

# Introduction to Web Scraping with R

Good practice



Simon Munzert | IPSDS

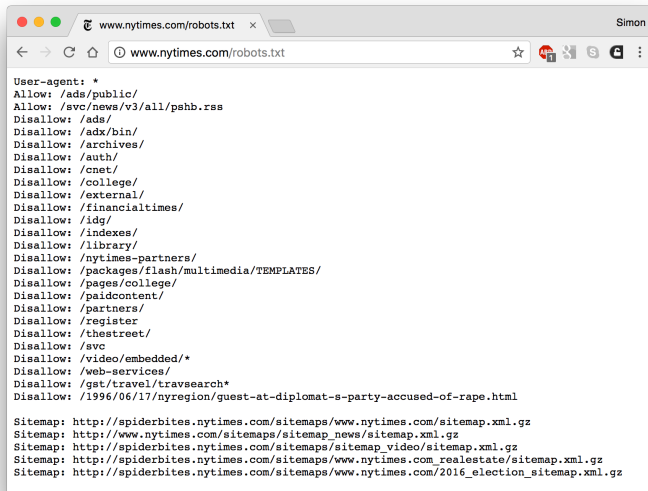
# Understanding robots.txt

# robots.txt

## What's robots.txt?

- 'Robots Exclusion Protocol', informal protocol to prohibit web robots from crawling content
- located in the root directory of a website (e.g., <http://www.google.com/robots.txt>)
- documents which bot is allowed to crawl which resources (and which not)
- not a technical barrier, but a sign that asks for compliance

# Example: NY Times's robots.txt

A screenshot of a web browser window displaying the robots.txt file for the New York Times website. The browser's address bar shows the URL 'www.nytimes.com/robots.txt'. The page content lists various disallowed paths and sitemaps. The text is as follows:

```
User-agent: *
Allow: /ads/public/
Allow: /svc/news/v3/all/pshb.rss
Disallow: /ads/
Disallow: /adx/bin/
Disallow: /archives/
Disallow: /auth/
Disallow: /cnet/
Disallow: /college/
Disallow: /external/
Disallow: /financialtimes/
Disallow: /idg/
Disallow: /indexes/
Disallow: /library/
Disallow: /nytimes-partners/
Disallow: /packages/flash/multimedia/TEMPLATES/
Disallow: /pages/college/
Disallow: /paidcontent/
Disallow: /partners/
Disallow: /register
Disallow: /thestreet/
Disallow: /svc
Disallow: /video/embedded/*
Disallow: /web-services/
Disallow: /gst/travel/travsearch*
Disallow: /1996/06/17/nyregion/guest-at-diplomat-s-party-accused-of-rape.html

Sitemap: http://spiderbites.nytimes.com/sitemaps/www.nytimes.com/sitemap.xml.gz
Sitemap: http://www.nytimes.com/sitemaps/sitemap_news/sitemap.xml.gz
Sitemap: http://spiderbites.nytimes.com/sitemaps/sitemap_video/sitemap.xml.gz
Sitemap: http://spiderbites.nytimes.com/sitemaps/www.nytimes.com_realestate/sitemap.xml.gz
Sitemap: http://spiderbites.nytimes.com/sitemaps/www.nytimes.com/2016_election_sitemap.xml.gz
```

# Syntax in robots.txt

## Syntax

- not an official W3C standard, partly inconsistent syntax
- rules listed bot by bot
- general, bot-independent rules under '\*' (most interesting entry for R-based crawlers)
- directories/folders listed separately

```
1 User-agent: Googlebot
2 Disallow: /images/
3 Disallow: /private/
```

```
1 User-agent: *
2 Disallow: /private/
```

# Syntax in robots.txt

## Universal ban

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

## Separation of bots by empty line

```
1 User-agent: Googlebot  
2 Disallow: /images/  
  
4 User-agent: Slurp  
5 Disallow: /images/
```

## Allow declaration

```
1 User-agent: *  
2 Disallow: /images/  
3 Allow: /images/public/
```

# Syntax in robots.txt

## Crawl-delay (in seconds)

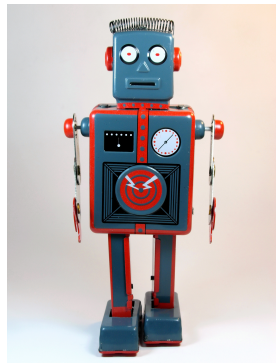
```
1 User-agent: *  
2 Crawl-delay: 2  
3 User-Agent: Googlebot  
4 Disallow: /search/
```

## Robots `<meta>` tag

```
1 <meta name="robots" content="noindex,nofollow" />
```

# How to deal with robots.txt?

- not clear if robots.txt is legally binding or not, and if yes for which activities
- originally not thought of as protection against small-scale web scraping applications, but against large-scale indexing bots
- guide to a webmaster's preferences with regards to visibility of content
- my advice: take robots.txt into account! If the data you are interested in are excluded from crawling: contact webmaster
- there is even an R package, `robotstxt`, that helps you parse robots.txt documents and stick to the rules. It's available here: <https://github.com/ropenscilabs/robotstxt>



By D J Shin - My Toy Museum,  
[https://commons.wikimedia.org/wiki/File:QSH\\_Tin\\_Wind\\_Up\\_Mechanical\\_Robot\\_\(Giant\\_Easelback\\_Robot\)\\_Front.jpg](https://commons.wikimedia.org/wiki/File:QSH_Tin_Wind_Up_Mechanical_Robot_(Giant_Easelback_Robot)_Front.jpg)



# Scraping etiquette

# Some advice for your work

# Some advice for your work

1. You take all the responsibility for your web scraping work.

# Some advice for your work

1. You take all the responsibility for your web scraping work.
2. Take all copyrights of a country's jurisdiction into account.

# Some advice for your work

1. You take all the responsibility for your web scraping work.
2. Take all copyrights of a country's jurisdiction into account.
3. If you publish data, do not commit copyright fraud.

# Some advice for your work

1. You take all the responsibility for your web scraping work.
2. Take all copyrights of a country's jurisdiction into account.
3. If you publish data, do not commit copyright fraud.
4. If possible, stay identifiable.

# Some advice for your work

1. You take all the responsibility for your web scraping work.
2. Take all copyrights of a country's jurisdiction into account.
3. If you publish data, do not commit copyright fraud.
4. If possible, stay identifiable.
5. If in doubt, ask the author/creator/provider of data for permission—if your interest is entirely scientific, chances aren't bad that you get data.

# Some advice for your work

1. You take all the responsibility for your web scraping work.
2. Take all copyrights of a country's jurisdiction into account.
3. If you publish data, do not commit copyright fraud.
4. If possible, stay identifiable.
5. If in doubt, ask the author/creator/provider of data for permission—if your interest is entirely scientific, chances aren't bad that you get data.
6. Consult current jurisdiction, e.g. on <http://blawgsearch.justia.com> or from a lawyer specialized on internet law.



# Scraping etiquette

