Digital Objects and Research Practices

Helge Moes

11348801

Work Group 2

Davide Beraldo

Research Project 2 – capturing data

Word count: 1414

Introduction

Nowadays, social media has been mentioning movies multiple times. Movies have been subject matter to discuss about for several decennia already. Many people have different opinions and views on certain movies. Moreover, not only people, but also web pages such as https://www.imdb.com/. The webpage gives certain information on movies. Furthermore, the webpage categorizes movies and submits them to certain lists such as "top rated movies" and "most popular movies". However, what makes a certain movie to be a top rated movie or a most popular movie. In addition, what is the difference between these two categories? This paper will assess these questions and will try to clarify the grey area between the two lists.

Moreover, movies have been criticized and categorized since its existence. The criticism consists out of subjective and objective features. What makes some movies to be considered as masterpieces and what makes certain movies to be disregarded in their existence. In the case of IMDB, it is based off of ratings by people who visit the page. The difference between the two categories seem to be difficult to differ from one another. Since top rated movies are popular movies, as they generated a lot of revenue and success. Therefore, the research will be clarifying the difference between both lists as presented on the IMDB web page. Moreover, both lists will also be compared to the most successful movies displayed on the box office list that is represented on the IMDB web page.

Methodology

Dataminer is an extension of Google Chrome that scrapes data from web pages. Scraping is the act of collecting data. After the data has been scraped, data miner allows the user to convert the collected data to an excel sheet. For this research, data miner has to be downloaded to the google chrome browser.

The scraping of the data did not go as planned initially. Before utilizing the dataminer extension, the research created a different google chrome account to prevent biased external factors. IMDB is a site that does make their data publicly accessible. Therefore, the validity of the research was based on the data presented on the web page and the data was not restricted by any limitations. The research started off only using the "Capturing Data" (Peeters & Borra n.d.) worksheet as a means to scrape the data from the IMDB web page. However, selecting the specific rows and columns made it difficult to generate an excel sheet that presented the findings. Mainly, it only showed the titles that were given to the rows.

In addition, after a while of figuring out how the dataminer tool worked, the research managed to present the findings of the data in a single excel cell. This made the data irrelevant and difficult to read.

Eventually, the data was easily to scrape, since the whole page got selected and through the dataminer extension the other rows were removed from the excel program. However, there were a few missing ratings that were not able to include in the tables. This problem consisted mainly in the "most popular list" of IMDB. Another problem that came up was the "box office" list. This list was not complete and it was not able to generate a list that was complete from IMDB. However, the list does provide insight to which movies are represented the most from the lists of "top rated movies" and "most popular movies."

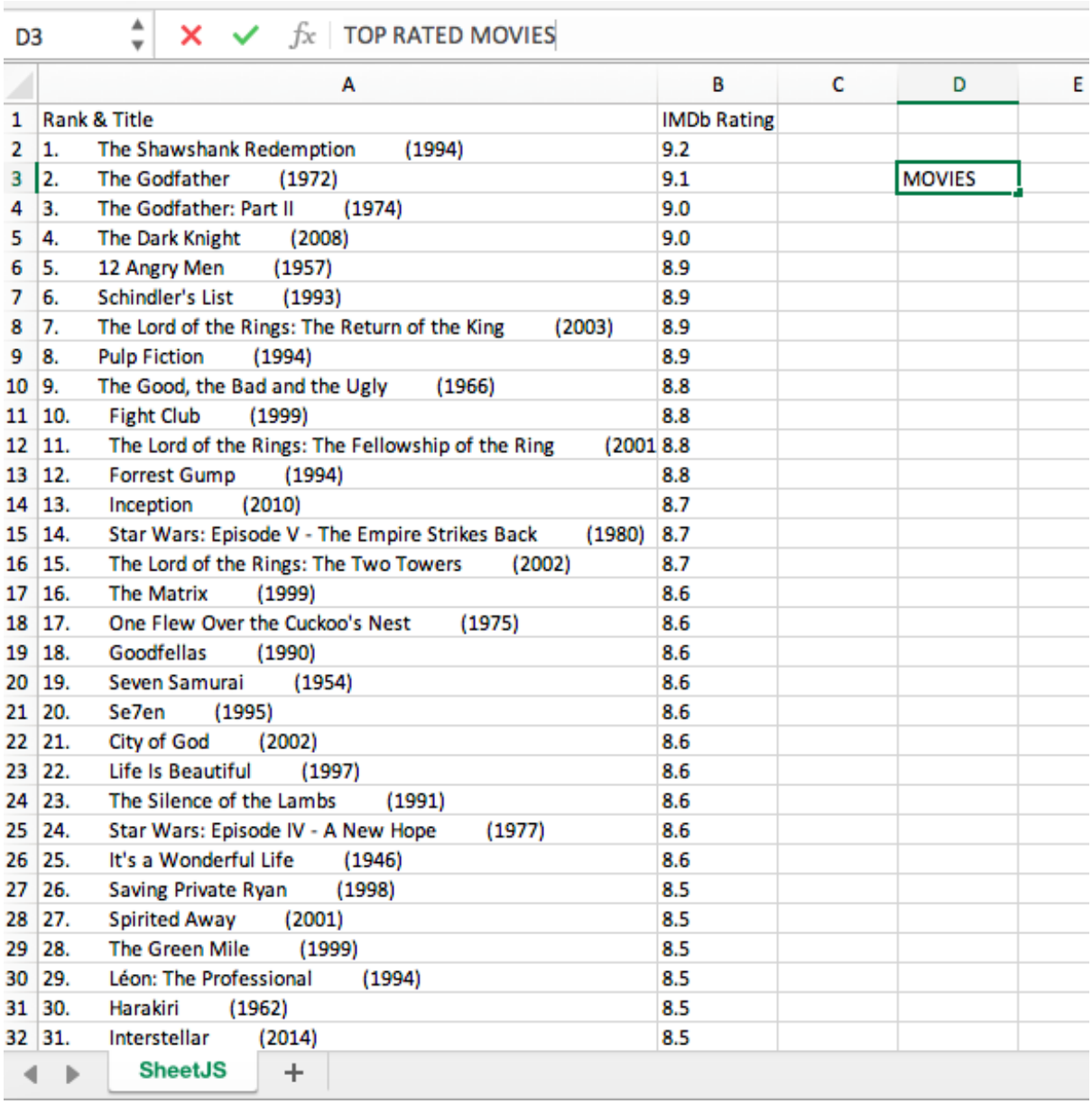
Analysis

By using Dataminer, the following findings were scraped from the IMDB site and transferred to excel.

D1 | MOST POPULAR MOVIES

	A	B	C	D	E	F
1	Rank & Title	IMDb Rating		MOVIES		
2	It Chapter Tv	7.0				
3	Joker (2	9.5				
4	Hustlers	6.6				
5	Once Upon a	8.0				
6	It (2017	7.4				
7	Doctor Sleep					
8	Downton Abl	7.8				
9	The Goldfinc	6.3				
10	Ad Astra	7.3				
11	Anna (2	6.6				
12	Avengers: En	8.6				
13	Spider-Man:	7.8				
14	Ready or Not	7.2				
15	Dark Phoenix	5.8				
16	Rambo: Last	7.2				
17	Midway					
18	John Wick: C	7.7				
19	Aladdin	7.1				
20	Angel Has Fa	6.7				
21	Star Wars: T					
22	Yesterday	7.0				
23	Terminator:					
24	Jojo Rabbit	7.2				
25	The Dead Do	5.6				
26	The Fanatic	3.8				
27	Child's Play	6.0				
28	Chhichhore	8.6				
29	Men in Black	5.6				
30	Good Boys	6.9				
31	Rocketman	7.5				
32	The Dark Cry	7.2				

Figure 1: List of the "most popular movies" on the IMDB web page



D3 TOP RATED MOVIES

	A	B	C	D	E
1	Rank & Title	IMDb Rating			
2	1. The Shawshank Redemption (1994)	9.2			
3	2. The Godfather (1972)	9.1		MOVIES	
4	3. The Godfather: Part II (1974)	9.0			
5	4. The Dark Knight (2008)	9.0			
6	5. 12 Angry Men (1957)	8.9			
7	6. Schindler's List (1993)	8.9			
8	7. The Lord of the Rings: The Return of the King (2003)	8.9			
9	8. Pulp Fiction (1994)	8.9			
10	9. The Good, the Bad and the Ugly (1966)	8.8			
11	10. Fight Club (1999)	8.8			
12	11. The Lord of the Rings: The Fellowship of the Ring (2001	8.8			
13	12. Forrest Gump (1994)	8.8			
14	13. Inception (2010)	8.7			
15	14. Star Wars: Episode V - The Empire Strikes Back (1980)	8.7			
16	15. The Lord of the Rings: The Two Towers (2002)	8.7			
17	16. The Matrix (1999)	8.6			
18	17. One Flew Over the Cuckoo's Nest (1975)	8.6			
19	18. Goodfellas (1990)	8.6			
20	19. Seven Samurai (1954)	8.6			
21	20. Se7en (1995)	8.6			
22	21. City of God (2002)	8.6			
23	22. Life Is Beautiful (1997)	8.6			
24	23. The Silence of the Lambs (1991)	8.6			
25	24. Star Wars: Episode IV - A New Hope (1977)	8.6			
26	25. It's a Wonderful Life (1946)	8.6			
27	26. Saving Private Ryan (1998)	8.5			
28	27. Spirited Away (2001)	8.5			
29	28. The Green Mile (1999)	8.5			
30	29. Léon: The Professional (1994)	8.5			
31	30. Harakiri (1962)	8.5			
32	31. Interstellar (2014)	8.5			

SheetJS

Figure 2: List of the "top rated movies" on the IMDB web page

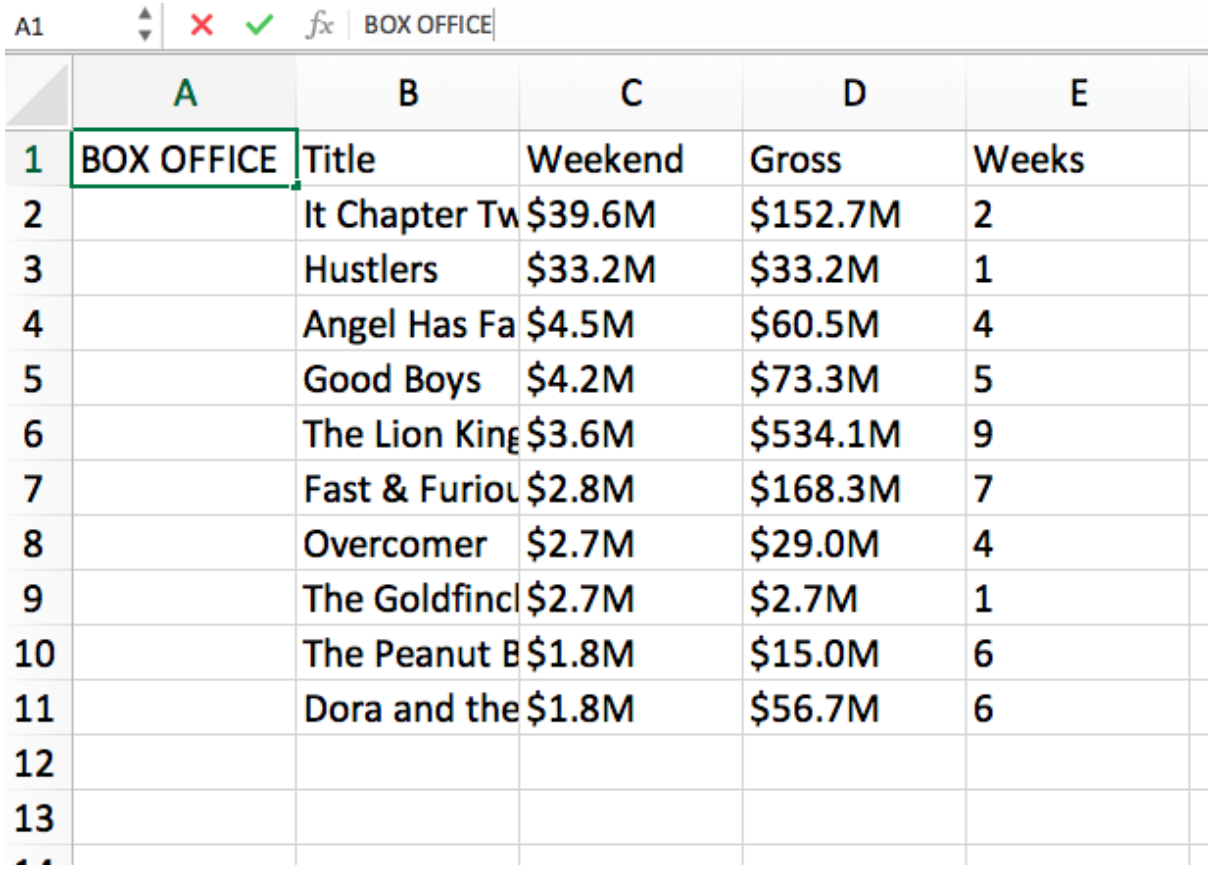
Based on these lists, the “top rated movies” list is arranged in an order of most rated movie to less rated movies. Furthermore, the top rated movies are mentioning all movies ever produced. The list is reduced to main stream cinema and does not consider art house or independently produced movies.

The list of the “most popular movies” is limited to the most recent produced movies. The ratings are not arranged in an order. However, the movie titles are ordered from most recent published movie to older published movies.

Moreover, the “most popular movie” list consisted out of more Hollywood produced movies, since there are the biggest studios that create the most blockbusters. Consequently, based on the “most popular movie” list the research can conclude that it consists out of primarily blockbusters. These blockbusters consist out of mainly the following genres: comedy, action and horror.

On the other hand, “the top rated movies” list consists out of different productions, for example, independent productions and other studios outside of Hollywood. Therefore, the “top rated movies” list has a broader point of view than the “most popular movie” list, since it considers movies produced from all over the world. The movies on this list consists out of the following genres: art house, drama and documentary. Furthermore, the “top rated movies” list has mainly movies that have withstood the test of time and developed themselves as classics.

In addition to comparing both the lists, the research also included a “box office” list as to provide an insight whether which list is more represented in the top earning movies at the moment.



A1 | BOX OFFICE

	A	B	C	D	E
1	BOX OFFICE	Title	Weekend	Gross	Weeks
2		It Chapter Tw	$39.6M	$152.7M	2
3		Hustlers	$33.2M	$33.2M	1
4		Angel Has Fa	$4.5M	$60.5M	4
5		Good Boys	$4.2M	$73.3M	5
6		The Lion King	$3.6M	$534.1M	9
7		Fast & Furiou	$2.8M	$168.3M	7
8		Overcomer	$2.7M	$29.0M	4
9		The Goldfinch	$2.7M	$2.7M	1
10		The Peanut B	$1.8M	$15.0M	6
11		Dora and the	$1.8M	$56.7M	6
12					
13					

Figure 3: List of the "box office" on the IMDB web page

Figure 3 shows the list of the top earning movies at the moment. Since the list could only be continued to another list that has been published by mojo, it was not able to get the full list of movies. However, the list on the IMDB web page still showed that the "most popular movies" were represented more than the "top rated movies." In this case it is expected, since the "most popular list" consists out of the movies that have been published recently. Furthermore, the list also shows the weeks of how long it took until a movie generated a certain sum of revenue. Therefore, the data did not show a fixed time for all movies, for example only considering the profit a movie generated after the first week being shown in the cinema. Moreover, movies have been earning more nowadays compared to twenty years ago. Therefore, the data of this list is rather biased and does not represent a valid picture of best earning movies.

In conclusion, the "most popular movies" list is focused on the most recent published movies and disregards the ratings. In addition, the "top rated movies" is based on the ratings of all movie blockbusters that have been published. The "box office" list did not provide a liable insight to whether which list has been more successful to the other.

Conclusion

In summary, the research is lacking definitive data, since the data is based on subjective ratings by people, which presents biased data. The lists of "top rated movies" and "most popular movies" were limited to blockbuster movies and disregarded the arthouse and independent produced publications. However, the "top rated movies" list did represent more alternative movies in the list compared to the "most popular movies" list. Nevertheless, the list of "top rated movies" was not complete, as some ratings of movies did not seem scrapable. Therefore, leaving gaps in the excel sheet where a rating is supposed to be.

Nevertheless, the data scraping, although it was not perfect, still accomplished the means to find an evident difference. The data displayed a clear difference between both lists. The "box office" list was not able to continue the list properly, as it went over to another site. Therefore, exceeding the boundaries that were able to scrape with dataminer. The list on the mojo site seemed to be restricted, as it was not able to extract the data from the page. For that reason, the "box office" list was not complete and did not display enough data to support a valid argument.

However, despite the fact that some data was not scrapable, the difference between both the lists has been evident in the excel sheets. The data of "top rated movies" puts the emphasis on ratings, where the data of "most popular movies" arranges the movies from most recent to oldest publication. This research also displayed lists that were not sufficient to support the argument of the difference between both lists, in this case it was the "box office" list. Nevertheless, it did provide the fact that the list of "most popular movies" was more represented than the "top rated movies."

References

Peeters, Stijn, Erik Borra. “Capturing Data.” n.d. Google Docs. Accessed September 20, 2019. https://docs.google.com/document/d/10Pvry6SgaybdLBnYxUu0G45qF0upOW0sHjfCJOvpNTg/edit?usp=embed_facebook.

“Ratings and Reviews for New Movies and TV Shows.” n.d. IMDb. Accessed September 20, 2019. http://www.imdb.com/.