WebScraping

Rotten Tomatoes WebScraper

Project purpose: Web Scraping and Social Media Scraping Project prepared at University of Warsaw, Data Science and Business Analytics Master's Degree Program in Faculty of Economic Sciences.

Introduction

The internet is an absolutely huge source of data that we can collect in many ways and further analyze, but in reality, it turns out often that Web Scraping is the only way to access data. There is a lot of information that is not available in convenient CSV format to export or easy-to-connect APIs. Stock prices, product details, sports stats, company contacts, movie reviews are the most popular data to scrap and thus we decided to build automated tools to scrap details of "Top 100 movies of 2020" at Rotten Tomatoes - a website presenting information, reviews and news from the world of film. Our WebScrapers were built with the use of `BeautifulSoup` , `Scrapy` and `Selenium` frameworks. Using the above-mentioned mechanisms, we obtained the same movie details, like genre, rating, original language, Director, Producer etc. that we later shortly analyzed.

GitHub repository structure

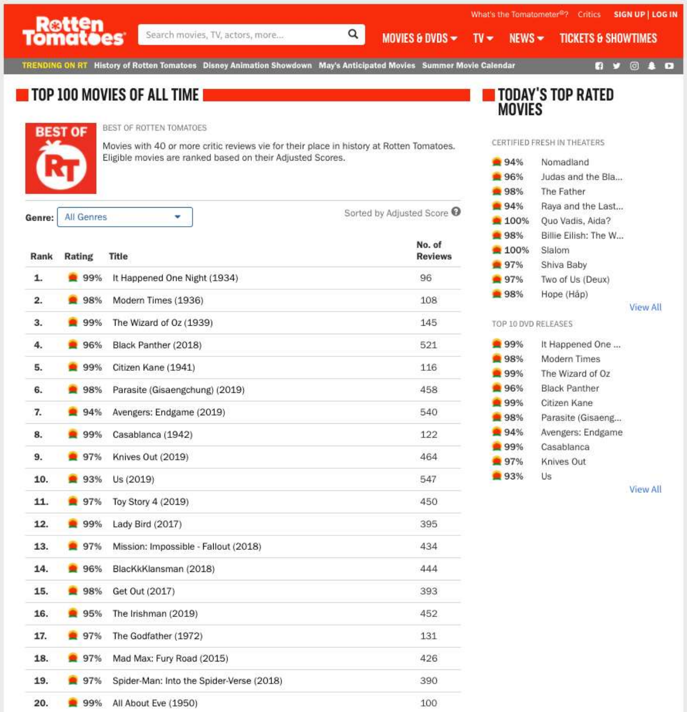
Our project repository includes all the required files arranged in a logical structure including the the imposed requirements according to which the project was prepared saved as `project_rules.pdf` , described source codes for each of the individual WebScraper, the obtained output and a report detailing the mechanisms, their comparison and analysis of the results.

WebScraper mechanisms

Beautiful Soup

The detailed described code for the first WebScraper can be seen in the file `soup.py` . The tool uses libraries such as: `BeautifulSoup` , `Requests` , `Pandas` , `Time` and regular expression package `Re` thus it is important to download them before running the code using `pip install` command. Libraries will be described in detail later, taking into account the functions in the tool that were obtained with their usage.

First, our BeautifulSoup WebScraper downloads `Rotten Tomatoes` `https://www.rottentomatoes.com/top/bestofrt/` with the list of TOP 100 movies of all time page using the Python `Requests` library. By `GET` request to a web server WebScraper downloads the HTML contents and then using `BeautifulSoup` library parses the document and extracts the text.



TOP 100 MOVIES OF ALL TIME

BEST OF ROTTEN TOMATOES

Movies with 40 or more critic reviews vie for their place in history at Rotten Tomatoes. Eligible movies are ranked based on their Adjusted Scores.

Genre: All Genres

Sorted by Adjusted Score

Rank	Rating	Title	No. of Reviews
1.	99%	It Happened One Night (1934)	96
2.	98%	Modern Times (1936)	108
3.	99%	The Wizard of Oz (1939)	145
4.	96%	Black Panther (2018)	521
5.	99%	Citizen Kane (1941)	116
6.	98%	Parasite (Gisaengchung) (2019)	458
7.	94%	Avengers: Endgame (2019)	540
8.	99%	Casablanca (1942)	122
9.	97%	Knives Out (2019)	464
10.	93%	Us (2019)	547
11.	97%	Toy Story 4 (2019)	450
12.	99%	Lady Bird (2017)	395
13.	97%	Mission: Impossible - Fallout (2018)	434
14.	96%	BlacKkKlansman (2018)	444
15.	98%	Get Out (2017)	393
16.	95%	The Irishman (2019)	452
17.	97%	The Godfather (1972)	131
18.	97%	Mad Max: Fury Road (2015)	426
19.	97%	Spider-Man: Into the Spider-Verse (2018)	390
20.	99%	All About Eve (1950)	100

TODAY'S TOP RATED MOVIES

CERTIFIED FRESH IN THEATERS

94%	Nomadland
96%	Judas and the Bla...
98%	The Father
94%	Raya and the Last...
100%	Quo Vadis, Aida?
98%	Billie Eilish: The W...
100%	Slalom
97%	Shiva Baby
97%	Two of Us (Deux)
98%	Hope (Håp)

View All

TOP 10 DVD RELEASES

99%	It Happened One ...
98%	Modern Times
99%	The Wizard of Oz
96%	Black Panther
99%	Citizen Kane
98%	Parasite (Gisaeng...
94%	Avengers: Endgame
99%	Casablanca
97%	Knives Out
93%	Us

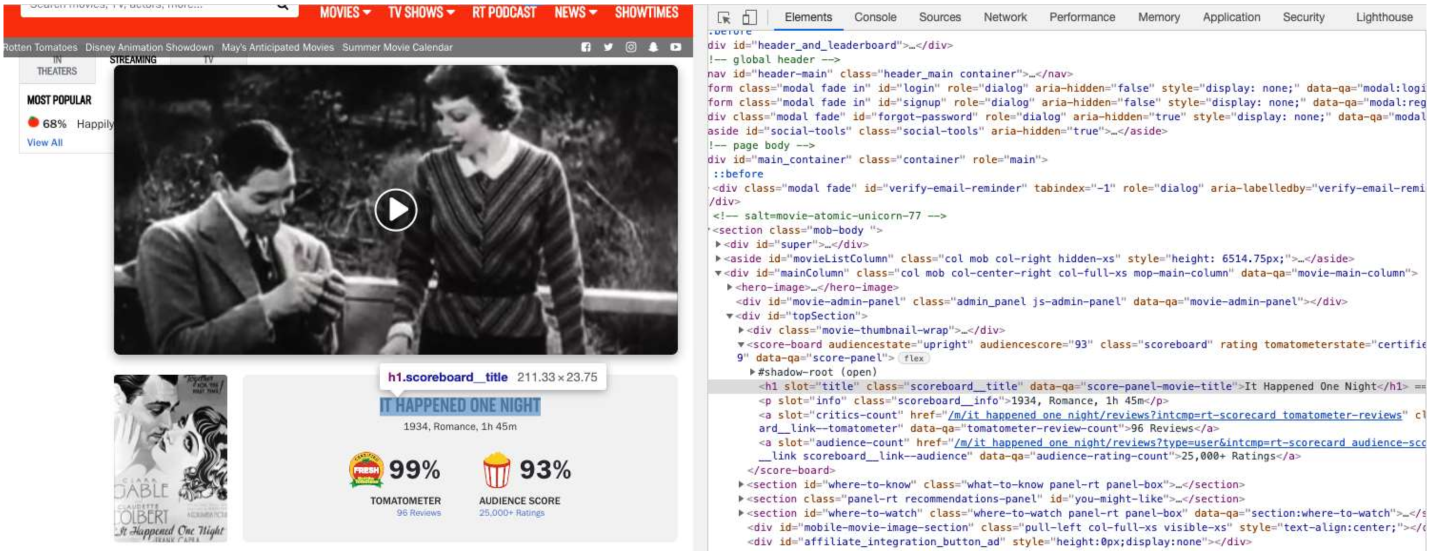
View All

```
url = "https://www.rottentomatoes.com/top/bestofrt/"
data = requests.get(url).text
soup = BeautifulSoup(data, 'html.parser')
```

Using `find_all` method WebScraper navigates a page , finds all the instances of a `td` tag and extracts all the `href` for each movie appending a previously created list.

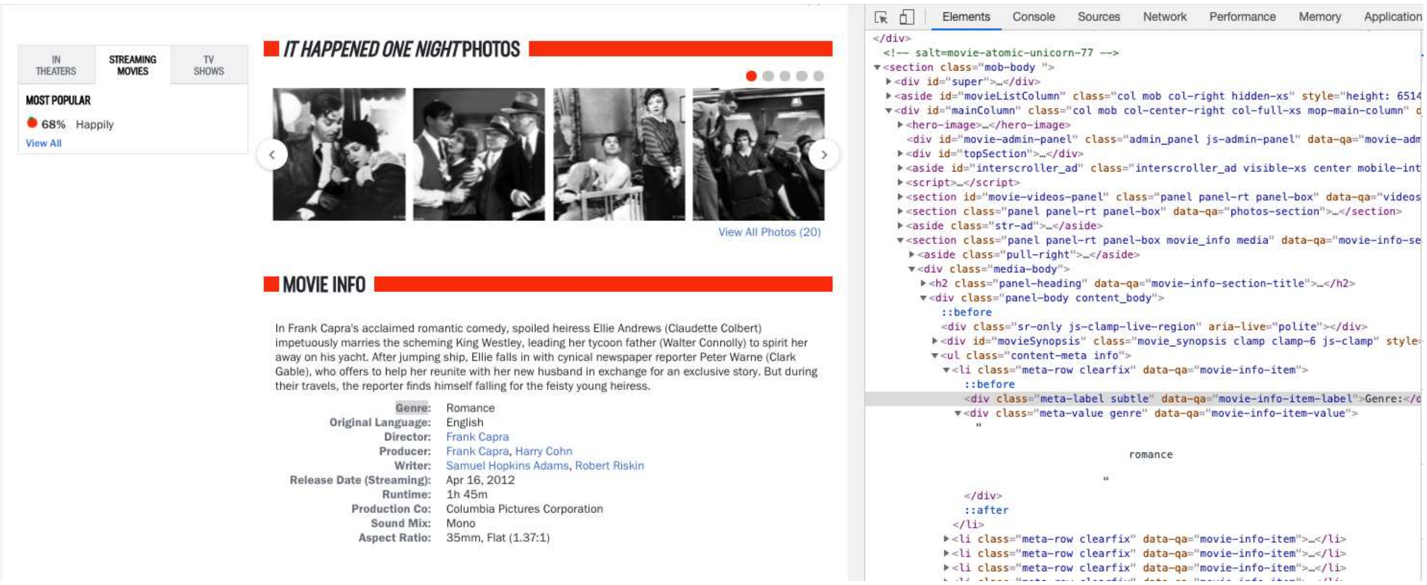
```
temp = [td.find('a', {'class': 'unstyled articleLink'}) for td in soup.find_all('td')]
for i in temp:
    if i is not None:
        href.append(i.get('href'))
```

Each link to a particular movie page has a similar structure, `url = "https://www.rottentomatoes.com" + href"` and thus `href` is used finding the links to the movie's page. For each link defined in this way, first an html copy of the page is requested using the link, then parsed and saved into previously created dictionary with movie names scrapped from `Score Board` with `h1` tag and attribute `data-qa = "score-panel-movie-title"` .



```
:before
div id="header_and_leaderboard">…</div>
!-- global header -->
nav id="header-main" class="header_main container">…</nav>
form class="modal fade in" id="login" role="dialog" aria-hidden="false" style="display: none;" data-qa="modal:logi
form class="modal fade in" id="signup" role="dialog" aria-hidden="false" style="display: none;" data-qa="modal:reg
div class="modal fade" id="forgot-password" role="dialog" aria-hidden="true" style="display: none;" data-qa="modal
aside id="social-tools" class="social-tools" aria-hidden="true">…</aside>
!-- page body -->
div id="main_container" class="container" role="main">
 ::before
 <div class="modal fade" id="verify-email-reminder" tabindex="-1" role="dialog" aria-labelledby="verify-email-remi
/div>
 <!-- salt=movie-atomic-unicorn-77 -->
 <section class="mob-body ">
  ▶<div id="super">…</div>
  ▶<aside id="movieListColumn" class="col mob col-right hidden-xs" style="height: 6514.75px;">…</aside>
  ▼<div id="mainColumn" class="col mob col-center-right col-full-xs mop-main-column" data-qa="movie-main-column">
    ▶<hero-image>…</hero-image>
     <div id="movie-admin-panel" class="admin_panel js-admin-panel" data-qa="movie-admin-panel"></div>
    ▼<div id="topSection">
      ▶<div class="movie-thumbnail-wrap">…</div>
      ▼<score-board audiencestate="upright" audiencescore="93" class="scoreboard" rating tomatometerstate="certifie
       9" data-qa="score-panel"> flex
        ▶#shadow-root (open)
         <h1 slot="title" class="scoreboard__title" data-qa="score-panel-movie-title">It Happened One Night</h1> ==
         <p slot="info" class="scoreboard__info">1934, Romance, 1h 45m</p>
         <a slot="critics-count" href="/m/it_happened_one_night/reviews?intcmp=rt-scorecard_tomatometer-reviews" cl
         ard__link--tomatometer" data-qa="tomatometer-review-count">96 Reviews</a>
         <a slot="audience-count" href="/m/it_happened_one_night/reviews?type=user&intcmp=rt-scorecard_audience-sco
         __link scoreboard__link--audience" data-qa="audience-rating-count">25,000+ Ratings</a>
       </score-board>
      ▶<section id="where-to-know" class="what-to-know panel-rt panel-box">…</section>
      ▶<section class="panel-rt recommendations-panel" id="you-might-like">…</section>
      ▶<section id="where-to-watch" class="where-to-watch panel-rt panel-box" data-qa="section:where-to-watch">…</s
       <div id="mobile-movie-image-section" class="pull-left col-full-xs visible-xs" style="text-align:center;"></c
       <div id="affiliate_integration_button_ad" style="height:0px;display:none"></div>
```

Using loop WebScraper is visiting each link in the Top 100 movies and extracting details such as `Genre` , `Original Language` , `Director` , `Producer` etc. which will later be used for elementary analysis. WebScraper navigates `Movie Info` section for each movie by its `ul` tag with class `content-meta info` and then iterates over each label and its value having `data-qa` attribute to finally saves extracted values and append dictionary to the `Pandas` dataframe. In the end dataframe is saved as `CSV` file and the time it tooks to scrape all data is displayed in the terminal window by using `Time` library.



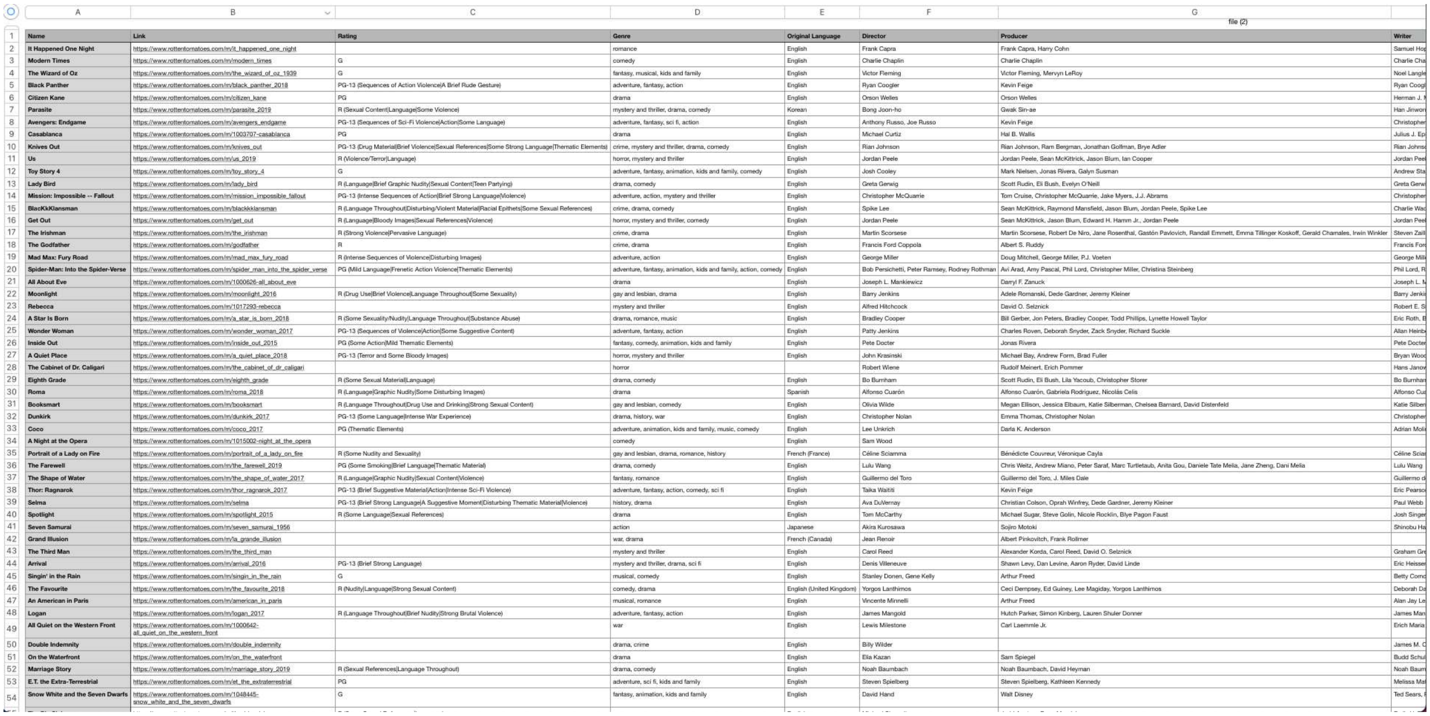
MOVIE INFO

In Frank Capra's acclaimed romantic comedy, spoiled heiress Ellie Andrews (Claudette Colbert) impetuously marries the scheming King Westley, leading her tycoon father (Walter Connolly) to spirit her away on his yacht. After jumping ship, Ellie falls in with cynical newspaper reporter Peter Warne (Clark Gable), who offers to help her reunite with her new husband in exchange for an exclusive story. But during their travels, the reporter finds himself falling for the feisty young heiress.

Genre:	Romance
Original Language:	English
Director:	Frank Capra
Producer:	Frank Capra, Harry Cohn
Writer:	Samuel Hopkins Adams, Robert Riskin
Release Date (Streaming):	Apr 16, 2012
Runtime:	1h 45m
Production Co:	Columbia Pictures Corporation
Sound Mix:	Mono
Aspect Ratio:	35mm, Flat (1.37:1)

```
</div>
 <!-- salt=movie-atomic-unicorn-77 -->
▼<section class="mob-body ">
  ▶<div id="super">…</div>
  ▶<aside id="movieListColumn" class="col mob col-right hidden-xs" style="height: 6514
  ▼<div id="mainColumn" class="col mob col-center-right col-full-xs mop-main-column" c
    ▶<hero-image>…</hero-image>
     <div id="movie-admin-panel" class="admin_panel js-admin-panel" data-qa="movie-adm
    ▶<div id="topSection">…</div>
    ▶<aside id="interscroller_ad" class="interscroller_ad visible-xs center mobile-int
    ▶<script>…</script>
    ▶<section id="movie-videos-panel" class="panel panel-rt panel-box" data-qa="videos
    ▶<section class="panel panel-rt panel-box" data-qa="photos-section">…</section>
    ▶<aside class="str-ad">…</aside>
    ▼<section class="panel panel-rt panel-box movie_info media" data-qa="movie-info-se
      ▶<aside class="pull-right">…</aside>
      ▼<div class="media-body">
        ▶<h2 class="panel-heading" data-qa="movie-info-section-title">…</h2>
        ▼<div class="panel-body content_body">
           ::before
           <div class="sr-only js-clamp-live-region" aria-live="polite"></div>
          ▶<div id="movieSynopsis" class="movie_synopsis clamp clamp-6 js-clamp" style=
          ▼<ul class="content-meta info">
            ▼<li class="meta-row clearfix" data-qa="movie-info-item">
               ::before
               <div class="meta-label subtle" data-qa="movie-info-item-label">Genre:</d
              ▼<div class="meta-value genre" data-qa="movie-info-item-value">
                 "

                                 romance

                             "
               </div>
               ::after
             </li>
            ▶<li class="meta-row clearfix" data-qa="movie-info-item">…</li>
            ▶<li class="meta-row clearfix" data-qa="movie-info-item">…</li>
            ▶<li class="meta-row clearfix" data-qa="movie-info-item">…</li>
```

Obtained output in a `.csv` format.



file (2)

	A	B	C	D	E	F	G	
1	Name	Link	Rating	Genre	Original Language	Director	Producer	Writer
2	It Happened One Night	https://www.rottentomatoes.com/m/it_happened_one_night		romance	English	Frank Capra	Frank Capra, Harry Cohn	Samuel Hop
3	Modern Times	https://www.rottentomatoes.com/m/modern_times	G	comedy	English	Charlie Chaplin	Charlie Chaplin	Charlie Cha
4	The Wizard of Oz	https://www.rottentomatoes.com/m/the_wizard_of_oz_1939	G	fantasy, musical, kids and family	English	Victor Fleming	Victor Fleming, Mervyn LeRoy	Noel Langle
5	Black Panther	https://www.rottentomatoes.com/m/black_panther_2018	PG-13 (Sequences of Action Violence|A Brief Rude Gesture)	adventure, fantasy, action	English	Ryan Coogler	Kevin Feige	Ryan Coogl
6	Citizen Kane	https://www.rottentomatoes.com/m/citizen_kane	PG	drama	English	Orson Welles	Orson Welles	Herman J. M
7	Parasite	https://www.rottentomatoes.com/m/parasite_2019	R (Sexual Content|Language|Some Violence)	mystery and thriller, drama, comedy	Korean	Bong Joon-ho	Gwak Sin-ae	Han Jinwon
8	Avengers: Endgame	https://www.rottentomatoes.com/m/avengers_endgame	PG-13 (Sequences of Sci-Fi Violence|Action|Some Language)	adventure, fantasy, sci fi, action	English	Anthony Russo, Joe Russo	Kevin Feige	Christopher
9	Casablanca	https://www.rottentomatoes.com/m/1003707-casablanca	PG	drama	English	Michael Curtiz	Hal B. Wallis	Julius J. Ep
10	Knives Out	https://www.rottentomatoes.com/m/knives_out	PG-13 (Drug Material|Brief Violence|Sexual References|Some Strong Language|Thematic Elements)	crime, mystery and thriller, drama, comedy	English	Rian Johnson	Rian Johnson, Ram Bergman, Jonathan Golfman, Brye Adler	Rian Johnso
11	Us	https://www.rottentomatoes.com/m/us_2019	R (Violence/Terror|Language)	horror, mystery and thriller	English	Jordan Peele	Jordan Peele, Sean McKittrick, Jason Blum, Ian Cooper	Jordan Peel
12	Toy Story 4	https://www.rottentomatoes.com/m/toy_story_4	G	adventure, fantasy, animation, kids and family, comedy	English	Josh Cooley	Mark Nielsen, Jonas Rivera, Galyn Susman	Andrew Sta
13	Lady Bird	https://www.rottentomatoes.com/m/lady_bird	R (Language|Brief Graphic Nudity|Sexual Content|Teen Partying)	drama, comedy	English	Greta Gerwig	Scott Rudin, Eli Bush, Evelyn O'Neill	Greta Gerw
14	Mission: Impossible -- Fallout	https://www.rottentomatoes.com/m/mission_impossible_fallout	PG-13 (Intense Sequences of Action|Brief Strong Language|Violence)	adventure, action, mystery and thriller	English	Christopher McQuarrie	Tom Cruise, Christopher McQuarrie, Jake Myers, J.J. Abrams	Christopher
15	BlacKkKlansman	https://www.rottentomatoes.com/m/blackkklansman	R (Language Throughout|Disturbing/Violent Material|Racial Epithets|Some Sexual References)	crime, drama, comedy	English	Spike Lee	Sean McKittrick, Raymond Mansfield, Jason Blum, Jordan Peele, Spike Lee	Charlie Wac
16	Get Out	https://www.rottentomatoes.com/m/get_out	R (Language|Bloody Images|Sexual References|Violence)	horror, mystery and thriller, comedy	English	Jordan Peele	Sean McKittrick, Jason Blum, Edward H. Hamm Jr., Jordan Peele	Jordan Peel
17	The Irishman	https://www.rottentomatoes.com/m/the_irishman	R (Strong Violence|Pervasive Language)	crime, drama	English	Martin Scorsese	Martin Scorsese, Robert De Niro, Jane Rosenthal, Gastón Pavlovich, Randall Emmett, Emma Tillinger Koskoff, Gerald Chamales, Irwin Winkler	Steven Zaill
18	The Godfather	https://www.rottentomatoes.com/m/godfather	R	crime, drama	English	Francis Ford Coppola	Albert S. Ruddy	Francis Ford
19	Mad Max: Fury Road	https://www.rottentomatoes.com/m/mad_max_fury_road	R (Intense Sequences of Violence|Disturbing Images)	adventure, action	English	George Miller	Doug Mitchell, George Miller, P.J. Voeten	George Mill
20	Spider-Man: Into the Spider-Verse	https://www.rottentomatoes.com/m/spider_man_into_the_spider_verse	PG (Mild Language|Frenetic Action Violence|Thematic Elements)	adventure, fantasy, animation, kids and family, action, comedy	English	Bob Persichetti, Peter Ramsey, Rodney Rothman	Avi Arad, Amy Pascal, Phil Lord, Christopher Miller, Christina Steinberg	Phil Lord, R
21	All About Eve	https://www.rottentomatoes.com/m/1000626-all_about_eve		drama	English	Joseph L. Mankiewicz	Darryl F. Zanuck	Joseph L. M
22	Moonlight	https://www.rottentomatoes.com/m/moonlight_2016	R (Drug Use|Brief Violence|Language Throughout|Some Sexuality)	gay and lesbian, drama	English	Barry Jenkins	Adele Romanski, Dede Gardner, Jeremy Kleiner	Barry Jenki
23	Rebecca	https://www.rottentomatoes.com/m/1017293-rebecca		mystery and thriller	English	Alfred Hitchcock	David O. Selznick	Robert E. S
24	A Star Is Born	https://www.rottentomatoes.com/m/a_star_is_born_2018	R (Some Sexuality/Nudity|Language Throughout|Substance Abuse)	drama, romance, music	English	Bradley Cooper	Bill Gerber, Jon Peters, Bradley Cooper, Todd Phillips, Lynette Howell Taylor	Eric Roth, E
25	Wonder Woman	https://www.rottentomatoes.com/m/wonder_woman_2017	PG-13 (Sequences of Violence|Action|Some Suggestive Content)	adventure, fantasy, action	English	Patty Jenkins	Charles Roven, Deborah Snyder, Zack Snyder, Richard Suckle	Allan Heinb
26	Inside Out	https://www.rottentomatoes.com/m/inside_out_2015	PG (Some Action|Mild Thematic Elements)	fantasy, comedy, animation, kids and family	English	Pete Docter	Jonas Rivera	Pete Docter
27	A Quiet Place	https://www.rottentomatoes.com/m/a_quiet_place_2018	PG-13 (Terror and Some Bloody Images)	horror, mystery and thriller	English	John Krasinski	Michael Bay, Andrew Form, Brad Fuller	Bryan Wood
28	The Cabinet of Dr. Caligari	https://www.rottentomatoes.com/m/the_cabinet_of_dr_caligari		horror		Robert Wiene	Rudolf Meinert, Erich Pommer	Hans Janow
29	Eighth Grade	https://www.rottentomatoes.com/m/eighth_grade	R (Some Sexual Material|Language)	drama, comedy	English	Bo Burnham	Scott Rudin, Eli Bush, Lila Yacoub, Christopher Storer	Bo Burnham
30	Roma	https://www.rottentomatoes.com/m/roma_2018	R (Language|Graphic Nudity|Some Disturbing Images)	drama	Spanish	Alfonso Cuarón	Alfonso Cuarón, Gabriela Rodriguez, Nicolás Celis	Alfonso Cu
31	Booksmart	https://www.rottentomatoes.com/m/booksmart	R (Language Throughout|Drug Use and Drinking|Strong Sexual Content)	gay and lesbian, comedy	English	Olivia Wilde	Megan Ellison, Jessica Elbaum, Katie Silberman, Chelsea Barnard, David Distenfeld	Katie Silber
32	Dunkirk	https://www.rottentomatoes.com/m/dunkirk_2017	PG-13 (Some Language|Intense War Experience)	drama, history, war	English	Christopher Nolan	Emma Thomas, Christopher Nolan	Christopher
33	Coco	https://www.rottentomatoes.com/m/coco_2017	PG (Thematic Elements)	adventure, animation, kids and family, music, comedy	English	Lee Unkrich	Darla K. Anderson	Adrian Moli
34	A Night at the Opera	https://www.rottentomatoes.com/m/1015002-night_at_the_opera		comedy	English	Sam Wood		
35	Portrait of a Lady on Fire	https://www.rottentomatoes.com/m/portrait_of_a_lady_on_fire	R (Some Nudity and Sexuality)	gay and lesbian, drama, romance, history	French (France)	Céline Sciamma	Bénédicte Couvreur, Véronique Cayla	Céline Scia
36	The Farewell	https://www.rottentomatoes.com/m/the_farewell_2019	PG (Some Smoking|Brief Language|Thematic Material)	drama, comedy	English	Lulu Wang	Chris Weitz, Andrew Miano, Peter Saraf, Marc Turtletaub, Anita Gou, Daniele Tate Melia, Jane Zheng, Dani Melia	Lulu Wang
37	The Shape of Water	https://www.rottentomatoes.com/m/the_shape_of_water_2017	R (Language|Graphic Nudity|Sexual Content|Violence)	fantasy, romance	English	Guillermo del Toro	Guillermo del Toro, J. Miles Dale	Guillermo d
38	Thor: Ragnarok	https://www.rottentomatoes.com/m/thor_ragnarok_2017	PG-13 (Brief Suggestive Material|Action|Intense Sci-Fi Violence)	adventure, fantasy, action, comedy, sci fi	English	Taika Waititi	Kevin Feige	Eric Pearso
39	Selma	https://www.rottentomatoes.com/m/selma	PG-13 (Brief Strong Language|A Suggestive Moment|Disturbing Thematic Material|Violence)	history, drama	English	Ava DuVernay	Christian Colson, Oprah Winfrey, Dede Gardner, Jeremy Kleiner	Paul Webb
40	Spotlight	https://www.rottentomatoes.com/m/spotlight_2015	R (Some Language|Sexual References)	drama	English	Tom McCarthy	Michael Sugar, Steve Golin, Nicole Rocklin, Blye Pagon Faust	Josh Singer
41	Seven Samurai	https://www.rottentomatoes.com/m/seven_samurai_1956		action	Japanese	Akira Kurosawa	Sojiro Motoki	Shinobu Ha
42	Grand Illusion	https://www.rottentomatoes.com/m/la_grande_illusion		war, drama	French (Canada)	Jean Renoir	Albert Pinkovitch, Frank Rollmer	
43	The Third Man	https://www.rottentomatoes.com/m/the_third_man		mystery and thriller	English	Carol Reed	Alexander Korda, Carol Reed, David O. Selznick	Graham Gre
44	Arrival	https://www.rottentomatoes.com/m/arrival_2016	PG-13 (Brief Strong Language)	mystery and thriller, drama, sci fi	English	Denis Villeneuve	Shawn Levy, Dan Levine, Aaron Ryder, David Linde	Eric Heisser
45	Singin' in the Rain	https://www.rottentomatoes.com/m/singin_in_the_rain	G	musical, comedy	English	Stanley Donen, Gene Kelly	Arthur Freed	Betty Comd
46	The Favourite	https://www.rottentomatoes.com/m/the_favourite_2018	R (Nudity|Language|Strong Sexual Content)	comedy, drama	English (United Kingdom)	Yorgos Lanthimos	Ceci Dempsey, Ed Guiney, Lee Magiday, Yorgos Lanthimos	Deborah Da
47	An American in Paris	https://www.rottentomatoes.com/m/american_in_paris		musical, romance	English	Vincente Minnelli	Arthur Freed	Alan Jay Le
48	Logan	https://www.rottentomatoes.com/m/logan_2017	R (Language Throughout|Brief Nudity|Strong Brutal Violence)	adventure, fantasy, action	English	James Mangold	Hutch Parker, Simon Kinberg, Lauren Shuler Donner	James Man
49	All Quiet on the Western Front	https://www.rottentomatoes.com/m/1000642-all_quiet_on_the_western_front		war	English	Lewis Milestone	Carl Laemmle Jr.	Erich Maria
50	Double Indemnity	https://www.rottentomatoes.com/m/double_indemnity		drama, crime	English	Billy Wilder		James M. C
51	On the Waterfront	https://www.rottentomatoes.com/m/on_the_waterfront		drama	English	Elia Kazan	Sam Spiegel	Budd Schul
52	Marriage Story	https://www.rottentomatoes.com/m/marriage_story_2019	R (Sexual References|Language Throughout)	drama, comedy	English	Noah Baumbach	Noah Baumbach, David Heyman	Noah Baum
53	E.T. the Extra-Terrestrial	https://www.rottentomatoes.com/m/et_the_extraterrestrial	PG	adventure, sci fi, kids and family	English	Steven Spielberg	Steven Spielberg, Kathleen Kennedy	Melissa Mat
54	Snow White and the Seven Dwarfs	https://www.rottentomatoes.com/m/1048445-snow_white_and_the_seven_dwarfs	G	fantasy, animation, kids and family	English	David Hand	Walt Disney	Ted Sears, F

Scrapy

The Spider classes are the most important part of the Scrapy WebScraper that define the scraper-crawler mechanism as it passes through the web page. Similar to the other WebScrapers described source code was also saved in the separate files `spider_1.py` and `spider_2.py` To run code it is necessary to import `scrapy` and `re` libraries .

First, the Spider class allows us to define the start URLs from which, by default, our crawler will start browsing websites. They are defined as a list in the `start_urls` variable. In our case it will be `https://www.rottentomatoes.com/top/bestofrt/` . Then `parse` function defines how data is extracted from the page - links of each movie are navigated by defined `XPath` and then extracted to a `links.csv` file used by `spider_2` .

```
def parse(self, response):
       xpath = "//td[3]/a[@class = 'unstyled articleLink']/@href"
       selection = response.xpath(xpath)
       for s in selection:
           l = Link()
           l['link'] = 'https://www.rottentomatoes.com/' + s.get()
           yield l
```

In the case of Scrapy, the downloaded data is called `Items` and by declaring `Movies class` we determine what data from the page will be scrape:

```
# Declaring Movies class
class Movies(scrapy.Item):
    Name = scrapy.Field()
    Link = scrapy.Field()
    Rating = scrapy.Field()
    Genre = scrapy.Field()
    Original_Language = scrapy.Field()
    Director = scrapy.Field()
    Producer = scrapy.Field()
    Writer = scrapy.Field()
    Release_Date_Streaming = scrapy.Field()
    Runtime = scrapy.Field()
    Production_Co = scrapy.Field()
    Aspect_Ratio = scrapy.Field()
    Sound_Mix = scrapy.Field()
    Release_Date_Theaters = scrapy.Field()
    Box_Office = scrapy.Field()
```

Then, after defining `LinksSpider` , it will be opening links for each individual movie, previously stored in the `links.csv` file and scrape detailed data navigated by `XPaths` .

Selenium

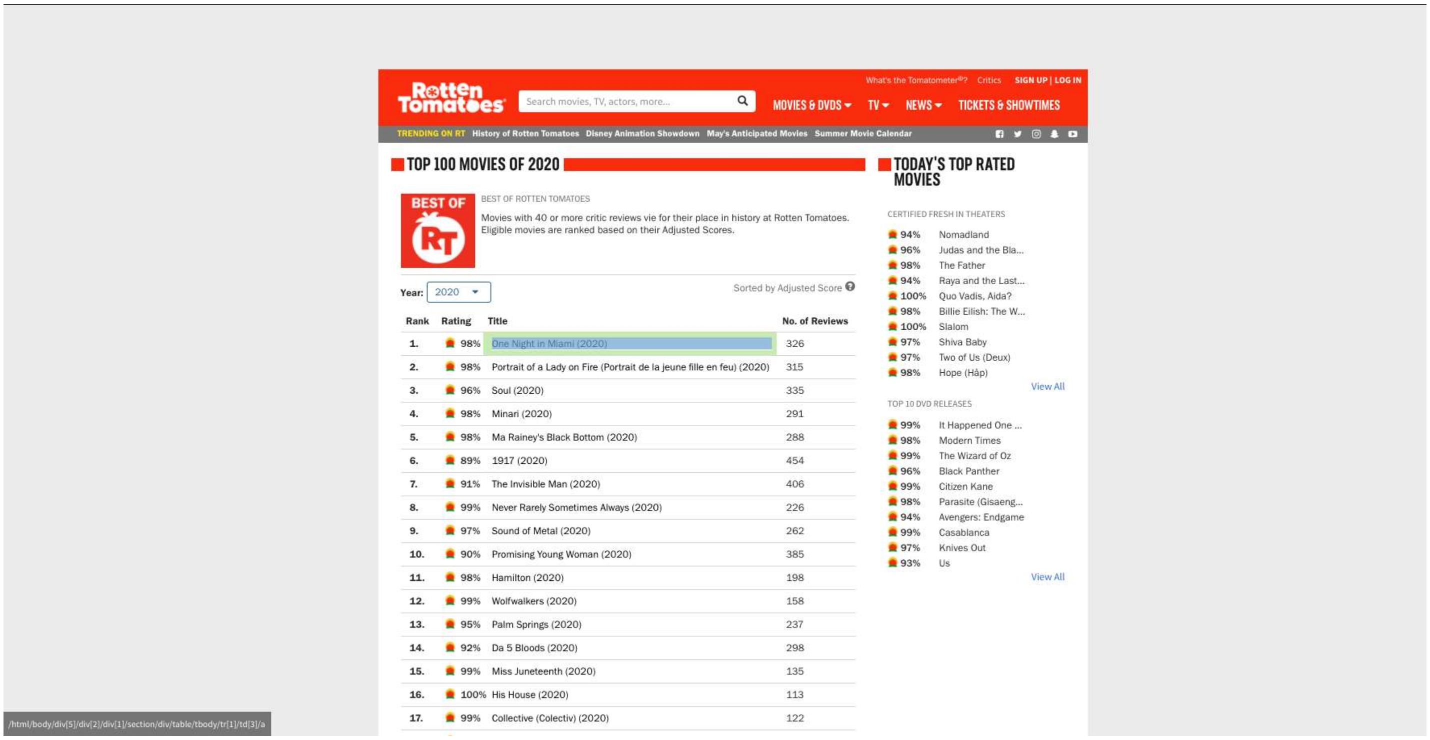
Similar to the previous WebScraper, for Selenium, described source code was also saved in the file `bot.py` . The design of this tool uses `Selenium` framework from which `Webdriver` for Chrome is imported, as well as `ActionChains` to automate low-level interactions such as mouse movements, `NoSuchElementException` to handle with data which may be not present on a page and dict subclass `defaultdict` from `Colections` that calls a factory function to supply missing values . Additionally, libraries such as `Time` , `Pandas` and `Re` are used which are necessary to run the code.

The tool starts with creating a browser control object `Webdriver` and then enters the page using the `GET` method.

```
# Chromedriver
gecko_path = '...'

# Rotten Tomatoes URL
url = 'https://www.rottentomatoes.com/'

# Setting driver
options = webdriver.chrome.options.Options()
options.headless = False
driver = webdriver.Chrome(options=options, executable_path=gecko_path)

# Actual Program
driver.get(url)
print(driver.page_source)

time.sleep(2)
```

Then using the webdriver and clicking on buttons defined by `XPaths` tool will locate `WebElements` on the page: `Movies >> Top Movies >> View All` respectively and get all the movie links by accessing their `href` attribute and store in the predefined list.

```
r = []
for i in range(100):
    path = "/html/body/div[5]/div[2]/div[1]/section/div/table/tbody/tr[" + str(i + 1) + "]/td
[3]/a"
    d = driver.find_element_by_xpath(path).get_attribute("href")
    r.append(d)
```


Rotten Tomatoes
Search movies, TV, actors, more...
What's the Tomatometer®? Critics SIGN UP | LOG IN
MOVIES & DVDS
TV
NEWS
TICKETS & SHOWTIMES
TRENDING ON RT History of Rotten Tomatoes Disney Animation Showdown May's Anticipated Movies Summer Movie Calendar
BEST OF RT
View All
View All

TOP 100 MOVIES OF 2020

BEST OF ROTTEN TOMATOES

Movies with 40 or more critic reviews vie for their place in history at Rotten Tomatoes. Eligible movies are ranked based on their Adjusted Scores.

Year: 2020

Sorted by Adjusted Score

Rank	Rating	Title	No. of Reviews
1.	98%	One Night in Miami (2020)	326
2.	98%	Portrait of a Lady on Fire (Portrait de la jeune fille en feu) (2020)	315
3.	96%	Soul (2020)	335
4.	98%	Minari (2020)	291
5.	98%	Ma Rainey's Black Bottom (2020)	288
6.	89%	1917 (2020)	454
7.	91%	The Invisible Man (2020)	406
8.	99%	Never Rarely Sometimes Always (2020)	226
9.	97%	Sound of Metal (2020)	262
10.	90%	Promising Young Woman (2020)	385
11.	98%	Hamilton (2020)	198
12.	99%	Wolfwalkers (2020)	158
13.	95%	Palm Springs (2020)	237
14.	92%	Da 5 Bloods (2020)	298
15.	99%	Miss Juneteenth (2020)	135
16.	100%	His House (2020)	113
17.	99%	Collective (Colectiv) (2020)	122

TODAY'S TOP RATED MOVIES

CERTIFIED FRESH IN THEATERS

94%	Nomadland
96%	Judas and the Bla...
98%	The Father
94%	Raya and the Last...
100%	Quo Vadis, Aida?
98%	Billie Eilish: The W...
100%	Slalom
97%	Shiva Baby
97%	Two of Us (Deux)
98%	Hope (Håp)

TOP 10 DVD RELEASES

99%	It Happened One ...
98%	Modern Times
99%	The Wizard of Oz
96%	Black Panther
99%	Citizen Kane
98%	Parasite (Gisaeng...
94%	Avengers: Endgame
99%	Casablanca
97%	Knives Out
93%	Us

/html/body/div[5]/div[2]/div[1]/section/div/table/tbody/tr[1]/td[3]/a

Next, our WebScraper using loop is visiting each link in the Top 100 movies, extracts movie names navigating them by `XPaths` and store together with links in the dictionary which will be appended to the dataframe with previously extracted movie details. Finally dataframe is saved as `.csv` file and the time it tooks to scrape all data is displayed in the terminal window.

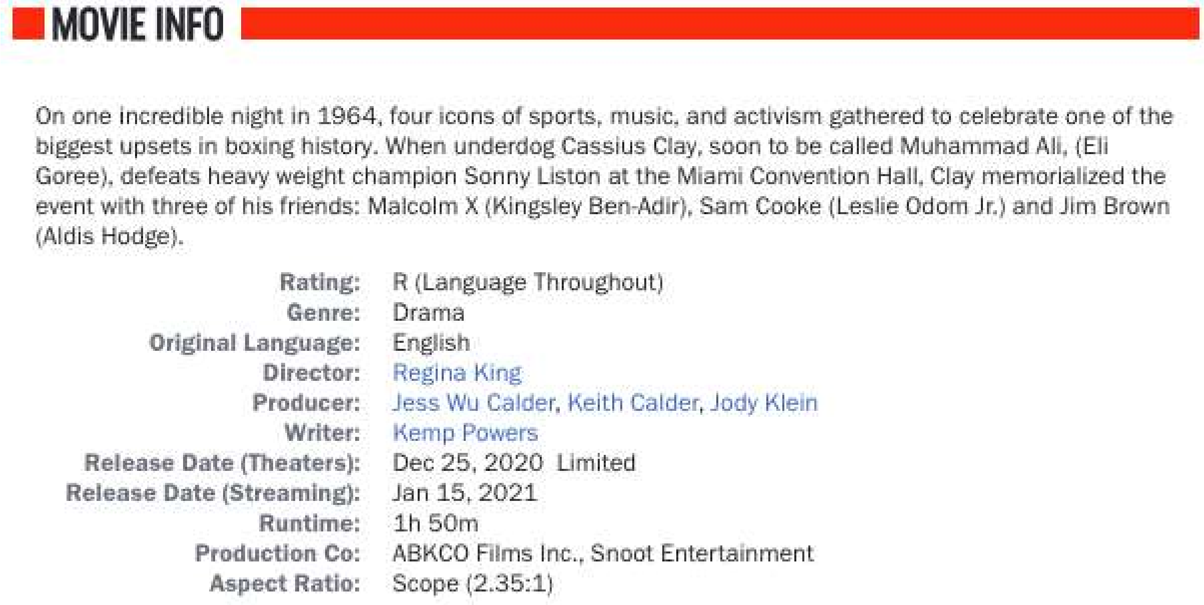
WebScrapers comparison

All three WebScrapers: `Beautiful Soup`, `Scrapy` and `Selenium` scraped the same data stored in a `.csv` file which contains 100 rows with observations for each movie and 15 columns with variables such as:

```
> ls(file)
 [1] "Aspect Ratio"              "Box Office (Gross USA)"   "Director"
 [4] "Genre"                     "Link"                     "Name"
 [7] "Original Language"         "Producer"                 "Production Co"
[10] "Rating"                    "Release Date (Streaming)" "Release Date (Theaters)"
[13] "Runtime"                   "Sound Mix"                "Writer"
```

Thus the goal of the project that all of the scrapers should scrap the same information from the domain of our choice was fulfilled. It is also worth to mention that we did not include a boolean parameter to True at the beginning of codes because by the definition our webpage conatins list of Top100 movies of 2020.

Down below there is a sample table of scraped data attached for each movie of the `Top100 movies of 2020` list.



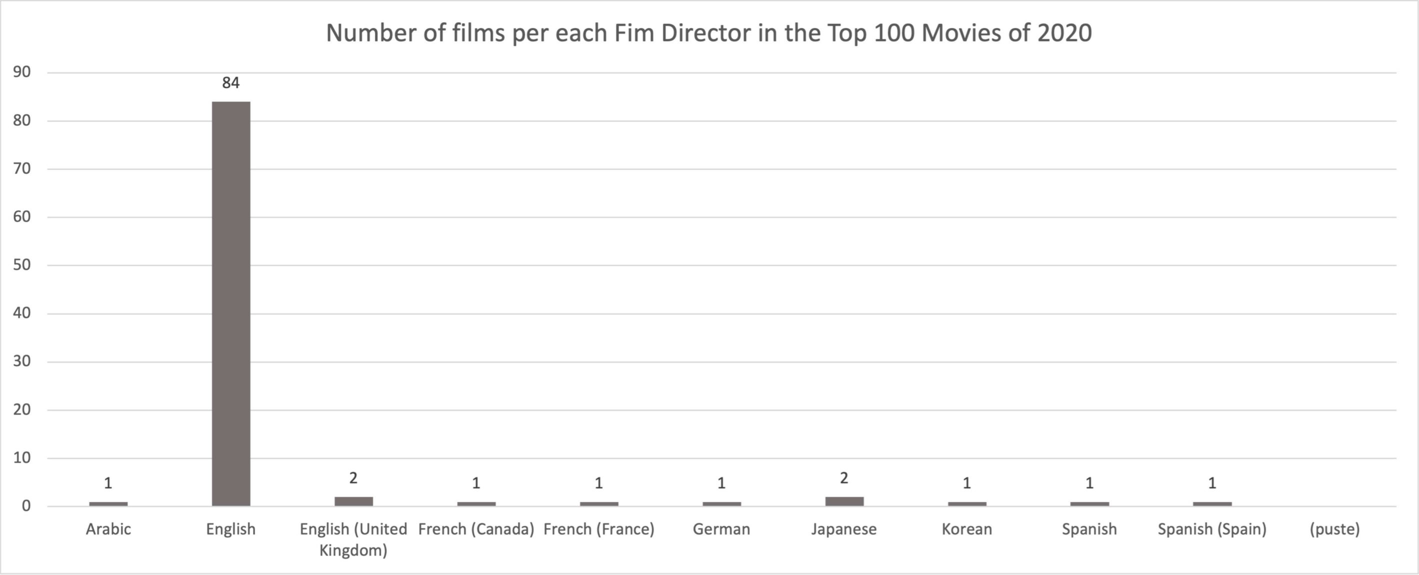
MOVIE INFO

On one incredible night in 1964, four icons of sports, music, and activism gathered to celebrate one of the biggest upsets in boxing history. When underdog Cassius Clay, soon to be called Muhammad Ali, (Eli Goree), defeats heavy weight champion Sonny Liston at the Miami Convention Hall, Clay memorialized the event with three of his friends: Malcolm X (Kingsley Ben-Adir), Sam Cooke (Leslie Odom Jr.) and Jim Brown (Aldis Hodge).

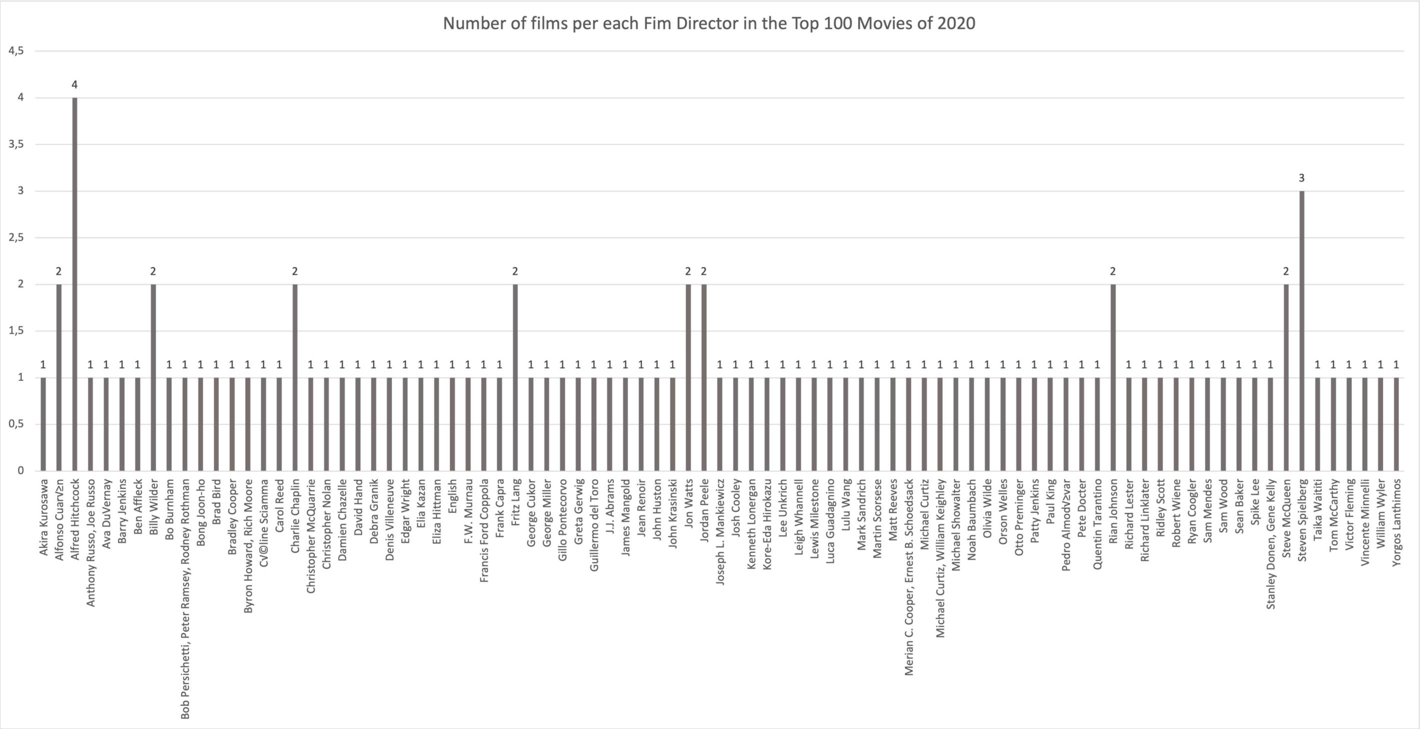
Rating:	R (Language Throughout)
Genre:	Drama
Original Language:	English
Director:	Regina King
Producer:	Jess Wu Calder, Keith Calder, Jody Klein
Writer:	Kemp Powers
Release Date (Theaters):	Dec 25, 2020 Limited
Release Date (Streaming):	Jan 15, 2021
Runtime:	1h 50m
Production Co:	ABKCO Films Inc., Snoot Entertainment
Aspect Ratio:	Scope (2.35:1)

Basic Data Analysis

The main aim of this section is to show how our collected data of Top100 movies of 2020 from Rotten Tomatoes can be further analyzed.In the following bar charts showing based on scrapped data, we can see that the vast majority of the Top 100 movies in 2020 are originally in English.



The next bar chart shows the number of films of Top100 movies of 2020 per each Film Director and as we can see Alfred Hitchcock and Steven Spielberg leads the list with 4 and 3 films respectively.



Division of work

Group Members: Andrea Furmanek (345813), Divij Pherwani (430990)

Task	Contributors
Beautiful Soup	Divij, Andrea
Scrapy	Divij
Selenium	Divij, Andrea
Github	Andrea
Project Report	Andrea, Divij

Instruction how to run scrapers:

BeautifulSoup Scraper

Run command: python soup.py

Total time taken is: 158

Scrapy Scraper

Spider 1: Run command: scrapy crawl links -o links.csv

Time to run (seconds): 1.11

Spider 2: Run command: scrapy crawl movies -o movies.csv

Time to run (seconds): 14.20

Selenium Scraper

Run command: python selenium.py

Total time taken is: 521

Run Time Analysis

Method	Time (in seconds)
Beautiful Soup	~ 160
Scrapy	~ 16
Selenium	~ 520

Scrapy is the fastest method for extracting movie information. It is followed by Beautiful Soup which takes 10 times more time. Selenium is the slowest method for such type of scraping requirements.