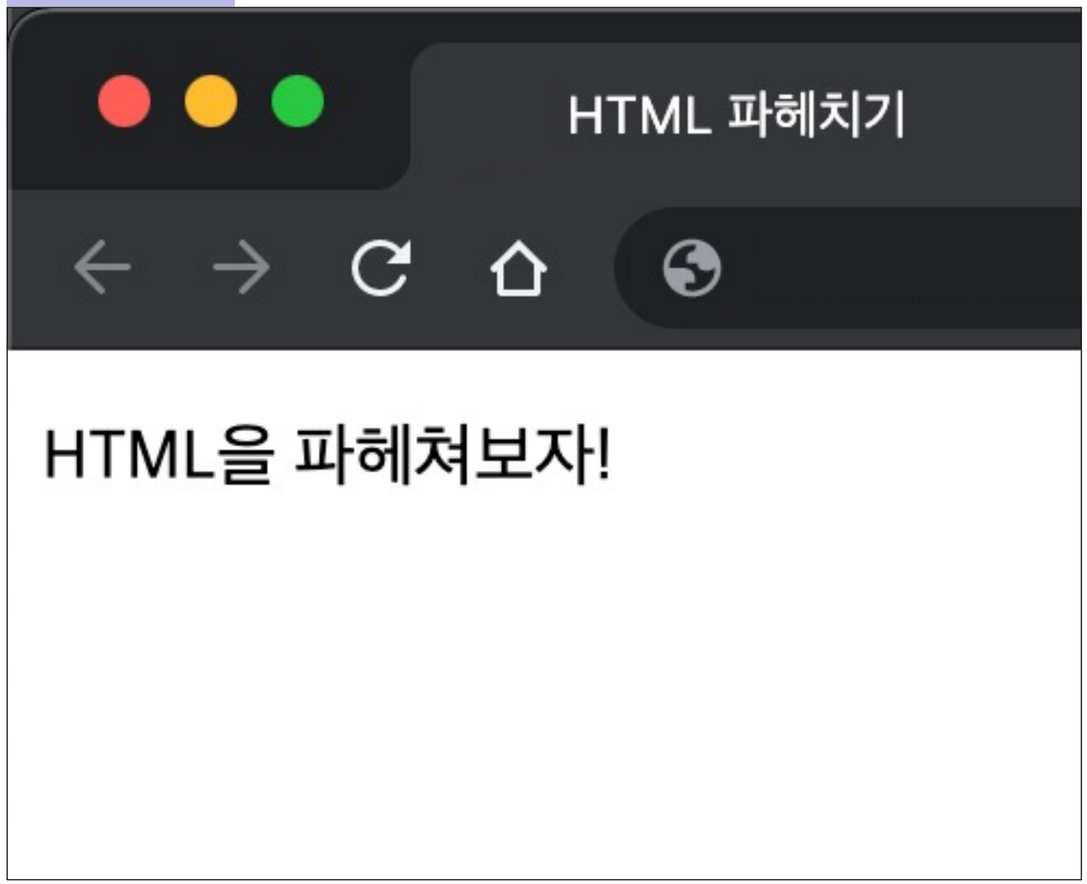
: 기능과 효과

간단한 웹 페이지 예시

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <title>HTML 파헤치기</title>
  </head>
  <body>
    <p>HTML을 파헤쳐보자!</p>
  </body>
</html>
```

실행결과



태그 1, 2

요소명	내용
!DOCTYPE	html 문서임을 선언하는 요소
html	html 문서의 최상단 요소 (root 요소)
head	문서의 요약 정보가 담김
body	웹 페이지에 표시될 내용이 담김
meta	문서를 설명하는 정보
title	웹 페이지의 제목
p	문단
h1 ~ h6	각 구획의 제목
a	하이퍼링크
img	이미지

태그 3, 전역 속성

요소명	내용
div	여러 요소를 한 구역에 묶음
span	요소 중간에서 꾸며줌
li	리스트의 항목 (ul, ol 안에서 사용)
ul	순서 없는 리스트 (비정렬 목록)
ol	순서 있는 리스트 (정렬 목록)

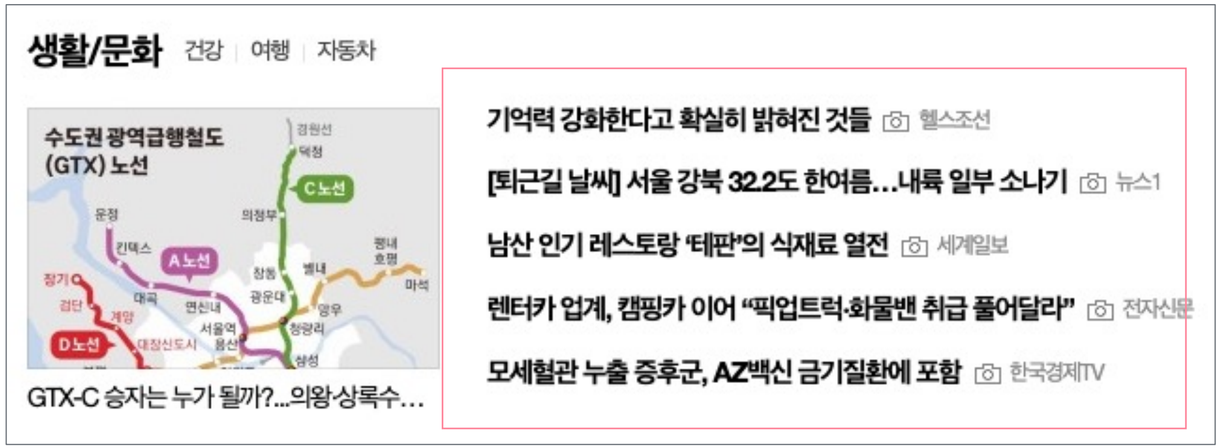
속성명	내용	
style	CSS 문법 적용, 테스트 용도 적합	
class	스타일, 스크립트를 특정 요소에 적용할 수 있게 해줌	여러 개 적용 가능
id	스타일, 스크립트를 특정 요소에 적용할 수 있게 해줌	한 개만 적용 가능
hidden	요소가 웹에서 보이지 않게 함	
title	툴팁 제공	

다음 수업 예고

Python

Selenium

다음 수업 예고

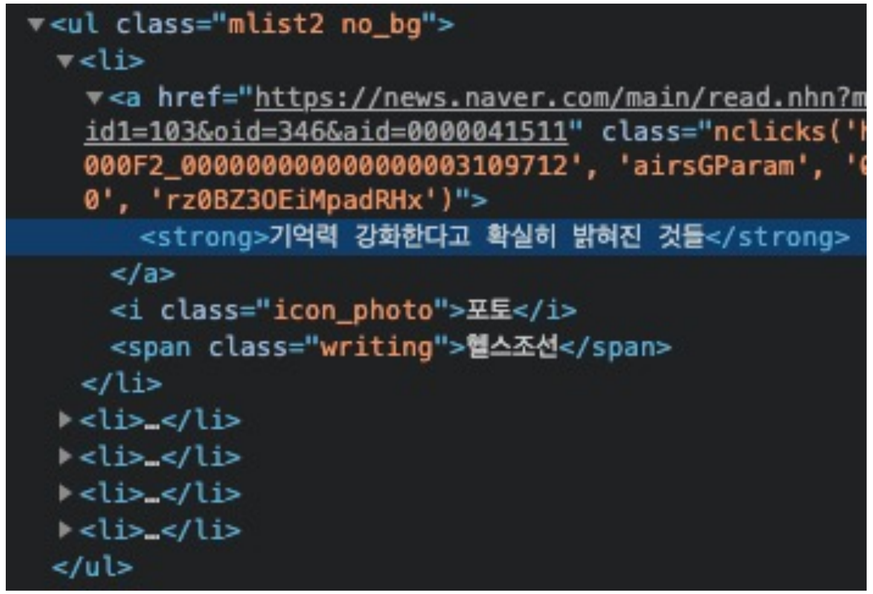


네이버 뉴스 헤드라인

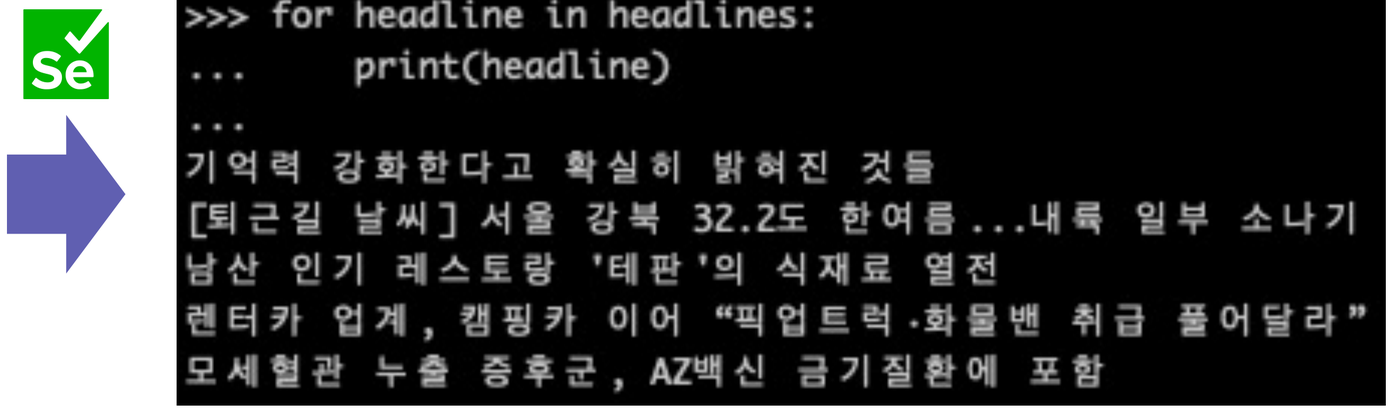
```
▼<ul class="mlist2 no_bg">
  ▼<li>
    ▼<a href="https://news.naver.com/main/read.nhn?m
    id1=103&oid=346&aid=0000041511" class="nclicks('
    000F2_00000000000000003109712', 'airsGParam', '
    0', 'rz0BZ3OEiMpadRHx')">
        <strong>기억력 강화한다고 확실히 밝혀진 것들</strong>
      </a>
      <i class="icon_photo">포토</i>
      <span class="writing">헬스조선</span>
    </li>
  ▶<li>…</li>
  ▶<li>…</li>
  ▶<li>…</li>
  ▶<li>…</li>
</ul>
```

헤드라인 구역의 html

다음 수업 예고



네이버 뉴스 헤드라인



스크래핑해온 데이터

크레딧

/* elice */

코스 매니저
임승연

콘텐츠 제작자
신용기

강사
신용기

감수자
장석준

디자이너
강혜정

연락처

TEL
070-4633-2015

WEB
https://elice.io

E-MAIL
contact@elice.io