Web Data Collection with R

robots.txt

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT

teaser

- robots.txt files are part of many webpages
- the idea is ask bots (e.g. googlebot, etc.) t adhere to those rules and not cause any trouble
- for us this is valuable information about what might be unwanted behaviour as well

exmaple and syntax :

https://www.researchgate.net/robots.txt

```
User-agent: *
Allow: /
Disallow: /connector/
Disallow: /deref/
Disallow: /plugins.
Disallow: /firststeps.
Disallow: /publicliterature.PublicLiterature.search.html

Sitemap: https://www.researchgate.net/sitemaps/sitemap-inde
```

there is a package for that

```
library(robotstxt)
??robotstxt

rtxt <-
  robotstxt$new(
    domain="https://www.researchgate.net"
  )
```

```
## No encoding supplied: defaulting to UTF-8.
```

there is a package for that

```
rtxt
```

```
## <robotstxt>
##   Public:
##     bots: *
##     check: function (paths = "/", bot = "*", permission
##     clone: function (deep = FALSE)
##     comments: data.frame
##     crawl_delay: data.frame
##     domain: https://www.researchgate.net
##     host: data.frame
##     initialize: function (domain, text)
##     other: data.frame
##     permissions: data.frame
##     sitemap: data.frame
##     text: User-agent: *
##     Allow: /
##     Disallow: /connector/
```

there is a package for that

```
rtxt$check(c("/connector/index.html","index.html"))
```

```
## /connector/index.html                index.html
##                 FALSE                      TRUE
```

But always have a look at the ToS

https://www.researchgate.net/application.TermsAndConditions.html

ARTICLE 5: MISUSE OF THE SERVICE

- Users must not misuse the Service.
- Misuse of the Service includes, without limitation:
  - insults to other Users;
  - automated or massive manual retrieval of other Users' profile data ("data harvesting");
  - advertising for commercial products or services of all kinds;
  - unsolicited job offers and business proposals;
  - all kinds of technical attacks on the servers.
- All aforementioned behaviors in this article are strictly forbidden, unless the User has obtained prior written permission by the Provider.