Scraping with R

Cincinnati R Users Group
June 21, 2016

Eugene Pyatigorsky

This presentation and supporting materials available at:
https://github.com/epspi/Rscraping

Agenda

Overview of packages
A look at how to scrape
Working example
Best practices

Overview of packages

rvest

Most of the work will be done by Hadley's package `rvest`

> Based on Python's `beautifulsoup`
> Extracts elements from the dom using CSS or XPath

e.g.

```
rvest::read_table()
```

httr

This is (Hadley's) wrapper for `curl`

Really useful for making customized calls to APIs
Can also be used for writing your own APIs

e.g.

```
httr::GET("some_endpoint",
config)
```

How to Scrape: An Example

Let's ask Bing about the R Users Group

```
lnk <- 'http://www.bing.com/search?q=Cincinnati+R+users+group&go=Submit&qs=n&form=

read_html(lnk) %>%
    html_nodes("h2 a") %>%
    html_text
```

```
## [1] "Cincinnati UC Users Group (Cincinnati, OH) - Meetup"
## [2] "Local R User Group Directory - Revolutions"
## [3] "New R User Group in Cincinnati / Dayton - Revolutions"
## [4] "Cincinnati Sharepoint User Group - Facebook"
## [5] "Cincinnati .Net Users Group"
## [6] "CincyPowerShell | PowerShell Community Groups"
## [7] "Reinaldo R. - Cincinnati UC Users Group (Cincinnati, OH ..."
## [8] "Group: Cincinnati |Tableau Support Community"
```

Common CSS Selectors

# for "id="

. for "class="

OR you can use SelectorGadget for Chrome
https://chrome.google.com/webstore/detail/selectorgadget/

A Working Site

Cincinnati Foreclosures - A Real Estate Scraper

Best Practices

Authentication

Use APIs instead of scraping whenever possible. There isn't a lot of documentation for `rvest` and cookie-based authentication can be tricky.

Automation

The real power of `R` and `rvest` shines when used with `shiny` (npi).
Put your scraping code in a standalone R script and automate with `cron`.

End