Web Scraping in the Statistics Curricula: Challenges and Opportunities

Mine Dogucu
Denison University
2018-11-14
@MineDogucu
бит.ly/scrape_duke
Have you ever scraped data from the web?
Web Scraping Protocol

CSV
The 'Bohemian Rhapsody' Cast Will Rock You
IMDb feature of the day: [an error occurred while processing this directive]

New to the IMDb? Check out the site introduction or take our tour.
The most comprehensive free source of movie information on the Internet.
Sponsorship and advertising details.

Bandwidth provided by Exec-PC's FilePile

Search - Help - Index - Tour - Linking

Copyright © 1990-1997 The Internet Movie Database Ltd

AmeriSites logo has provided the graphics for this site.

This site is powered by FreeBSD and Apache
Movie credits

As of November 1996 there were over 85,000 titles in the Internet Movie Database, ranging from some of the earliest moving pictures such as Express Train on a Railway Cutting (1898) from last century to several currently in production (we can't show you those, because it'll mean updating this text on a regular basis.. trust us.) and many recent releases.

Each week the number of titles in the database grows to cover almost all the new releases, as well as to fill some holes from earlier days.

All types of movie are covered, for example

- blockbusters - Sleepless in Seattle (1993),
- classics - Citizen Kane (1941),
- cult following - Rocky Horror Picture Show, The (1975),
- silent - Kid, The (1921),
- bad - Plan 9 from Outer Space (1958),
- animated - Wrong Trousers, The (1993),
- Hollywood - Fugitive, The (1993),
- world-wide - Belle Epoque (1992),
- and much, much more.
Titles: **5,310,913** (Year Range: 1874 - 2025)

- Titles w/ primary image 1,058,871
- Reviews 3,793,406
- Plots 1,807,393
- Trivia 964,144
- Quotes 907,178
- Parental guides 416,387
- Release dates 6,026,114
- Certificates 1,103,092
- Genres 2,863,475
- Keywords 8,150,624
- Running times 1,687,265
- Soundtrack 916,062
Top 250 movies
# Hand Scraping

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank &amp; Title</td>
<td>IMDb Rating</td>
<td>Your Rating</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1. The Shawshank Redemption (1994)</td>
<td>9.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2. The Godfather (1972)</td>
<td>9.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4. The Dark Knight (2008)</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5. 12 Angry Men (1957)</td>
<td>8.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### .xls

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank &amp; Title</td>
<td>IMDb Rating</td>
<td>Your Rating</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1. The Shawshank Redemption</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2. The Godfather</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3. The Godfather: Part II</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4. The Dark Knight</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5. 12 Angry Men</td>
<td>8.9</td>
<td></td>
</tr>
</tbody>
</table>

### .csv

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank &amp; Title</td>
<td>IMDb Rating</td>
<td>Your Rating</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1. The Shawshank Redemption</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2. The Godfather</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3. The Godfather: Part II</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4. The Dark Knight</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5. 12 Angry Men</td>
<td>8.9</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>
Feature Films Released in 2018

1. **Bohemian Rhapsody** (2018)
   - PG-13 | 134 min | Biography, Drama, Music
   - Rating: 8.4
   - Director: Bryan Singer
   - Stars: Rami Malek, Lucy Boynton, Gwilym Lee, Ben Hardy
   - Votes: 72,662 | Gross: $69.16M

2. **Halloween** (I) (2018)
   - R | 106 min | Horror, Thriller
   - Rating: 7.1
   - Director: David Gordon Green
   - Stars: Jamie Lee Curtis, Judy Greer, Andi Macdowell, James Jude Courtney
   - Votes: 44,006 | Gross: $152.97M

   - R | 136 min | Drama, Music, Romance

1-50 of 14,485 titles.
Six recommendations from GAISE

1. Teach statistical thinking.
   - Teach statistics as an investigative process of problem-solving and decision-making.
   - Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. **Integrate real data with a context and purpose.**
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.
Desirable Characteristics of Class Activities

Choosing data...

- Relevance – The activity should involve data about topics that interest students. Using real data makes data relevant to a wide variety of student majors. If real data are not used, then the activity should mimic a real-world situation. It should not seem like “busywork” to students. For example, if you use coins or cards to conduct a binomial experiment, explain real-world binomial experiments they could represent.
Statistics Education Literature

• Nature of statistics is changing. Statistics curricula must adapt to this change. One change needed is inclusion of computational skills (Nolan & Lang, 2010).

• Instructors and institutions that have computational skills will be in demand and students who have these skills will get the jobs (Horton, 2015).

• Some schools have already adopted “Data Science in Statistics Curricula” and they included web scraping (Hardin et. al, 2015).
Keynote by Chief Statistician of the United States, Nancy Potok

What skills are needed

- **Web Scraping**
  - Structured Data (different ways of collecting data)

- **Text Analysis**
  - Text data (making use of new types of data)
  - Relationship data (creating new units of economic activity)

- **Networks**
  - Output data (creating innovation measures)

- **Link Statistical and Other Data**
  - Data Analysis
    - Sense making (from the population)

- **Statistical Inference**
  - Analysis (capturing social and economic dynamics)

- **Longitudinal Database**
  - Data Dissemination
    - Conveying information (in statistically valid context)

- **Visualization**

Photo: Kim Flagg Sellers
Data Science Internship, Spring 2019
Weber Shandwick - Washington, DC 20005

General Responsibilities:

- Understand client background and needs, including general business strategy, industry issues, products and services, key customers and competitors in the marketplace
- Participate in strategic brainstorming sessions when invited by account leads or supervisors
- Create and maintain optimal data pipelines and data integration solutions
- **Assemble large, complex data sets that meet business requirements through web scraping or API’s**
- Identify, design, and implement internal process improvements: automating manual processes, optimizing data delivery, re-designing infrastructure for greater scalability, etc.
- Mine and analyze data to drive optimization and improvement of campaign management, marketing techniques and business strategies
Digital Product Data Analyst
Galco Industrial Electronics - Madison Heights, MI 48071

TASKS, DUTIES, AND RESPONSIBILITIES

- Write & optimize database queries to transform raw, disparate datasets into our standard
- Process data from a range of formats, including CSV, JSON, XML
- Write basic scripts for data manipulation, data fetching, web scraping, and simple automation
- Drive continuous improvement of our processes and internal tools
- Drive automation capabilities with data entry and data extraction
- Working knowledge of SQL or Progress is a plus
- Ability to write and optimize queries
- Joins, aggregation, indexes, and transactions * Ability to write and automate basic scripts for data manipulation and retrieval including CSV, JSON, XML data formats
- General understanding of APIs, command line, client server architecture, networks, and (S)FTP
- Use Excel to standardize and structure product data into our format
- Familiar with load sheets, appropriate keywords, effective copy, and product classifications
- Research products to ensure data accuracy
- Work with internal and external resources (ex. Manufacturers) to ensure products are accurately described
- Maintain departmental records of progress and completed work
Research Analyst
Decision Resources Group - Parsippany, NJ

RESPONSIBILITIES

- Research, review, and enter pharmaceutical product management data efficiently and accurately for all therapeutic drug classes and for all payer segments: Commercial, PBM, Commercial Medicaid, State Medicaid and Medicare plans
- Work with the Analyst team to create best practices for capturing plan changes and management updates
- **Leverage a set technology tools, including proprietary web scraping and natural language processing software to drive efficiency**
- Collaborate with other DRG teams to facilitate a variety of data sets
- Support the timely and appropriate response to client inquiries regarding the data
- Serve as key support for the delivery of data, reports, and ad hoc research assignments
Data Analysis Librarian
UNC-Chapel Hill - Chapel Hill, NC

Minimum Qualifications:

Required ALA-accredited master’s degree in Library or Information Science, or related advanced degree. Coursework in statistics, and knowledge of statistical methods. Strong customer service orientation and excellent communication and interpersonal skills. **Demonstrated advanced data skills, including data cleaning/wrangling/normalization, using regular expressions, and web scraping.** Demonstrated experience with data analysis tools such as R, STATA, SPSS, and SAS. Proficiency with at least one programming language (such as Python, Java, or R). Demonstrated aptitude for quickly learning new tools and technologies. Experience working effectively with a team to plan and complete projects. Demonstrated ability to work with diverse populations as well as demonstrated commitment to diversity, inclusion, and accessibility. Preferred Advanced degree in statistics. Experience providing data-related services in a library or research setting. Experience teaching technology, either one-on-one or in a classroom setting. Proficiency using tools and programming libraries to support text analysis. Geospatial technology skills. Data visualization skills. ALA-accredited master’s degree in Library or Information Science, or related advanced degree.
1. Request
2. HTML document
3. Render

Client → Server
<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Shawshank Redemption</td>
<td>1994</td>
</tr>
<tr>
<td>2</td>
<td>The Godfather</td>
<td>1972</td>
</tr>
<tr>
<td>3</td>
<td>The Godfather: Part II</td>
<td>1974</td>
</tr>
<tr>
<td>4</td>
<td>The Dark Knight</td>
<td>2008</td>
</tr>
<tr>
<td>5</td>
<td>12 Angry Men</td>
<td>1957</td>
</tr>
</tbody>
</table>
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
Hyper Text Markup Language tags

<html>
</html>
HTML tags

<html>
<head>
</head>
<body>
</body>
</html>
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.

What we would like

**Duke University** is a private research university located in Durham, NC. I am currently visiting Duke University.

**Denison University** is a private liberal arts college located in Granville, OH.
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.
<a href="https://www.duke.edu">Duke University</a>
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
Duke University is a private university located in Durham, North Carolina. I am currently visiting Duke University.
Duke University is a private university located in Durham, North Carolina. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
**Duke University** is a private research university located in Durham, NC. I am currently visiting Duke University.

**Denison University** is a private liberal arts college located in Granville, OH.
Duke University is a private university located in Durham, North Carolina. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
Duke University is a private university located in Durham, North Carolina. I am currently visiting Duke University.
Duke University is a private university located in Durham, North Carolina. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.
rvest::read_html(document)

rvest::html_nodes("a")

rvest::html_text()
Top 250 movies
SelectorGadget
Offered by: selectorgadget.com

Overview | Reviews | Support | Related

[Image of Chrome Web Store page for SelectorGadget]

bit.ly/SelectorGadget
<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>IMDb Rating</th>
<th>Your Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>The Shawshank Redemption (1994)</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>The Godfather (1972)</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>The Dark Knight (2008)</td>
<td>9.0</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>12 Angry Men (1957)</td>
<td>8.9</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>The Lord of the Rings: The Return of the King (2003)</td>
<td>8.9</td>
<td></td>
</tr>
</tbody>
</table>
## HTML Nodes

<table>
<thead>
<tr>
<th>Our Interest</th>
<th>Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td><code>.titleColumn a</code></td>
</tr>
<tr>
<td>Year</td>
<td><code>.secondaryInfo</code></td>
</tr>
<tr>
<td>Rating</td>
<td><code>.ratingColumn.imdbRating</code></td>
</tr>
<tr>
<td>Title: The Shawshank Redemption (1994)</td>
<td>Year: 9.2</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>The Shawshank Redemption</td>
<td>1994</td>
</tr>
<tr>
<td>9.2</td>
<td></td>
</tr>
</tbody>
</table>
Step 1 - \texttt{rvest::read\_html()}

Download the HTML file

\begin{verbatim}
> read_html("http://www.imdb.com/chart/top")
{xml_document}
[1] <head>
<meta http-equiv="Content-Type" content=" ... 
[2] <body id="styleguide-v2" class="fixed">\n\n ... 
>
\end{verbatim}
Step 2 - `rvest::html_nodes()`

Extract specific nodes

```r
> read_html("http://www.imdb.com/chart/top") %>%
+ html_nodes(".titleColumn a")
{xml_nodeset (250)}
```
Step 3 - `rvest::html_text()`

Extract text

```r
> read_html("http://www.imdb.com/chart/top") %>%
+  html_nodes(".titleColumn a") %>%
+  html_text()
[1] "The Shawshank Redemption"
[4] "The Dark Knight"
[5] "12 Angry Men"
```
Top 250 movies

1. The Shawshank Redemption (1994) - Rating: 9.2
2. The Godfather (1972) - Rating: 9.2
4. The Dark Knight (2008) - Rating: 9.0
5. 12 Angry Men (1957) - Rating: 8.9
<table>
<thead>
<tr>
<th>Our Interest</th>
<th>Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>.titleColumn a</td>
</tr>
<tr>
<td>Year</td>
<td>.secondaryInfo</td>
</tr>
<tr>
<td>Rating</td>
<td>.ratingColumn.imdbRating</td>
</tr>
<tr>
<td></td>
<td>Year1</td>
</tr>
<tr>
<td>---</td>
<td>-------</td>
</tr>
</tbody>
</table>
> read_html("http://www.imdb.com/chart/top") %>%
+ html_nodes(".secondaryInfo") %>%
+ html_text() %>%
+ str_remove("\((") %>% # remove (
+ str_remove("\)") %>% # remove )
+ as.numeric()

Web scraping

```r
> read_html("http://www.imdb.com/chart/top") %>%
+ html_nodes(".secondaryInfo") %>%
+ html_text() %>%
```

String manipulation

```r
+ str_remove("\('(')") %>%
+ str_remove("\(\)") %>%
+ as.numeric()
# remove ( )
```
# Read the entire page ----
page <- read_html("http://www.imdb.com/chart/top")

# Scrape titles ----
titles <- page %>%
  html_nodes(".titleColumn a") %>%
  html_text()

# Scrape years ----
years <- page %>%
  html_nodes(".secondaryInfo") %>%
  html_text() %>%
  str_remove("\\(\)") %>%
  str_remove("\\\") %>% # remove ( # remove )
as.numeric()

# Scrape ratings ----
ratings <- page %>%
  html_nodes("#main strong") %>%
  html_text() %>%
as.numeric()

# Save titles, years, and rating together in a tibble ----
imdb_top_250 <- tibble(
  title = titles,
  year = years,
  rating = ratings
)
<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>
Prior Knowledge (Before Learning Web Scraping)

Data types – character, factor, integer, double

Experience working with rectangular data

Familiarity with R packages, functions, arguments
Knowledge Building (While Learning Web Scraping)

HTML 5
CSS 3
Optional
Optional
Optional

rvest
stringr
Subsequent Activities (After Having Learned Web Scraping)

STATISTICS
# Predicting Movie Ratings

```r
> tidy(full_model)
# A tibble: 5 x 5
  term        estimate std.error statistic p.value
   <chr>      <dbl>     <dbl>       <dbl>    <dbl>
1 (Intercept)  4.74      0.178       26.7  1.60e-93
2 meta_score   0.0243    0.00130      18.7  1.51e-57
3 run_time     0.00492   0.00142      3.46  5.85e- 4
4 vote         0.00000115 0.00000231  4.99  8.86e- 7
5 gross        -0.000151  0.000201   -0.751 4.53e- 1
```
```r
> reduced_model <- step(full_model, direction = "backward")
> tidy(reduced_model)

# A tibble: 4 x 5
  term         estimate std.error statistic  p.value
  <chr>        <dbl>     <dbl>      <dbl>     <dbl>
1 (Intercept)  4.74      0.178     26.7     1.14e-93
2 meta_score   0.0246    0.00125   19.7     2.96e-62
3 run_time     0.00477   0.00141   3.39     7.52e- 4
4 vote        0.00000103 0.00000170 6.07     2.80e- 9
```
> glance(full_model)$AIC
[1] 496.9882

> glance(reduced_model)$AIC
[1] 495.5577
#model.script <- "
  + data {
  +   int N;
  +   real sigma;
  +   real rating[N];
  +   real meta_score[N];
  + }
  + transformed data {
  +   real m_meta_score;
  +   m_meta_score = mean(meta_score);
  + }
  + parameters {
  +   real alpha;
  +   real beta;
  + }
  + model {
  +   alpha ~ normal(0,1/.001);
  +   beta ~ normal(0,1/.001);
  +   for (i in 1:N)
  +     rating[i] ~ normal(alpha +
  +                         beta*(meta_score[i]- m_meta_score),
  +                         sigma);
  + }"
Posterior Distribution of Slope
When?

- After prior topics have been mastered
- After some data maturity has been gained
- Before the desired subsequent activity
- In introductory and/or advanced courses
Making use of the written code
StatSci @ Duke

The Department of Statistical Science is nationally ranked in the top 5 research departments and as a top 10 graduate program. Recognized as the world's leading center for Bayesian statistics and its interdisciplinary applications, the Department is a Duke campus hub for statistical and computational research. The Department administers and teaches a broad range of undergraduate statistics courses, including introductory courses and more advanced courses for the Statistical Science major, as well as many courses in statistics and related areas at the M.S. and Ph.D. level.

Learn More
Hrm.

The Wayback Machine has not archived that URL.
Website server can be down

404
Data entry

$00 / 2br - Won't Last! Most Requested 2 Bedroom with Loft in Columbus! 1 Left (Columbus)

Luxury lakeside community just 9 minutes from The Ohio State University featuring 1, 2 and 3 bedroom floor plans!

Link – Retrieved 2018-11-11
★ $500 / 1250ft² - 2 Bed, 2 Bath, 2 Parking Stall

Come see this Portland Tower 2BR, 2BA with 2 indoor heated parking stalls and huge walkout patio with southern exposure and gas grill hook up! Located just one block from Skyway, 3 blocks from Downtown East Commons, Light Rail Station and US Bank Stadium, this downtown condo is an amazing location!

Link – Retrieved 2018-11-11
Feature Films Released in 2018– Missing Nodes

PG | 90 min | Animation, Comedy, Family
⭐️ 6.1 ⭐️ Rate this | 50 Metascore
A grumpy Grinch plots to ruin Christmas for the village of Whoville.

Directors: Yarrow Cheney, Scott Mosier | Stars: Benedict Cumberbatch, Cameron Seely, Rashida Jones, Pharrell Williams
Votes: 1,484

PG-13 | Drama, Sport | Completed

Under the tutelage of Rocky Balboa, newly crowned light heavyweight champion Adonis Creed faces off against Viktor Drago, the son of Ivan Drago.

Director: Steven Caple Jr. | Stars: Tessa Thompson, Sylvester Stallone, Michael B. Jordan, Dolph Lundgren

Not Rated | 121 min | Action, Horror, Thriller
⭐️ 6.7 ⭐️ Rate this | 81 Metascore
The enchanted lives of a couple in a secluded forest are brutally shattered by a nightmarish hippie cult and their demon-biker henchmen, propelling a man into a spiraling, surreal rampage of vengeance.

Director: Panos Cosmatos | Stars: Nicolas Cage, Andrea Riseborough, Linus Roache, Ned Dennehy
Votes: 22,867 | Gross: $1.21M
PG | 90 min | Animation, Comedy, Family

🌟 6.1 Rate this 🟠 50 Metascore
A grumpy Grinch plots to ruin Christmas for the village of Whoville.

Directors: Yarrow Cheney, Scott Mosier | Stars: Benedict Cumberbatch, Cameron Seely, Rashida Jones, Pharrell Williams

Votes: 1,484

30. Creed II (2018)
PG-13 | Drama, Sport | Completed

Under the tutelage of Rocky Balboa, newly crowned light heavyweight champion Adonis Creed faces off against Viktor Drago, the son of Ivan Drago.

Director: Steven Caple Jr. | Stars: Tessa Thompson, Sylvester Stallone, Michael B. Jordan, Dolph Lundgren

31. Mandy (I) (2018)
Not Rated | 121 min | Action, Horror, Thriller

🌟 6.7 Rate this 🟠 81 Metascore
The enchanted lives of a couple in a secluded forest are brutally shattered by a nightmarish hippie cult and their demon-biker henchmen, propelling a man into a spiraling, surreal rampage of vengeance.

Director: Panos Cosmatos | Stars: Nicolas Cage, Andrea Riseborough, Linus Roache, Ned Dennehy

Votes: 22,867 | Gross: $1.21M

32. Apostle (2018)
TV-MA | 120 min | Drama, Fantasy, Horror
PG | 90 min | Animation, Comedy, Family

6.1 Rate this  
50 Metascore

A grumpy Grinch plots to ruin Christmas for the village of Whoville.

Directors: Yarrow Cheney, Scott Mosier  
Stars: Benedict Cumberbatch, Cameron Seely, Rashida Jones, Pharrell Williams

Votes: 1,484

---

30. **Creed II** (2018)
PG-13 | Drama, Sport  
Completed

Under the tutelage of Rocky Balboa, newly crowned light heavyweight champion Adonis Creed faces off against Viktor Drago, the son of Ivan Drago.

Director: Steven Caple Jr  
Stars: Tessa Thompson, Sylvester Stallone, Michael B. Jordan, Dolph Lundgren

---

31. **Mandy** (I) (2018)
Not Rated | 121 min | Action, Horror, Thriller

6.7 Rate this  
81 Metascore

The enchanted lives of a couple in a secluded forest are brutally shattered by a nightmarish hippie cult and their demon-biker henchmen, propelling a man into a spiraling, surreal rampage of vengeance.

Director: Panos Cosmatos  
Stars: Nicolas Cage, Andrea Riseborough, Linus Roache, Ned Dennehy

Votes: 22,867  
Gross: $1.21M

---

32. **Apostle** (2018)
TV-MA | 120 min | Drama, Fantasy, Horror

---

89
<table>
<thead>
<tr>
<th>Challenge</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reproducibility</td>
<td>Remains unresolved: nature of the work (Partial solutions: Internet Archive + Local downloads)</td>
</tr>
</tbody>
</table>
WHAT?  WHY?  HOW?  WHEN?
Variety of sources for data

IMDb

craigslist

g

National Weather Service

amazon

United States of America
Not only for the learners
Ethics
I'm not a robot
> library(robotstxt)
> paths_allowed("http://www.imdb.com")

[1] TRUE
> paths_allowed("http://www.facebook.com")
> www.facebook.com

[1] FALSE
TERMS OF USE
“I feel like I can rule the world with web scraping”
meaningful COMPUTING + endless STATISTICS
Dr. Mine Çetinkaya-Rundel
ecOTS
Manuscript
QUESTIONS?

✉️ dogucum@denison.edu

🐦 @MineDogucu

🔗 bit.ly/scrape_duke
References


