Teaching web scraping: Integrating data science into statistics

Mine Dogucu
New College of Florida
@MineDogucu
Mine Çetinkaya-Rundel
Duke University
@minebocek
2018-05-21
What is Web Scraping?

Web scraping protocol

CSV
# Hand Scraping

<table>
<thead>
<tr>
<th>Rank &amp; Title</th>
<th>IMDb Rating</th>
<th>Your Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The Shawshank Redemption (1994)</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>2. The Godfather (1972)</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>4. The Dark Knight (2008)</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank &amp; Title</th>
<th>IMDb Rating</th>
<th>Your Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The Shawshank Redemption (1994)</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>2. The Godfather (1972)</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>4. The Dark Knight (2008)</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>
Web scraping

<table>
<thead>
<tr>
<th>title</th>
<th>year</th>
<th>rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Shawshank Redemption</td>
<td>1994</td>
<td>9.2</td>
</tr>
<tr>
<td>The Godfather</td>
<td>1972</td>
<td>9.2</td>
</tr>
<tr>
<td>The Godfather: Part II</td>
<td>1974</td>
<td>9.0</td>
</tr>
<tr>
<td>The Dark Knight</td>
<td>2008</td>
<td>9.0</td>
</tr>
<tr>
<td>12 Angry Men</td>
<td>1957</td>
<td>8.9</td>
</tr>
<tr>
<td>Schindler’s List</td>
<td>1993</td>
<td>8.9</td>
</tr>
<tr>
<td>The Lord of the Rings: The Return of the King</td>
<td>2003</td>
<td>8.9</td>
</tr>
<tr>
<td>Pulp Fiction</td>
<td>1994</td>
<td>8.9</td>
</tr>
<tr>
<td>The Good, the Bad and the Ugly</td>
<td>1966</td>
<td>8.8</td>
</tr>
<tr>
<td>Fight Club</td>
<td>1999</td>
<td>8.8</td>
</tr>
<tr>
<td>The Lord of the Rings: The Fellowship of the Ring</td>
<td>2001</td>
<td>8.8</td>
</tr>
</tbody>
</table>
Example 2

Most Popular Feature Films Released 2017-01-01 to 2017-12-31

   - PG-13 | 130 min | Action, Adventure, Comedy
   - **Rating:** 7.9 | **Metascore:** 74
   - Thor is imprisoned on the planet Sakaar, and must race against time to return to Asgard and stop Ragnarok, the destruction of his world, at the hands of the powerful and ruthless villain Hela.
   - Director: Taika Waititi | Stars: Chris Hemsworth, Tom Hiddleston, Cate Blanchett, Mark Ruffalo
   - Votes: 326,235 | Gross: $315.06M

   - PG | 105 min | Biography, Drama, Musical
   - **Rating:** 7.8 | **Metascore:** 65
   - Celebrates the birth of show business, and tells of a visionary who rose from nothing to create a spectacle that became a worldwide sensation.
   - Director: Michael Gracey | Stars: Hugh Jackman, Michelle Williams, Zac Efron, Zendaya
   - Votes: 120,071 | Gross: $173.88M
Hypertext Markup Language (HTML) Nodes

```html
<html>
<head>
<title>Page Title</title>
</head>
<body>
<h1>My First Heading</h1>
<p>My first paragraph.</p>
</body>
</html>
```

https://www.w3schools.com/html/tryit.asp?filename=tryhtml_intro
Things to Consider

I'm not a robot

reCAPTCHA
Privacy - Terms
```r
library(robotstxt)

paths_allowed("http://www.imdb.com")

#> www.imdb.com
#> [1] TRUE
```
Things to Consider
```r
library(rvest)
library(tidyverse)

page <- read_html("http://www.imdb.com/chart/top")
titles <- page %>%
  html_nodes(".titleColumn a") %>%
  html_text()

head(titles)
#> [1] "The Shawshank Redemption" "The Godfather"
#> [3] "The Godfather: Part II" "The Dark Knight"
#> [5] "12 Angry Men" "Schindler's List"
```

Created on 2018-05-20 by the reprex package (v0.2.0).
bit.ly/ecots2018

Click on Projects

Make your own copy of the project called Examples
Benefits

- Students get exposed to non-standard (non-rectangular) data format.
- Students get large amounts of data in a short span of time and in a tidy format.
- Students get exposure to working with strings.
- Students can have more diverse sources of data for statistics projects.
- Web scraping can bring computing topics (e.g. HTML, functions, loops) into the statistics classroom.
- Instructors can use web scraping to curate datasets for classroom use.
Potential Problems

- Website can be down
- NA values
Notes

- More complex scraping is possible
- Timing in the semester
- Web APIs
- Terms of Use
QUESTIONS?

Mine Dogucu
mdogucu@ncf.edu
@MineDogucu

Mine Çetinkaya-Rundel
mine@stat.duke.edu
@minebocek

bit.ly/ecots2018-webscraping
https://github.com/mdogucu/eCOTS2018