Nutzung von GeoDaten in den Sozialwissenschaften - Webscraping

Jan-Philipp Kolb

07 April 2016

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.bergziege-owl.de/deutschland-
bundeslander-und-ihre-hauptstadte/"
BLA_tab <- readHTMLTable(BLA_link)
```

Überblick und Daten bearbeiten

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

V1	V2
Bundesländer	Hauptstädte
Schleswig-Holstein	Kiel
Mecklenburg-Vorpommern	Schwerin
Hamburg	Hamburg
Bremen	Bremen
Niedersachsen	Hannover

Daten bearbeiten

Die erste Zeile ist die Überschrift:

```
colnames(BLA_tab) <- c("Bundesländer","Hauptstädte")
```

Die erste Zeile entfernen:

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

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/Fl%C3%BCchtlingskrise_in_Europa_ab_2015"

link_data <- read_html(link)
```

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

Alle Tabellen bekommen:

```
mps.tabs <- readHTMLTable(mps.doc)
```

Daten zu Asylanträgen bearbeiten

```
mps.tabs <- mps.tabs[[2]]
mps.tabs <- mps.tabs[-c(1,34),]
```

	V1	V2	V3
2	Belgien Belgien	44.665	3,97
3	Bulgarien Bulgarien	20.375	2,83
4	Danemark Dänemark	20.940	3,70
5	Deutschland Deutschland	476.510	5,87
6	Estland Estland	230	0,18
7	Finnland Finnland	32.345	5,91

Geburtenrate

```
link2 <- "http://www.laenderdaten.de/bevoelkerung/
geburtenrate.aspx"
link_data2 <- read_html(link2)
doc <- htmlParse(link_data2)
tab <- readHTMLTable(doc)
tab1 <- tab[[2]]
```

Land	Â	Geburtenrate	Weltrang
	Niger	45,45	1
	Mali	44,99	2
	Uganda	43,79	3
	Sambia	42,13	4
	Burkina Faso	42,03	5
	Burundi	42,01	6

Fertility Rate

```
link3 <- "https://en.wikipedia.org/wiki/List_of_sovereign_s
link_data3 <- read_html(link3)
doc3 <- htmlParse(link_data3)
tab3 <- readHTMLTable(doc3)
tab4 <- tab3[[2]]
```

- Die US Flughäfen und Fluglinien

http://www.sasanalysis.com/2013/06/
the-us-airports-with-most-flight-routes.html

- Mehr Daten - http://openflights.org/data.html

```
link1 <- "http://openflights.svn.sourceforge.net/viewvc/ope
openflights/data/airports.dat"
airport <- read.csv(link1, header = F)

link2 <- "http://openflights.svn.sourceforge.net/viewvc/ope
openflights/data/routes.dat"
route <- read.csv(link2, header = F)
```

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},
##   }
```