★ Movie-stars - Success Factors and Recommender System

Introduction

The Internet has become more and more a platform for user-generated content. Especially the movie industry is interesting for many users and there are different platforms which offer exchange of information about movies or rating systems, e.g. IMDb. The aim of this project is to use this user generated-content to build a recommender system and to figure out which factors lead to a good user rating.

Data

Our project is based on datasets from 2 sources. The first part is gathered from kaggle-website. It contains metadata about movies, actors, genres, keywords and single user ratings. The size is **about 900 MB** and it has over **45.000** movies. Additionally we collected data from IMDb and got rating data with more than **45.000** entries.

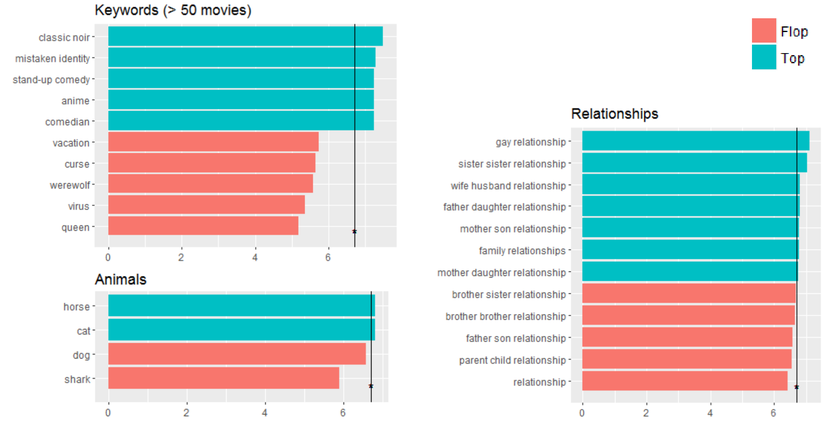
All in one, our dataset is an ensemble of data collected from **TMDB**, **IMDb** and **GroupLens**.

Success factors

To figure out which factors might implicate good or outstanding user ratings, first of all the mean rating of the different rating systems had to be calculated for each area such as keywords, actors etc.. Therefore the average rating from all movies where e.g. a single keyword appears was calculated. Afterwards the mean rating was used to create a ranking of top 5 and worse 5 ratings.

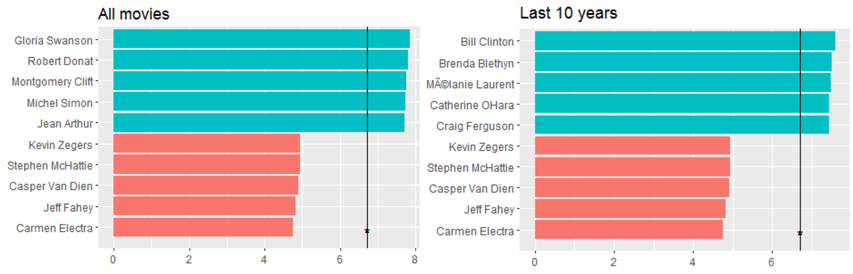
Keywords

For the evaluation, we just took keywords, which occurs in more than 50 movies. We also used them to generate a top/flop ranking of relationships and animals.



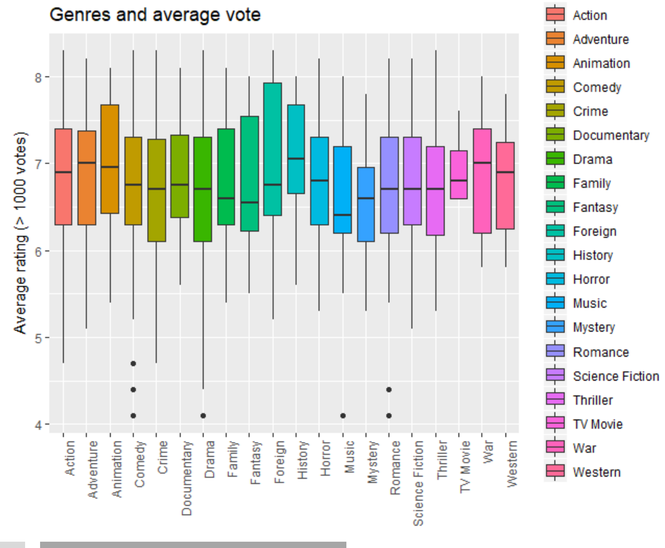
Actor/Actress

For determining top actors and actresses, we took these, who appear in more than 10 movies. The top 5 actors of all movies had already been dead for several years. That's the reason we made another ranking considering just movies of the last 10 years.



Genre

There are 20 genres in total inside the dataset. For this analysis we just used movies with over 1000 votes to get a more evaluable result. The following plot shows, that the median of history movies is the highest and the median of the music genre the lowest. Foreign movies have a wide range which reaches high average ratings, but the median is quite low.



Conclusion

Movies containing the keyword "classic noir" got the best mean rating result, which is reinforced by the fact that all top 5 actors are already dead. Movies with horses and cats got outstanding rating results, while movies with dogs and sharks got under-average ratings. Gay and sister-sister relationships might be more interesting for users than parent-child relationships.

Recommender system

We have built 5 separate recommender systems, which use different parts from our data. They base on self declared distance measures. It's possible to find more information about each system below. There's also one additional system, which connects all the other systems and it should give the best recommendations in general.

Most of the systems are using information, how you have persnonally rated seen movies. In the right you can see sample movie ratings that are used by the systems. Below you can also see, systems's recommendations.

title	rating
The Dark Knight	10
Forrest Gump	6
Star Wars	4
The Terminator	8
Minions	9
I Love You, Man	7
The Hunger Games: Mockingjay - Part 1	10

System 1 – It's the only system that isn't using your ratings. It recommends movies that have recieved good ratings from many people, many critics, many reviews and are popular in general.

title	prediction_score
The Dark Knight	9.359429
The Shawshank Redemption	9.110090
Pulp Fiction	8.770164
Inception	8.768091
Fight Club	8.746026

System 2 – It's recommending movies based on actors. It gives higher scores to the movies that have actors, whose movies you have liked before.

title	prediction_score
The Hunger Games	2.0
The Hunger Games: Catching Fire	2.0
The Hunger Games: Mockingjay - Part 2	2.0
Batman Begins	1.5
The Dark Knight Rises	1.5

System 3 – It's recommending movies based on genres. Movies with genres that you have given better ratings than other genres will come up in the top with this system.

title	prediction_score
The Tie That Binds	5
Dream Man	5
The Convent	5
Baton Rouge	5
The Innocent Sleep	5

System 4 – it's recommending movies based on it's story. In the data, there are avilable keywords for every movie and this system recommends movies with similar keywords that have been liked by the user before.

title	prediction_score
The Tie That Binds	5
Dream Man	5
The Convent	5
Baton Rouge	5
The Innocent Sleep	5

System 5 – This recommender system will find other people who have been rating movies in commn way like you and then recommends movies that the similar persons have liked.

title	predicton_score
The Million Dollar Hotel	507.75
Terminator 3: Rise of the Machines	491.20
Once Were Warriors	441.60
Sleepless in Seattle	418.60
The 39 Steps	410.95

General system uses all previous systems with self defined weights for every systems's score.

title	score
Terminator 3: Rise of the Machines	14.96198
The Dark Knight Rises	14.71908
Batman Begins	14.53097
The Hunger Games: Catching Fire	14.11390
Once Were Warriors	14.09601
The Million Dollar Hotel	13.96342
The Hunger Games	13.55163
The Hunger Games: Mockingjay - Part 2	13.52493

Marielle Egert, Andreas Bau *https://github.com/mariellee/datamining2017*