

# Introduction to Web Scraping with R

## Summary



Simon Munzert | IPSDS

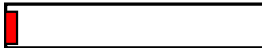
# Level Up!

## Skills

HTML Nerdery



Regex Magic



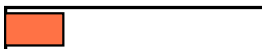
XPath Mastery



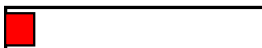
Selenium Wizardry



Ethical Awareness



API Brilliancy



Congrats!

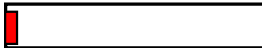
# Level Up!

## Skills

HTML Nerdery



Regex Magic



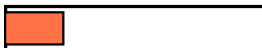
XPath Mastery



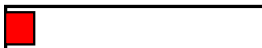
Selenium Wizardry



Ethical Awareness



API Brilliancy



Congrats!

# Level Up!

## Skills

HTML Nerdery



Regex Magic



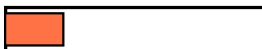
XPath Mastery



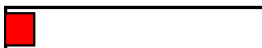
Selenium Wizardry



Ethical Awareness



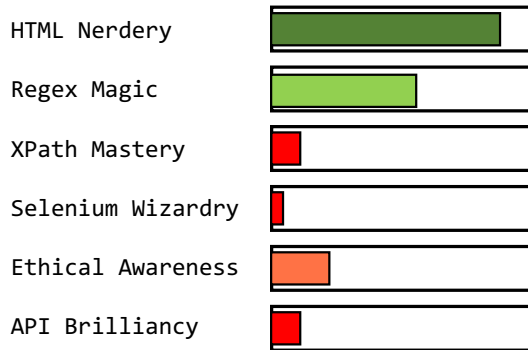
API Brilliancy



Congrats!

# Level Up!

## Skills



Bonus skill: String Manipulation Geek ★



Congrats!

# Lessons learned

You...

## Lessons learned

You...

- got an overview of what's possible in terms of web scraping with R.

## Lessons learned

You...

- got an overview of what's possible in terms of web scraping with R.
- know how HTML documents are basically structured.



## Lessons learned

You...

- got an overview of what's possible in terms of web scraping with R.
- know how HTML documents are basically structured.
- are able to use regular expressions to identify patterns in text data.

# Lessons learned

You...

- got an overview of what's possible in terms of web scraping with R.
- know how HTML documents are basically structured.
- are able to use regular expressions to identify patterns in text data.
- learned how to manipulate string data.

R packages encountered

## R packages encountered

- the `stringr` package for string manipulation with regular expressions

Inspiration given

## Inspiration given

- Inspect the HTML code of your favorite web pages

## Inspiration given

- Inspect the HTML code of your favorite web pages
- Write your own HTML document

# Inspiration given

- Inspect the HTML code of your favorite web pages
- Write your own HTML document
- Check out some of the following online resources:
  - <http://w3schools.com> – a great tutorial playground to learn web technologies interactively
  - <http://www.w3.org/> – base of the World Wide Web Consortium (W3C)
  - <http://regexone.com/> – an interactive regex tutorial (not R-based)