Webscraping

Jan-Philipp Kolb

Wed Oct 14 11:19:52 2015

Das R-Paket XML

Gaston Sanchez - Basics of XML and HTML

```
# install.packages("XML")
library(XML)
```

Einlesen einer Tabelle mit der Funktion `readHTMLTable`:
Beispieldaten für die Bundesländer

```
BLA_link <- "http://www.bernhard-gaul.de/wissen/bundeslaend
BLA_tab <- readHTMLTable(BLA_link)
```

Überblick über die Daten

```
BLA_tab <- BLA_tab[[1]]
```

Wappen	Land	FlÃ¤che (kmÂ²)	Einwohner
Wappen	Land	Fläche (km²)	Einwohner
	Baden-Württemberg	35 751,36	10 569 111
	Bayern	70 550,23	12 519 571
	Berlin	891,70	3 375 222
	Brandenburg	29 485,63	2 449 511
	Bremen	419,24	654 774
	Hamburg	755,30	1 734 272
	Hessen	21 114,93	6 016 481
	Mecklenburg-Vorpommern	23 210.55	1 600 327
	Niedersachsen	47 613,78	7 778 995
	Nordrhein-Westfalen	34 109,70	17 554 329
	Rheinland-Pfalz	19 854,10	3 990 278
	Saarland	2 568,70	994 287
	Sachsen	18 420,01	4 050 204

Daten bearbeiten

Die erste und 18. Zeile entfernen:

```
BLA_tab <- BLA_tab[-c(1,18),-1]
```

Die Daten editieren

```
mean(BLA_tab[,2])
```

```
## Warning in mean.default(BLA_tab[, 2]): argument is not n
## returning NA
```

```
## [1] NA
```

um dies zu ändern ist ein wenig Kosmetik notwendig:

```
BLA_tab[,2] <- gsub(" ","",BLA_tab[,2])
BLA_tab[,2] <- gsub(",",".",BLA_tab[,2])
BLA_tab[,2] <- as.numeric(BLA_tab[,2])
```

```
mean(BLA_tab[,2])
```

```
## [1] 22323
```

Die weiteren Spalten bearbeiten

```
BLA_tab[,3] <- as.numeric(gsub(" ","",BLA_tab[,3]))
BLA_tab[,4] <- as.numeric(gsub(" ","",BLA_tab[,4]))
```

Das Editing ist also aufwändiger als das eigentliche Scraping

Die Daten speichern

```
write.csv(BLA_tab,file="BLA_tab.csv")
```

Das Paket rvest

```
install.packages("rvest")
```

```
library("rvest")

link <- "https://de.wikipedia.org/wiki/Liste_der_St%C3%A4dt

link_data <- read_html(link)

# parse the document for R representation:
mps.doc <- htmlParse(link_data)

# get all the tables in mps.doc as data frames
mps.tabs <- readHTMLTable(mps.doc)
```

Links

Beispiel: GiventheData
Five easy steps for webscraping

Reference XML

```
citation("XML")
```

```
##
## To cite package 'XML' in publications use:
##
##   Duncan Temple Lang and the CRAN Team (2016). XML: Tool
##   Parsing and Generating XML Within R and S-Plus. R pack
##   version 3.98-1.4. https://CRAN.R-project.org/package=X
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {XML: Tools for Parsing and Generating XML W
##     author = {Duncan Temple Lang and the CRAN Team},
##     year = {2016},
##     note = {R package version 3.98-1.4},
##     url = {https://CRAN.R-project.org/package=XML},
##   }
```

Reference rvest

```
citation("rvest")
```

```
##
## To cite package 'rvest' in publications use:
##
##   Hadley Wickham (2015). rvest: Easily Harvest (Scrape)
##   R package version 0.3.1.
##   https://CRAN.R-project.org/package=rvest
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {rvest: Easily Harvest (Scrape) Web Pages},
##     author = {Hadley Wickham},
##     year = {2015},
##     note = {R package version 0.3.1},
##     url = {https://CRAN.R-project.org/package=rvest},
##   }
```