Web Scraping with R

Eduard Szöcs

Institute for Environmental Sciences, University of Koblenz-Landau

Landau, 25.02.2016

About me

- Phd-Student @Uni Koblenz-Landau
  - Environmental monitoring data
- Freelance R Consultant
  - R Courses
  - Data sourcing, cleaning & analysis

About me

- Phd-Student @Uni Koblenz-Landau
  - Environmental monitoring data
- Freelance R Consultant
  - R Courses
  - Data sourcing, cleaning & analysis

R packages:

taxize Taxonomic Information from Around the Web

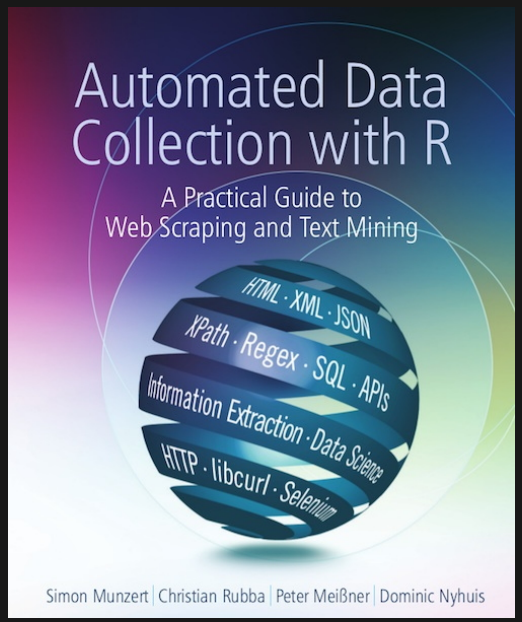
webchem Chemical Information from the Web

I - Packages (selected)

xml2	Parsing XML & HTML
rvest	parse common html structures (e.g. tables)
xmlview	View pretty HTML/XML, explore XPaths
httr*	Working with APIs / http protocol
jsonlite*	Parse JSON

I - Packages (selected)

xml2	Parsing XML & HTML
rvest	parse common html structures (e.g. tables)
xmlview	View pretty HTML/XML, explore XPaths
httr*	Working with APIs / http protocol
jsonlite*	Parse JSON



* I won't cover APIs here.

II - Scraping structured web-pages

What is structured?

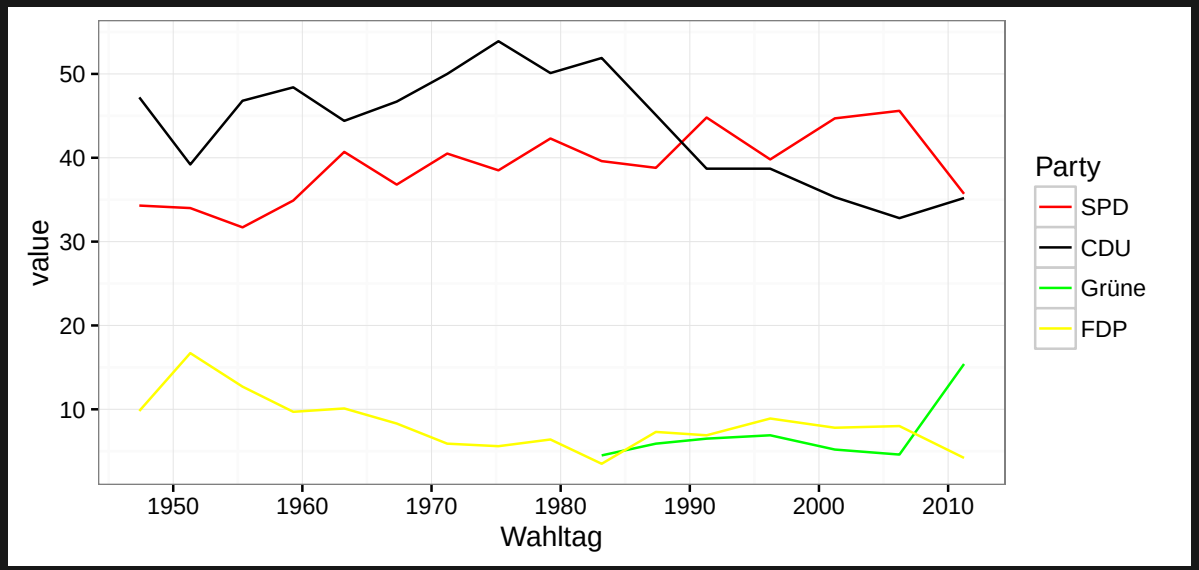
```
<table class="wikitable">
  <tbody>
    <tr>
      <th>First name</th>
      <th>Last name</th>
      <th>Age</th>
    </tr>
    <tr>
      <td>Tinu</td>
      <td>Elejogun</td>
      <td>14</td>
    </tr>
    <tr>
      <td>Blaszczyk</td>
      <td>Kostrzewski</td>
      <td>25</td>
    </tr>
    [...]
  </tbody>
</table>
```

2D representation of data -> data.frame

II - Demo

Extract election results from wikipedia

`https://de.wikipedia.org/wiki/Landtagswahlen_in_Rheinland-Pfalz`



III - Scraping unstructured web-pages

- Not all data is stored in <table>
- scattered on the web page
- Need to find and extract the parts we need.

```
<h1 id="Top">Formaldehyde</h1>
<ul>
<li>
<strong>
<a class="external" href="http://goldbook.iupac.org/R05271.html" title="IUPAC definition of relative molec
:
</strong>
30.0260
</li>
<li>
<strong>IUPAC Standard InChIKey:</strong>
<span style="font-family: monospace;">WSFSSNUMVMOOMR-UHFFFAOYSA-N</span>
</li>
<li>
<strong>CAS Registry Number:</strong>
50-00-0
</li>
```

III - XPath

XML Path Language (XPath) is a query language for selecting nodes

IV - Remarks

- scraping is easy.
- being user-friendly is not.
- be nice to the servers!
  - scrape slowly, time-outs
  - Sys.sleep(rgamma(1, shape = 30, scale = 1/10))
- illegal?
  - even slower
  - anonymous EC2
  - TOR

Web Scraping with R

Eduard Szöcs

Institute for Environmental Sciences, University of Koblenz-Landau

http://edild.github.io/

https://github.com/edild/talk_webscrapingr

@EduardSzoecs

szoecs@uni-landau.de