DIGITAL DEMOGRAPHY: ANALYZING WEB AND SOCIAL MEDIA DATA

WORKING WITH WEB-APIS 1

EDSD DECEMBER 2024 PARIS

TOM THEILE

DEPARTEMENT OF DIGITAL AND COMPUTATIONAL DEMOGRAPHY

DISADVANTAGES OF SCRAPING

- Websites are finicky and prone to change
- CSS-selectors that work today might not work tomorrow
- Webpages are made to be read by humans, not R-scripts

You are not alone if you experienced problems while webscraping.

APIS

APIs!

→ Application Programming Interfaces

→ Interfaces that are meant to be used by machines

APIS

APIs!

→ Application Programming Interfaces

→ Interfaces that are meant to be used by machines

There are different classes of APIs:

- Operating System APIs (e.g.: write a file to a folder)
- R-package APIs (e.g.: read a dataframe from a csv-file with read.csv())
- Web-API
- == webpage for machines

WEB-APIS

Web APIs are "websites for machines". They provide machine-readable access to to data and services.

While websites are designed to be read by humans, APIs are designed to be read by machines.

That means that the answers of APIs don't need to be formatted in a fancy style (like a flashy website), but must follow predictable rules.

FROM SCRAPING TO WEB-APIS

Web APIs use (mostly) the same protocols

APIs are structured and machine readable

RESTful APIs - most common nowadays

GraphQL - can make complex queries, newer, not as common

APIs can respond with any data type, but most often: JSON

RESTFUL WEB-APIS

- Uniform interface
- Client-server decoupling
- Statelessness
- Cacheability
- Layered system architecture

FROM SCRAPING TO WEB-APIS

APIs can be accessed through the browser:

https://official-joke-api.appspot.com/jokes/random

You can pass variables to the server (just like r-functions):

https://official-joke-api.appspot.com/jokes/random/2

https://official-joke-api.appspot.com/jokes/random/5

JSON DATA

Most web-APIs respond with JSON:

Key:value

Keys can be strings or numbers,

Keys must be unique

Values can be strings, numbers, objects or lists

Lists can contain strings, numbers, objects or lists

```
{
  "employees":[
    {"firstName":"John", "lastName":"Doe"},
    {"firstName":"Anna", "lastName":"Smith"},
    {"firstName":"Peter", "lastName":"Jones"}
  ]
}
```

JSON DATA

Very versatile for structured data.

Can contain table-like data,

but also hierarchical data

```
{
  "employees":[
    {"firstName":"John", "lastName":"Doe"},
    {"firstName":"Anna", "lastName":"Smith"},
    {"firstName":"Peter", "lastName":"Jones"}
  ]
}
```

JSON DATA

Can be queried effectively in

most modern programming languages

J1$employees[[0]]$firstName

→ ???

https://cran.r-project.org/web/packages/jsonlite/vignettes/json-aaquickstart.html

```
J1= {
  "employees":[
    {"firstName":"John", "lastName":"Doe"},
    {"firstName":"Anna", "lastName":"Smith"},
    {"firstName":"Peter", "lastName":"Jones"}
  ]
}
```

EXAMPLES OF WEB-APIS

https://open-meteo.com/en

https://archive-api.open-meteo.com/v1/era5?latitude=41.39&longitude=2.16&start_date=1985-10-24&end_date=1985-10-26&daily=temperature_2m_max,temperature_2m_min&timezone=auto

https://docs.openalex.org/about-the-data/work

https://api.openalex.org/works?filter=author.id:A773951722

https://api.openalex.org/authors?filter=display_name.search:tom+theile as webpage

WEB-APIS

Please open
03_02_using_simple_web_APIs.R

WEB-APIS

Authentication

Most APIs are not open, but only available to authenticated accounts.

Methods:

* Authentication token: Log in, get a token, append token to every request

  Long string, e.g.: „SDFI93KOEM2M40b01AaeopfE320MOkk2qQP“

* Username/Password: add username/password to every request

THANK YOU FOR YOUR ATTENTION!

Tom Theile

Research Software Engineer

theile@demogr.mpg.de