Exercise 4: Data Collection & Quality

Exercise 4 for the lecture ‘Foundations of Data Science’

Prof. Dr. Karsten Donnay, Assistant: Marcel Blum

This session covers

- Scraping websites
- Legal issues
- Reproducibility

University of Zurich^UZH

Department of Political Science

Scraping websites

Scraping websites

- Websites might provide data that can be used for research.
- Basic scraping workflow for static websites:

Step	Task	R commands

1	Download the website	`download.file()`
		`RCurl::getUrl()`
		`RSelenium::remoteDriver()$getPageSource()`

2	Read document into R ('parse')	`xml2::read_html()`
3	Identify XPath to information of interest	Use e.g. Browser + Inspect or Selector Gadget.

4	Extract information of interest	`rvest::html_nodes()`
		`rvest::html_children()`
		`rvest::html_names()`

Scraping websites

- Some websites only provide information after an interaction with the websites. Examples are:
  - Information that requires a log in to the website
  - Content that loads only after we scrolled down far enough.
  - …
- To access dynamic websites we use Selenium WebDrivers to interact with a website and then download it.

Legal issues

Legal issues

- Is it legal to scrape websites?
- No clear “Yes” or “No”: If there was any legal action it was mostly about
  - Privacy concerns
  - Commercial damage
  - Large data crawled

Legal issues

- For your own work:
  - Respect copyrights and abide by national law
  - If in doubt: get the confirmation of the website provider
  - In the end, you are the one who is responsible for any infringements!

  - One indicator are *robots.txt*-files on the websites.

Legal issues

- Robots.txt
  - Documentation of permissions and restrictions of bots to content on a website.
  - Usually accessible in the root directory of a website (e.g. www.karstendonnay.net/robots.txt)
  - Robots.txt - files are not some kind of firewall but only recommendations.

  - Most importantly: have a look at the basic rules (‘*’)

```
User-agent: *
Disallow: /
```

would mean a general ban of everything.

Legal issues

- Scraping etiquette:
  - Identify yourself
  - Only make meaningful requests and not too frequently
  - Consider other data resources: is there an API? Has there at anytime been a complete download of a website/database?

Reproducibility

Reproducibility

- Online data is subject to frequent changes
  - Websites change their structure
  - Old content does not get archived (publicly)
  - Comments get deleted

- As researchers, we need to document our work and make it accessible to others
  - Save local copies of scraped websites
  - Keep track of the date of the download
  - Check if you are allowed to publish the content
    - Anonymize personal information before publication
    - Remove copyrighted content

Reproducibility

- Additionally:
  - EUGDPR specifies how to save data:
    - Location of hosting server might be important
    - Access to data needs to be limited physically and with passwords
    - Careful when saving data (e.g. in your Dropbox folder or adding it temporarily to a github repository)