

Introduction to Web Scraping with R

Summary



Simon Munzert | IPSDS

Level Up!

Skills

HTML Nerdery



Regex Magic



XPath Mastery



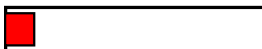
Selenium Wizardry



Ethical Awareness



API Brilliancy



Congrats!

Level Up!

Skills

HTML Nerdery



Regex Magic



XPath Mastery



Selenium Wizardry



Ethical Awareness



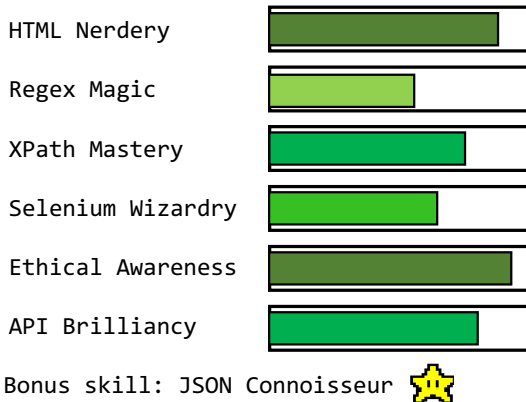
API Brilliancy



Congrats!

Level Up!

Skills



Congrats!

Lessons learned

You...

Lessons learned

You...

- learned about the logic of web APIs.

Lessons learned

You...

- learned about the logic of web APIs.
- got an overview of the universe of existing R API clients and how to find them.

Lessons learned

You...

- learned about the logic of web APIs.
- got an overview of the universe of existing R API clients and how to find them.
- know the basics of JSON, a standard web data interchange format.

Lessons learned

You...

- learned about the logic of web APIs.
- got an overview of the universe of existing R API clients and how to find them.
- know the basics of JSON, a standard web data interchange format.
- learned how to access APIs using R from scratch, even if this involves authentication.

R packages encountered

R packages encountered

- the `stringr` package for string manipulation with regular expressions
- the `xml2` package for working with XML/HTML files
- the `rvest` package for downloading and manipulating HTML files
- the `RSelenium` package to connect to the Selenium API and scrape dynamic webpages
- the `jsonlite` package for working with JSON data
- lots of other packages that serve as clients for web APIs

Inspiration given

Inspiration given

- Start experimenting with little APIs and API clients

Inspiration given

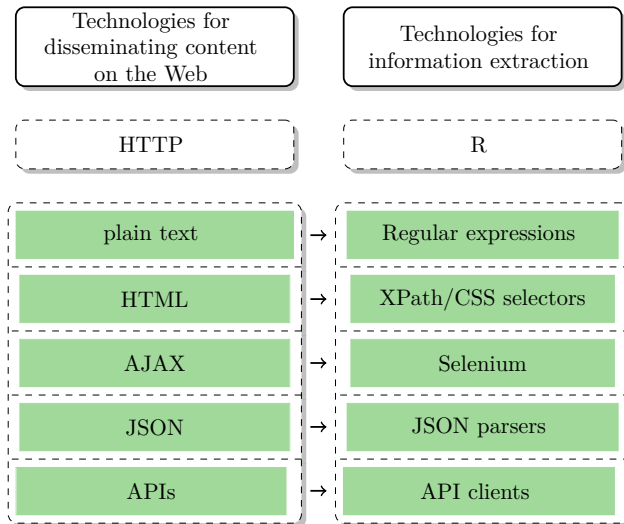
- Start experimenting with little APIs and API clients
- Are you on social media? Try to read out your data using one of the existing R social media clients!

Inspiration given

- Start experimenting with little APIs and API clients
- Are you on social media? Try to read out your data using one of the existing R social media clients!
- Check out some of the following online resources:
 - <http://www.json.org/> – everything you need to know about JSON
 - <https://ropensci.org/packages/> – a great resource for R web clients relevant for the scientific community and beyond

Outlook

Outlook



Outlook

- Keep on scraping the web!

Outlook

- Keep on scraping the web!
- Integrate more packages in your scraping toolbox!

Outlook

- Keep on scraping the web!
- Integrate more packages in your scraping toolbox!
- Get inspired by cutting-edge social science research that exploits web data!

Outlook

- Keep on scraping the web!
- Integrate more packages in your scraping toolbox!
- Get inspired by cutting-edge social science research that exploits web data!
- **And finally:** Stay safe during your scraping adventures!