

Web Scraping with R

Eduard Szöcs

Institute for Environmental Sciences, UnVersity of Koblenz-Landau

Landau, 07.04.2016



https://github.com/edild/talk_webscrapingr

I - About
oo

II - Tools
o

III - Structured data
oo

IV - Unstructured data
oooo

V - Automatisations
oo

VI - Conclusions
ooo

About me

- ▶ Phd-Student @Uni Koblenz-Landau
 - ▶ Environmental data
- ▶ Freelance (R) Consultant
 - ▶ Data sourcing, cleaning & analysis
 - ▶ Chemoinformatics
 - ▶ R Courses

About me

- ▶ Phd-Student @Uni Koblenz-Landau
 - ▶ Environmental data
- ▶ Freelance (R) Consultant
 - ▶ Data sourcing, cleaning & analysis
 - ▶ Chemoinformatics
 - ▶ R Courses

R packages:

taxize Taxonomic Information from Around the Web

webchem Chemical Information from the Web



I What is web scraping?

Getting data from the internet

I What is web scraping?

Getting data from the internet

(in a structured automated way)

II - R Packages for web scraping (selected)

xml2 Parsing HTML & XML

rvest parse common html structures (e.g. tables)

xmlview View pretty HTML/XML, explore XPath

httr* Working with APIs / http protocol

jsonlite* Parse JSON

II - R Packages for web scraping (selected)

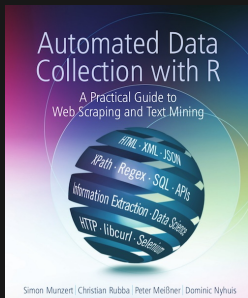
xml2 Parsing HTML & XML

rvest parse common html structures (e.g. tables)

xmlview View pretty HTML/XML, explore XPath

httr* Working with APIs / http protocol

jsonlite* Parse JSON



* I won't cover APIs here.

III - Scraping structured web-pages

What is structured data?

```
<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

III - Demo

Extract election results from wikipedia

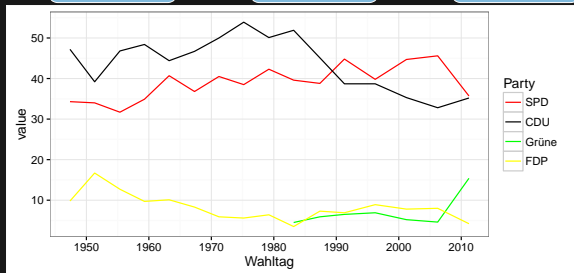
https://de.wikipedia.org/wiki/Landtagswahlen_in_Rheinland-Pfalz



III - Demo

Extract election results from wikipedia

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



IV - 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.

<http://webbook.nist.gov/cgi/cbook.cgi?ID=50-00-0&Units=SI>

```
<h1 id="Top">Formaldehyde</h1>
<ul>
<li>
<strong>
<a title="IUPAC definition of relative molecular mass (molecular weight)">Molecular weight</a>
:
</strong>
30.0260
</li>
<li>
<strong>IUPAC Standard InChIKey:</strong>
<span style="font-family: monospace;">WSFSSNUMVMOMR-UHFFFAOYSA-N</span>
</li>
<li>
<strong>CAS Registry Number:</strong>
50-00-0
</li>
```

IV - XPath

- ▶ XPath is a query language for selecting parts (nodes).

`\\a` Select all a elements (links)

`\\li\\a` Select all links within li

`\\li\\a[3]` Select third link within li

- ▶ Inspectors can export paths (fragile)
- ▶ Build **robust** Xpaths
- ▶ http://www.w3schools.com/xsl/xpath_intro.asp

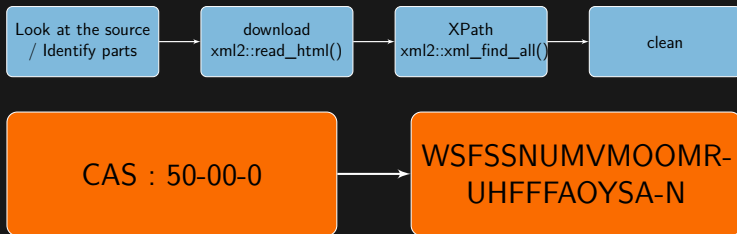
IV - RegEx

Regular Expressions

- ▶ For residual cleaning of characters
- ▶ E.g. Split Value and Unit
- ▶ gsub() & Co
- ▶ `http://www.regular-expressions.info`

IV - Demo

Extract Inchikeys from `webbook.nist.gov`



V - Automatisisation

- ▶ Wrap into function (Input = URL)
- ▶ Function should return vector or list
- ▶ Build URLs
- ▶ Iterate with l/sapply

V - Demo Automatisisation



CAS	inchikey	mw
50-00-0	WSFSSNUMVMOOMR-UHFFFAOYSA-N	30.026
126-86-3	LXOFYPKXCSULTL-UHFFFAOYSA-N	226.355
28159-98-0	HDHLIWCXDDZUFH-UHFFFAOYSA-N	253.367
1461-25-2	AFCAKJKUYFLYFK-UHFFFAOYSA-N	347.167
120-18-3		
25637-99-4		

VI - Remarks

- ▶ Error handling?
- ▶ Robustness? (e.g. Change in website)
- ▶ APIs! (functions provided by servers to query data)
- ▶ ROpenSci (e.g. webchem) provides good functionalities
- ▶ `forum.r-statistik.de`

VI - Lessons learned

- ▶ 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))`
- ▶ legal?
 - ▶ even slower
 - ▶ anonymous EC2
 - ▶ TOR

Web Scraping with R

Eduard Szöcs

Institute for Environmental Sciences, UnVersity of Koblenz-Landau



<http://edild.github.io/>



https://github.com/edild/talk_webscrapingr



@EduardSzoecs



szoecs@uni-landau.de

