Movie Recommendation System

IST 718: Big Data Analytics
Instructor: Prof. Daniel Acuna
Team Members: Adesh Gadge, Aniruddh Garge, Parshva Shah, Tanushree Shetty

THE iSCHOOL Syracuse University

1. Problem and Objective

The world of entertainment is blooming with movies being the chief source of entertainment in the modern world. The main aim of this project is to build a movie recommendation system that gives movie suggestions based on the kind of movies the users have watched and rated. The recommendation system will use 2 algorithms: ALS Matrix Factorization and KNN to recommend movies to users based on various parameters such as tags, genres and ratings. This recommendation system could be widely used by streaming companies, TV channels and even production houses to understand the customer's data, increase in revenue, business growth and the user experience.

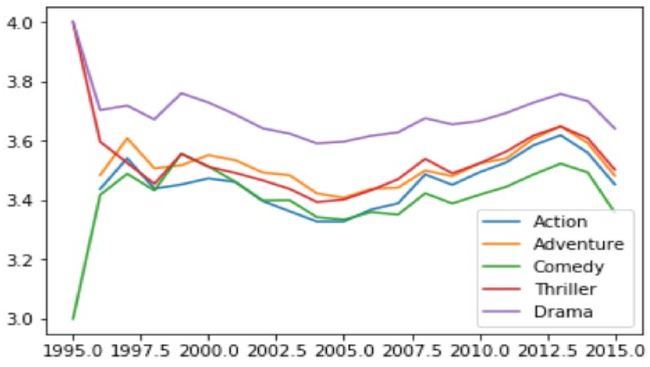
2. Goals

1. Recommend movies to a user based on tags of movies (Similarity Distance Measure and PCA)
2. Recommend movies to a user based on ratings (ALS Matrix Factorization)

3. Data Description

The dataset describes ratings and free-text tagging activities from MovieLens, a movie recommendation service. It contains 20M ratings and 465K tag applications across 27K movies. This data was created by 138K users between January 09, 1995 and March 31, 2015. This dataset was generated on October 17, 2016.

The data contains four files which consists of the following csv's: links.csv, movies.csv, ratings.csv and tags.csv. We only used the files movies.csv, ratings.csv and tags.csv to build the recommendation system.



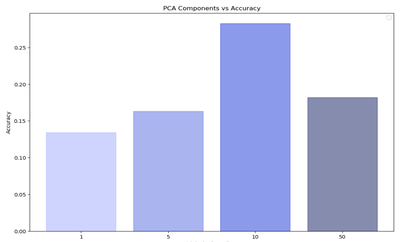
4. Model Description

Model	Features	Description	Evaluation
Distance-based Recommendation (PCA)	Tags, Tags-TFIDF, Ratings, MovieID, Movies, UserID	Recommends 5 movies from pca space when compared with the "movie" pca vector.	Accuracy*
ALS Matrix Factorization	Utility Matrix based on: • MovieID • Ratings • UserID	Recommends movies based on previous user rating data. we'll use utility matrix to fit our data and find similarities	RMSE

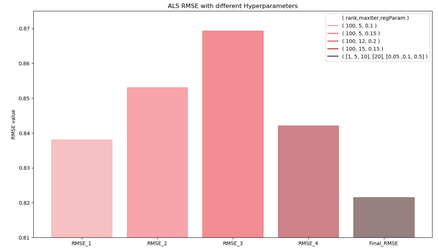
*Accuracy = $\frac{recommendations\ match\ with\ person's\ watched\ movies\ having\ \boldsymbol{high}\ rating}{recommendations\ match\ with\ person's\ watched\ movies\ having\ \boldsymbol{any}\ rating}$

5. Model Comparison

PCA



ALS Matrix Factorization



6. Performance and Interpretations

PCA Based Method

Small PCA Loading

Tag	PC1
Juliette Lewis	1.00947E-06
children	3.78264E-06
electronic soundtrack	4.35381E-06

Tag	PC2
food	2.48877E-07
DVD Collection	2.93517E-07
nihilism	5.72091E-07

Large PCA Loading

Tag	PC2
time travel	0.509965
Post-apocalyptic	0.363844
Kevin Spacey	0.346801

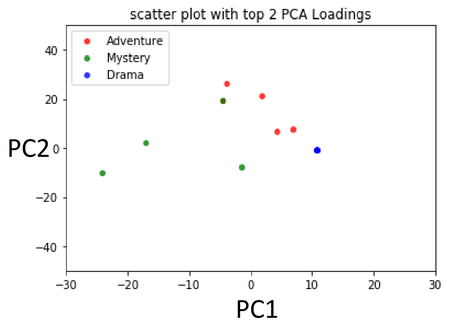
Tag	PC1
twist ending	0.602452
Kevin Spacey	0.346489
Brad Pitt	0.274587

Recommendation for Movie: **Seven**

MovieID	Title
50.0	The Usual Suspects
70.0	From Dusk Till Dawn
111.0	Taxi Driver
22.0	Copycat
29.0	The City of Lost Children

Movies recommended with distance based PCA method recommends movies which are similar to each other but fails at predicting if the person will like the movie, which could be seen through the recommendations and lower accuracy in the user interest prediction.

Genre Plot based on PCA values



ALS Matrix Factorization

Movies watched by a user

MovieID	UserID	Title
1	12	Toy Story
5	12	Father of the Bride
6	12	Heat
7	12	Sabrina
17	12	Sense and Sensibility

Movie Recommendations

MovieID	UserID	Prediction	Title
32	12	3.7293124	Twelve Monkeys
36	12	3.7077885	Dead Man Walking
34	12	3.4906259	Babe
380	12	3.381249	True Lies
590	12	3.3618577	Dances with Wolves

7. Conclusion

Recommender systems open new opportunities of retrieving personalized information on the Internet. We come up with a strategy that focuses on dealing with users' (tags) interests and based on previous reviews (ratings), movies are recommended to users. This strategy helps in improving accuracy of the recommendations making the system responsive.

By using such models, production houses and other entertainment businesses could learn about users' likes and recommend them movies accordingly. This would not only make the users stay online and watch more movies resulting in more sales, but also contribute in the dataset bringing about better understanding of users' choices.

Data Source: https://www.kaggle.com/grouplens/movielens-20mdataset