

PRESENTATION

# Check in



**개강이 약 일주일 남았는 데,  
남은 일주일을 어떻게 보낼 예정인가요?**



20210220

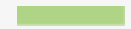
# Crawling



이윤정

PRESENTATION

# CONTENT



01. Internet의 원리

04. Robots.txt

02. Crawling 이란

05. 정적 수집

03. HTML 구조

06. 동적 수집



PRESENTATION

# Internet의 원리

- 01 Client가 Server에게 Contents를 요청한다. ( 예 : Url을 클릭하는 행위 )
- 02 Server는 요청 받은 Contents를 Client에게 건네준다.
- 03 Browser는 Server에게 받은 HTML을 해석하여 화면에 보여준다.



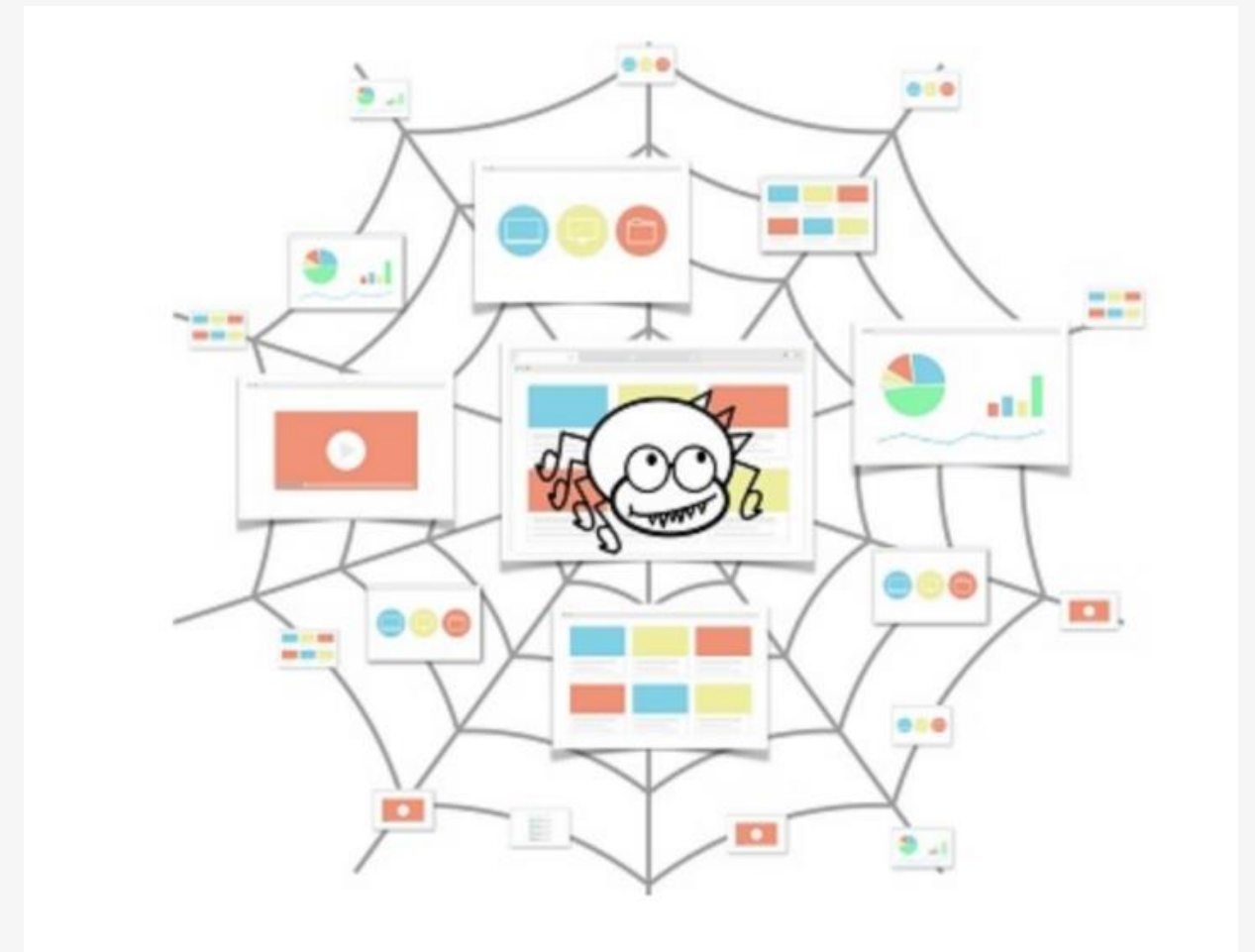
✓ Http : Client가 Server에 요청을 보내면, Server가 응답하는 방식 ( = 프로토콜 )

PRESENTATION

# Crawling 이란

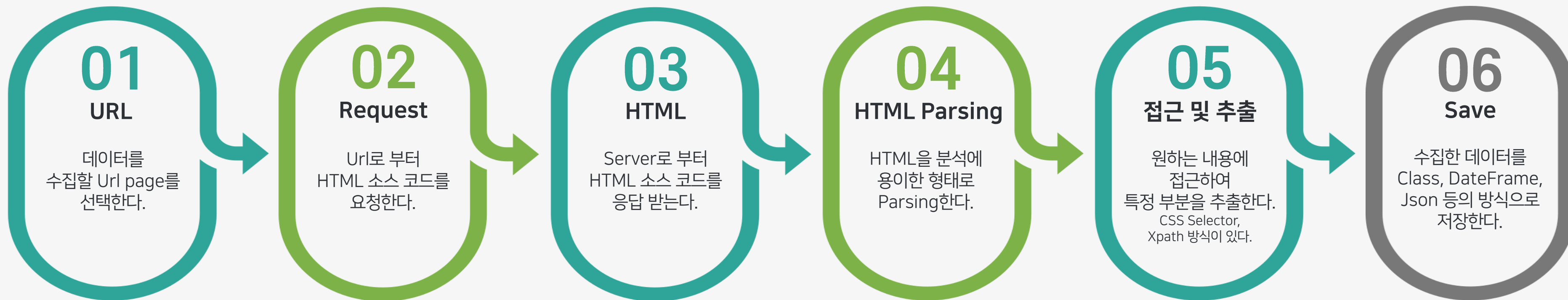
자동화된 방법으로 탐색하며 웹 상에 존재하는 정보들을 수집하는 작업

- 01 우리가 가지고 오고 싶은 부분을 눈으로 확인하고
- 02 HTML 구조를 분석해서 선택자를 선별하고 (태그는 div, class는 age구나)
- 03 선택자를 컴퓨터가 이해하도록 변환하는 작업 (div.age)



PRESENTATION

# Crawling의 과정



PRESENTATION

# Crawling의 종류



PRESENTATION

# HTML 이란

01 제목, 단락, 목록 등의 구조적 문서를 표현하기 위한 마크업 언어 ( : HyperText Markup Language)

02 이를 위해 <태그></태그>가 사용됨

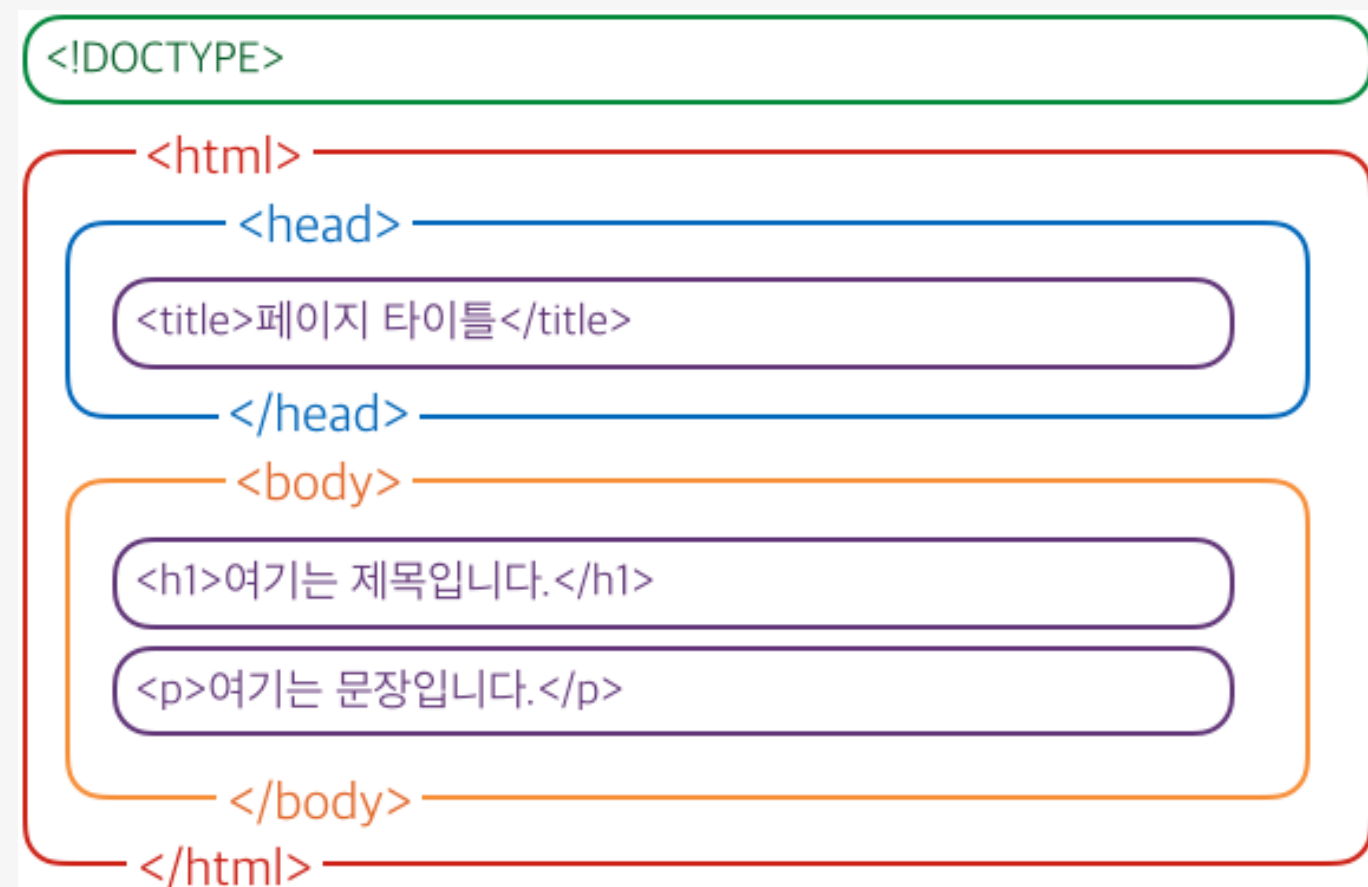
03 행동에 영향을 주는 JavaScript, 외관 및 배치를 정의하는 CSS 등의 스크립트 사용



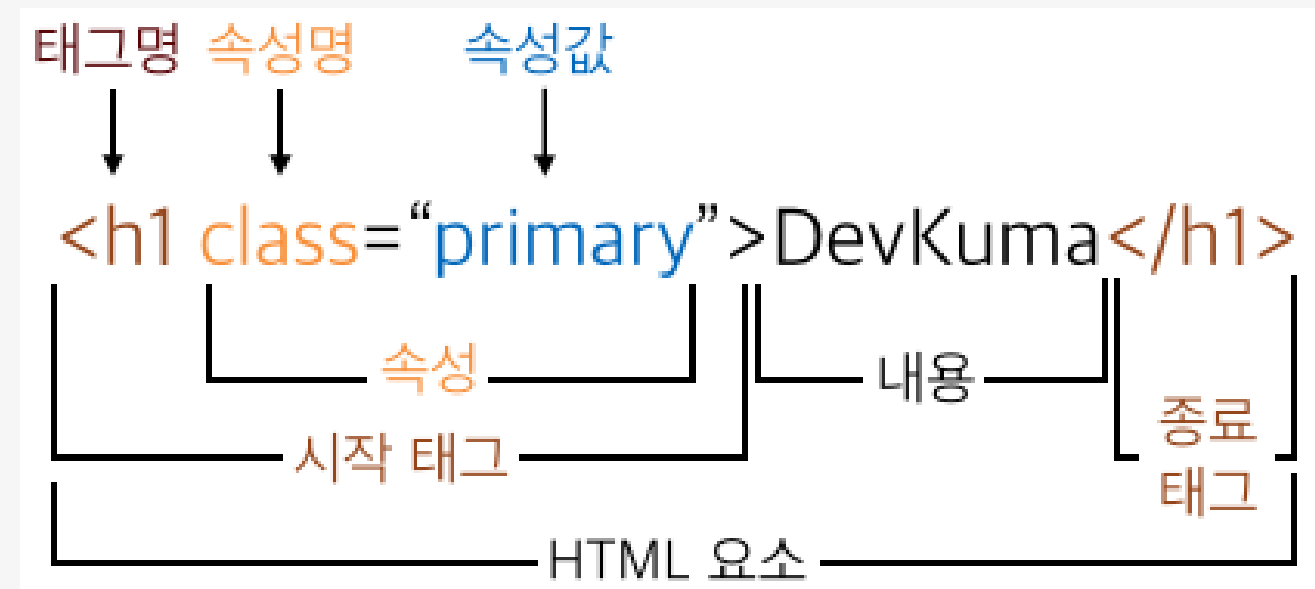


PRESENTATION

# HTML 구조



HTML 기본구조



문법 : <태그명 속성명="속성값">내용</태그명>

HTML 요소구조

✓ Div 태그 : 영역을 구분해주는 구역 구분자로 하나의 완벽한 독립된 영역을 가지는 성질을 지님

PRESENTATION

# Robots.txt



웹 로봇에서 크롤링을 할 수 있는 방법과 대상에 대한 검색 엔진 로봇과 같은 웹 로봇에 대한 접근 관련 내용을 나열한 파일  
크롤링 허가 · 불허의 여부 및 크롤링 허용 범위



웹 사이트의 루트 폴더에 텍스트 형식 (확장명 .txt)으로 저장



도메인을 입력하고 /robots.txt를 입력하여 접근 가능



다만, 명시 해놓지 않은 사이트도 많고 권고사항임을 유의할 것

```
User-agent: *
Disallow: /search
Allow: /search/about
Allow: /search/static
Allow: /search/howsearchworks
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*&
Allow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&*&gws_rd=ssl
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
Disallow: /imgres
Disallow: /u/
```

PRESENTATION

# 정적 수집

## 원리

01

Requests 라이브러리를 이용해서 사이트 요청

02

저장된 HTML 을 BeautifulSoup 라이브러리로 파싱

03

Beautifulsoup 객체에서 원하는 태그 검색

## 조건

1. 별다른 절차 없이 주소로 바로 접속해도 데이터를 볼 수 있다.
2. 새로 고침을 하지 않으면 페이지 안의 데이터가 변하지 않는다. - 정적 페이지

## 특징

1. 주소를 통해 단발적으로 접근한다.
2. 수집 대상에 한계가 있으나, 속도가 빠르다.



PRESENTATION

# Requests

HTTP 요청을 보내 원하는 HTML 정보를 가져오는 모듈

1. Web Server로부터 초기 HTML만 받을 뿐, 추가적으로 CSS/JavaScript 처리 X
2. 거의 모든 플랫폼에서 구동 가능

요청

1. Get
2. Post



PRESENTATION

# 동적 수집

## 원리

- 01 Webdriver를 통해 driver 객체 생성
- 02 Selenium 라이브러리를 이용해서 특정 사이트에 브라우저 접근
- 03 저장된 HTML 을 BeautifulSoup 라이브러리로 파싱
- 04 BeautifulSoup 객체에서 원하는 태그 검색

## 조건

1. 사용자의 액션에 따라 화면에 표현 » 무한 스크롤
2. JavaScript 등이 사용됨 - 동적 페이지

## 특징

1. 실제 브라우저를 띄어 이용하는 것과 동일 » 입력, 클릭 등 가능
2. 수집 대상에 한계가 없으나, 속도가 매우 느림



PRESENTATION

# Selenium

## 브라우저를 조작하면서 웹을 테스트하는 용도로 사용하는 프레임워크

1. 브라우저를 원격 컨트롤하는 테스트 라이브러리 - 브라우저 X
2. 클릭하거나 입력하는 작업을 할 수 있음 » 로그인 기능도 가능!
3. 브라우저를 띄워서 직접 스크롤을 내려주며 크롤링하기 위함

## 특징

1. Chrome Webdriver 필요 - 크롬 드라이버 저장 경로를 인자로 주어 웹드라이버 객체를 생성
2. 실제로 브라우저를 띄우기 때문에 로딩 시간을 고려하여 sleep문 사용 » 컴퓨터마다 구동 환경에 따라 달라질 수 있음



PRESENTATION

# BeautifulSoup

가공되지 않은 문자열에서 필요한 부분을 추출하여 의미있는 데이터로 만드는 과정

HTML 소스코드를 태그 기준으로 parsing 해주는 모듈

: 정규식을 작성할 필요 없이 tag, id, class 등으로 parsing 가능

## Parsing

- CSS Selector

1. HTML 문서의 서식을 지정하기 위해서 사용
2. 짧고 간단하게 특정 부분을 추출할 수 있음

- Xpath

1. 직접 접근하는 경로(path)를 지정
2. 복잡한 경로도 찾아갈 수 있음

# BeautifulSoup



PRESENTATION


# TIP



## 수집하고자 하는 정보가 어떤 태그를 가진 정보인지 확인하는 법

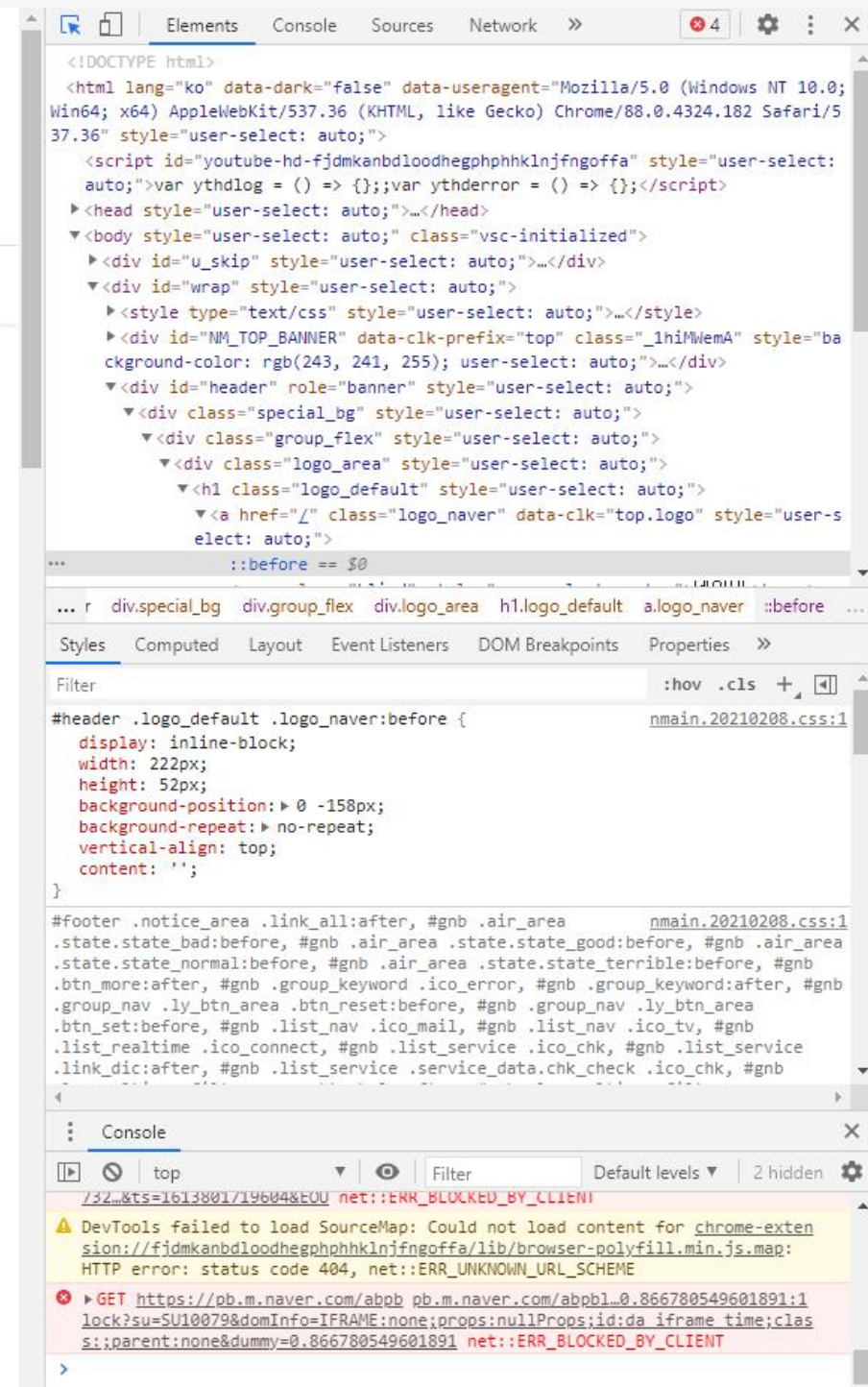
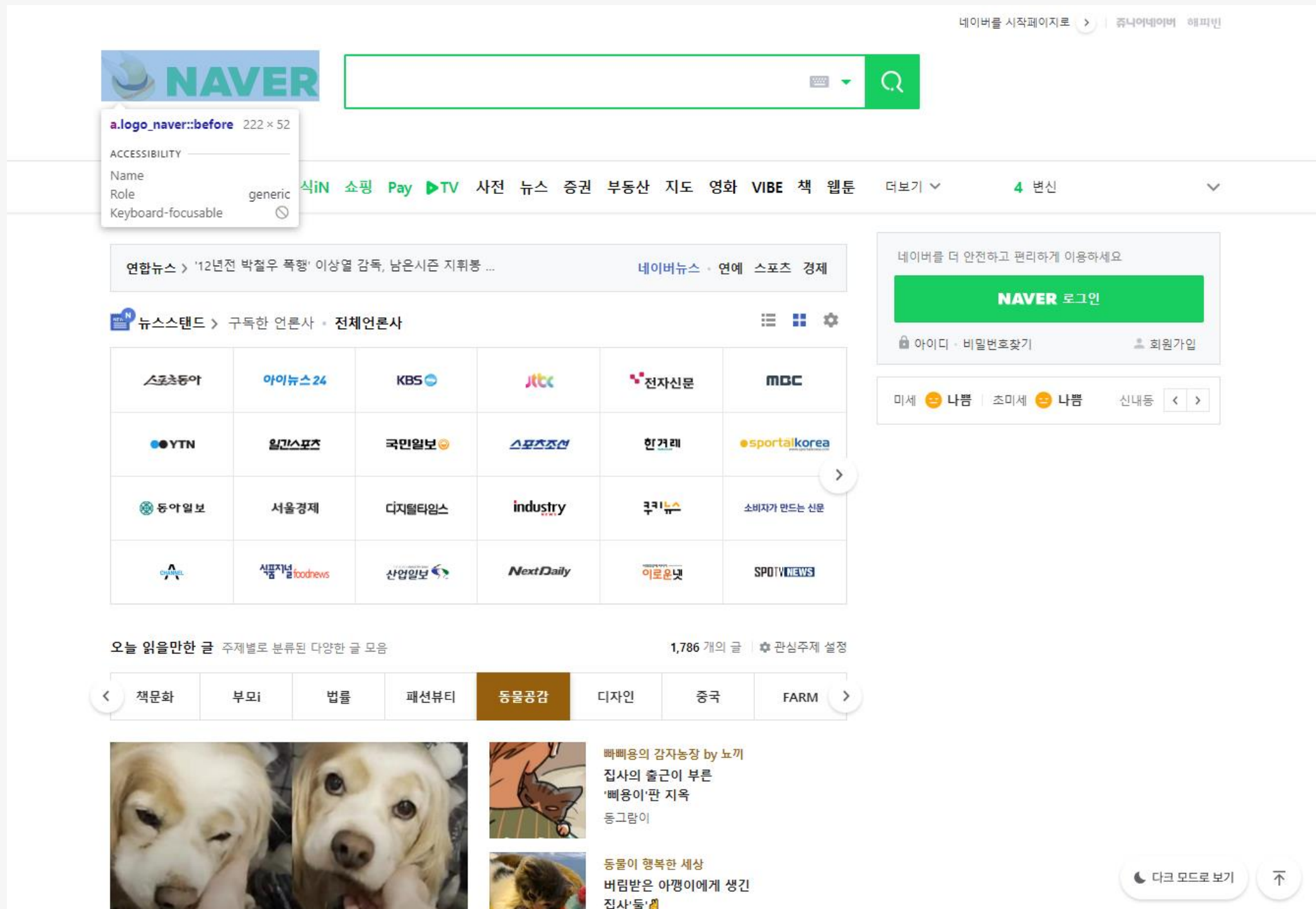
1. Windows : F12(개발자 도구)를 누른 다음, 패널 메뉴의 검사도구(Inspect)를 눌러 태그 확인
2. Mac : option + cmd + I 를 누른 다음, 패널 메뉴의 검사도구(Inspect)를 눌러 태그 확인

## 정적 웹 페이지, 동적 웹페이지 구분하기

1. 개발자 도구로 들어가기
  2. 오른쪽 상단 메뉴의 'Setting'으로 들어가기
  3. Preferences - Debugger - Disable JavaScript를 설정하기
  4. 이후 새로 고침 시 웹브라우저 상에서 동일한 증상이 보인다면 ≫ 동적 웹페이지!
- 



# PRESENTATION TIP



PRESENTATION

# 코드 비교



Github로 이동



**감사합니다!**

