



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, often, it turns out 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 use with 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 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.

TBD ADD PHOTO of a github structure and file names in the text!!!!!!

WebScraper mechanisms

Beautiful Soup

The detailed described code for the first WebScraper can be seen in the file BeautifulSoup.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.

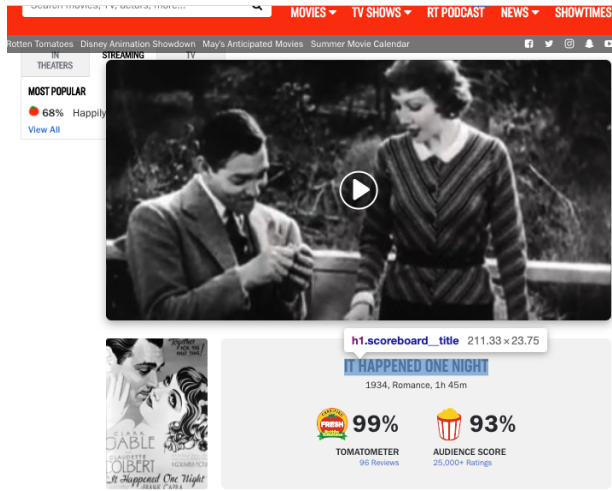
Rank	Rating	Title	No. of Reviews
1.	99%	It Happened One Night (1934)	96
2.	98%	Modern Times (1936)	108
3.	99%	The Wizard of Oz (1939)	145
4.	96%	Black Panther (2018)	521
5.	99%	Citizen Kane (1941)	116
6.	98%	Parasite (Gisaengchung) (2019)	458
7.	94%	Avengers: Endgame (2019)	540
8.	99%	Casablanca (1942)	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%	BlackKkklansman (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

```
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.

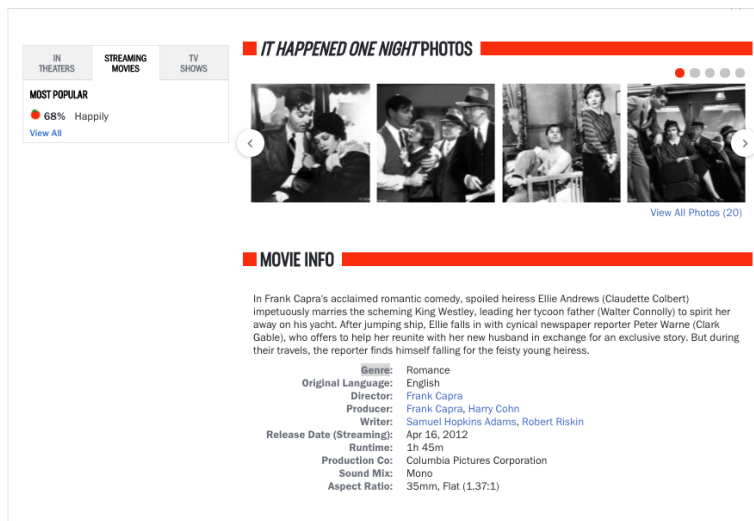
```
temp = [td.find('a', {'class': 'unstyled articleLink'}) for td in soup.find_all('td')]
for i in temp:
    if i is not None:
        href.append(i.get('href'))
```

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-ga = "score-panel-movie-title"`.



The screenshot shows the Rotten Tomatoes page for the movie "It Happened One Night". The page features a movie poster, a play button, and a score board with a 99% Tomatometer and 93% Audience Score. The right sidebar shows the movie's title and a brief synopsis.

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 `u1` tag with class `content-meta info` and then iterates over each label and its value having `data-ga` 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.



The screenshot shows the Rotten Tomatoes page for the movie "It Happened One Night". The page features a movie poster, a play button, and a score board with a 99% Tomatometer and 93% Audience Score. The right sidebar shows the movie's title and a brief synopsis.

Obtained output in a .csv format.

	A	B	C	D	E	F	G	H
	Movie	URL	Rating	Genre	Original Language	Director	Producers	Writer
1	It Happened One Night	https://www.rottentomatoes.com/m/it_happened_one_night		commedia	English	Frank Capra	Frank Capra, Harry Gray	Bertual Hill
2	Modern Times	https://www.rottentomatoes.com/m/modern_times	G	comedy	English	Charlie Chaplin	Charlie Chaplin	Charlie Chaplin
3	The Wizard of Oz	https://www.rottentomatoes.com/m/the_wizard_of_oz_1939		fantasy, musical, kids and family	English	Victor Fleming	Victor Fleming, Mervyn LaRoff	Noel Langley
4	Black Panther	https://www.rottentomatoes.com/m/black_panther_2018		adventure, fantasy, action	English	Ryan Coogler	Koeny Feggs	Ryan Coogler
5	Citizen Kane	https://www.rottentomatoes.com/m/citizen_kane	PG	drama	English	Orson Welles	Orson Welles	Herman J. Mank
6	Parasite	https://www.rottentomatoes.com/m/parasite_2019	R (Sexual Content, Language, Some Violence)	mystery and thriller, drama, comedy	Korean	Bong Joon-ho	Danah Kim	Han Joon-won
7	Avengers: Endgame	https://www.rottentomatoes.com/m/avengers_endgame	PG-13 (Sequences of Sex, F, Violence, Alcohol, Strong Language)	adventure, fantasy, sci-fi, action	English	Anthony Russo, Joe Russo	Kevin Feige	Christopher Yost
8	Goodwillie	https://www.rottentomatoes.com/m/goodwillie		drama	English	Michael Curtiz	Hal B. Wallis	Julius J. Egg
9	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 Goldstein, Bryce Alder	Rian Johnson
10	Toy Story	https://www.rottentomatoes.com/m/toy_story_1995	G	home, mystery and thriller	English	Jordan Peele	Jonas Pank, Sean McElrhain, Jason Blum, Jon Cooper	Jordan Peele
11	Toy Story 4	https://www.rottentomatoes.com/m/toy_story_4	G	adventure, fantasy, animation, kids and family, comedy	English	Josh Cooley	Mark Nelson, Jason Rouse, Galyn Susman	Andrew Sta
12	Lady Bird	https://www.rottentomatoes.com/m/lady_bird	R (Language, Brief Graphic Nudity, Sexual Content, Clean Language)	drama, comedy	English	Greta Gerwig	Scott Rudin, Eli Bush, Evelyn O'Hall	Greta Gerwig
13	Mission: Impossible - Fallout	https://www.rottentomatoes.com/m/mission_impossible_fallout	PG-13 (Some Sequences of Action, Brief Strong Language, Intense Violence)	adventure, action, mystery and thriller	English	Christopher McQuarrie	Tom Cruise, Christopher McQuarrie, John Nurn, J.J. Abrams	Christopher McQuarrie
14	Blockbuster	https://www.rottentomatoes.com/m/blockbuster	R (Language Through, Drug Use, Strong Language, Thematic Elements, Sexual References)	crime, drama, comedy	English	Stephen Lee	Sean McElrhain, Raymond McElrhain, Jason Blum, Jordan Peele, Spike Lee	Charles Wal
15	Get Out	https://www.rottentomatoes.com/m/get_out	R (Language, Blood, Images, Sexual References, Intense Violence)	horror, mystery and thriller, comedy	English	Jordan Peele	Sean McElrhain, Jason Blum, Edward H. Hannon Jr., Jordan Peele	Jordan Peele
16	The Irishman	https://www.rottentomatoes.com/m/the_irishman	R	crime, drama	English	Martin Scorsese	Martin Scorsese, Robert De Niro, Jane Rosenthal, Gaetano Pastorelli, Harold Ernsdorf, Emma Tillinger Koskoff, Gerald Chonacas, Ivan Witzler	Steven Zaif
17	The Godfather	https://www.rottentomatoes.com/m/the_godfather	R	crime, drama	English	Francis Ford Coppola	Albert R. Broccoli	Francis Ford
18	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. Vanston	George Miller
19	Spider-Man Into the Spider-Verse	https://www.rottentomatoes.com/m/spider-man_into_the_spider_verse	PG (Brief Language, Fantasy Action Violence, Thematic Elements)	adventure, fantasy, animation, kids and family, action, comedy	English	Bob Persichetti, Peter Ramsey, Rodney Rothman	Avi Arad, Amy Poehler, Phil Lord, Christopher Miller, Christina Steinberg	Phil Lord, R
20	Al About Eve	https://www.rottentomatoes.com/m/al_about_eve		drama	English	Joseph L. Mankiewicz	Joseph L. Mankiewicz	Joseph L. M
21	Moulin Rouge	https://www.rottentomatoes.com/m/moulin_rouge_2001	R (Drug Use, Brief Violence, Language Through, Sexual References)	gay and lesbian, drama	English	Baz Luhrmann	Adrian Romo, David Gindoff, Jeremy Kleiner	Baz Luhrmann
22	Reference	https://www.rottentomatoes.com/m/reference	PG-13 (Some Sequences of Action, Brief Strong Language, Thematic Elements)	mystery and thriller	English	Alfred Hitchcock	David O. Selznick	Robert E. O
23	A Star Is Born	https://www.rottentomatoes.com/m/a_star_is_born_2018	R (Some Sexual Nudity, Language Through, Thematic Elements)	drama, romance, music	English	Bradley Cooper	BM Cooper, Jon Peters, Bradley Cooper, Todd Phillips, Lynette Howell Taylor	Eric Roth, B
24	Wonder Woman	https://www.rottentomatoes.com/m/wonder_woman_2017	PG-13 (Sequences of Violence, Alcohol, Sexual References, Content)	adventure, fantasy, action	English	Patty Jenkins	Charles Roven, Deborah Snyder, Zack Snyder, Richard Suckle	Alan Horn
25	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
26	A Quiet Place	https://www.rottentomatoes.com/m/a_quiet_place_2018	PG-13 (Some and Some Blood, Images)	horror, mystery and thriller	English	John Krasinski	Michael Bay, Andrew Form, Brad Fuller	Bryan Woods
27	The Cabinet of Dr. Caligari	https://www.rottentomatoes.com/m/the_cabinet_of_dr_caligari		horror	English	Rudolf Wenz	Rudolf Wenz, Erich Pommer	Hans Janes
28	Eighth Grade	https://www.rottentomatoes.com/m/eighth_grade	R (Some Sexual Material, Language)	drama, comedy	English	Bryl Sunam	Scott Rudin, Eli Bush, Lisa Niren, Christopher Stone	Bryl Sunam
29	Reus	https://www.rottentomatoes.com/m/reus_2018	R (Language, Graphic Nudity, Sexual References, Thematic Elements)	drama	Spanish	Alfonso Cuarón	Alfonso Cuarón, Gabriela Rodríguez, Nicolás Gale	Alfonso Cuar
30	Booksmart	https://https://www.rottentomatoes.com/m/booksmart	R (Language Through, Drug Use and Drinking, Strong Sexual Content)	gay and lesbian, comedy	English	Olivia Wilde	Megan Ellison, Jessica Elbaum, Katie Silverman, Chelsea Bennett, David O'Connell	Katie Silver
31	Dunkirk	https://www.rottentomatoes.com/m/dunkirk_2017	PG-13 (Some Language, Intense War Experiences)	drama, history, war	English	Christopher Nolan	Emma Thomas, Christopher Nolan	Christopher Nolan
32	Good	https://www.rottentomatoes.com/m/good	PG (Thematic Elements)	adventure, animation, kids and family, music, comedy	English	Lee Unkrich	Dan A. Anderson	Adrian Maki
33	A Night of the Opera	https://www.rottentomatoes.com/m/a_night_of_the_opera		comedy	English	Ben Hurst	Ben Hurst	Ben Hurst
34	Portrait of a Lady on Fire	https://www.rottentomatoes.com/m/portrait_of_a_lady_on_fire	R (Some Nudity and Sexual)	gay and lesbian, drama, romance, history	French (France)	Céline Sciamma	Bénédicte Couvreur, Vitorique Caple	Céline Sciam
35	The Favourite	https://www.rottentomatoes.com/m/the_favourite_2018	PG (Some Sexual Material, Language, Thematic Elements)	drama, comedy	English	Luke Wragg	Chris Weitz, Andrew Mann, Peter Saraf, Matt Turfitt, Andie Cox, Darlene Tan Mello, Jane Zheng, Dan Mello	Luke Wragg
36	The Shape of Water	https://www.rottentomatoes.com/m/the_shape_of_water_2017	R (Language, Graphic Nudity, Sexual Content, Thematic Elements)	fantasy, romance	English	Guillermo del Toro	Guillermo del Toro, J. Miles Dale	Guillermo de
37	Thor: Ragnarok	https://www.rottentomatoes.com/m/thor_ragnarok_2017	PG-13 (Brief Sexual Material, Action Violence, Some F, Violence)	adventure, fantasy, action, comedy, sci-fi	English	Taika Waititi	Kevin Feige	Eric Pearson
38	Belva	https://www.rottentomatoes.com/m/belva	PG-13 (Brief Strong Language, Suggestive Material, Disturbing Thematic Material, Violence)	history, drama	English	Alex Giblin	Christian Colson, Oprah Winfrey, David Gindoff, Jeremy Kleiner	Paul Weitz
39	Springsteen	https://www.rottentomatoes.com/m/springsteen_2015	R (Some Language, Sexual References)	drama	English	Tom McCarthy	Michael Sugar, Steve Golin, Nicole Radloff, Ryan Pickett	John Singer
40	Seven Samurai	https://www.rottentomatoes.com/m/seven_samurai_1954		action	Japanese	Akira Kurosawa	Suifu Matsui	Shirofuku Hig
41	Grand Hustle	https://www.rottentomatoes.com/m/grand_hustle		war, drama	French (Canada)	Jean Renoir	Albert Prévost, Frank Ruhlman	Orson Wel
42	The Third Man	https://www.rottentomatoes.com/m/the_third_man		mystery and thriller	English	Carol Reed	Alexander Korda, Carol Reed, David O. Selznick	Orson Wel
43	Amel	https://www.rottentomatoes.com/m/amel_2018	PG-13 (Brief Strong Language)	mystery and thriller, drama, sci-fi	English	David Villeneuve	Shawn Levy, Dan Levine, Adam Ryder, David Linde	Eric Heine
44	Single in the Rain	https://www.rottentomatoes.com/m/single_in_the_rain	G	musical, comedy	English	Stanley Donen, Gene Kelly	Arthur Freed	Betty Comp
45	The Favourite	https://www.rottentomatoes.com/m/the_favourite_2018	R (Nudity, Language, Strong Sexual Content)	comedy, drama	English (United Kingdom)	Yorgos Lanthimos	Clare Corbett, El Guany, Lee Maglay, Yorgos Lanthimos	Delaney O
46	An American in Paris	https://www.rottentomatoes.com/m/an_american_in_paris		musical, romance	English	Vicente Minnelli	Arthur Freed	Alan Jay L
47	Logan	https://www.rottentomatoes.com/m/logan_2017	R (Language Through, Brief Nudity, Strong Sexual Violence)	adventure, fantasy, action	English	James Mangold	Hugh Parker, Simon Kitting, Lauren Shuler Donner	James Man
48	All Quiet on the Western Front	https://www.rottentomatoes.com/m/all_quiet_on_the_western_front		war	English	Levins Mestow	Carl Laemmle Jr.	John Singer
49	Double Intensity	https://www.rottentomatoes.com/m/double_intensity		drama, crime	English	Billy Wilder	Stanley Donen, Gene Kelly	James M. C
50	On the Waterfront	https://www.rottentomatoes.com/m/on_the_waterfront		drama	English	Elia Kazan	Stanley Donen, Gene Kelly	Betty Comp
51	On the Waterfront	https://www.rottentomatoes.com/m/on_the_waterfront		drama	English	Elia Kazan	Stanley Donen, Gene Kelly	Betty Comp
52	Marriage Story	https://www.rottentomatoes.com/m/marriage_story_2019	R (Sexual References, Language Through)	drama, comedy	English	Noah Baumbach	Noah Baumbach, David Heyman	Noah Baumb
53	E.T. the Extra-Terrestrial	https://www.rottentomatoes.com/m/e_t_the_extra_terrestrial	PG	adventure, sci-fi, kids and family	English	Steven Spielberg	Steven Spielberg, Kathleen Kennedy	Melissa Mat
54	Some White and the Seven Dwarfs	https://www.rottentomatoes.com/m/some_white_and_the_seven_dwarfs	G	fantasy, animation, kids and family	English	Walt Disney	Walt Disney	Walt Disney

Scrapy

TBC

Selenium

Similar to the previous WebScraper, for Selenium, described source code was also saved in the file `scrapy.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 `collections` 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.headless = False
driver = webdriver.Chrome(options=options, executable_path=gecko_path)

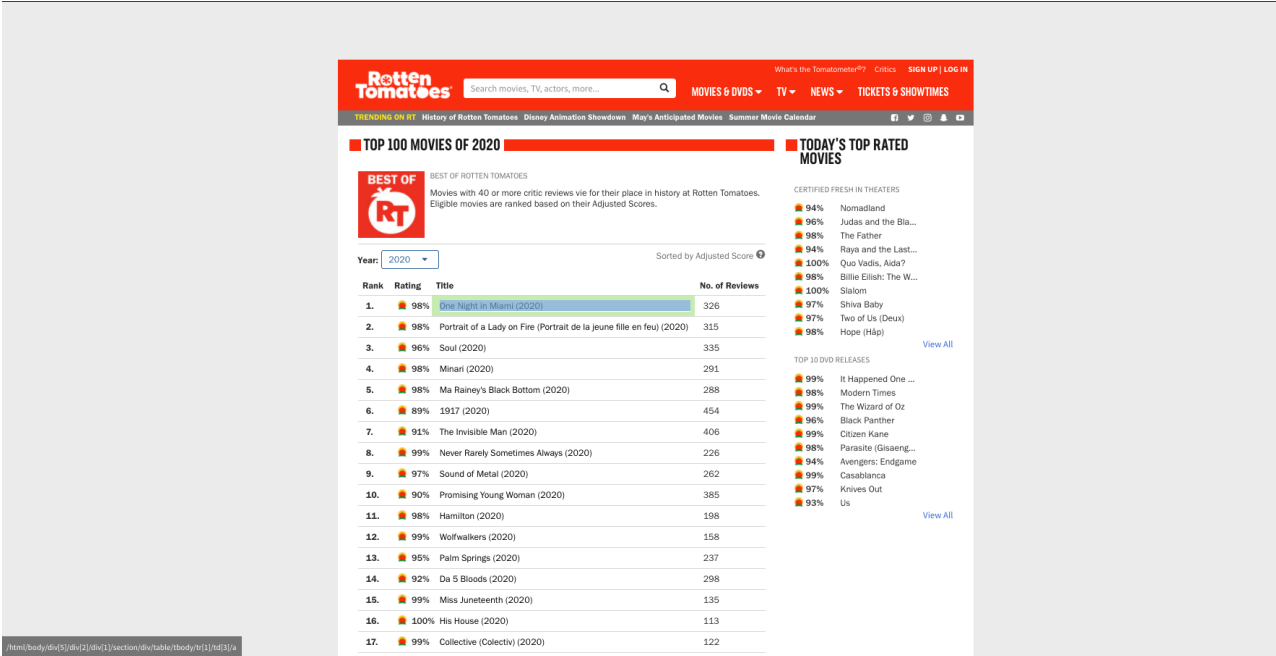
# Actual Program
driver.get(url)
print(driver.page_source)

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)
```



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 16 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"             "View the collection"
[16] "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. Down below there is sample table of scraped data attached for each movie of Top100 movies of 2020.

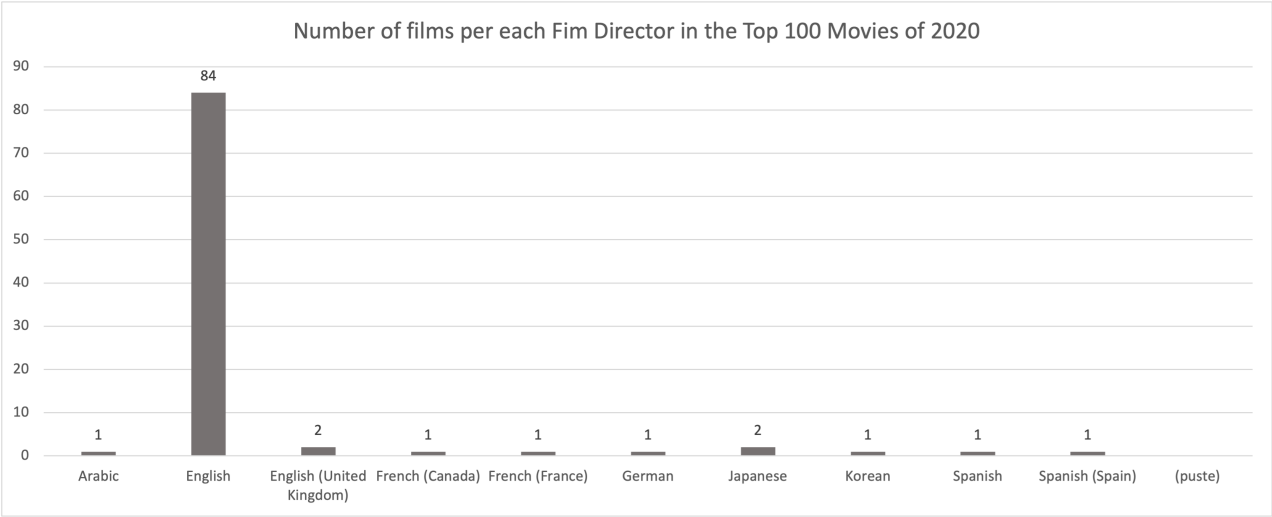
MOVIE INFO

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).

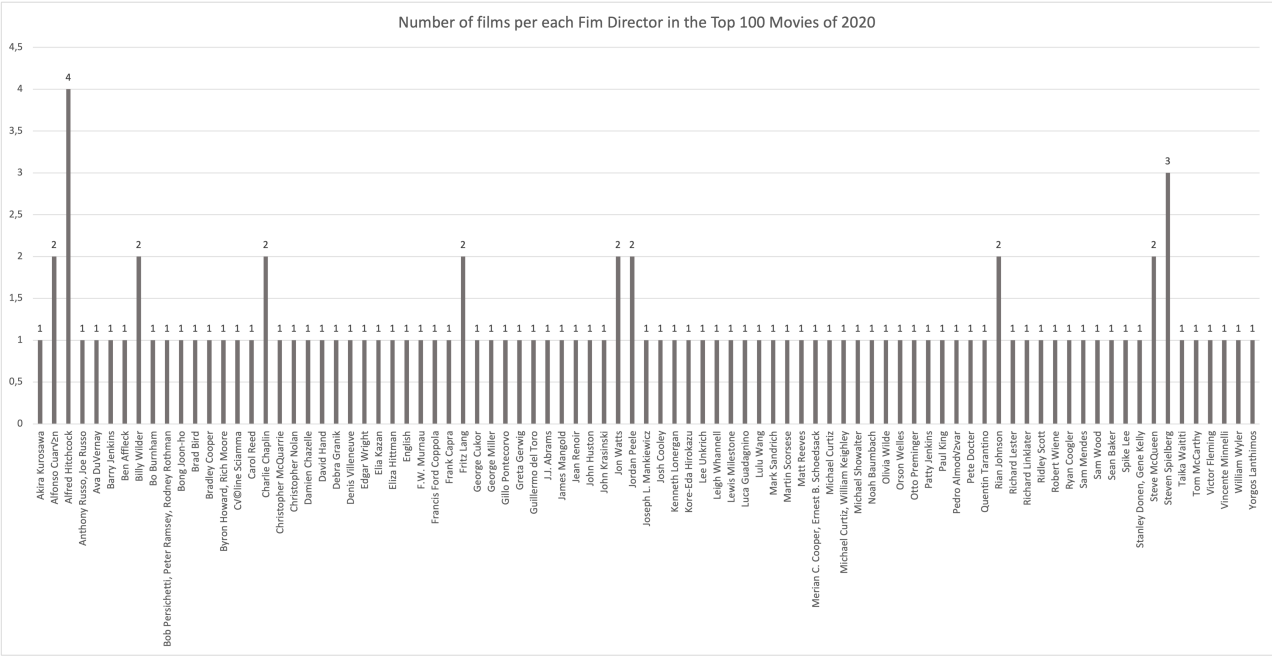
Rating:	R (Language Throughout)
Genre:	Drama
Original Language:	English
Director:	Regina King
Producer:	Jess Wu Calder , Keith Calder , Jody Klein
Writer:	Kemp Powers
Release Date (Theaters):	Dec 25, 2020 Limited
Release Date (Streaming):	Jan 15, 2021
Runtime:	1h 50m
Production Co:	ABKCO Films Inc.,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