Webscraping

Jan-Philipp Kolb

9 Mai 2017

Notwendige Pakete

```
install.packages("tidyverse")
```

```
library(tidyverse)
```

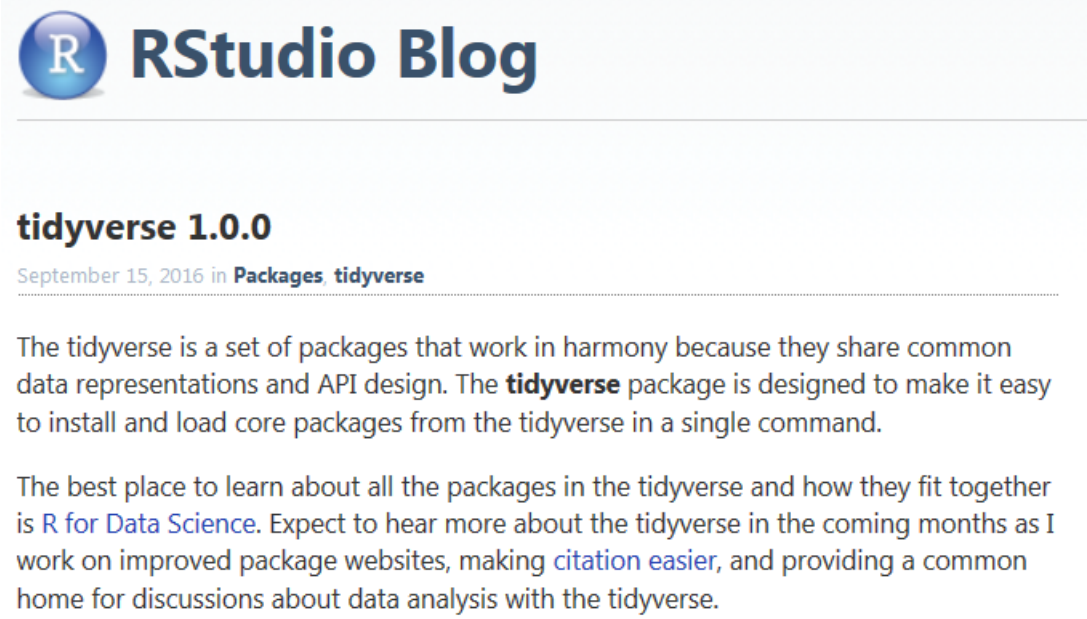

Figure 1:

- R für DataScience

Weitere benötigte Pakete

- Das Paket `stringr`

```
library(stringr)
```

```
library(forcats)
library(ggmap)
library(rvest)
```

Daten von Wikipedia einsammeln

```
html.world_ports <- read_html("https://en.wikipedia.org/wiki/List_of_busiest_container_ports")
df.world_ports <- html_table(html_nodes(html.world_ports, "table")[[2]], fill = TRUE)
```

```
library(DT)
datatable(df.world_ports)
```

Die Daten anschauen

```
glimpse(df.world_ports)
```

```
## Observations: 50
## Variables: 15
## $ Rank     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ Port     <chr> "Shanghai", "Singapore", "Shenzhen", "Ningbo-Zhoushan...
## $ Economy  <chr> "China", "Singapore", "China", "China", "Hong Kong", ...
## $ 2015[1]  <chr> "36,516", "30,922", "24,142", "20,636", "20,073", "19...
## $ 2014[2]  <chr> "35,268", "33,869", "23,798", "19,450", "22,374", "18...
## $ 2013[3]  <chr> "33,617", "32,240", "23,280", "17,351", "22,352", "17...
## $ 2012[4]  <chr> "32,529", "31,649", "22,940", "16,670", "23,117", "17...
## $ 2011[5]  <chr> "31,700", "29,937", "22,570", "14,686", "24,384", "16...
## $ 2010[6]  <chr> "29,069", "28,431", "22,510", "13,144", "23,532", "14...
## $ 2009[7]  <chr> "25,002", "25,866", "18,250", "10,502", "20,983", "11...
## $ 2008[8]  <chr> "27,980", "29,918", "21,414", "11,226", "24,248", "13...
## $ 2007[9]  <chr> "26,150", "27,932", "21,099", "9,349", "23,881", "13,...
## $ 2006[10] <chr> "21,710", "24,792", "18,469", "7,068", "23,539", "12,...
## $ 2005[11] <chr> "18,084", "23,192", "16,197", "5,208", "22,427", "11,...
## $ 2004[12] <chr> "14,557", "21,329", "13,615", "4,006", "21,984", "11,...
```

Das Paket `rvest`

```
library(rvest)
ht <- read_html('https://www.google.co.in/search?q=guitar+repair+workshop')
links <- ht %>% html_nodes(xpath='//h3/a') %>% html_attr('href')
gsub('/url\\?q=','',sapply(strsplit(links[as.vector(grep('url',links))],split='&'),'[',1))
```

```
## [1] "http://theguitarrepairworkshop.com/"
## [2] "http://www.guitarservices.com/"
## [3] "http://www.guitarrepairbench.com/guitar-building-projects/guitar-workshop/guitar-workshop-proje
## [4] "https://www.taylorguitars.com/dealer/guitar-repair-workshop-ltd"
## [5] "https://www.facebook.com/The-Guitar-Repair-Workshop-847517635259712/"
## [6] "http://www.laweekly.com/music/10-best-guitar-repair-shops-in-los-angeles-4647166"
## [7] "http://guitarworkshopglasgow.com/pages/repairs-1"
## [8] "https://www.justdial.com/Mumbai/Guitar-Repair-Services/nct-10988623"
## [9] "https://www.justdial.com/Delhi-NCR/Guitar-Repair-Services/nct-10988623"
```

Links

- How to really do an analysis in R (part 1, data manipulation)
- Read CSV From The Web

- Scraping CRAN with rvest