Scraping the internet with R

Gilian Ponte

31 august 2018

About me

* Gilian Ponte, 24 years, Groningen.

* MSc Marketing Intelligence & Management.

* ex-Web Analyst at Nextail.

* R, Python and JavaScript.

Scraping the internet with R. Why?

* Scraping is a time-saving skill.

* Replace expensive scrape tooling.

* Compare prices with competitors.

* Analyse a competitors' pricing strategy over a longer time period.

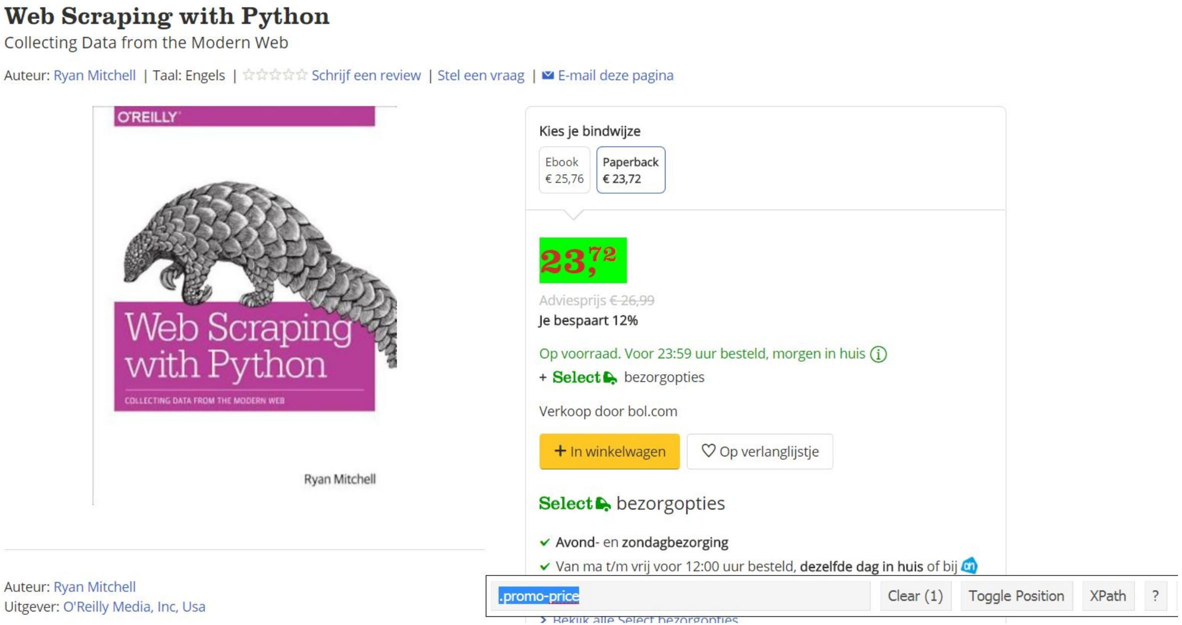
* Content collection.

Before we get to scraping...

* RStudio

* R

* Selector Gadget



Web Scraping with Python

Collecting Data from the Modern Web

Auteur: Ryan Mitchell | Taal: Engels | ☆☆☆☆☆ Schrijf een review | Stel een vraag | E-mail deze pagina

Kies je bindwijze

Ebook € 25,76

Paperback € 23,72

23,72

Adviesprijs ~~€ 26,99~~

Je bespaart 12%

Op voorraad. Voor 23:59 uur besteld, morgen in huis

+ Select bezorgopties

Verkoop door bol.com

Select bezorgopties

✓ Avond- en zondagbezorging

✓ Van ma t/m vrij voor 12:00 uur besteld, **dezelfde dag in huis** of bij

› Bekijk alle Select bezorgopties

Auteur: Ryan Mitchell

Uitgever: O'Reilly Media, Inc, Usa

.promo-price

Who already scraped in R before?

Let’s start scraping!

* Scrape some prices from Bol.com, Amazon.de and AH.nl.

- Scrape one price.
- Build an automatic scraper and compare.
- How to deal with JavaScript.

* **BUT**, this technique is applicable to **ALL** content.

* You should be able to scrape (almost) everything.

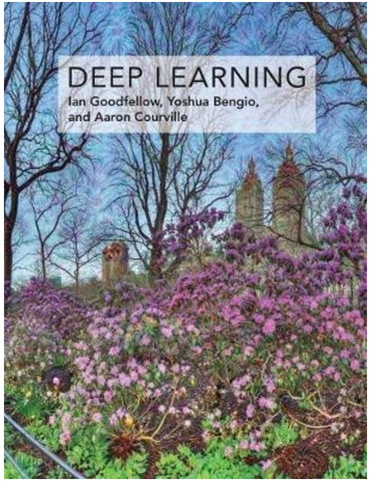
* Download R scripts

Scrape one price from Bol.com.

Let’s focus our attention on Bol.com and RStudio . . .

Build an automatic scraper.

Build an automatic scraper



* Scraping one price or element is cool. But..

  - Compare prices?

  - Long trends?

  - Store multiple prices?

* Start of the academic year. So we are going to compare the prices of books.

Build an automatic scraper

Let’s focus our attention on Github and RStudio . . .

How to deal with JavaScript? (I’m sorry Mac users)

How to deal with JavaScript?

* Some websites use JavaScript to load elements. For example: lazy loading by scrolling. For example: AH.nl.

* How to deal with this in R?

    - Phantom.js

    - To simulate a browser and wait until the page is finished loading.

How to deal with JavaScript?

Let’s focus our attention on RStudio . . .

Want to know more? Or just connect?

Gilian Ponte
gilianponte@gmail.com
https://nl.linkedin.com/in/gilianponte
Full code and slides available at: https://github.com/GilianPonte/R-workshop