Web Data Collection with R

How browsing and scraping works

Peter Meißner / 2016-02-29 - 2016-03-01 / ECPR WSMT

How browsing works . . .

How scraping works . . .

How browsing works . . .

How browsing works . . .

1. start up browser
2. type in *URL*
3. browser asks your computer (client) to connect to another computer (server) via *TCP/IP*
4. once the connection is established the browser can send and receive messages through that connection
5. via some protocol, e.g. *HTTP* (*POP3*, *IMAP*, *FTP*, . . .)
6. and exchange information via
   - request(s) send from client to server and
   - response(s) send from server to client

How browsing works . . .

7. requests and responses might entail
   - all kind of information [head]
     - version of protocol used
     - status codes
     - who is asking
     - who is responding
     - further details
   - as well as further (file) formats, e.g. [body]
     - HTML
     - XML, JSON
     - PNG, JPEG, . . .

How browsing works . . .

8. responses and data are then
   - interpreted by our browser
   - displayed to us
   - *and if need be, new requests are made*

How scraping works . . .

How scraping works . . .

1. we establish a **connection** to server
2. make a **request**
3. get back **responses**
4. try to make sense of **content**
5. *clean, munge, save, analyse, visualize data*
6. *make new request if need be*

When we want to do scraping . . .

. . . we should get . . .

- connection and protocols work
- pose valid requests to server
- get data received from server into R

. . . but most of the stuff is take care off by . . .

- packages for connections and protocols (RCurl, httr, curl) and
- packages for parsing/extraction (XML, jsonlite, rvest, stringr) as well as
- special purpose packages (e.g. twitteR, wikipediatrend)

When we want to do scraping . . .

. . . still we should have some knowledge about . . .

- **connections and protocolls**
  - HTTP (HTTPS, FTP, . . .)
  - URL
  - cookies
- **content**
  - HTML, XML
  - JSON
  - . . .
- **extraction**
  - Regular Expressions
  - Xpath, CSS-Selectors