Rotten Tomatoes Sentiment and Network Analysis

Contents

Part 1

Introduction

This report aims to take advantage of the rotten tomatoes data that contains interesting details about movies by getting intuitive ideas or explanations after in-depth understanding of the data.
In order not to dig into the details, I should define the direction of the research with a clear and explicit research question, therefore, all the variables required to answer the question are present, or rather they can be created through some processing or manipulation of these variables.
I will start by exploring the data so that the picture becomes clearer.
After loaded the data, I found that there are two CSVs, the first file is rotten_tomatoes_ movies.csv which contains 22 columns and 17,712 rows. Of course, we do not need to deal with all the variables, but we can focus only on what we can benefit from as follows:
movie title
genres
actors
audience_rating
tomatometer_rating

I'll explain later why I rely on these columns specifically.
the second file is rotten_tomatoes_meta.csv which contains 3 columns and 7290 rows, I will use the three columns later in my analysis as follows:
title
text
top critic

I figured out also the data that I have is a good candidate for sentiment analysis as we have textual data. In addition, the data could be used for network analysis due to the relations between the columns as we will see later.

Part 2

Research Question

As a data scientist in a company like amazon or Netflix, you must be well aware of the evidence and the market, as well as the social and psychological factors of the consumer or recipient must be aware of, and it must be known that there is no magic or radical solution that can be extracted from the data, but some quality or service improvement can be added, and so on, and from this Starting off, I thought my question would be who are the actors and influencers in the success of the films.
In order to answer this question, it is necessary to define what is a successful movie.
I explored the data and figured out that audience_rating and tomatometer_rating could be used to represent the successful movie.
We can use them together because they are the summation of webmasters and movie viewers
And I assumed that the successful film had a rating of at least 50, whether it is the website rating or the audience's rating.

Here it should be noted that this does not mean that the evaluation is perfect or expressive in an absolute way about the reality of the film, but this is the closest way from my point of view based on the available data.

Part 2

Methodology:

In order to carry out the research, the data must be prepared in the appropriate manner, which requires deletion of some columns that will not be used and will not be relied on, and it is one of the important steps, especially in the case that the data is huge, because this saves time and processing capacity, and therefore a lot of money.
AS shown Table 1, I will depend on those columns from the first cvs file to conduct the first part of the research to generate the significant actors of successful movies.

movie_title	actors	tomatometer_rating	audience_rating	genres
Bitch Slap	Julia Voth, Erin Cumming...	29	29	Action & Adventure
Breakout	Charles Bronson, Robert ...	40	41	Action & Adventure
Commando	Arnold Schwarzenegger, ...	71	67	Action & Adventure
Missing in Action	Chuck Norris, M. Emmet ...	19	42	Action & Adventure
Nighthawks	Sylvester Stallone, Lindsa...	70	54	Action & Adventure
Raw Deal	Arnold Schwarzenegger, ...	23	28	Action & Adventure
Red Dawn	Patrick Swayze, C. Thoma...	46	65	Action & Adventure
Tarzan, the Ape Man	Bo Derek, Miles O'Keeffe,...	10	19	Action & Adventure
The Hitman	Chuck Norris, Michael Pa...	13	45	Action & Adventure
Sniper	Tom Berenger, Billy Zane,...	38	55	Action & Adventure
Gunmen	Christopher Lambert, Ma...	15	35	Action & Adventure
Street Fighter	Jean-Claude Van Damme...	10	20	Action & Adventure
Bad Boys	Martin Lawrence, Will Smi...	42	78	Action & Adventure
The Quest	Jean-Claude Van Damme...	14	36	Action & Adventure

Showing 1 to 14 of 17,712 entries, 5 total columns

Table 1

From the previous table, I will separate each actor from the rest of the actors so that I can link each actor and evaluate the film separately in a separate row as shown in Table 2.

	movie_title	actors	rating
1	Percy Jackson & the Olympians: The Lightning Thief	Logan Lerman	51.0
2	Percy Jackson & the Olympians: The Lightning Thief	Brandon T. Jackson	51.0
3	Percy Jackson & the Olympians: The Lightning Thief	Alexandra Daddario	51.0
4	Percy Jackson & the Olympians: The Lightning Thief	Jake Abel	51.0
5	Percy Jackson & the Olympians: The Lightning Thief	Sean Bean	51.0
6	Percy Jackson & the Olympians: The Lightning Thief	Pierce Brosnan	51.0
7	Percy Jackson & the Olympians: The Lightning Thief	Steve Coogan	51.0
8	Percy Jackson & the Olympians: The Lightning Thief	Rosario Dawson	51.0
9	Percy Jackson & the Olympians: The Lightning Thief	Melina Kanakaredes	51.0
10	Percy Jackson & the Olympians: The Lightning Thief	Catherine Keener	51.0
11	Please Give	Catherine Keener	75.5
12	Please Give	Amanda Peet	75.5
13	Please Give	Oliver Platt	75.5

Table 2

In terms of ratings, I took the average of Rotten Tomato's rating and audience ratings in order to account for all available ratings.
Also, I took the average rating according to the number of movies the star participates in, which makes sense because with more movies, the rating gets closer to the truth.
The results are shown in Table 3.

	actors	Number_of_Movies	mean_rating
1	Robert Duvall	45	77.70000
2	Robert De Niro	49	75.96939
3	Samuel L. Jackson	57	73.70175
4	Meryl Streep	48	71.22917
5	Liam Neeson	47	69.93617

Table 3

The order of the data was taken into account in the previous table, and from the table as shown, Robert Duvall is the best actor that guarantees a successful movie in average, De Niro is number two and so on as seen in the table 3.

In this way, in short, we are able to show the most influential stars in the success of the films, in addition to the previous analysis, I will dig more through the other file rotten_tomatoes_ meta.csv that was referred to at the beginning.

I will apply Natural Language Processing (NLP) to the audiences' review of the stars selected from the previous table as influential film stars. I will implement the Bag of Words algorithm.

Before implementing the algorithm there are many steps for data preprocessing to prepare the reviews column for further analysis because in text there are many mentions, website URLs,

hashtags, punctuations, numbers, stop words. Such objects in the data are misleading and should be removed before further analysis.

Figure 1 shows the output of NLP basic analysis that I have done to the rotten tomatoes data.

Note: the code is attached in a separate .R file.

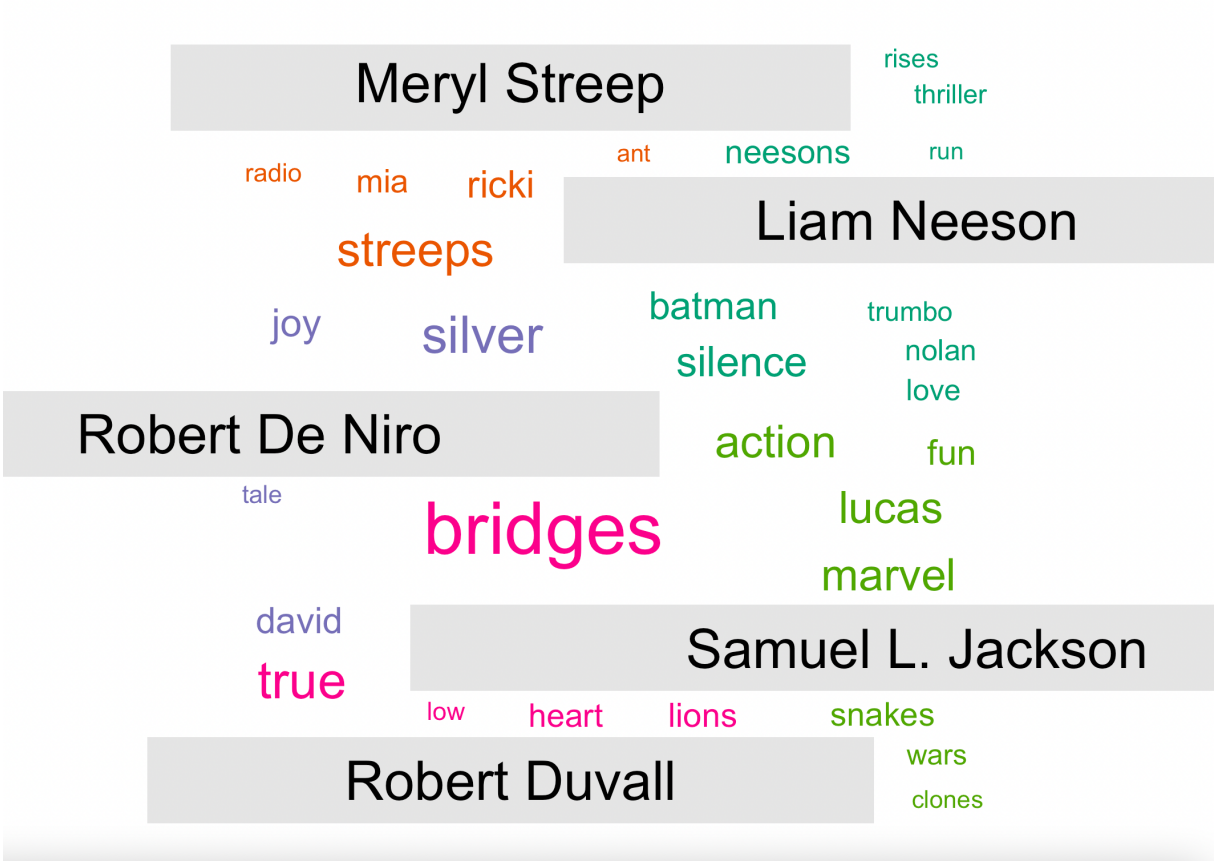

Figure 1

Figure 1 shows the result of the application of the bag of words algorithm.
The idea of the algorithm is simply to calculate the frequency of words in the reviews from rotten tomatoes, and the meaning of repeating a word in a large way with a specific actor name means that there is a close association between the two, and this was displayed in the map in proportion to the size of the word to indicate the extent of the association.

From the map we can infer the following ideas:

- Liam Neeson success is connected with the famous director Nolan.
- Liam Neeson is very connected to the batman movie.
- Liam Neeson is very proficient when it comes to thriller or action movies.

- Merl Streep is very well known for mama mia movie.
- The radio show of A Prairie Home Companion is a key point in Streep career and this is from the word radio.
- Ricki and the Flash, the drama movie got a lot of attention and the word Ricki repeated a lot associated with Merl Streep.

- Samuel L. Jackson is so connected with Marvel entertainment company.
- Samuel L. Jackson is associated to the audiences with Star Wars movies.

- A Bronx Tale is so repeated in the reviews of the audiences to indicate how this movie is so important in the career of Robert de Niro.

As a data scientist at a company like Netflix, I noticed from the previous analysis and the use of natural language processing that many ideas can be drawn, especially. For example, it is possible to reproduce movies or series similar to ones that succeeded.

It should be noted that it is not possible to depend entirely on the data without understanding the nature of the cinematic field. For example, if we look at the words that echo with the star de Niro, we will not find expected words such as God Father or the director Tarantino which means that the results, of course, do not reflect reality. So, the reasons must be studied because there is a problem in the data or in its processing.

In the next part I will apply Network Analysis on the data that I referred to in the introduction (rotten_tomatoes_movies.csv).

I have already previously reduced the data to make it pertain to the special movie stars only or those who have been classified as contributing strongly to the success of the movies. I created some figures representing each of the five actors and the content (genres) associated with this actor in most of his works, for example in the next figure I am visualizing the network analysis of Robert De Niro vs genres of his movies.

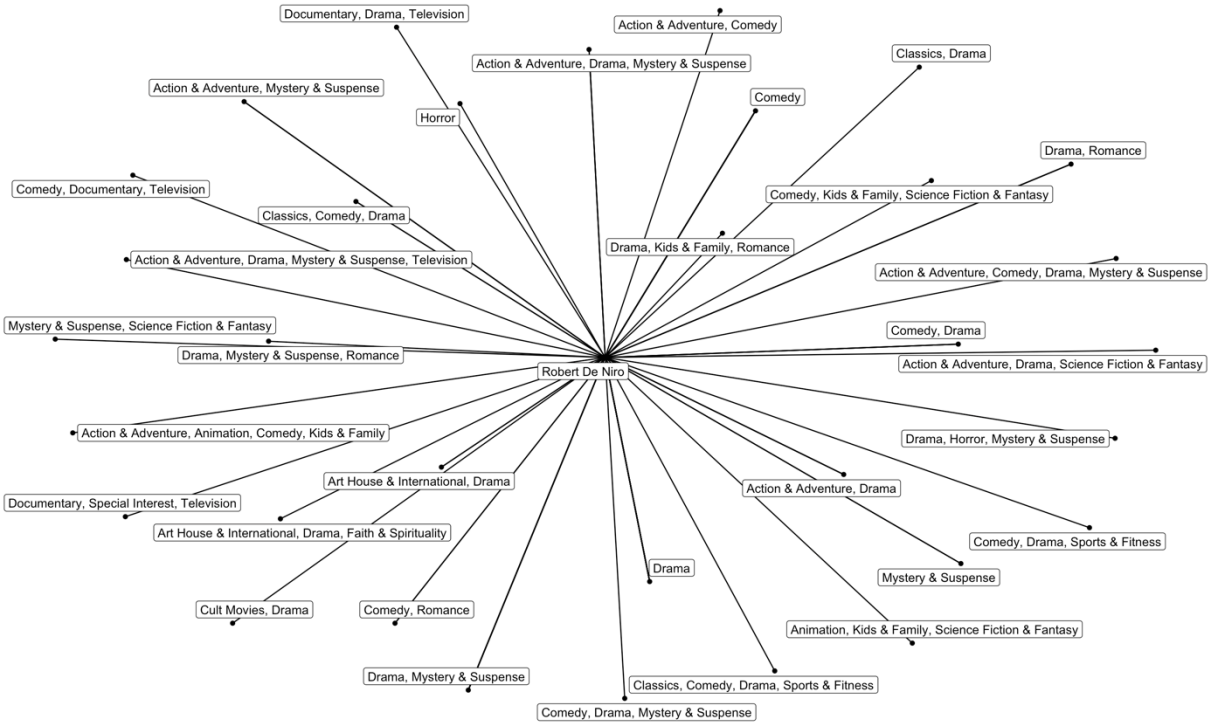

Figure 2

From the previous figure we can see that the dominant genre of De Niro is Drama but combined with many other categories like Adventure, Comedy, Romance and others which indicate how De Niro is really special because he proved success in many areas.

In the same manner in figure 3 I visualize Network Diagram for Robert Duvall.

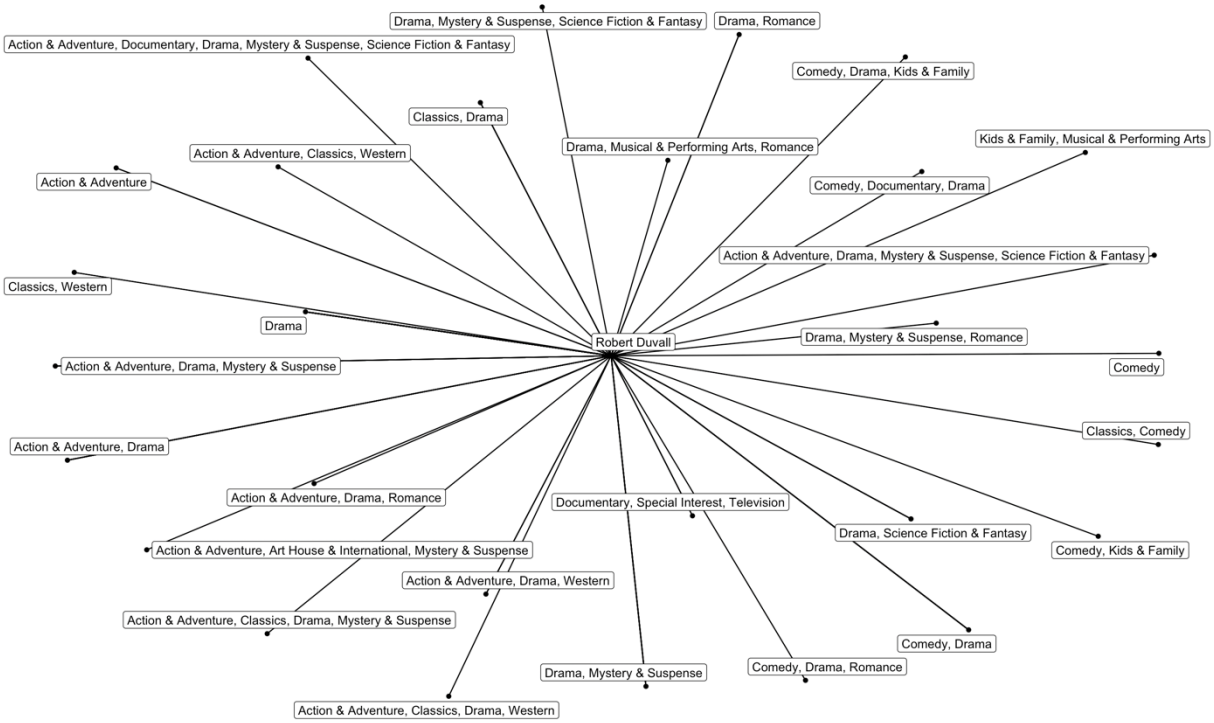

Figure 3

From figure 3 I can say Robert Duvall shares the same point as De Niro that he is proficient whatever the genre of the movie is which is intuitive that historical figures in cinema like him and De Niro are super stars in many kinds of movies and the success not just for certain director or production company but, their success is beyond that.

Part 4

Conclusion

I would say as data scientist in a company specialized in films and cinema, I must be very well acquainted with this field, or at least I have real sources of information and reference so that I can interpret some of the outputs of data analysis as we saw in natural Language processing part. The timestamp of dataset observations must be taken into account because they are important in interpreting what you see as outputs of the analysis, because what is required or trend today as movies may not be a trend tomorrow.
It is important, as a data scientist, to take into account the social and psychological dimensions in the data collection phase, if possible, or to try to remedy that during data processing before extracting outputs, maps, etc. For example, in the rotten tomatoes data that I was working on, the nationality of the participant or the age group was omitted, however that sometimes reveals the opinions of the audiences, and of course, the subject is much more

complex than that in terms of social science, and it may not be limited to age, nationality, political and social status, or his position on homosexuality, for example. Here it should be noted that data science does not give the desired results in light of not including all the inputs before processing.