소프트웨어 나눔 축제

AnA - Node.js로 크롤링해보기

크롤링의 이해

1. 크롤링이란?
2. 크롤링 이야기
3. 크롤링 과정

크롤링이란?

크롤링(Crawling)은 웹 콘텐츠를 검색하고
해당 콘텐츠(이미지, 텍스트 등)를 추출하는 행위입니다.

크롤링의 이해

1. 크롤링이란?
2. 크롤링 이야기
3. 크롤링 과정

크롤링 이야기

기술적 측면

보안 측면

윤리적 측면

크롤링의 이해

1. 크롤링이란?
2. 크롤링 이야기
3. 크롤링 과정

크롤링 이야기

기술적 측면

크롤링은 통신 프로토콜을 활용한 콘텐츠 수집 방식입니다.
기술적으로는 정상적인 콘텐츠 요청과 특별히 다르지않습니다.

보안 측면
윤리적 측면

통신 프로토콜 : 컴퓨터 사이에서 메시지를 주고 받는 규칙 체계

크롤링의 이해

크롤링 이야기

기술적 측면

보안 측면
공개된 자료에서, 특정한 이용자만이 접근하기를 원한다면
서비스 제공자는 이에 대한 기술적 보호조치가 필요합니다.

윤리적 측면

크롤링의 이해

1. 크롤링이란?
2. 크롤링 이야기
3. 크롤링 과정

크롤링 이야기

기술적 측면
보안 측면

윤리적 측면
 각 사이트의 특정 루트에서 규칙을 살펴야합니다.

크롤링의 이해

1. 크롤링이란?
2. 크롤링 이야기
3. 크롤링 과정

윤리적 측면
 각 사이트의 특정 루트에서 규칙을 살펴야합니다.

/robots.txt를 통하여 규칙을 확인할 수 있습니다.
ex) https://www.google.com/robots.txt

크롤링의 이해

1. 크롤링이란?
2. 크롤링 이야기
3. 크롤링 과정

/robots.txt

robots.txt는 웹사이트에서 웹 크롤러 같은 로봇드르이 접근을 제어하기 위한 규약입니다.

규칙

User-agent : 규칙이 적용되는 크롤러 식별
Allow : 크롤링할 수 있는 URL 경로
Disallow : 크롤링할 수 없는 URL 경로

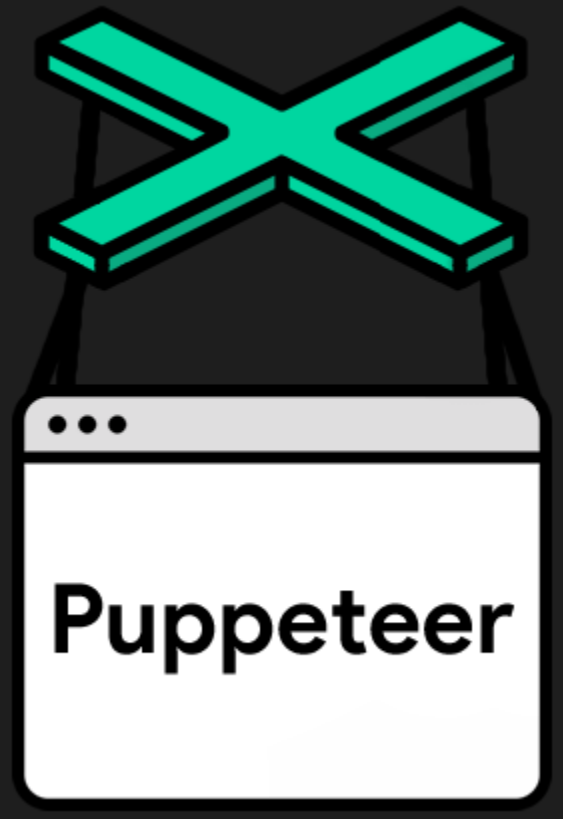
크롤링의 이해

1. 크롤링이란?
2. 크롤링 이야기
3. 크롤링 과정

크롤링 과정

미리보기

Puppeteer



미리보기

시연 – 검색어책 추천 받기

생각나는 검색어를 입력해보세요! 책을 추천해 줄거예요!