Web Scraping 기초

3-1. 동적 웹 페이지와의 만남

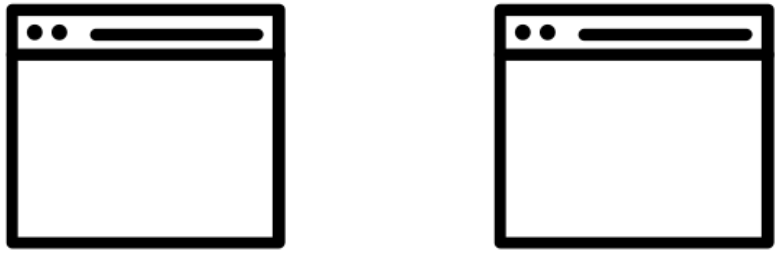
programmers

Contents

정적 웹 사이트와 동적 웹 사이트

웹 페이지는 크게 2가지!

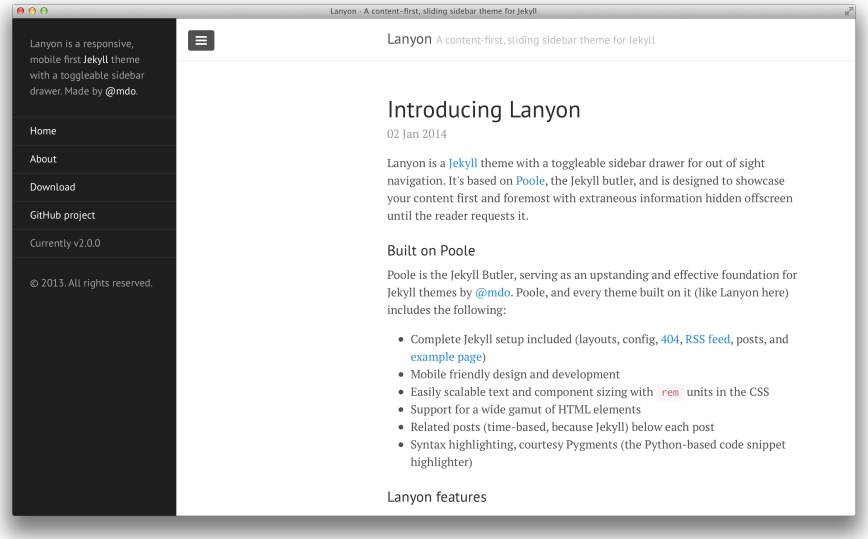
정적 웹 사이트와 동적 웹 사이트



웹 페이지는 어떻게 생성되냐에 따라 크게 2가지로 구분

웹 페이지는 크게 2가지!

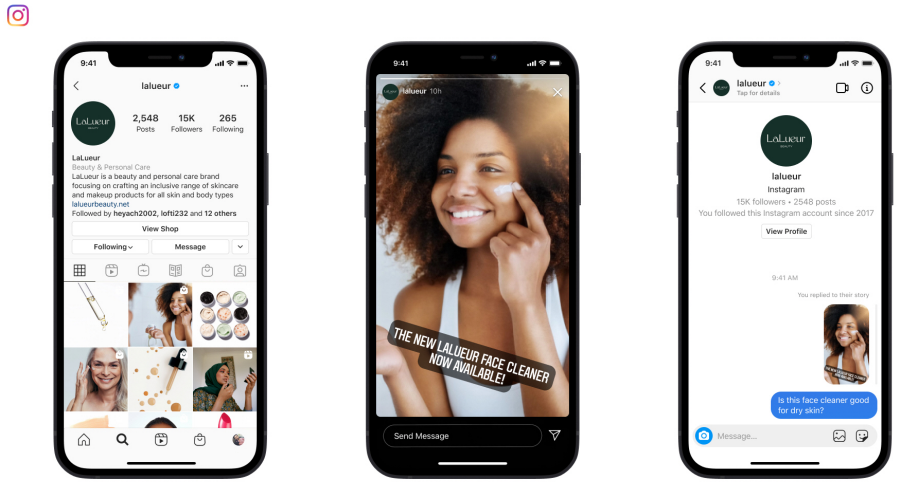
정적 웹 사이트와 동적 웹 사이트



HTML 내용이 고정된 **정적(static) 웹 사이트**

웹 페이지는 크게 2가지!

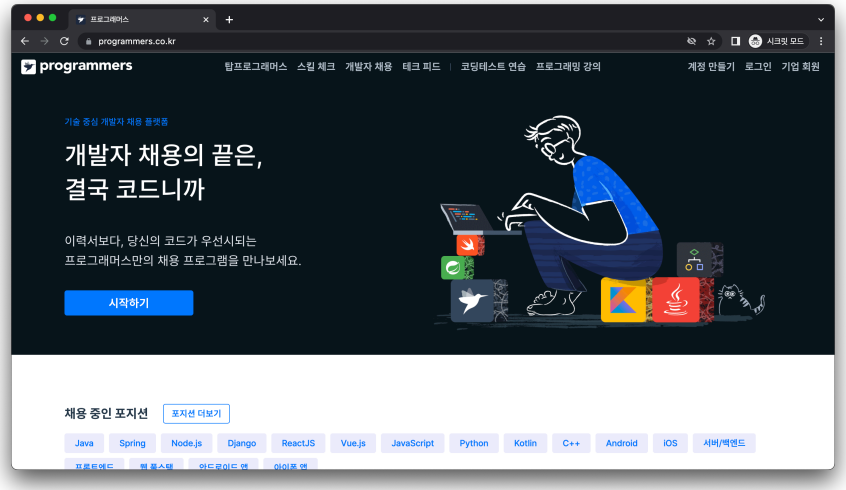
정적 웹 사이트와 동적 웹 사이트



HTML 내용이 변하는 **동적(dynamic) 웹 사이트**

각각의 특징은?

핵심은 정보의 완성 타이밍



정적 웹 사이트는 HTML 문서가 완전하게 응답된다

각각의 특징은?

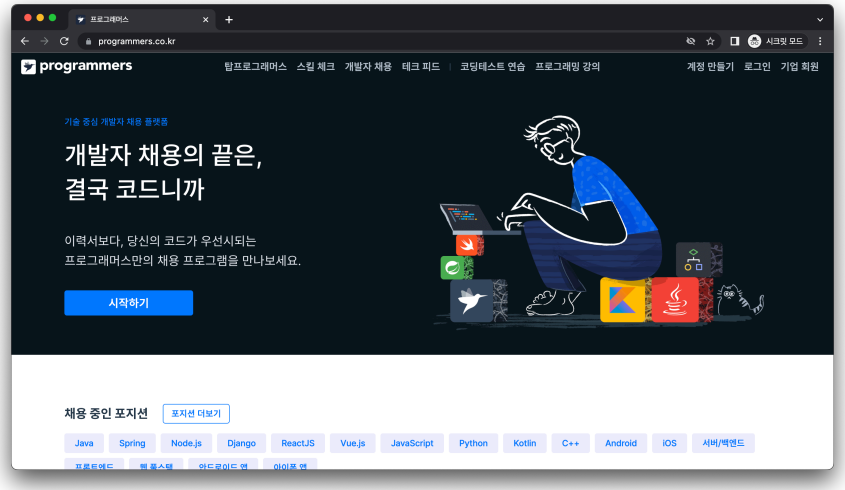
핵심은 정보의 완성 타이밍

동적 웹 사이트는 응답 후 HTML이 렌더링이 될 때까지의 **지연시간**이 존재!

동적 웹사이트의 동작 방식

동적 웹 사이트의 동작 방식

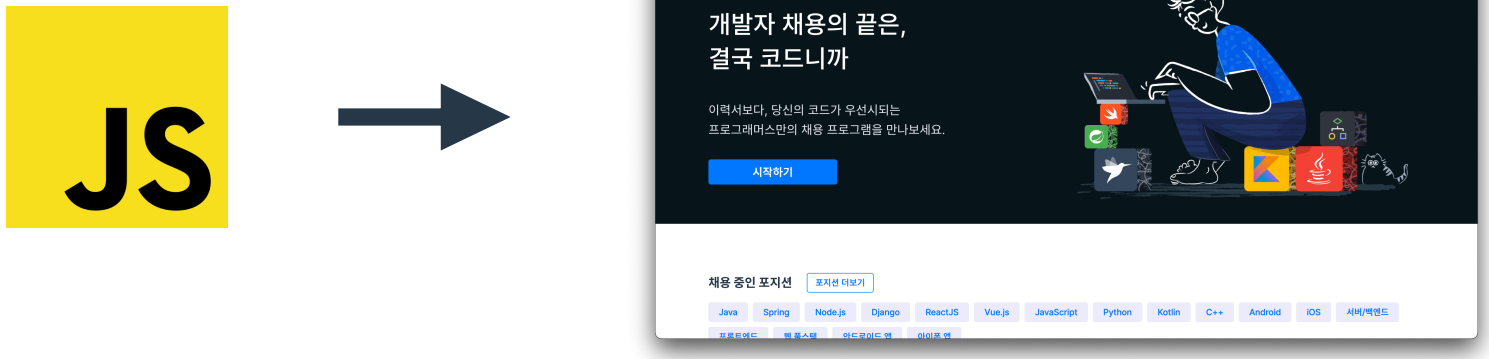
동적 웹 사이트는 어떻게 동작할까요?



웹 브라우저에선 **JavaScript**라는 프로그래밍 언어가 동작한다

동적 웹 사이트의 동작 방식

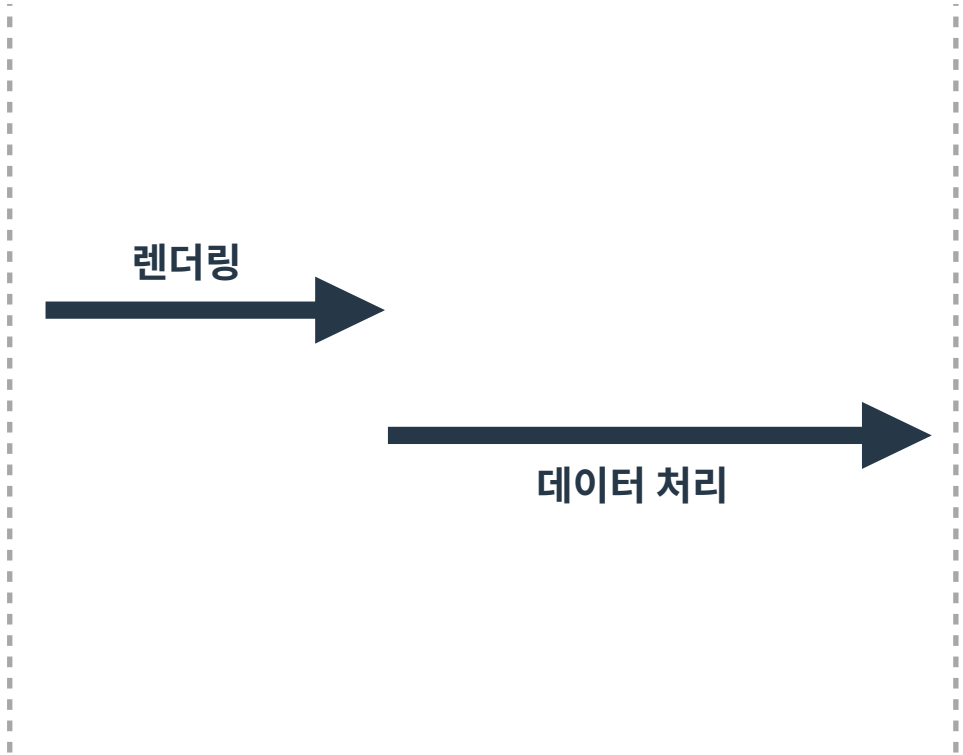
동적 웹 사이트는 어떻게 동작할까요?



비동기 처리를 통해서 필요한 데이터를 채운다

비동기 처리

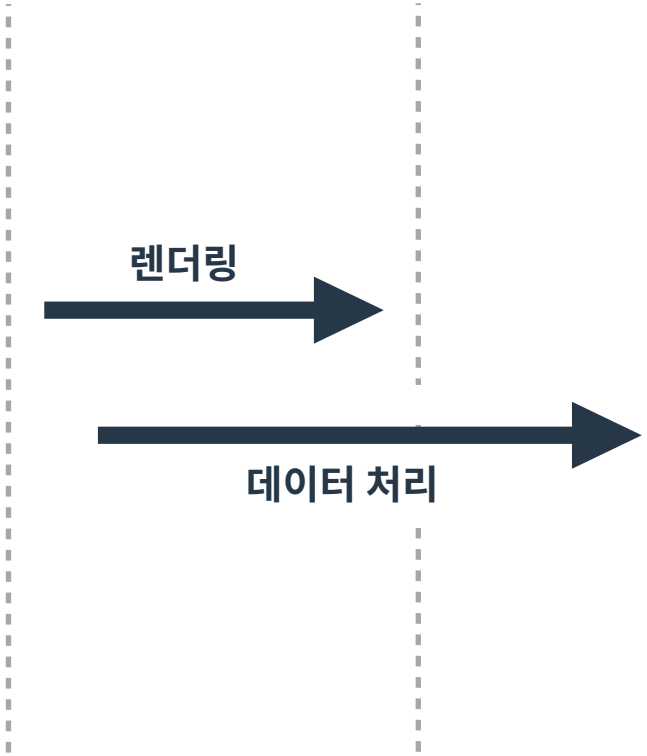
너는 일해, 나는 다른거 할게



동기 처리 : 요청에 따른 응답을 **기다린다**

비동기 처리

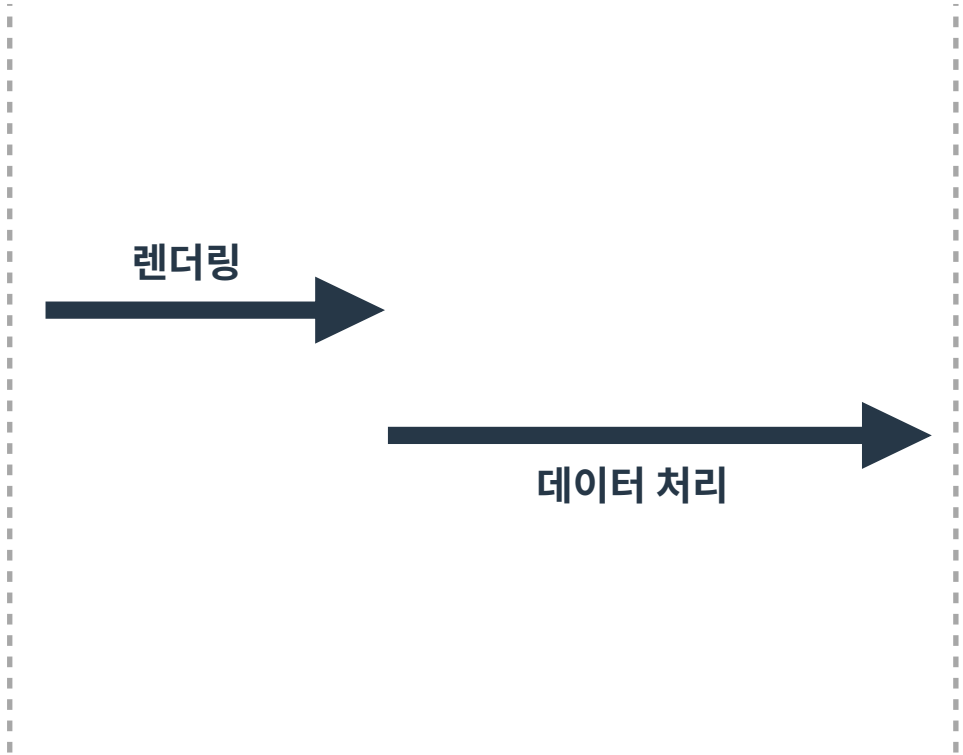
너는 일해, 나는 다른거 할게



비동기 처리 : 요청에 따른 응답을 **기다리지 않는다**

비동기 처리

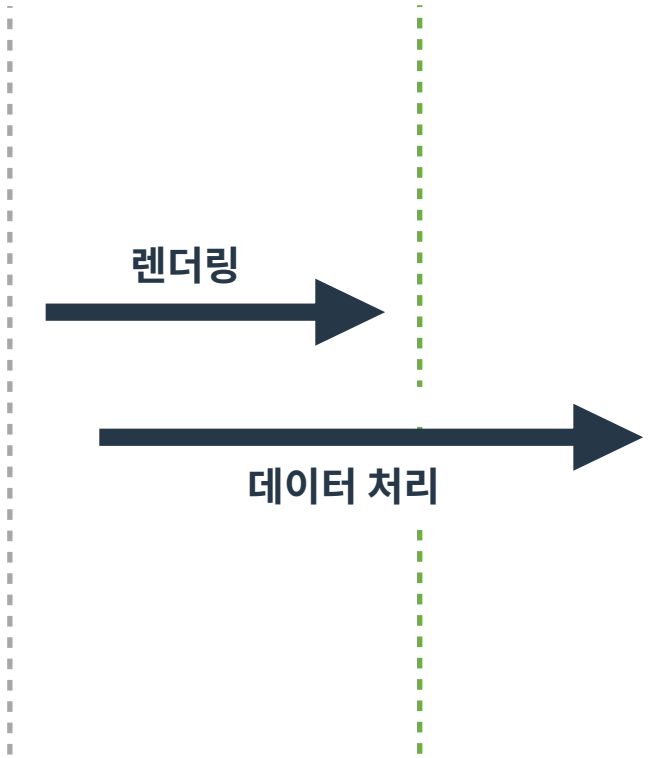
너는 일해, 나는 다른거 할게



동기 처리된 경우, HTML 로딩에 문제가 없다

비동기 처리

너는 일해, 나는 다른거 할게

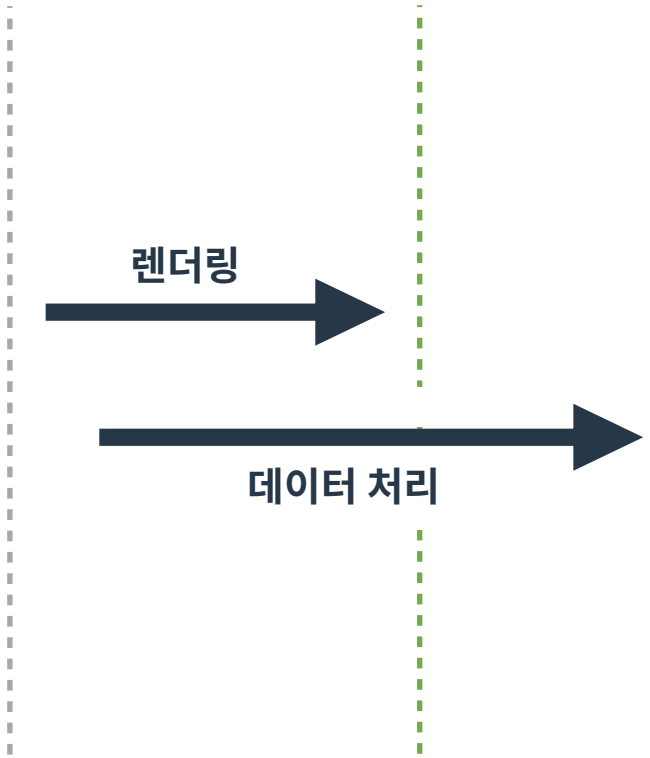


비동기 처리된 경우, 상황에 따라서 데이터가 **완전하지 않은 경우가 발생**한다!

지금까지의 스크래퍼의 문제점

requests로 요청 시 발생하는 문제점

동적 웹사이트에 적용이 어려움



이 상황에서 요청을 보내면 **불완전한 응답**을 받게된다

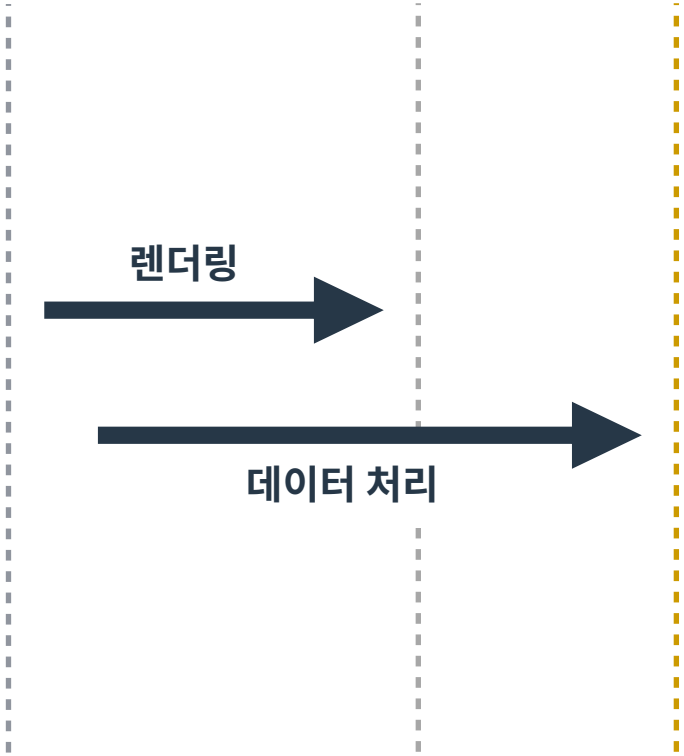
requests로 요청 시 발생하는 문제점

UI 상호작용의 어려움

키보드 입력, 마우스 클릭 등을 requests로는 진행하기 어렵다

이를 해결하려면…

정보를 추출하는 시간을 임의로 지연시킨다면?



임의로 시간을 **지연**한 후, 데이터 처리가 끝난 후 정보를 가져오면 된다

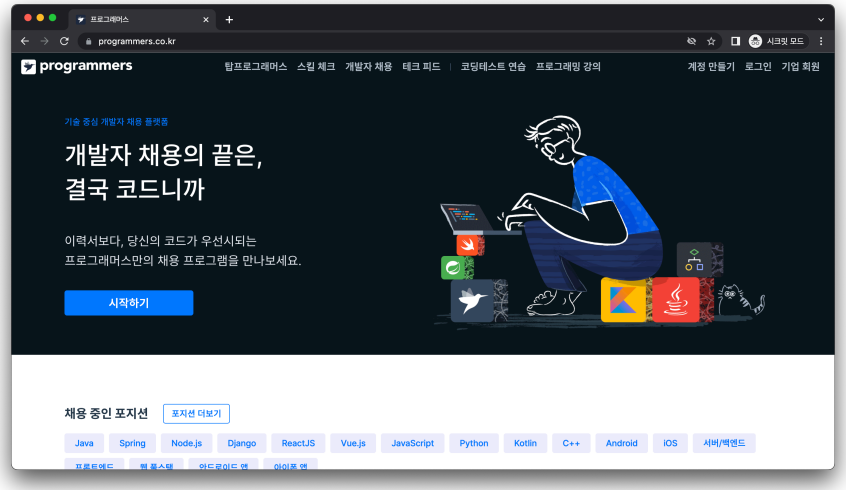
이를 해결하려면…

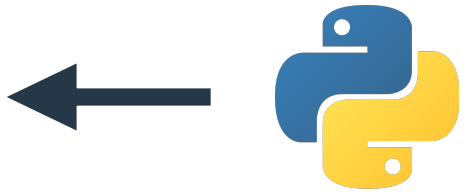
UI Action을 프로그래밍을 통해 명령을 내린다면?

키보드 입력, 마우스 클릭 등을 **프로그래밍**할 순 없을까?

이를 해결해보자!

웹 브라우저와 파이썬의 만남





이 버튼 클릭해

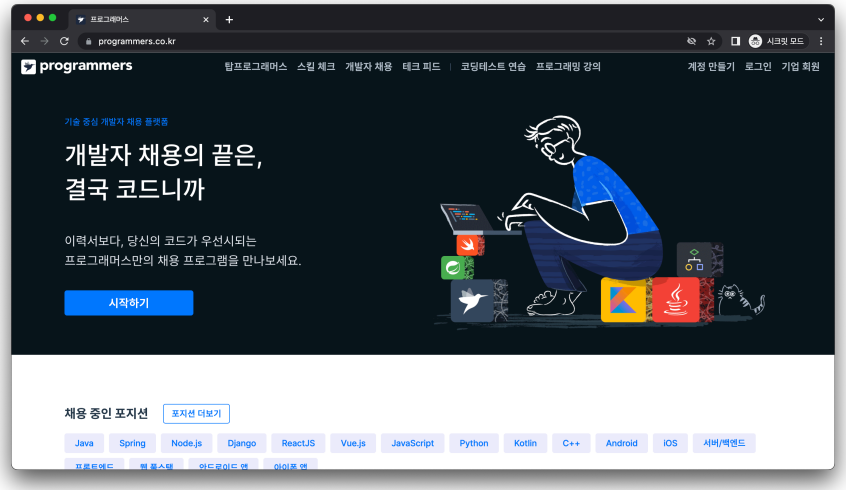
이 입력창에 “파이썬”을 입력해

…

웹 브라우저를 **파이썬으로 조작**하자!

이를 해결해보자!

웹 브라우저와 파이썬의 만남



웹 브라우저를 자동화하는 라이브러리 **Selenium**

이를 해결해보자!

웹 브라우저와 파이썬의 만남

```
from selenium import webdriver

driver = webdriver.Chrome()
driver.implicitly_wait(10)
driver.get("http://www.example.com")
```

응답 후 시간을 **지연**시킬 수 있다

이를 해결해보자!

웹 브라우저와 파이썬의 만남

```
from selenium import webdriver

elem = driver.find_element_by_tag_name("hello-input")
elem.send_keys("Hello!")
```

UI와의 **상호작용**이 가능하다

지금까지의 이야기 요약

웹 브라우저와 파이썬을 함께 사용한다는 아이디어 도출!

동적 웹사이트는 응답 후 바로 정보를 추출하기 **어렵다**

또한, 다양한 키보드 입력과 마우스 클릭 등의 **상호작용**이 존재한다

이런 상황을 해결하기 위해, 웹 브라우저를 **파이썬으로 조작하는** 전략을 취하자

End of Contents

Thank You! :)