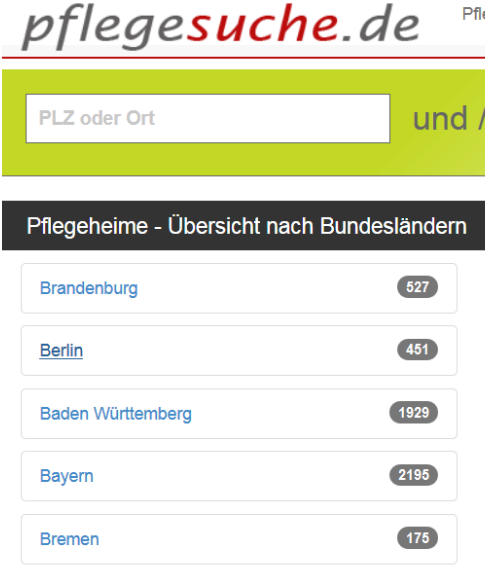
Web Scraping

Jan-Philipp Kolb

Wed Oct 14 11:19:52 2015

Target: Get data on nursing homes



pflegesuche.de

Pfl

PLZ oder Ort

und /

Pflegeheime - Übersicht nach Bundesländern

Brandenburg	527
Berlin	451
Baden Württemberg	1929
Bayern	2195
Bremen	175

The XML library

Getting Data from the Web with R

Part 4: Parsing XML/HTML Content

Gaston Sanchez

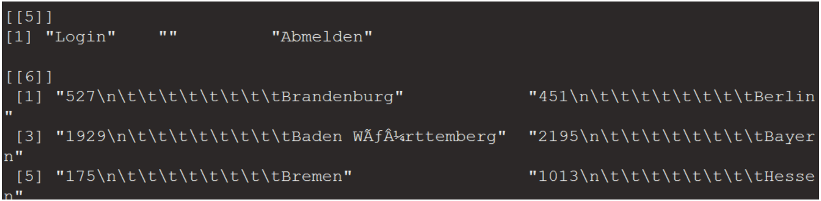
April-May 2014

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

Import the data to R

```r
link <- "http://www.pflegesuche.de/bundesland_pflegeheime.h
ab <- readHTMLList(link)
```

The result is quite unstructured:



```
[[5]]
[1] "Login"    ""          "Abmelden"

[[6]]
 [1] "527\n\t\t\t\t\t\t\t\tBrandenburg"           "451\n\t\t\t\t\t\t\t\tBerlin
"
 [3] "1929\n\t\t\t\t\t\t\t\tBaden WÃƒÂ¼rttemberg"  "2195\n\t\t\t\t\t\t\t\tBayer
n"
 [5] "175\n\t\t\t\t\t\t\t\tBremen"                "1013\n\t\t\t\t\t\t\t\tHesse
n"
```

Structuring data

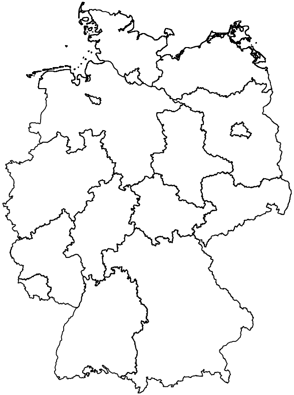
```
tab <- ab[[6]]
tab1 <- gsub("\t","",tab)
tab2 <- strsplit(tab1,"\n")
dat <- unlist(lapply(tab2,function(x)x[1]))
bla <- unlist(lapply(tab2,function(x)x[2]))
df_pflege <- data.frame(bla,
          Anz=as.numeric(as.character(dat)))
```

Overview of data

bla	Anz
Brandenburg	527
Berlin	451
Baden WÃfÂ¼rttemberg	1929
Bayern	2195
Bremen	175
Hessen	1013

Get a map

```
library(raster)
DEU1 <- getData('GADM', country='DEU', level=1)
plot(DEU1)
```



Combine map and data

```r
ind <- match(DEU1@data$NAME_1,df_pflege$bla)
ind
```

```
##  [1] NA  4  2  1  5  7  6 NA  9 10 11 13 NA 14 NA NA
```

Match the missing entries

```
DEU1@data$NAME_1[is.na(ind)]
```

```
## [1] "Baden-Württemberg"      "Mecklenburg-Vorpommern"
## [3] "Sachsen-Anhalt"         "Schleswig-Holstein"
## [5] "Thüringen"
```

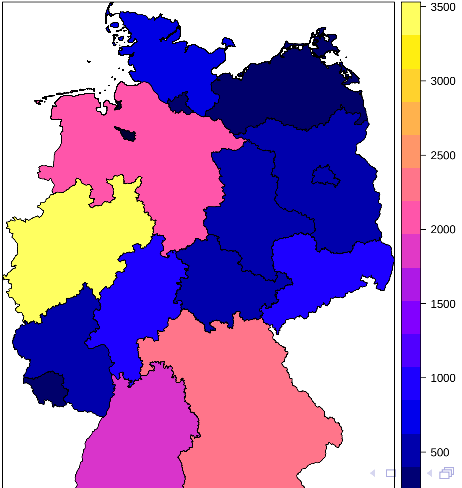
```
which(is.na(ind))
```

```
## [1]  1  8 13 15 16
```

```
ind[1] <- agrep("rttemberg",df_pflege$bla)
ind[8] <- agrep("pommern",df_pflege$bla)
ind[13] <- agrep("Anhalt",df_pflege$bla)
ind[15] <- agrep("Holstein",df_pflege$bla)
ind[16] <- agrep("ringen",df_pflege$bla)
```

Combine data and map

```r
DEU1@data$nursingHomes <- df_pflege$Anz[ind]
library(sp)
spplot(DEU1,"nursingHomes")
```


3500
3000
2500
2000
1500
1000
500