Introduction to Web Scraping with R

Dynamic Webpages

Simon Munzert | IPSDS

Static webpages

The simple world of static webpages

Static webpages

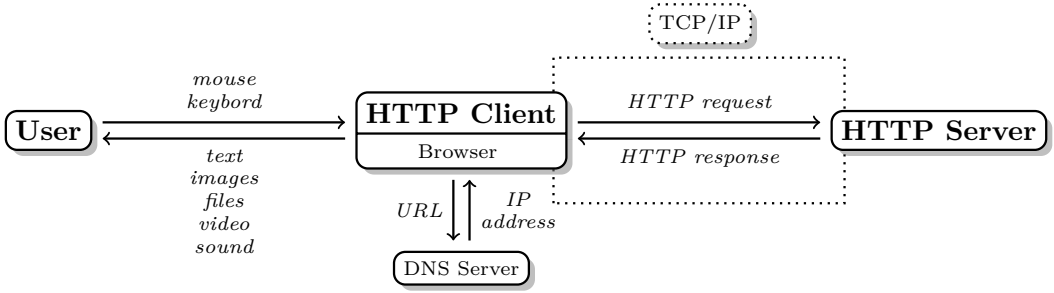
- every user who visits the site gets the same content (unless the developer edits the source code)
- example: https://www.jstatsoft.org
- can contain "dynamic features", such as video or sound, but the page itself is not dynamic

Screenshot of Amazon.com in 1997

A tiny bit of HTTP

Client-server communication with HTTP

- HTTP, the **H**yper**t**ext **T**ransfer **P**rotocol, is a stateless protocol
- no information is retained by either sender or receiver
- makes interaction with websites straightforward, but not very exciting



A tiny bit of HTTP

Client-server communication with HTTP

1. Establishing connection

```
About to connect() to www.r-datacollection.com port 80 (#0)
  Trying 173.236.186.125... connected
Connected to www.r-datacollection.com (173.236.186.125) port 80 (#0)
Connection #0 to host www.r-datacollection.com left intact
```

2. HTTP request

```
GET /index.html HTTP/1.1
Host: www.r-datacollection.com
Accept: */*
```

A tiny bit of HTTP

Client-server communication with HTTP

3. HTTP response

```
1  HTTP/1.1 200 OK
2  Date: Thu, 27 Feb 2014 09:40:35 GMT
3  Server: Apache
4  Vary: Accept-Encoding
5  Content-Length: 131
6  ...

8  <!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
9  <html> <head>
10 <title></title>
11 </head>
12 ...
```

4. Closing connection

```
1  Closing connection #0
```

The "problem" of static webpages

Static webpages reconsidered

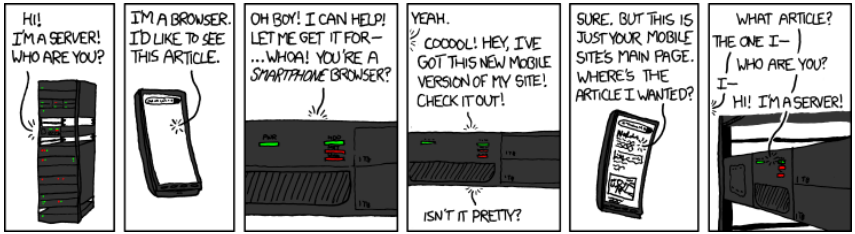
- HTML/HTTP are used for static display of content → same content for every visitor
- in order to display dynamic content, they lack
  1. mechanisms to detect user behavior in the browser (and not only on the server)
  2. a scripting engine that reacts on this behavior
  3. a mechanism for asynchronous queries

Dynamic webpages

The not so simple world of dynamic webpages

Dynamic webpages

- with dynamic webpages, the displayed content can differ between users, even if the source code is the same
- things that can cause the display of different content:
  - operating system, browser, device
  - user actions on the page (mouse movements, scrolling, clicks, keyboard strokes)
  - conditions on the server-side



Source: https://xkcd.com/869/ (Randall Munroe)

The not so simple world of dynamic webpages

Dynamic webpages

- massively used in modern webpage design and architecture
- (are thought to) enhance the user experience
- allow for much more ways to interact with webpage content

The not so simple world of dynamic webpages

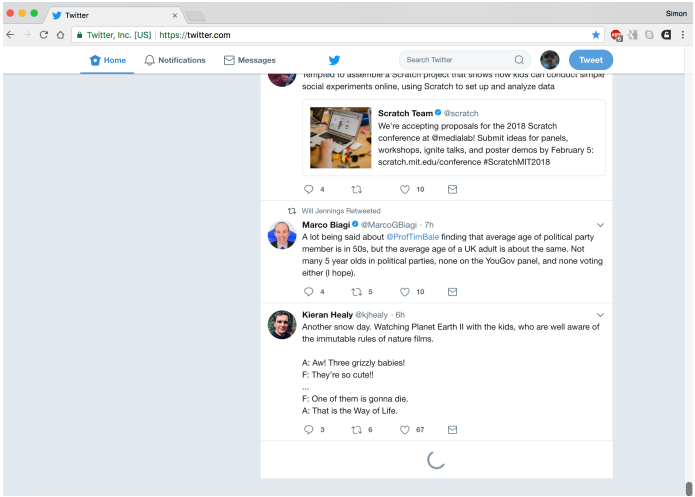
Dynamic webpages

- massively used in modern webpage design and architecture
- (are thought to) enhance the user experience
- allow for much more ways to interact with webpage content

Technologies

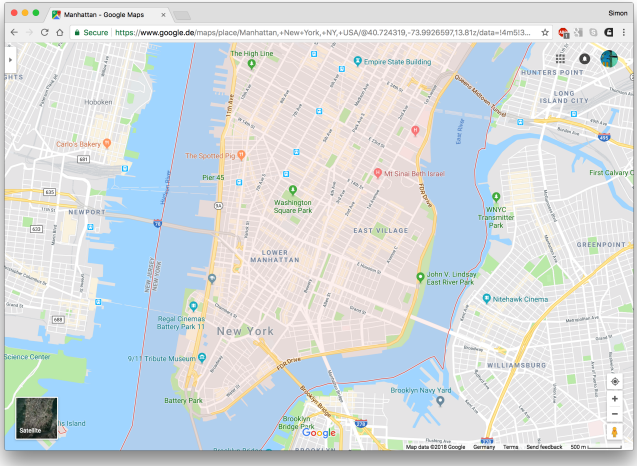
- in the case of **server-side** dynamic webpages, scripts on the server control webpage construction (e.g., PHP script reacts to user input to a form and returns subsets of a database)
- in the case of **client-side** dynamic webpages, (typically) JavaScript embedded in HTML determine what is displayed and how the DOM is changed
- a combination of these technologies is **A**synchronous **J**avaScript **a**nd **X**ML (AJAX)

Examples of dynamic webpages



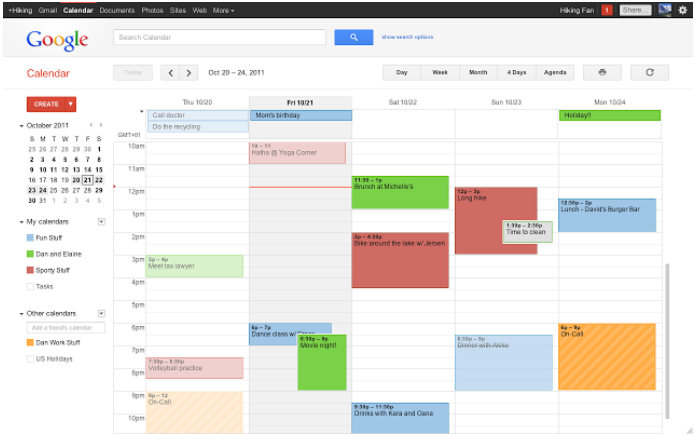
Your Twitter or Facebook feed that automatically gets updated when you scroll down

Examples of dynamic webpages



The map application that automatically loads new content when you zoom in

Examples of dynamic webpages



The calendar app that displays personal information and lets you interact a lot

The problem of dynamic webpages

The problem of dynamic webpages

The problem of dynamic webpages

- the tools you have encountered so far operated on the static source code: you access a page, download it, parse it

The problem of dynamic webpages

- the tools you have encountered so far operated on the static source code: you access a page, download it, parse it
- but what if content on the page changes, but the source code does not?

The problem of dynamic webpages

- the tools you have encountered so far operated on the static source code: you access a page, download it, parse it
- but what if content on the page changes, but the source code does not?
- dynamic webpages make classical screen scraping more difficult if not impossible

The problem of dynamic webpages

- the tools you have encountered so far operated on the static source code: you access a page, download it, parse it
- but what if content on the page changes, but the source code does not?
- dynamic webpages make classical screen scraping more difficult if not impossible
- and: dynamically rendered webpages become more and more common

The problem of dynamic webpages

- the tools you have encountered so far operated on the static source code: you access a page, download it, parse it
- but what if content on the page changes, but the source code does not?
- dynamic webpages make classical screen scraping more difficult if not impossible
- and: dynamically rendered webpages become more and more common
- before we learn about tools that can help us extract data in such scenarios, we will briefly consider the underlying processes, in particular AJAX technologies