

Web Scraping 기초

1-2. 웹 페이지와 HTML

웹 사이트와 웹 페이지

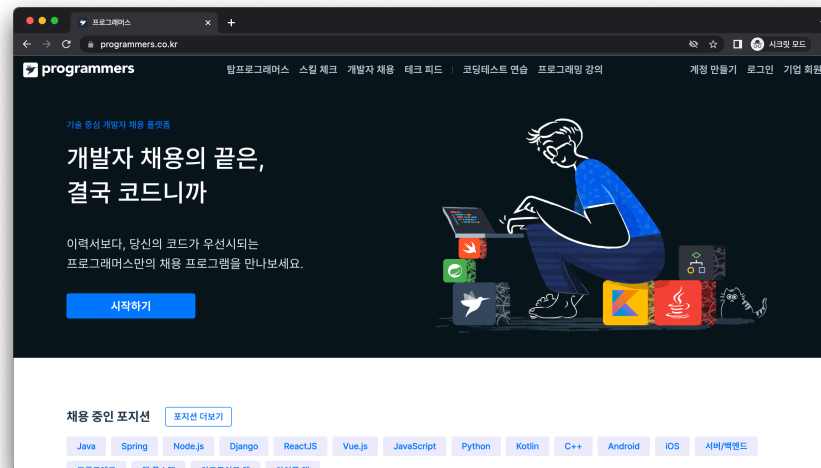
웹 페이지는 어떻게 만들까요?

HTML의 구조

웹 사이트와 웹 페이지

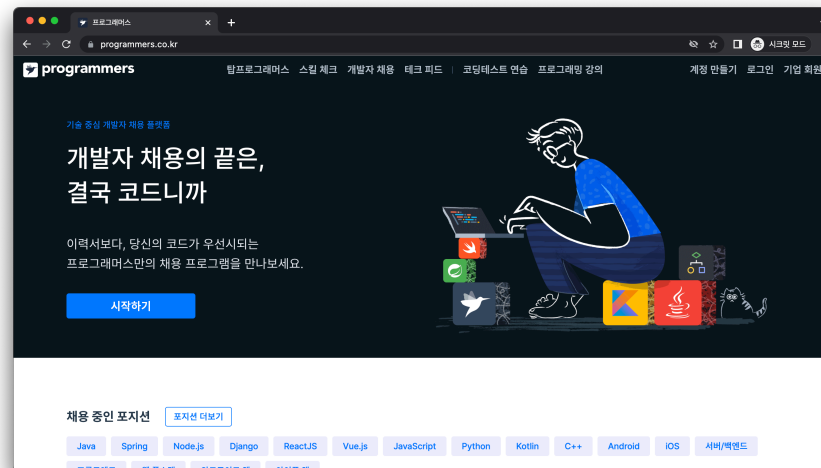
우리가 자주 사용하는 이것

프로그래머스 플랫폼의 정체는?



우리가 자주 사용하는 이것

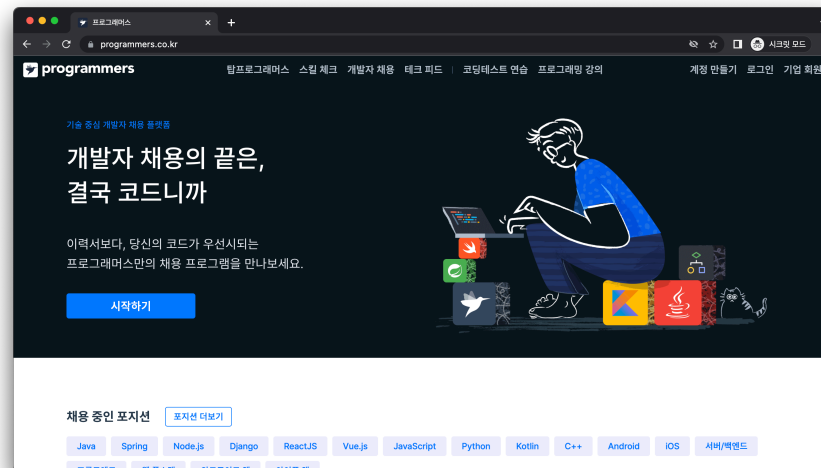
프로그래머스 플랫폼의 정체는?



웹 페이지? 웹 사이트?

우리가 자주 사용하는 이것

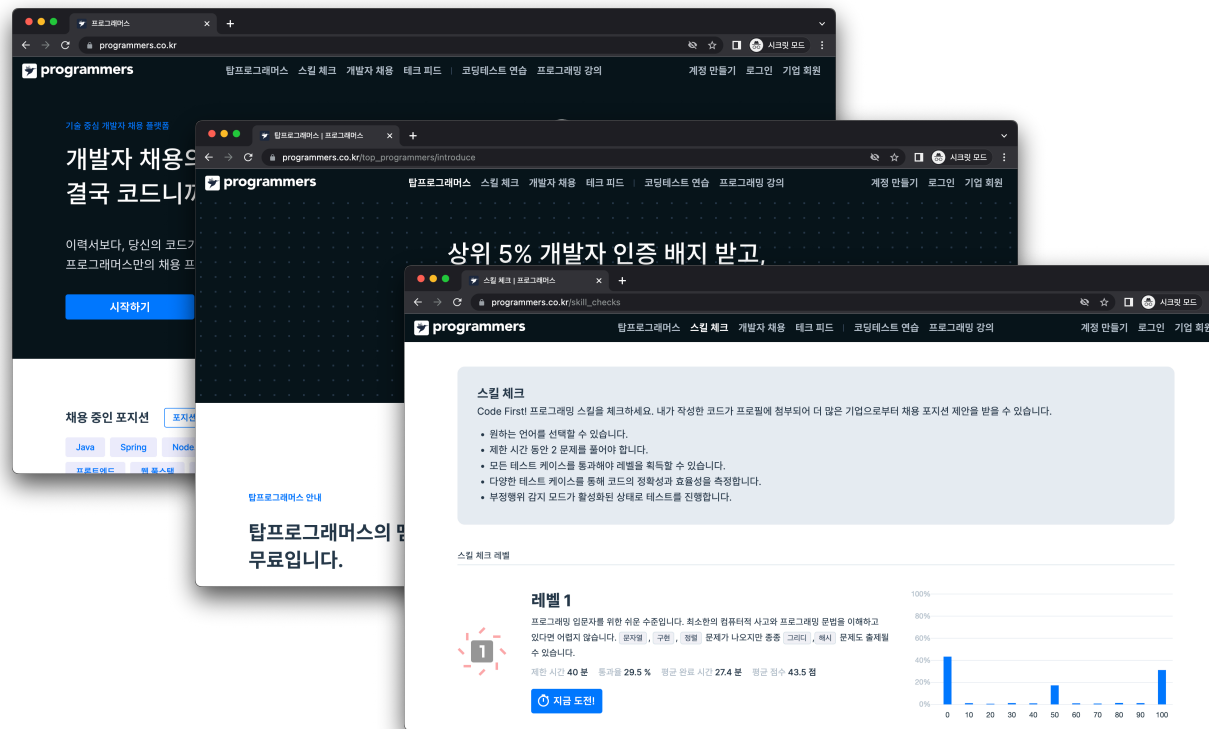
프로그래머스 플랫폼의 정체는?



웹 속에 있는 문서 하나는 웹 페이지

우리가 자주 사용하는 이것

프로그래머스 플랫폼의 정체는?

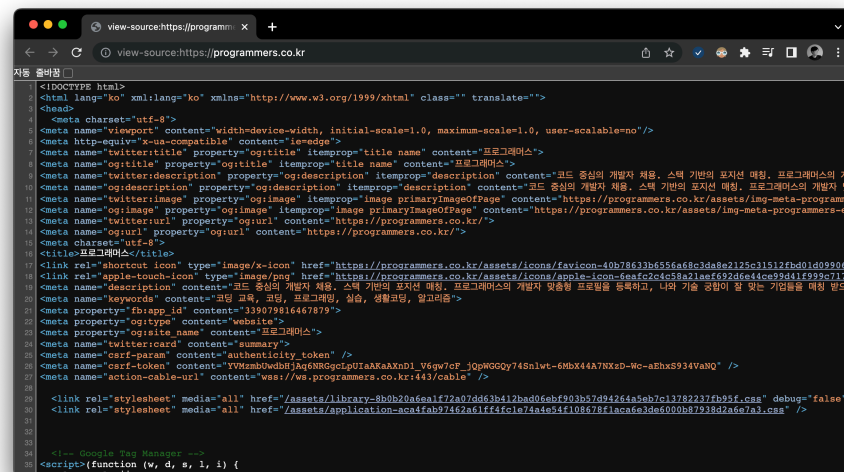
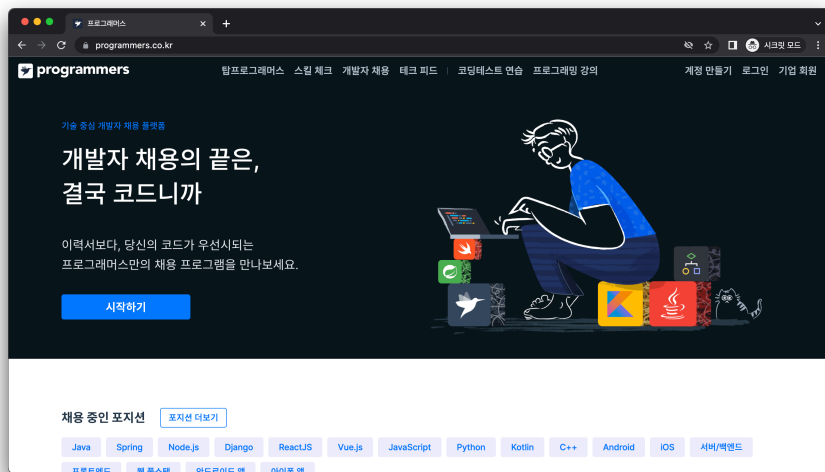


이런 웹 페이지의 모음은 웹 사이트

웹 페이지는 어떻게 만들까요?

HTTP와 이 이야기를 연결해보자!

웹 페이지와 웹 브라우저의 비밀



웹 페이지는 다음과 같은 엄청 복잡한 줄글로 되어있다

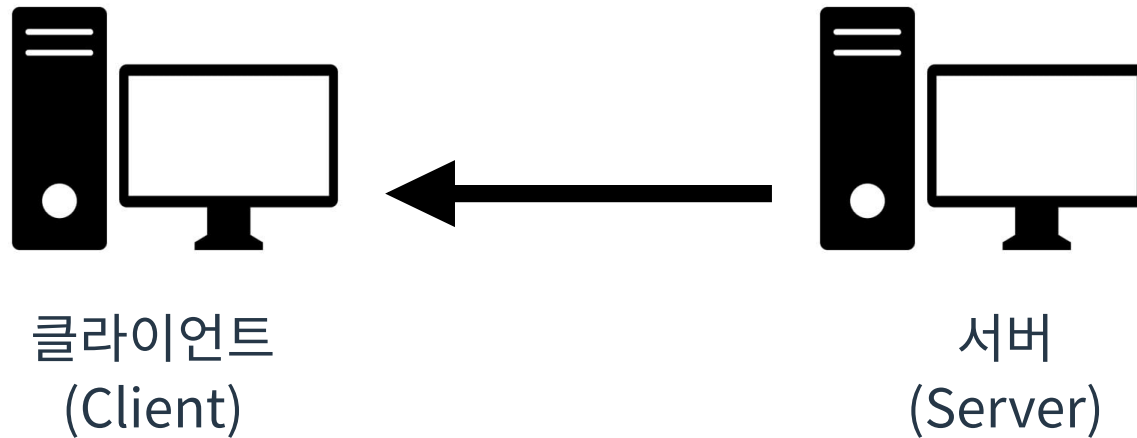
HTTP와 이 이야기를 연결해보자!

웹 페이지와 웹 브라우저의 비밀

HTTP/1.1 200 OK

...

<html>...</html>



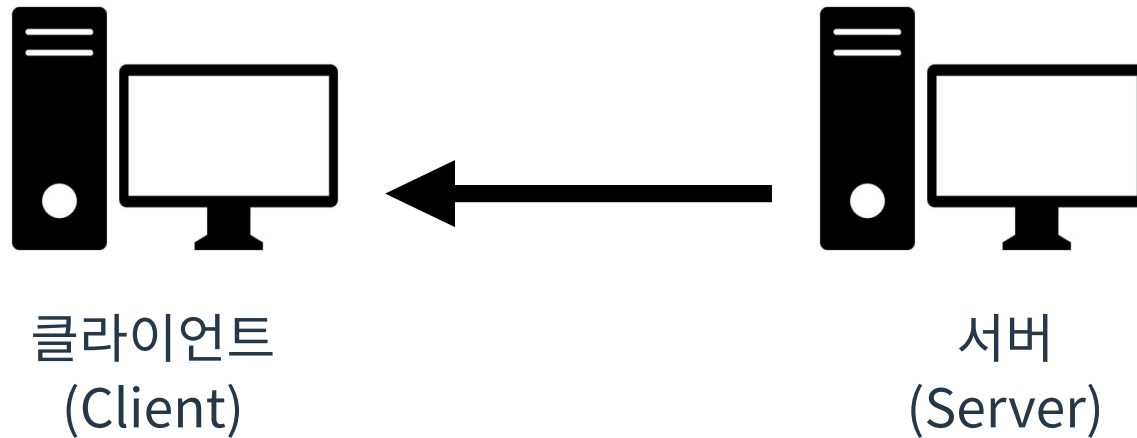
HTTP와 이 이야기를 연결해보자!

웹 페이지와 웹 브라우저의 비밀

HTTP/1.1 200 OK

...

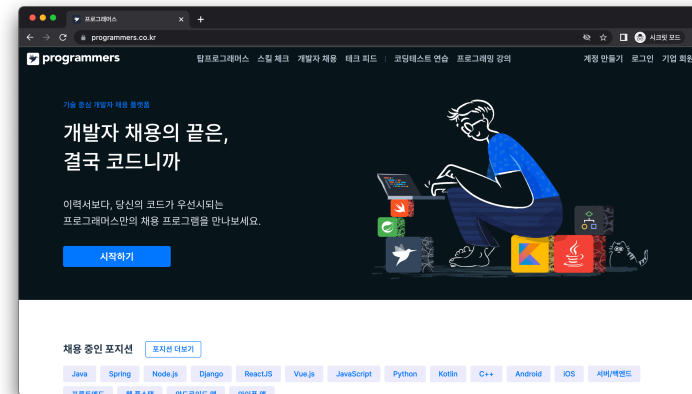
HTTP 응답의 Body! → `<html>...</html>`



HTTP와 이 이야기를 연결해보자!

웹 페이지와 웹 브라우저의 비밀

<html>...</html>



웹 브라우저는 HTML 요청을 보내고,
HTTP 응답에 담긴 HTML 문서를 우리가 보기 쉬운 형태로 **화면을 그려주는** 역할을 담당

웹 페이지와 웹 브라우저 이야기 요약

웹 페이지와 웹 브라우저의 비밀

웹 페이지는 HTML이라는 형식으로 되어있고,
웹 브라우저는 우리가 HTTP 요청을 보내고, 응답받은 HTML 코드를 렌더링 해주었습니다.

웹 페이지와 웹 브라우저 이야기 요약

웹 페이지와 웹 브라우저의 비밀

웹 페이지는 HTML이라는 형식으로 되어있고,
웹 브라우저는 우리가 HTTP 요청을 보내고, 응답받은 HTML 코드를 렌더링 해주었습니다.

저희는 지금부터 이 웹 브라우저의 역할을 코드로 대신 해보려고 합니다!

그 전에, HTML에 대해서 잘 알아야겠죠?

HTML의 구조

HTML?

웹 페이지 속 디자인은 어떻게 꾸밀까?



```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Document</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    <p>My name is Mussg!</p>
  </body>
</html>
```

HTML(HyperText Markup Language)

HTML 살펴보기

HTML의 여러 특징들

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Document</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    <p>My name is Mussg!</p>
  </body>
</html>
```

<!DOCTYPE html>를 통해 HTML5임을 명시

HTML 살펴보기

HTML의 여러 특징들

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Document</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    <p>My name is Mussg!</p>
  </body>
</html>
```

가장 바깥에 **<html>** 태그로 감싸져있다

HTML 살펴보기

HTML의 여러 특징들

여는 태그 `<...>` →

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Document</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    <p>My name is Mussg!</p>
  </body>
```

닫는 태그 `</...>` →

```
</html>
```

가장 바깥에 **<html>** 태그로 감싸져있다.

HTML 살펴보기

HTML의 여러 특징들

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Document</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    <p>My name is Mussg!</p>
  </body>
</html>
```

HTML 코드는 크게 **Head**와 **Body**로 나눌 수 있다

HTML 살펴보기

HTML의 여러 특징들

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Document</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    <p>My name is Mussg!</p>
  </body>
</html>
```

Head는 문서에 대한 정보 (제목, 언어 등)을 작성한다.

HTML 살펴보기

HTML의 여러 특징들

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Document</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    <p>My name is Mussg!</p>
  </body>
</html>
```

Body는 문서의 내용 (글, 이미지, 동영상 등)을 작성한다.

HTML 살펴보기

HTML의 여러 특징들

```
<tag>  
  Contents 1  
  Contents 2  
  ...  
</tag>
```

이렇게 HTML은 여러 **태그(Tag)**로 감싼 **요소(Element)**의 집합으로 이루어져있다!

HTML 살펴보기

HTML의 여러 특징들



한글, 워드 등은 상단바의 UI를 통해서 서식을 작성

HTML 살펴보기

HTML의 여러 특징들

```
<p>이것은 글씨입니다.</p>
```

```
<p><strong>진한 글씨는 이렇게 씁니다.</strong></p>
```

```
<h1>1번째 소제목은 이렇게 달 수 있죠</h1>  
<h2>2번째 소제목은 이렇게 달 수 있죠</h2>  
<h3>3번째 소제목은 이렇게 달 수 있죠</h3>
```

태그로 내용을 묶어 글의 형식을 지정

HTML 살펴보기

HTML의 여러 특징들

```
<p>이것은 글씨입니다.</p>
```

```
<p><strong>진한 글씨는 이렇게 씁니다.</strong></p>
```

```
<h1>1번째 소제목은 이렇게 달 수 있죠</h1>  
<h2>2번째 소제목은 이렇게 달 수 있죠</h2>  
<h3>3번째 소제목은 이렇게 달 수 있죠</h3>
```



이것은 글씨입니다.

진한 글씨는 이렇게 씁니다.

1번째 소제목은 이렇게 달 수 있죠

2번째 소제목은 이렇게 달 수 있죠

3번째 소제목은 이렇게 달 수 있죠

태그로 내용을 묶어 글의 형식을 지정

HTML 살펴보기

HTML의 여러 특징들

```
<p>이것은 글씨입니다.</p>
```

```
<p><strong>진한 글씨는 이렇게 씁니다.</strong></p>
```

```
<h1>1번째 소제목은 이렇게 달 수 있죠</h1>  
<h2>2번째 소제목은 이렇게 달 수 있죠</h2>  
<h3>3번째 소제목은 이렇게 달 수 있죠</h3>
```



이것은 글씨입니다.

진한 글씨는 이렇게 씁니다.

1번째 소제목은 이렇게 달 수 있죠






2번째 소제목은 이렇게 달 수 있죠

3번째 소제목은 이렇게 달 수 있죠

태그는 그에 맞는 속성(attribute)를 가지기도 한다

HTML 살펴보기

HTML의 여러 특징들

					
<a>	Yes	Yes	Yes	Yes	Yes
<u>download</u>	14.0	18.0	20.0	10.1	15.0
<u>href</u>	Yes	Yes	Yes	Yes	Yes
<u>hreflang</u>	Yes	Yes	Yes	Yes	Yes
<u>media</u>	Yes	Yes	Yes	Yes	Yes
<u>ping</u>	Yes	No	Yes	No	Yes
<u>referrerpolicy</u>	51.0	79.0	50.0	11.1	38.0
<u>rel</u>	Yes	Yes	Yes	Yes	Yes
<u>target</u>	Yes	Yes	Yes	Yes	Yes
<u>type</u>	Yes	Yes	Yes	Yes	Yes

참고: 웹 브라우저마다 지원하는 태그와 속성이 다르다...?!

웹 스크래핑 관점에서 HTML을 정리하면...

우리가 원하는 내용이 HTML 문서의
어디에 있지? 어떤 태그로 묶여있지?
를 관찰해야 한다.

End of Contents

Thank You! :)