Introduction to Web Scraping with R

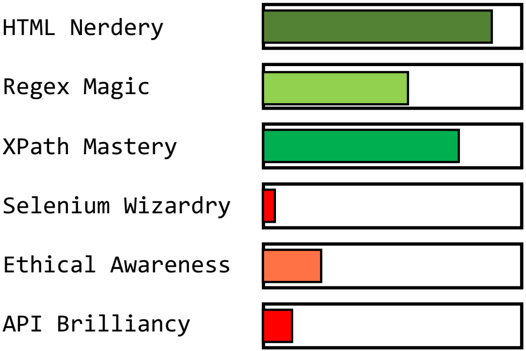
Summary

data
technologies
source
sources
quality
index
marginnote
text
ref
web
ajax
might
scraping
provide
part
collection
one
xml
the
chapter
html
use
example
information
json
techniques
documents
can
book
general

Simon Munzert | IPSDS

Level Up!

Skills



HTML Nerdery

Regex Magic

XPath Mastery

Selenium Wizardry

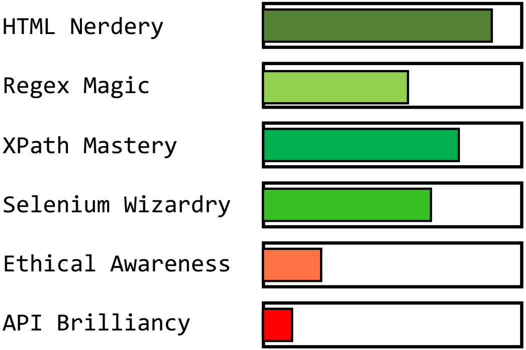
Ethical Awareness

API Brilliancy

Congrats!

Level Up!

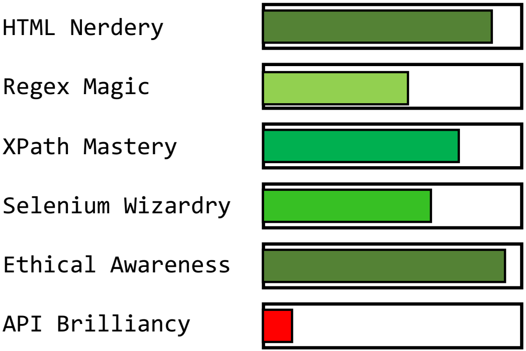
Skills



Congrats!

Level Up!

Skills



HTML Nerdery

Regex Magic

XPath Mastery

Selenium Wizardry

Ethical Awareness

API Brilliancy

Congrats!

Level Up!

Skills

HTML Nerdery

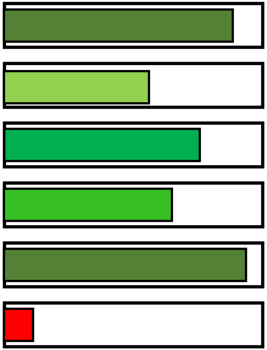
Regex Magic

XPath Mastery

Selenium Wizardry

Ethical Awareness

API Brilliancy



Bonus skill: robots.txt Veteran

Congrats!

Lessons learned

You...

Lessons learned

You...

- learned about the difference between static and dynamic webpages.

Lessons learned

You...

- learned about the difference between static and dynamic webpages.
- recognize JavaScript in the wild.

Lessons learned

You...

- learned about the difference between static and dynamic webpages.
- recognize JavaScript in the wild.
- know how to work with Selenium to gather information from Ajax-generated websites.

Lessons learned

You...

- learned about the difference between static and dynamic webpages.
- recognize JavaScript in the wild.
- know how to work with Selenium to gather information from Ajax-generated websites.
- got an overview of legal and ethical issues around web scraping

Lessons learned

You...

- learned about the difference between static and dynamic webpages.
- recognize JavaScript in the wild.
- know how to work with Selenium to gather information from Ajax-generated websites.
- got an overview of legal and ethical issues around web scraping
- got an overview of practices that help establish a server-friendly workflow

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

`Inspiration` given

Inspiration given

- Try to get RSelenium running on your machine (threads on StackOverflow might be of help).

Inspiration given

- Try to get RSelenium running on your machine (threads on StackOverflow might be of help).
- Try to re-build a typical browser session of yourself using Selenium.

Inspiration given

- Try to get RSelenium running on your machine (threads on StackOverflow might be of help).
- Try to re-build a typical browser session of yourself using Selenium.
- Read the robots.txt of your favorite websites.

Inspiration given

- Try to get RSelenium running on your machine (threads on StackOverflow might be of help).

- Try to re-build a typical browser session of yourself using Selenium.

- Read the robots.txt of your favorite websites.

- Study the Terms of Use / Terms of Service of your favorite websites.

Inspiration given

- Try to get RSelenium running on your machine (threads on StackOverflow might be of help).

- Try to re-build a typical browser session of yourself using Selenium.

- Read the robots.txt of your favorite websites.

- Study the Terms of Use / Terms of Service of your favorite websites.

- Check out some of the following online resources:
  - http://www.seleniumhq.org/ – the entire world of Selenium
  - http://decisionsandr.blogspot.de/2014/04/play-2048-using-r.html – an R script that uses RSelenium to let R automatically play the browser game 2048
  - https://gijn.org/2015/08/12/on-the-ethics-of-web-scraping-and-data-journalism/ – an article on the ethics of web scraping