PROJECT:

Movie Recommendation System

Submitted by:

Ruchit Patel

# 

# Table of Content

1. [Table of Content 1](#_Toc532225162)
2. [Introduction 2](#_Toc532225163)
3. [Dataset Overview 2](#_Toc532225164)
4. [Goals 3](#_Toc532225165)
5. [Research Question 3](#_Toc532225166)
6. [Methodology 3](#_Toc532225167)
7. [Data Processing 3](#_Toc532225168)

* [Single Value Decomposition 4](#_Toc532225169)
* [Predictive Modelling 4](#_Toc532225170)
* [Sentiment Analysis 4](#_Toc532225171)
* [Text Mining 5](#_Toc532225172)

1. [Visualizations and Conclusions 5](#_Toc532225173)
2. [DataCamp Certification 8](#_Toc532225174)
3. [Key Personnel 9](#_Toc532225175)
4. [References 9](#_Toc532225176)

# Introduction

The project is programmed in python and it will deliver recommendation and classification of movies for any existing customer based on the standard training dataset. The effort is involved by utilizing machine learning techniques, algorithms to build the movie recommendation system by using single value decomposition method and data processing by help of packages of Python like NumPy and Pandas. The project is visualized with the help of seaborn library. The data set for training and testing is taken from Movielens and purely based on user input (explicit user ratings of movies).

The recommendation engines are based on Content based and Collaborative Filtering methodology techniques. This methodology follows two different approach to pick movies for the user one is based on user search term and another recommend based on user past movie list and try to group alike users to provide each other recommendation list. The model is not limited to movie recommendation it can also works best for the online shopping companies, books recommendation and songs recommendation with little modification. The project is built on Python and the money spent on designing of the project is $0.

# Dataset Overview

The dataset taken to build this project is from the Movie lens dataset available on [MovieLens](https://grouplens.org/datasets/movielens/) website, for building a recommendation system. It is a stable benchmark dataset with 1 million ratings from 6000 users on 4000 movies and Released 2/2003. The data set contains variety of categorical and qualitative data set in different DAT file format. The data preprocessing part also involves converting the DAT file in CSV format with properly aligned column names.

The data set contains movies, users and records specific information in separate “DAT” files from the various sources and has been combine and processed in different ‘.csv’ files. The dataset ratings are measured on scale of 1-5, movie ids in numbers, movies in alphabets and numbers, and tags in alphabets. The table below shows the detail description of each CSV files.

Table 1: DATA TABLE DESCRIPTION

|  |  |  |  |
| --- | --- | --- | --- |
| CSV file | Cell size | File information | Columns information |
| movies.csv | 27.3k \* 3 | Contains types of information on movies, like assigned movie id number, movie title and type of movie (drama/animation/horror/ etc.) | Movie ID, Title, Genres |
| ratings.csv | 1.05m\* 4 | This file contains movie ratings from 7.12k user’s. | User Id, Movie Id, Rating, Time Stamp |
| users.csv | 466k\*4 | This file contains different tags (Bollywood/Chinese/Jesus) for each movie ids and have the count of number of hits based on movies and tags by the users | User Id, Movie Id, Tag, Time stamp |

# Goals

The goal of this project is to build a movie recommendation system using Python. The project will help each user by providing the proper suggestions of movie content based on their liking and genre. It will be productive for the content provider as it will keep viewers engaged and pull up total viewership time and hence revenue for the company.

# Research Question

1. Inspecting how different data tables (movies, ratings, and users) look like for MovieLens?
2. Testing visualization of movie titles using word cloud library and how well it fit?
3. Statistical analysis of ratings for evaluating mean, median, quartiles of explicit ratings.
4. What is the most active user group age and distribution analysis of age group?
5. What type of movie genres (thriller/comedy/horror/ thriller etc.) are available, how many movies are classified under each genre equally?
6. What is the most popular rated genre in the movie collection? What is the frequency of most popular genre?
7. What are some of the top-rated movies by user 420?
8. What are some of the movies that user 420 will like?

# Methodology

## Data Processing

The data is downloaded from MovieLens website for 1 Million sizes. The data existed in DAT format which is processed and converted in CSV format for reading in dataframe for performing operation.

The data is saved in three different files namely users, ratings, movies containing related relevant details. The user is assigned individual unique rating. All the files are assigned latin-1 encoding and assigned as comma separated files.

For creating word cloud the movies table is used and all the movie title columns are made space free to represent perfectly in figure while using word cloud.

The data is later blended in one data frame df based on movie, rating and associating user over top of it. The data on arranging in descending order will provide insight on top rated movies.

The data in movie table holds multiple genre for single movie, to classify this the data preprocessing split data in individual genre and counted it uniquely so as to classify each movie under different genre separately. The data is processed to sparse matrix by pivoting user, movie.

### Single Value Decomposition

The single value decomposition is matrix factorization method into 3 matrices, which help to reduce the big data set into small data set without losing its importance (Luboobi, 2018).

So, on movie recommendation let’s assume one user have matrix 0 to 5, in which 0 mean they don’t like that movie and 5 mean they like a lot. So, this metrics store the rating and further it will factorize each rating and score each genre with the weight of each user ratings. Further the different matrixes of user rating and compare with the other user to find the similarity with the other users. It calculates matrixes and match the similarity by comparing user’s score. At the end return the recommendation to each other with the movies, watched by another similar user.

### Predictive Modelling

Predictive modelling is process, which makes future assumption based on past records and probability. Majorly the prediction is made on predictors patterns, facts, correlations and trends (Momoh, 2018). All of this can be achieved by collecting data of all predictors that cause or effects of certain occurrence. Prediction modelling is usually designed using regression, Neural network for creating artificial intelligence, decision tree, time series and data mining techniques.

In this movie recommendation project, we have collected 7GB data from the user to build recommendation system by running statistical analysis to predict, what user will like to watch. To carry out prediction, we have used several techniques like single value decomposition, classification, collaborative and context-based modelling, where rating are the predictors and help us making the prediction of the movie that they will like to watch. The recommendation is based on complex statistical algorithm.

### Sentiment Analysis

Sentimental analysis is the computational identifying and classifying someone opinion or attitude towards something based on text and comment shared on online mention (brandwatch, 2016). Sentimental analysis is very helpful in building recommendation model because of it features. This analysis can help this project in many ways. Because with sentimental analyses, we can analyse, processing and summarizing the emotional text coming from the user over social media or anywhere.

They are both positive and many drawbacks of this analysis. This analysis in purely based on human tone and it gets harder to teach everything to the machine because limitation of languages, grammar errors by humans, slangs and differing cultural.

### Text Mining

Text mining is the technique and the process of collecting data from the documents and create those unstructured data in structured data. This helps in retrieving the facts, relationship, and correlation between things. For example, this mining call help company in gaining the knowledge about any movie reviews and ongoing trends by analysing user comments review and people sharing information about the specific on the internet.

# Visualizations and Conclusions

1. Inspecting how different data tables (movies, ratings, and users) look like for MovieLens?

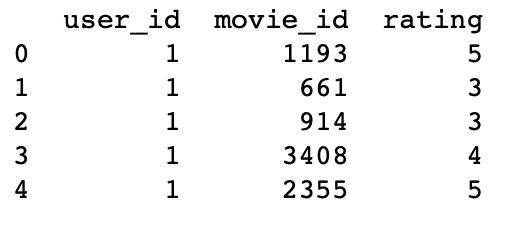


Figure 1: RATING TABLE DESCRIPTION

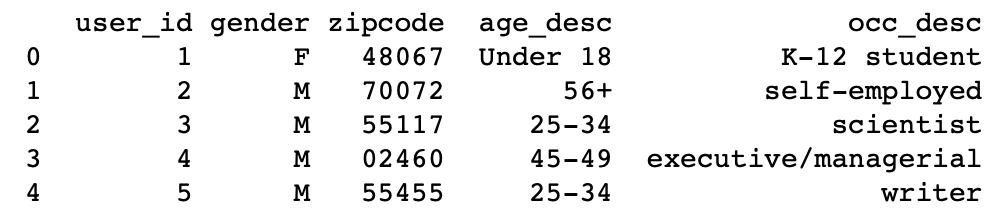


Figure 2: USER TABLE DESCRIPTION

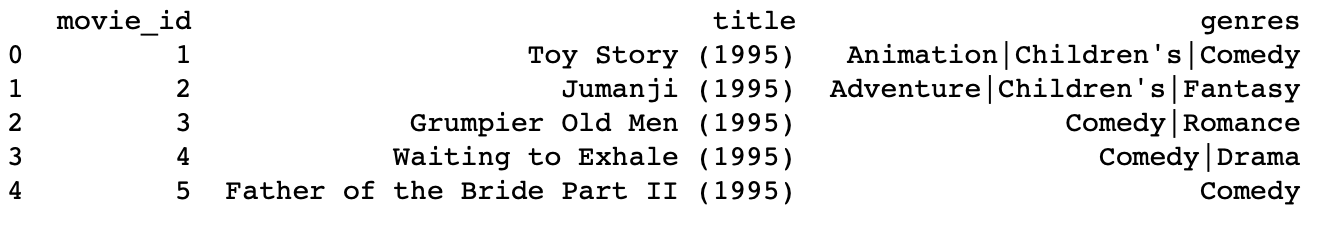


Figure 3: MOVIE TABLE DESCRIPTION

1. Testing visualization of movie titles using word cloud library and how well it fit?



Figure 4: WORDCLOUD FOR MOVIE TITLES

1. Statistical analysis of ratings for evaluating mean, median, quartiles of explicit ratings.

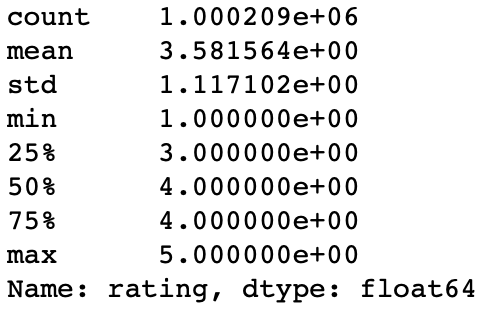


Figure 5: RATING STARS STATISTICAL ANALYSIS

1. What is the most active gender on user list of movie recommendation system?

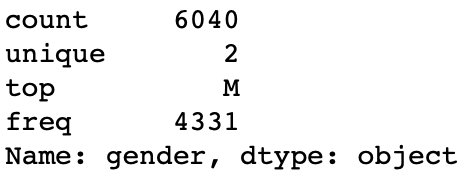


Figure 6: STATISTICAL ANALYSIS OF GENDER WITH MALE BEING MORE FREQUENT USER

1. What is the most active user group age and distribution analysis of age group?

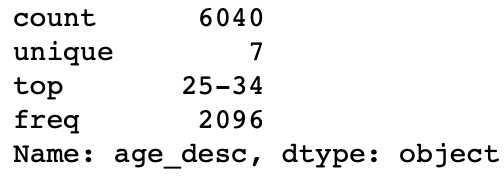


Figure 7: AGE ANALYSIS OF USER WITH 25-34 BEING MOST ACTIVE ONLINE USER GROUP

1. Statistical analysis of genres of movies for analysis of variety of genres, top rated genre and it’s frequency.

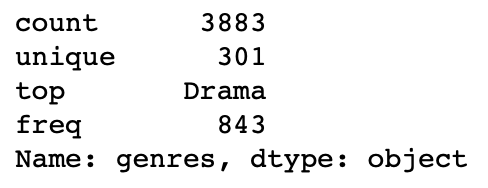


Figure 8: Drama is the most frequent genre

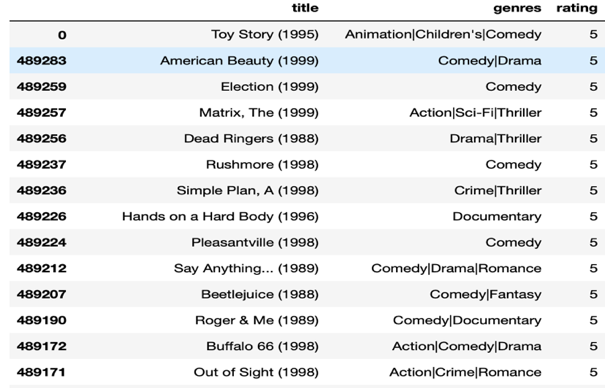


Figure 9: top 5 star rated movies

1. Which type of my movie genres (thriller/comedy/horror/ thriller etc.) are available, how many movies are classified under each genre equally?

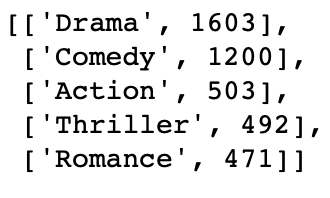


Figure 10: Genre and frequency

1. What are some of the Top recommended movies by user 420?

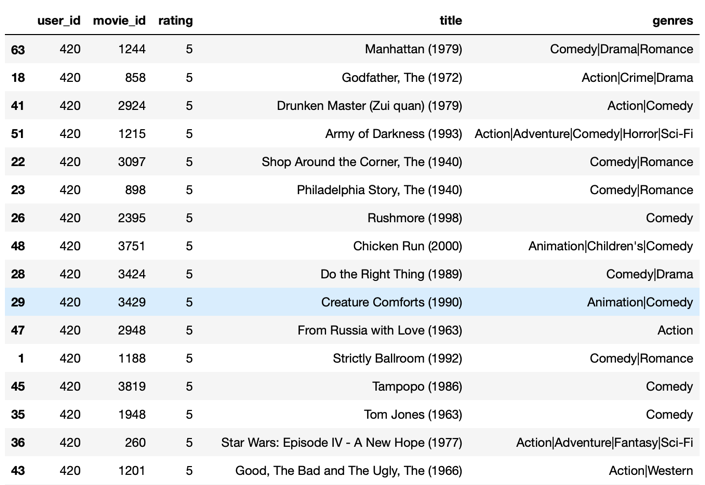


Figure 11: Top rated movies by user 420

1. What are some of the Top recommended movies for user 420?

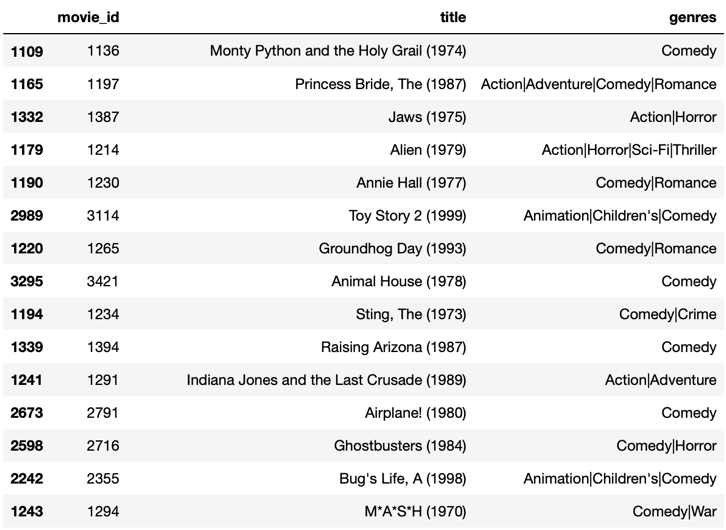


Figure 12: Top recommended movies for user 420 based on explicit ratings

## DataCamp Certification

The project was completed by following data camp coursework which awarded certification to the team once the course was completed. Both of the team members participated for completing and understanding the recommendation system using the Datacamp course work



## Key Personnel

For the entire project the work is divided equally between the team members

|  |  |
| --- | --- |
| Anupam Maheshwari | Data Preprocessing, Classification, Visualization |
| Ruchit Patel | Data Collection, Collaborative filtering, Documentation |
|  | | |

## References

Banik, R. (2017). The Movies Dataset. Retrieved November 10, 2018, from <https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv>

Brandwatch. (2018, February 26). Sentiment Analysis: How Does It Work? Why Should We Use It? Retrieved from <https://www.brandwatch.com/blog/understanding-sentiment-analysis/>

Cui, Bei-Bei. (2017). Design and Implementation of Movie Recommendation System Based on Knn Collaborative Filtering Algorithm. ITM Web of Conferences.

Elmoudni, S. (n.d.). Recommender System using Singular Value Decomposition. Retrieved December 9, 2018, from <http://rstudio-pubs-static.s3.amazonaws.com/335300_11d40bf12d8940f78d9661b3c63150dc.html>

<https://www.researchgate.net/publication/319487277_Design_and_Implementation_of_Movie_Recommendation_System_Based_on_Knn_Collaborative_Filtering_Algorithm>

Koren, “The BellKor solution to the Netflix grand prize,” [www.netflixprize.com/assets/GrandPrize2009 BPC BellKor.pd](http://www.netflixprize.com/assets/GrandPrize2009%20BPC%20BellKor.pd)

Linguamatics. (2018, November 28). What is NLP Text Mining?. Retrieved from <https://www.linguamatics.com/what-is-text-mining-nlp-machine-learning>

Luboobi, P. (2018, February 5). Foundations of Machine Learning : Singular Value Decomposition (SVD). Retrieved from <https://medium.com/the-andela-way/foundations-of-machine-learning-singular-value-decomposition-svd-162ac796c27d>

Marek Rei. (2014, October 18). MULTILINGUAL SEMANTIC MODELS [graph]. Retrieved from <http://www.marekrei.com/blog/multilingual-semantic-models/>

Momoh, O. (2016, October 30). Predictive Modeling. Retrieved from <https://www.investopedia.com/terms/p/predictive-modeling.asp>

Toscher, M. Jahrer, and R. Bell, “The BigChaos solution to the Netflix grand prize,”[www.netflixprize.com/assets/GrandPrize2009 BPC BigChaos.pdf](http://www.netflixprize.com/assets/GrandPrize2009%20BPC%20BigChaos.pdf)

Wang, Y. (2018, March 22). A Sentiment-Enhanced Hybrid Recommender System for Movie Recommendation: A Big Data Analytics Framework. Retrieved from <https://www.hindawi.com/journals/wcmc/2018/8263704/>