RECOMMENDATION SYSTEMS BASED ON MOVIES DATASETS ANALYSIS

Jakub Blaszczyk | Adam Dlubak | Aleksandra Orzechowska

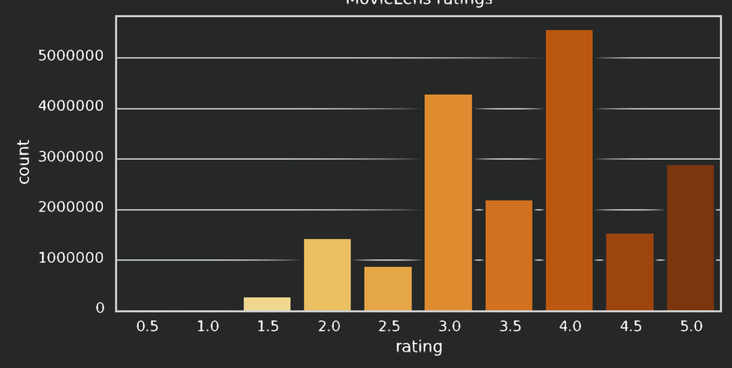
Wrocław University of Science and Technology

Movies are one of the strongest factor connecting people in social media. Our purpose was investigating two aspects of this phenomenon: **past and future model evolution.** We based our project on Full MovieLens Dataset obtained from Kaggle.com platform. For detailed analysis it was augmented by IMDB and TMDB sets, like also by movies' meta-informations. Raw dataset contains **26** million ratings from **270'000** users for over **45'000** described movies.

MAIN INSIGHT FROM DATA ANALYSIS

First approach was overall analysis which shows, that users give positive ratings definitely more frequent than negative: mean rating in range 0-5 with 0.5 step was **3.526.** General ratings distributions is shown below. This observation concluded with thesis, that users rate unfairly and they are more willing to recommend some production to others than discourage below the average. The mean of averages ratings per user was **3.627** in the same scoring 0-5.

MovieLens ratings



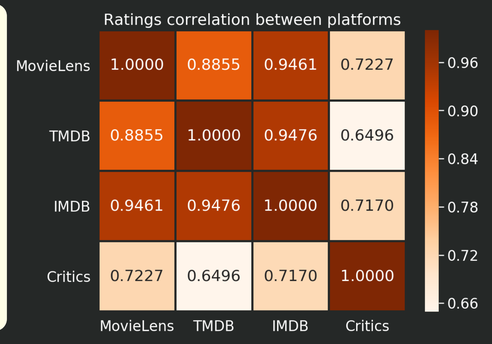
The most popular genres were **dramas** with **13340** tags and **comedies** with **8369** tags. Genre with the lowest number of tags was animations as the main watchers were children who are not the main target.

USERS ARE MORE WILLING TO RECOMMEND THAN DISCOURAGE

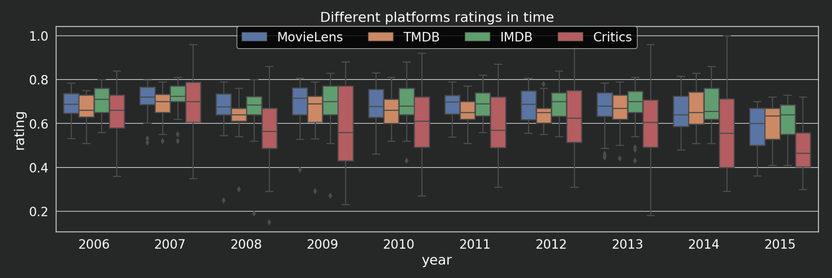
Most popular tags

Behaviors among platform are highly correlated, which gives opportunity to transform models pre-trained on one platform to be useable on a different one.

Moreover, the most different ratings are granted by critics, which strengthens thesis, that users rate movies they like and omit bad ones while critics have to fairly rate all of them.



RATINGS AMONG DIFFERENT SOCIAL MEDIAS ARE CORRELATED WITH EACH OTHER



Analysis in time shows common trends among groups, e.g. lower scores were observed in 2015 in relation to previous years for all groups. It could be used for modeling influence of social media and predicting further evolution of the net.

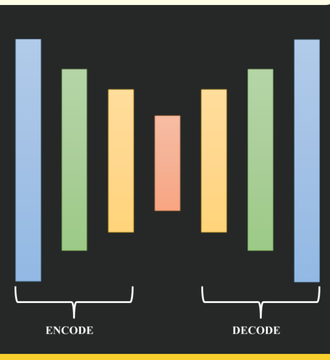
HOW THOSE INFORMATION CAN BE USED TO PREDICT (RECOMMEND) FURTHER MOVIES-NET EXPANSION?

We wanted to validate ours thesis about users' behavior and trends from analysis phase and check movies meta-informations similarities. To achieve that, we split recommendation system into 3 different approaches.

DEEP AUTOENCODER

The model is based on deep autoencoder with six layers. This approach is based on personalized recommendations that suggest movies to users using the **collaborative filtering**. User's interests are predicted by analysing others ratings and inferring similarities between them. The model consists of two transformations: encoder and decoder. The network was improved by applying dropout and output re-feeding. The final Root Mean Squarred Error was **0.92**.

AUTOENCODER CAN BE SUCCESSFULLY USED IN SYSTEMS BASED ON USERS' RATINGS SIMILARITIES



RECURRENT MODEL

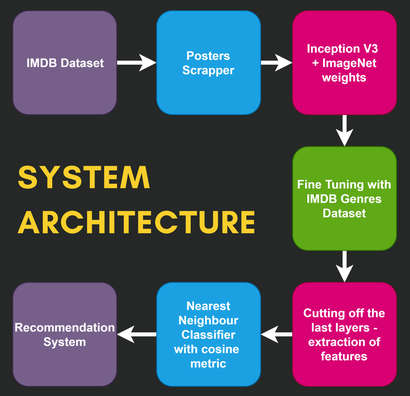
Second approach is intended to learn trends from user's movies history. Each movie - represented by 300 length vector - is its embedded description. Model - based on multiple GRU Cells and Residual Cell with fully connected layer at the top - generates 300 length vector as an output.

Model is fed with sequence of 5 vectors of positively rated movies in chronological order. An output is „idealistic" representation of recommended movie, which is unfolded to 5 nearest movies (known and not watched yet). The final system accuracy was low: **56.02%** for 49% verifiable ones.

DESCRIPTIONS ARE VARYING AMONG SIMILAR MOVIES AND THEY ARE NOT EFFICIENT EMBEDDING OF MOVIE

POSTERS RECOMMENDATIONS

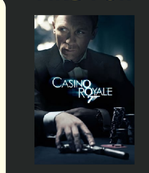
SYSTEM ARCHITECTURE



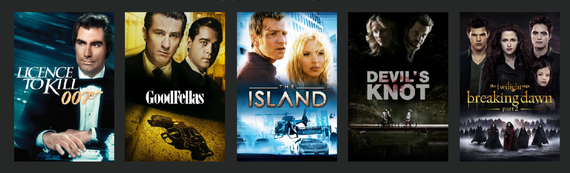
Due to use of transfer learning and fine tuning methods with the use of genre recognition while training phase, the system is based primarily on similarities of colouring scheme, poster elements geometry and movie genre.

POSTERS SIMILARITIES CORRESPONDS WITH MOVIES SIMILARITIES

FIND FOR

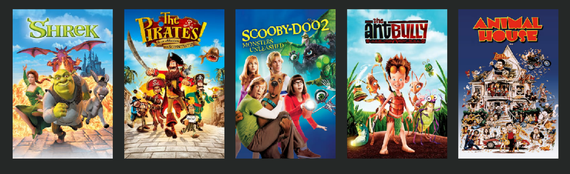


RECOMMENDATION



FIND FOR

RECOMMENDATION



Collaborative filtering and posters similarities can be used for predicting missing links of movies network in final recommendation system. Recurrent Neural Network approach requires exchanging movies vector embedding to less varying features set

Data Providers: kaggle, movielens, IMDb, POWERED BY THE MOVIE DB

Tools & Libraries: jupyter, TensorFlow, pandas, word_cloud, matplotlib, IMAGENET, NumPy, seaborn

Check our other project

Online version