Web Scraping 기초

1-4. 윤리적으로 웹 스크래핑/크롤링 진행하기

programmers

Contents

웹 크롤링와 웹 스크래핑

올바르게 HTTP 요청하기

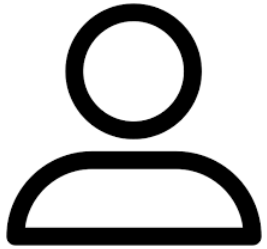
웹 크롤링과 웹 스크래핑

programmers

웹 스크래핑과 웹 크롤링

이 둘의 차이점은 무엇일까요?

웹 크롤링?

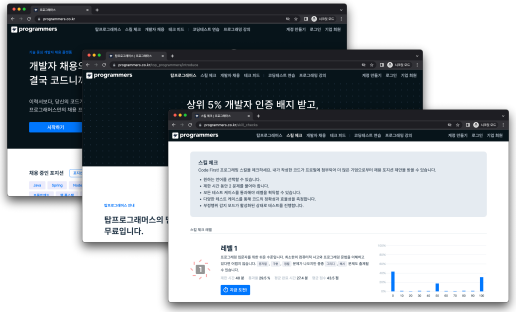


웹 스크래핑?

자주 혼용되는 두 단어, 차이는 무엇일까?

웹 스크래핑?

핵심은 '추출'



웹 페이지들로부터 우리가 원하는 정보를 **추출**

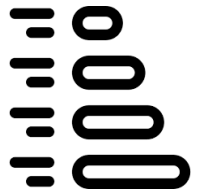
웹 스크래핑?

핵심은 '추출'

웹 페이지들로부터 우리가 원하는 정보를 **추출**

웹 스크래핑?

핵심은 '추출'



[Lv. 1] 계단 오르기

[Lv. 2] 2 * N 타일링

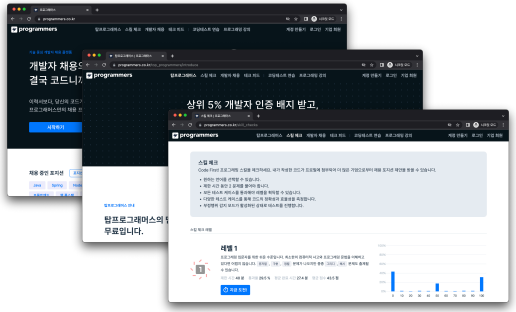
[Lv. 3] 단어 변환

…

웹 페이지들로부터 우리가 원하는 정보를 **추출**

웹 크롤링?

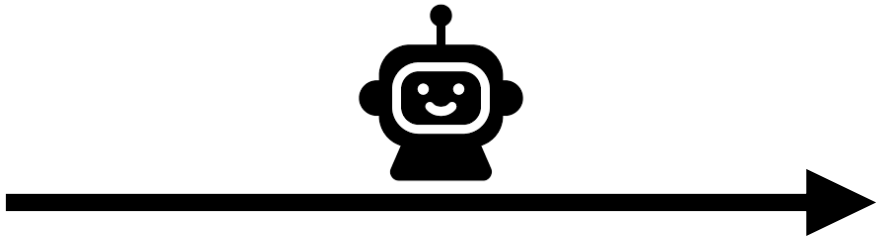
핵심은 '색인'



크롤러(Crawler)를 이용해서 웹 페이지의 정보를 **인덱싱**

웹 크롤링?

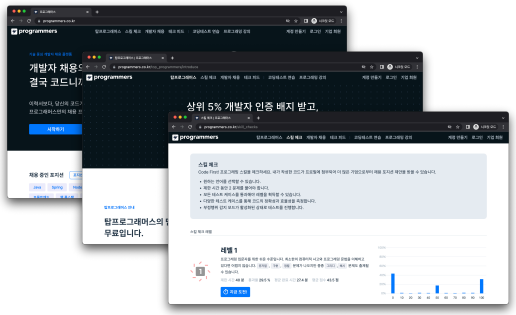
핵심은 '색인'

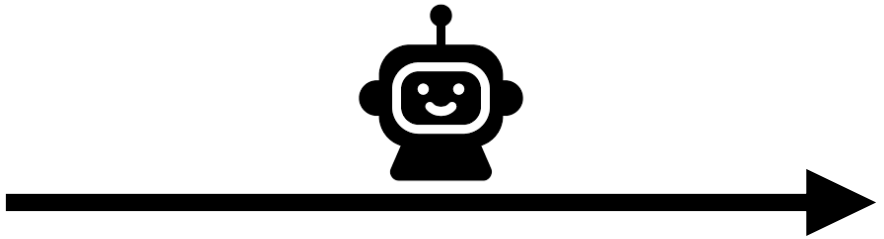


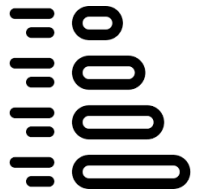
크롤러(Crawler)를 이용해서 웹 페이지의 정보를 **인덱싱**

웹 크롤링?

핵심은 '색인'







www.programmers.co.kr

www.programmers.co.kr/skill_checks

www.programmers.co.kr/job

...

크롤러(Crawler)를 이용해서 웹 페이지의 정보를 **인덱싱**

웹 크롤링과 웹 스크래핑 정리

웹 스크래핑 : 특정한 목적으로 특정 웹 페이지에서 데이터를 추출하는 것 - **데이터 추출**

e.g. 날씨 데이터 가져오기, 주식 데이터 가져오기, …

웹 크롤링 : URL을 타고다니며 반복적으로 데이터를 가져오는 과정 - **데이터 색인**

e.g. 검색 엔진의 웹 크롤러

올바르게 HTTP 요청하기

programmers

올바르게 HTTP 요청하기 위해 고려해야 할 것들

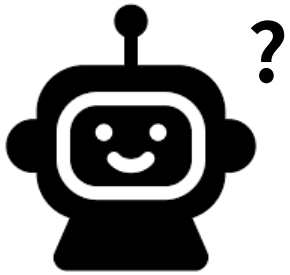
웹 스크래핑/크롤링을 통해 **어떤 목적**을 달성하고자 하는가?

나의 웹 스크래핑/크롤링이 **서버에 영향**을 미치지는 않는가?

로봇 배제 프로토콜(REP)

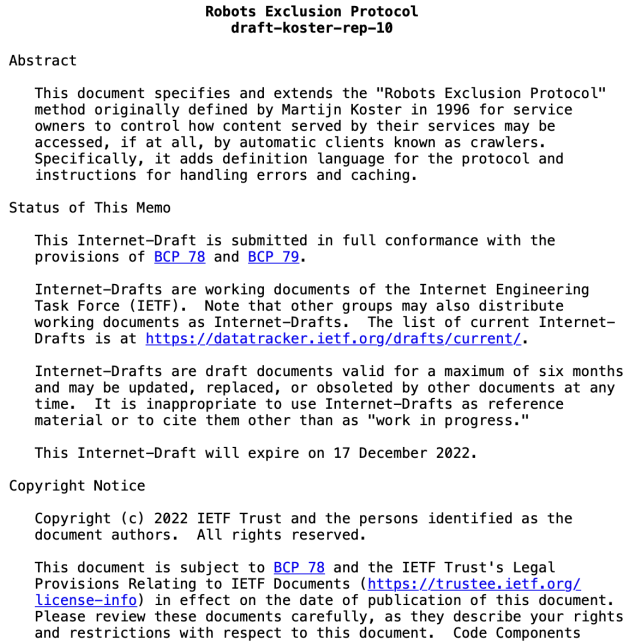
웹 브라우징은 사람이 아닌, **로봇**이 진행할 수 있다!

로봇 배제 프로토콜(REP)



그렇다면 무턱대고 **모든 사이트**에 대해 **모든 정보**를 취득하는 것이 정당할까?

로봇 배제 프로토콜(REP)



Robots Exclusion Protocol
draft-koster-rep-10

Abstract

This document specifies and extends the "Robots Exclusion Protocol" method originally defined by Martijn Koster in 1996 for service owners to control how content served by their services may be accessed, if at all, by automatic clients known as crawlers. Specifically, it adds definition language for the protocol and instructions for handling errors and caching.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 17 December 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components

1994년, **REP(Robot Exclusion Protocol)**의 탄생!

robots.txt 예시

User-agent, Disallow, Allow 등의 키워드를 통해 사용

```
User-agent: *
Disallow: /
```

웹 크롤러들은 이 규칙을 지키면서 크롤링을 진행

robots.txt 예시

User-agent, Disallow, Allow 등의 키워드를 통해 사용

```
User-agent: *
Disallow: /
```

robots.txt 예시

User-agent, Disallow, Allow 등의 키워드를 통해 사용

```
User-agent: *
Disallow: /
```

모든 user-agent에 대해서 접근을 **거부**

robots.txt 예시

User-agent, Disallow, Allow 등의 키워드를 통해 사용

```
User-agent: *
Allow: /
```

robots.txt 예시

User-agent, Disallow, Allow 등의 키워드를 통해 사용

```
User-agent: *
Allow: /
```

모든 user-agent에 대해서 접근을 **허가**

robots.txt 예시

User-agent, Disallow, Allow 등의 키워드를 통해 사용

```
User-agent: MussgBot
Disallow: /
```

robots.txt 예시

User-agent, Disallow, Allow 등의 키워드를 통해 사용

```
User-agent: MussgBot
Disallow: /
```

특정 user-agent에 대해서 접근을 불허

robots.txt 살펴보기

실습에서 계속…!

programmers

End of Contents

Thank You! :)