Datenaufbereitung

Jan-Philipp Kolb

07 April 2016

Die Daten editieren

```
load("data/refugeeTab.RData")
```

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

```
## Warning in mean.default(refugeeTab[, 2]): argument is no
## logical: returning NA
```

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

um dies zu ändern ist ein wenig Kosmetik notwendig:

```
refugeeTab[,2] <- as.numeric(refugeeTab[,2])
```

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

```
## [1] 17.16129
```

Die weiteren Spalten bearbeiten

- In R wird der Punkt als Dezimaltrennzeichen verwendet.
- Wenn ein Komma im Ausdruck ist, wird der Eintrag als `character` behandelt
- dann kann bspw. kein Mittelwert berechnet werden

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

Erste Spalte bearbeiten und Daten speichern

```
ab <- as.character(refugeeTab[,1])
info <- round(nchar(ab)/2)
Namen <- substr(ab,1,info)
Namen[1:29] <- gsub(" ","",Namen[1:29])
Namen[31]  <- "Zypern"
refugeeTab[,1] <- Namen
```

Spaltennamen verändern

```
colnames(refugeeTab) <- c("Land","2015",
                          "pro_tsd_Einwohner")
```

Die Daten abspeichern

```
save(refugeeTab,file="refugeeTab_final.RData")
```

Das Ergebnis

	Land	2015	pro_tsd_Einwohner
3	Bulgarien	14	2.83
4	Danemark	15	3.70
5	Deutschland	29	5.87
6	Estland	16	0.18
7	Finnland	20	5.91
8	Frankreich	30	1.14

Das Editing ist also aufwändiger als das eigentliche Scraping

CO2 Verbrauch

```
link <- "https://en.wikipedia.org/wiki/
List_of_countries_by_carbon_dioxide_
emissions_per_capita"
```

```
link_data <- read_html(link)
doc <- htmlParse(link_data)
tab <- readHTMLTable(doc)
```

```
save(tab,file="co2tab.RData")
```

```
str(tab)
```

```
## List of 20
##  $ NULL: NULL
##  $ NULL:'data.frame':    219 obs. of  25 variables:
##   ..$ V1 : Factor w/ 213 levels "","-","1.","10.",..: 3
##   ..$ V2 : Factor w/ 219 levels "Afghanistan",..: 157 19
##   ..$ V3 : Factor w/ 90 levels "","-","0","0.1",..: 48 2
```

Die Elemente des Objektes

```
tab[[1]]
```

```
## NULL
```

```
head(tab[[2]][,1:7])
```

```
##   V1                          V2   V3   V4   V5   V6   V7
## 1 1.                      Qatar 25.2 36.7 54.3 60.9 58.7
## 2 2.  Trinidad and Tobago 13.9 17.1 17.0 13.5 15.8
## 3 3. Netherlands Antilles 32.6 26.9 22.6 35.0 34.3
## 4 4.                    Kuwait 19.0  5.1 10.0 16.9 20.8
## 5 5.                    Brunei 25.0 22.0 21.4 20.0 18.8
## 6 6. United Arab Emirates 29.4 30.2 29.5 31.1 33.1
```

Auf die Daten schauen

```
tab2 <- tab[[2]]
Cnames <- c("Rank","Country",as.character(1990:2011))
colnames(tab2) <- Cnames
```

Haben die Daten die richtige Struktur?

```
mean(tab2[,3])
```

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

```
tab2[,3] <- as.numeric(as.character(tab2[,3]))
```

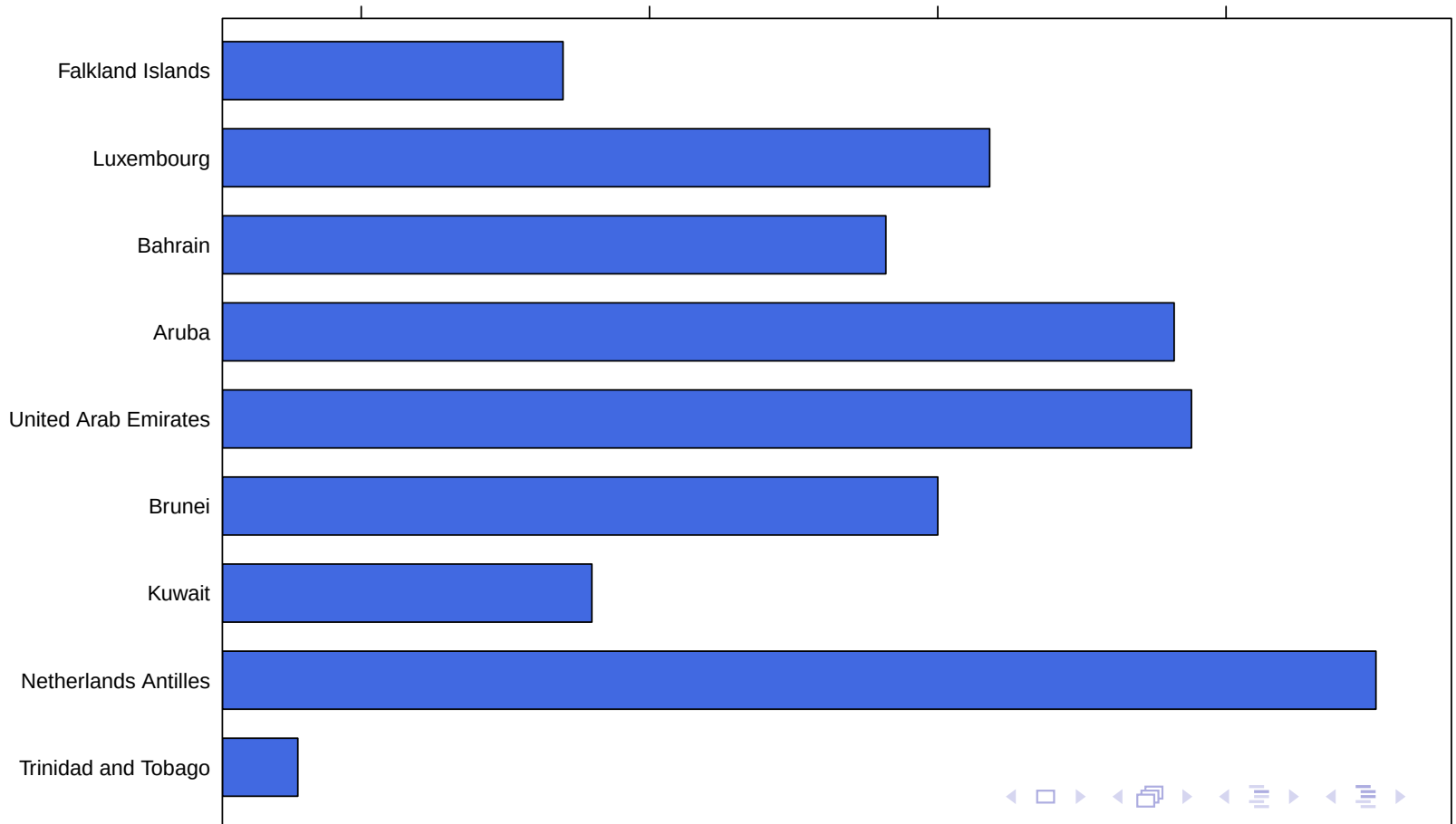
```
for (i in 3:ncol(tab2)){
  tab2[,i] <- as.numeric(as.character(tab2[,i]))
}
```

Daten speichern

```
save(tab2,file="CO2emissions.RData")
```

Eine Graphik

```
library(lattice)
emissions <- as.numeric(tab2[,3])
names(emissions) <- tab2[,2]
barchart(emissions[1:10],col="royalblue")
```



Take Home Message

- Mit Webscraping können sehr viele Daten gewonnen werden.
- Allerdings kann die Datenaufbereitung sehr aufwändig sein.
- Oftmals ist viel rumprobieren notwendig.