Automated Web Data Collection with R

Session 1

Theresa Gessler

last updated: 2020-02-14

Plan of the course

- if you have not, **install packages!!!**
- Session I (Friday, 13.00-15.00)
  - why learn scraping
  - HTML
- Session II (Friday, 15.30-18.00)
  - getting started
  - extracting content with CSS selectors
- Session III (Saturday, 9.00-12.30)
  - scraping multiple pages with loops and functions
  - overview of other techniques (RSS, APIs, Selenium, social media data)
- Session IV (Saturday, 14.00-17.30)
  - exercises

How this course works

- **learning by doing**
  - slides with 'lecture'
  - doing exercises together and alone
  - dedicated exercise session tomorrow
- hopefully, adapted to your level: https://tinyurl.com/r24xk7z

→ if you have specific plans, bring your own ideas tomorrow!

Introduction

About us: Dr. Theresa Gessler

- Who am I?
  - postdoc at Digital Democracy Lab, University of Zurich
  - co-organizer of the Zurich Summer School for Women in Political Methodology
- my research
  - (digital) democracy
  - immigration
  - political parties
- teaching
  - webscraping, text analysis, data journalism
  - (& substantive courses)
- contact
  - gessler@ipz.uzh.ch / www.theresagessler.github.io / @th_ges

About us: Felix Jäger

- Who am I?
  - doctoral student at the University of Bamberg
  - employed in the DFG project "Immigration, Integration, and Naturalization: New Immigrants, Policy Decisions and Citizens' Responses"

- my research
  - attitudes towards immigration
  - terrorism and public support for counter-terrorism

- webscraping experience
  - collecting meta data and images from political sciences journals
  - collecting tweets

- contact
  - felix.jaeger@uni-bamberg.de

About us: Ivo Bantel

- Who am I?
  - PhD student at the University of Zurich
  - Research area "Comparative Politics and Empirical Democracy Research"

- research interests
  - (perceptions of) terrorism
  - far-right politics and political space

- webscraping experience
  - Wikipedia table extraction
  - collecting Facebook posts via API (2018, python)

- contact
  - bantel@ipz.uzh.ch

Your turn

- previous experience: https://tinyurl.com/r24xk7z
- name (infront of you?)
- research interests
- why are you taking this course?
    - any plans that include webscraping?

What is webscraping?

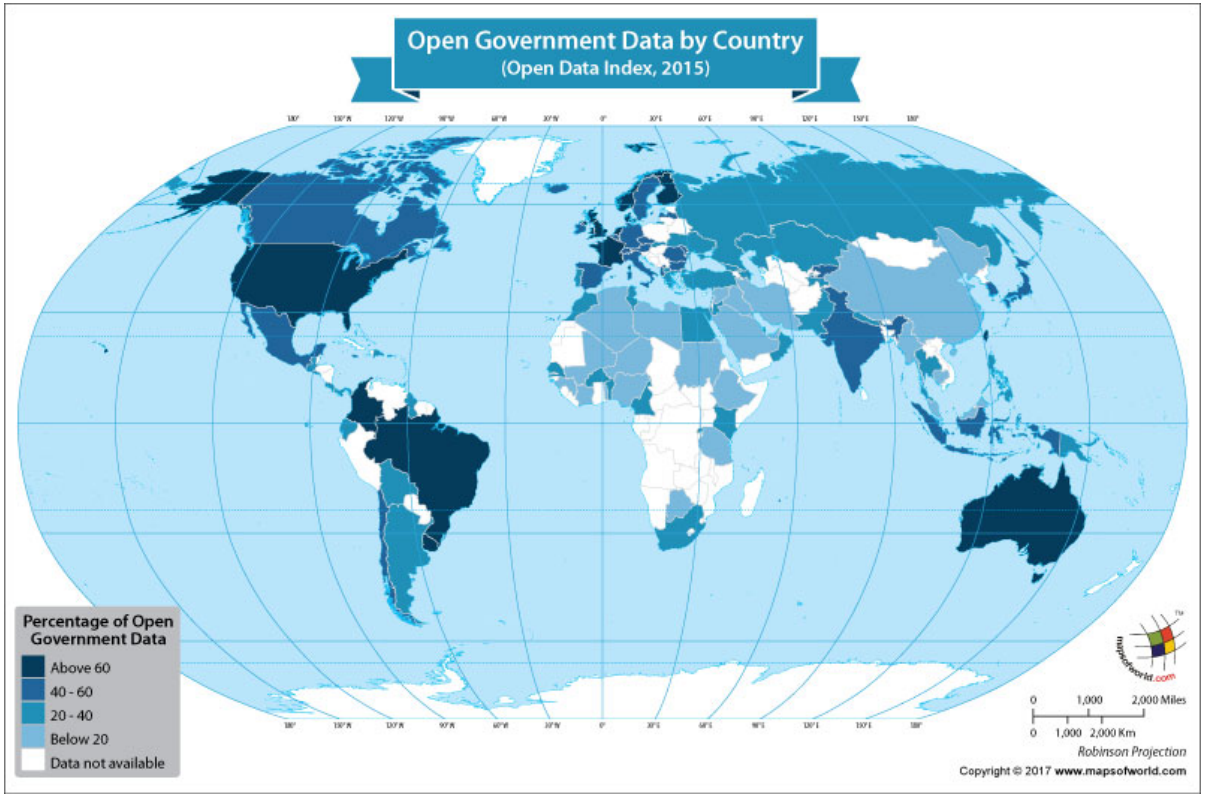
What is webscraping?

- extracting data from webpages
  - anything from university webpage to social media
  - lots of different techniques

- types of scraping
  - gathering as diverse information as possible from different pages vs. very specific scrapers
  - fully automated scrapers to half-automated scripts
  - single-use scraping vs. regular data collection

Why online data?

- increasing amount of public data online ('open government')

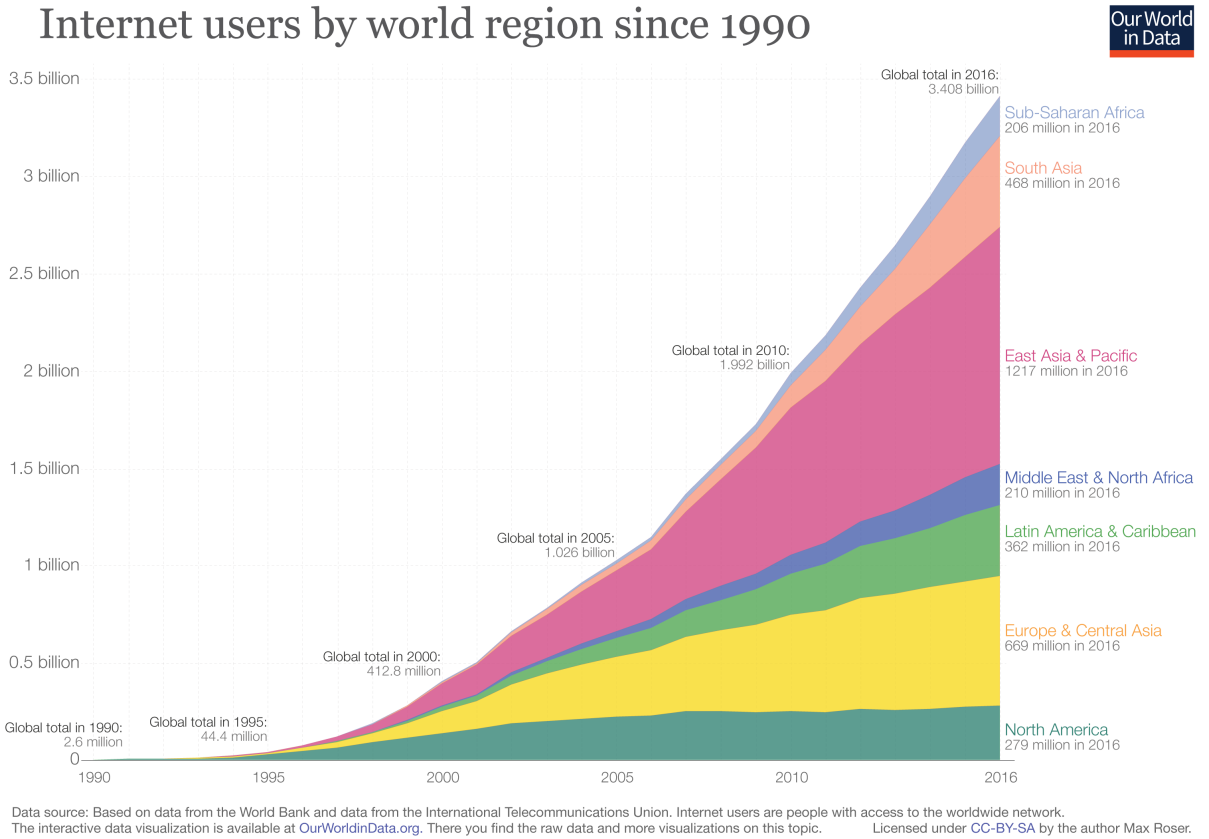


Why online data?

- increasing amount of politics happens online

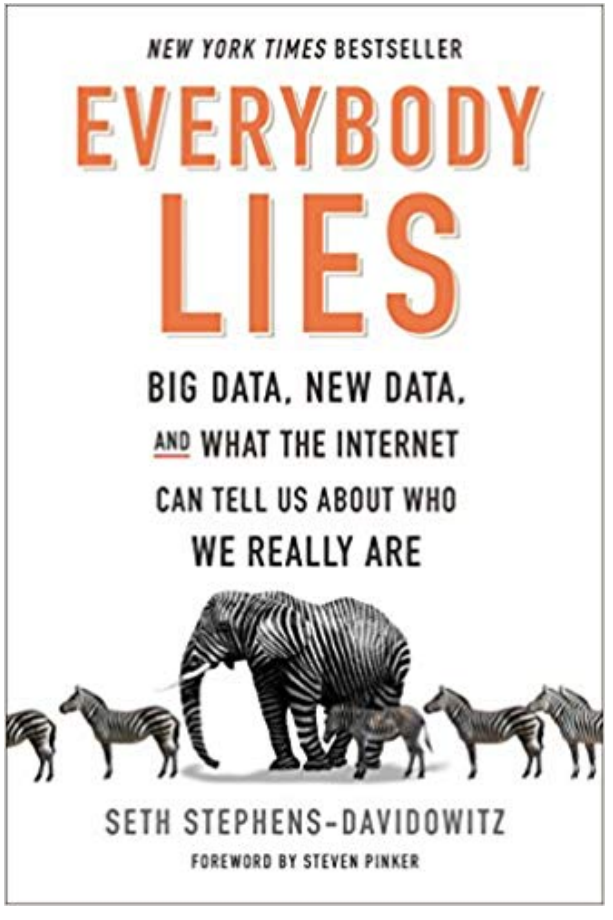
Why online data?

- increasing amount of people use the internet



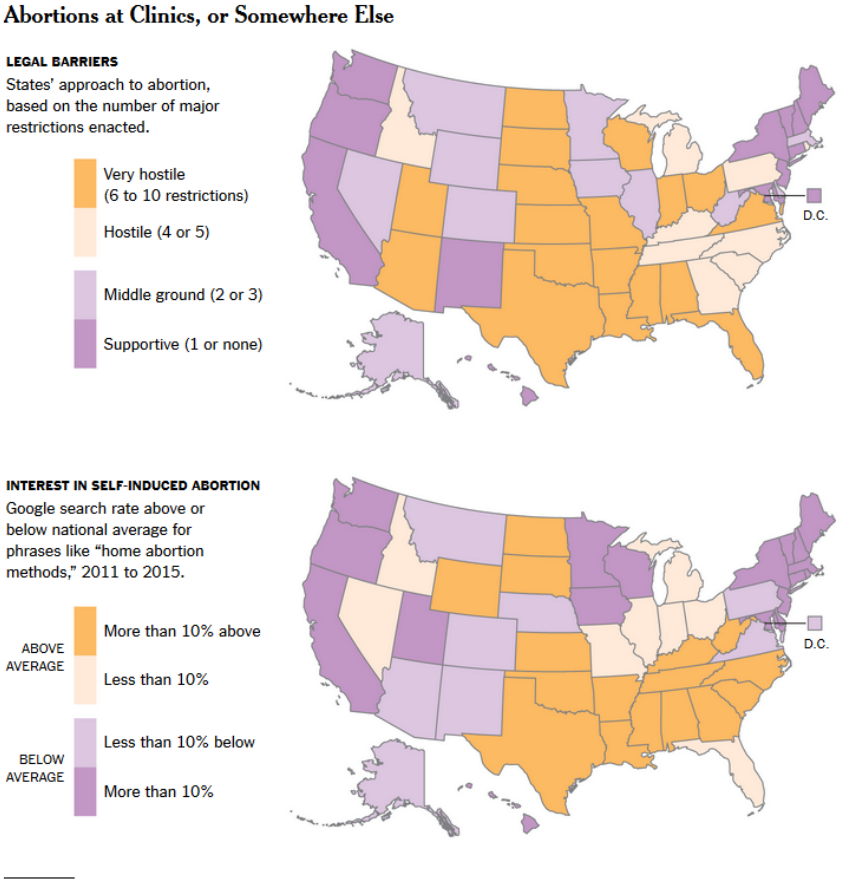
Why online data?

- we share everything online



Why online data?

- that makes real world phenomena more visible online



Why should you scrape?

- masses of data
- reproducible and renewable data collection
- once you learned it: simpler than manual data collection

Why should you stop scraping

- important to address viability of project
- less sustainable for large-scale projects
- changes to webpages over time

Why people stop scraping

- unclear legal situation
- data protection
- terms of service

Current debates about scraping

Current debates about scraping

Is scraping legal? Is scraping ethical?

- levels of regulation
  - 'terms of service'
  - technical measures that prevent scraping
  - legal regulations: copyright
  - legal regulations: data protection

- terms of service and technical measures *vary by page*
- legal regulations on copyright *vary by country* but are often unclear
  - e.g. recent 9th circuit ruling
- data protection regulations like GDPR often have *research exceptions*

Current debates about scraping

Is scraping legal? Is scraping ethical?

- companies prevent scraping
  - to protect their copyright
  - to reduce traffic to their pages
- terms of service *protect interest of company, not research subjects*
- good practices
  - reading terms of services and considering non-intrusive ways to gather data
  - economic considerations: reducing traffic
  - consider research subjects

Current debates about scraping

Is scraping legal? Is scraping ethical?

- clear ethical boundaries
  - data protection: data means traces of individuals
  - right to be forgotten
- good practices
  - secure storage vs. deletion of data
  - anonymization of users

Current debates about scraping

Do we need to scrape?

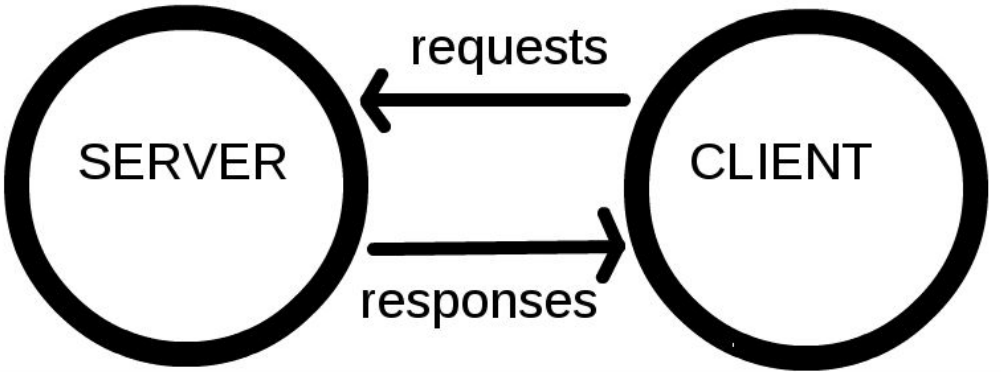
- 'post-API age' / APIcalypse
  - companies restrict data access and inhibit independent research
- commercial marketing of data access, e.g. 'firehose access' → researchers as watchdogs with limited resources

Three pieces of advice from Freelon (2018)

- use authorized methods whenever possible
- do not confuse terms of service compliance with data protection
- understand the risks of violating terms of service

HTML

Browsing vs. scraping



More info on How the web works

- browsing
  - you click on something
  - browser sends request to server that hosts webpage
  - server returns resource (e.g. HTML document)
  - browser interprets HTML and renders it in a nice fashion

→ First step is to learn to understand some HTML

HTML: The basics

- **H**yper **T**ext **M**arkup **L**anguage
  - *markup*: additional description of formatting beyond the content of the text
- language consists of **HTML tags** to specify character / behaviour of text
- HTML tags typically consist of a starting and an end tag (exceptions: images, line breaks etc.)
- they surround the text they are formatting

Example:

```
<tagname>Content goes here...</tagname>
```

- example page we will use: http://quotes.toscrape.com/

Example: In the browser

Example: HTML Code

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Quotes to Scrape</title>
    <link rel="stylesheet" href="/static/bootstrap.min.css">
    <link rel="stylesheet" href="/static/main.css">
</head>
<body>
    <div class="container">
        <div class="row header-box">
            <div class="col-md-8">
                <h1>
                    <a href="/" style="text-decoration: none">Quotes to Scrape</a>
                </h1>
            </div>
            <div class="col-md-4">
                <p>

                    <a href="/login">Login</a>

                </p>
            </div>
        </div>


<div class="row">
    <div class="col-md-8">

    <div class="quote" itemscope itemtype="http://schema.org/CreativeWork">
        <span class="text" itemprop="text">“The world as we have created it is a process of our thinking. It cannot be changed without
changing our thinking.”</span>
        <span>by <small class="author" itemprop="author">Albert Einstein</small>
        <a href="/author/Albert-Einstein">(about)</a>
        </span>
        <div class="tags">
            Tags:
            <meta class="keywords" itemprop="keywords" content="change,deep-thoughts,thinking,world" /    >

            <a class="tag" href="/tag/change/page/1/">change</a>

            <a class="tag" href="/tag/deep-thoughts/page/1/">deep-thoughts</a>

            <a class="tag" href="/tag/thinking/page/1/">thinking</a>

            <a class="tag" href="/tag/world/page/1/">world</a>

        </div>
    </div>
```

Basic HTML tags

```
<html>
    <head>
        <title>Title of your web page</title>
    </head>
    <body>
    HTML web page content
    </body>
</html>
```

- we are mostly interested in what is inside the **body**, that is, the content of a webpage
- **head** gives meta information, often used by search engines
- tags can be **nested**

Basic HTML Tags: Headings

Headings are defined by numbered h tags. Examples (with code and outcome):

```
<h1> your heading</h1>
```

your heading

```
<h2> a smaller heading</h2>
```

a smaller heading

```
<h3> an even smaller heading</h3>
```

an even smaller heading

```
<h4> an even smaller heading</h4>
```

Basic HTML Tags: Paragraphs

Paragraphs are defined by div or p tags.

Examples:

```
<p>this is a paragraph.</p><p>and this is the next.</p>
```

this is a paragraph.

and this is the next.

```
<div>this is a paragraph.</div><div>and this is the next.
</div>
```

this is a paragraph.
and this is the next.

Basic HTML Tags: Attributes

- All HTML elements can have attributes
- Attributes provide additional information about an element
  - they are included inside the tag

Usage

- they are always specified in the starting tag
  - e.g. `<tagname attribute="x"> Title </tagname>`
- Attributes usually come in name and value pairs
  - e.g. attributename="attributevalue"

Basic HTML Tags: Attributes - Links

- Most common case of attributes: **links**
  - text or images turned into a link by surrounding `<a>` tag (*anchor*)
  - link address specified as href attribute (*hyperreference*)

Example:

```
This is text <a href="http://quotes.toscrape.com/">with a
link</a>.
```

This is text with a link.

Basic HTML Tags: Attributes

- other examples of attributes
  - alt: descriptions, e.g. for images
    - for users with visual impairments, when image is missing

Examples:

```
<img src="no_smiley.jpg" alt="Image that does not exist.">

<p style="color:red">This is a paragraph.</p>
```

This is a paragraph.

Basic HTML Tags: Classes

- **Classes** are another special case of attributes that is used for formatting
  - usage within tags:

```
<div class="container"> This is the text</div>
```

This is the text

Styling with Classes

- common styling for repeated instances of elements across webpage
  - reduces formatting errors and repetitions
  - particularly frequent in blogs

```
<style>
p.error {
  color: red;   border: 1px solid red;
}
</style>
<p class="error">Red highlight</p>
```

Red highlight

Example: Quotes to scrape Webpage

Have another look at the webpage - do you understand more now?

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Quotes to Scrape</title>
    <link rel="stylesheet" href="/static/bootstrap.min.css">
    <link rel="stylesheet" href="/static/main.css">
</head>
<body>
    <div class="container">
        <div class="row header-box">
            <div class="col-md-8">
                <h1>
                    <a href="/" style="text-decoration: none">Quotes to Scrape</a>
                </h1>
            </div>
            <div class="col-md-4">
                <p>

                    <a href="/login">Login</a>

                </p>
            </div>
        </div>


<div class="row">
    <div class="col-md-8">

    <div class="quote" itemscope itemtype="http://schema.org/CreativeWork">
        <span class="text" itemprop="text">“The world as we have created it is a process of our thinking. It cannot be changed without
changing our thinking.”</span>
        <span>by <small class="author" itemprop="author">Albert Einstein</small>
        <a href="/author/Albert-Einstein">(about)</a>
        </span>
        <div class="tags">
            Tags:
            <meta class="keywords" itemprop="keywords" content="change,deep-thoughts,thinking,world" /    >

            <a class="tag" href="/tag/change/page/1/">change</a>

            <a class="tag" href="/tag/deep-thoughts/page/1/">deep-thoughts</a>

            <a class="tag" href="/tag/thinking/page/1/">thinking</a>

            <a class="tag" href="/tag/world/page/1/">world</a>

        </div>
    </div>
```