

Introduction to Web Scraping with R

Summary



Simon Munzert | IPSDS

Level Up!

Skills

HTML Nerdery



Regex Magic



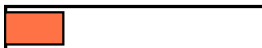
XPath Mastery



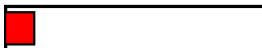
Selenium Wizardry



Ethical Awareness



API Brilliancy



Congrats!

Level Up!

Skills

HTML Nerdery



Regex Magic



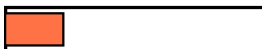
XPath Mastery



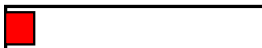
Selenium Wizardry



Ethical Awareness



API Brilliancy



Congrats!

Level Up!

Skills

HTML Nerdery



Regex Magic



XPath Mastery



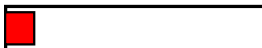
Selenium Wizardry




Ethical Awareness



API Brilliancy



Bonus skill: SelectorGadget Hero 



Congrats!

Lessons learned

You...

Lessons learned

You...

- know how to access and parse HTML source code.

Lessons learned

You...

- know how to access and parse HTML source code.
- have learned ways to identify and extract specific elements from HTML documents.

Lessons learned

You...

- know how to access and parse HTML source code.
- have learned ways to identify and extract specific elements from HTML documents.
- have learned XPath, an XML/HTML query language.

Lessons learned

You...

- know how to access and parse HTML source code.
- have learned ways to identify and extract specific elements from HTML documents.
- have learned XPath, an XML/HTML query language.
- never ever again have to copy and paste HTML tables by hand.

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

Inspiration given

Inspiration given

- Start scraping data from your favorite web pages

Inspiration given

- Start scraping data from your favorite web pages
- Force yourself to write XPath expressions from scratch and test them in the wild

Inspiration given

- Start scraping data from your favorite web pages
- Force yourself to write XPath expressions from scratch and test them in the wild
- Check out some of the following online resources:
 - <http://flukeout.github.io/> – a great interactive tutorial to learn how CSS selectors work
 - <https://cran.r-project.org/web/packages/rvest/rvest.pdf> – literally the documentation of the **rvest** package
 - <http://cran.r-project.org/web/views/WebTechnologies.html> – CRAN Task View on Web Technologies and Services - useful to stay in the loop of what's possible with R