







English

Preserve casing

✓ Keep spaces & non-letters

4

Training: WWW-Robots (HTTP, Training)

WWW-Robots

In this little training challenge, you are going to learn about the Robots exclusion standard.

The robots.txt file is used by web crawlers to check if they are allowed to crawl and index your website or only parts of it. Sometimes these files reveal the directory structure instead protecting the content from being crawled.

Enjoy!

© 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018 and 2019 by Gizmore

robots.txt는 웹사이트에 웹 크롤러같은 로봇들의 접근을 제어하기 위한 규약이다. 아직 권고안이라 꼭 지킬 의무는 없다.

크롤러들은 주로 검색엔진들의 인덱싱 목적으로 사용되는데, 웹사이트들 입장에서도 더 많은 검색 노출을 원하는게 일반적이므로 딱히 막을 이유는 없다. 다만 서버의 트레픽이 한정돼있거나 검색엔진에의 노출을 원하지 않는 경우, 이 robots.txt에 "안내문" 형식으로 특정 경로에 대한 크롤링을 자제해 줄 것을 권고하는 것이다. 지킬 의무가 없다고 하나 지켜주는게 상식이며, 마찬가지로 서버 주인 입장에서는 원치않는 크롤링이 들어오는데도 계속해서 서비스를 제공할 의무 또한 없으므로 크롤러의 아이피를 차단하면 그만이다.

robots.txt는 웹사이트의 최상위 경로(=루트)에 있어야 한다. 즉, 사이트를 치고 슬래시 후 바로 robots.txt를 넣으면 볼 수 있다는 것이다. ex) http://nam u.wiki/robots.txt

User-agent: *

Disallow:

User-agent: Yandex

Disallow: *





