An introduction to Web Scraping and Text Mining with R

Simon Munzert
University of Konstanz

October 2014
An introduction to Web Scraping with R

Simon Munzert
University of Konstanz

October 2014
### Session overview

<table>
<thead>
<tr>
<th>Session</th>
<th>Topics</th>
<th>Book chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fri, 10/03</td>
<td>Scraping static content using...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>... XML/HTML parsing</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>... XPath/SelectorGadget</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>... Regular expressions</td>
<td>8</td>
</tr>
<tr>
<td>Fri, 10/17</td>
<td>Scraping dynamic content + APIs using...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>... JSON</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>... APIs</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>... AJAX</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>... Selenium</td>
<td>9</td>
</tr>
</tbody>
</table>

What I won’t cover: internals of HTTP, complex parsing techniques, OAuth, databases, advanced workflow
First: ask questions! No matter what...
Web scraping. What? Why?

The World Wide Web is full of various kinds of new data, e.g.:

- open government data
- search engine data
- services that track social behavior

Web scraping

A.k.a. screen scraping, web harvesting. Computer-aided collection of predominantly unstructured data (e.g., from HTML code)

Practical arguments

- financial resources are sparse
- . . . and so is our time
- reproducibility
Real estate prices, London congestion charge

Data retrieved from http://www.zoopla.co.uk
Measuring issue saliency using Wikipedia page view data

Figure 1: Wikipedia article views for "Energiewende" from January 2008 - July 2013
The philosophy behind web data collection with R

- no point-and-click procedure
- automation of download, parsing, and data extraction procedures
- classical screen scraping
- tapping of web services and APIs
- post-processing of text data
- reproducibility
Technologies of the World Wide Web

Technologies for disseminating content on the Web
- HTTP
- XML/HTML
- JSON
- AJAX
- plain text

Technologies for information extraction
- R
- XPath
- JSON parsers
- Selenium
- Regular expressions

Technologies for data storage
- R
- SQL
- binary formats
- plain-text formats
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <title id=1>First HTML</title>
5   </head>
6 <body>
7   I am your first HTML file!
8 </body>
9 </html>
XML Parsing

Parsing

Syntactic analysis of text according to grammatical rules; analysis of the relationship between single parts of text. In programming, input has to be interpreted (e.g., by R) to process the command.
XML Parsing

- HTML/XML documents are human-readable
- HTML tags structure the document
- web user perspective: the browser interprets the code
- web scraper perspective: use the tags to locate information; document has to be parsed first

Parsing in R

- XML package to parse XML-style documents
- high-level functions: htmlParse(), xmlParse()
- other packages for other document types
- import via readLines() is not parsing - the document’s structure is not retained
**XPath**

**Definition**

- XML Path language, a W3C standard
- query language for XML-style documents
- used to locate and extract content

**Why XPath for web scraping?**

- information is structured by layout
- not only content, but context matters
- gold standard of classical screen scraping with R
XPath and R

**Definition**

- XML Path language, a W3C standard
- query language for XML-style documents
- used to locate and extract content

**Why XPath for web scraping?**

- information is structured by layout
- not only content, but context matters
- gold standard of classical screen scraping with R
**XPath and R**

**Procedure**
- load XML package
- parse document
- query document with XPath
- XML package can ‘speak’ XPath!

```r
R> library(XML)
R> parsed_doc <- htmlParse(file = "materials/fortunes.html")

R> xpathSApply(doc = parsed_doc, path = "/html/body/div/p/i")
[[1]]
<i>'What we have is nice, but we need something very different'</i>

[[2]]
<i>'R is wonderful, but it cannot work magic'</i>
```
Robert Gentleman

What we have is nice, but we need something very different

Source: Statistical Computing 2003, Reisensburg

Rolf Turner

R is wonderful, but it cannot work magic
answering a request for automatic generation of 'data from a known mean and 95% CI'

Source: R-help

The book homepage
Web Scraping with R

Robert Gentleman

What we... Statistical...

Robert Turner

R is... answering...

Source...

Source...

R-help

href: http...
R’s functionality for working with the Web

• managing file downloads
• import and parsing of XML and JSON content
• tapping REST-based web services
• authentication via OAuth
• communication via HTTP, HTTPS, FTP, ...
• automated browsing

For an extensive and up-to-date overview, see:
http://cran.r-project.org/web/views/WebTechnologies.html
Hands-on web scraping with R

You need

- R + Editor (RStudio)
- R packages: `RCurl`, `XML`, `stringr`, `plyr`, `ggplot2`
- R code and data from [https://github.com/simonmunzert/rscraping-intro-duke](https://github.com/simonmunzert/rscraping-intro-duke)
- Internet access
Web scraping etiquette

World Wide Web

Did you identify useful data on the Web? 

- yes
  - Is there an API which offers an interface to a relevant database? 
    - yes 
      - Is there an R package or project that provides a wrapper? 
        - yes 
          - Check out how it works and use it 
        - no 
          - Get familiar with API output and build your own wrapper 
      - no 
    - no 
      - Do you assume a database to exist behind the data? 
        - yes 
          - Is there someone who grants you access to the database? 
            - yes 
              - Retrieve the data from your personal contact and save a lot of time 
            - no 
              - Start scraping and consider all of the aspects on the right 
        - no 
          - Try harder... 

- no
  - Try harder... 

Does robots.txt permit bot action on files you are interested in? 

- yes
  - Is there a robots.txt? 
    - yes 
      - Reconsider your task. Speak to the owner of the data if possible. If you nevertheless start scraping, take into account the ‘Scraping dos and don’ts’ on the right. 
    - no 
      - Are there terms of use which explicitly deny the use of the webpage you have in mind? 
        - yes 
          - Start scraping and consider all of the aspects on the right 
        - no 
          - Try harder...
  - no 
    - Try harder...

Scraping dos and don’ts

- Stay identifiable with User-agent and From header fields, i.e. do not masquerade behind proxies or browser-like user-agents
- Reduce traffic: scrape as few as possible, use gzip if available, choose lightweight formats, monitor changes before scraping (Last-Modified header field) 
- Do not bombard the server with unnecessary requests