 **BIRSA INSTITUTE**

**OF TECHNOLOGY,**

**SINDRI**

MACHINE LEARNING WITH PYTHON

**NETFLIX-MOVIE RECOMMENDATION**

TRAINING PROJECT REPORT

NAME : UTKARSH RAJ

ROLL NO. :1709036

B.TECH IN COMPUTER SCIENCE AND ENGINEERING

**CANDIDATE’S DECLARATION**

I hereby declare that I have undertaken industrial training at “WEBTEK LABS” during a period from 7 JUNE to 22 JULY in partial fulfilment of requirements for the award of degree of B.TECH(COMPUTER SCIENCE AND ENGINEERING ) at BIRSA INSTITUTE OF TECHNOLOGY,SINDRI. The work which is being presented in the training report submitted to Department of COMPUTER SCIENCE AND ENGINEERING at BIRSA INSTITUTE OF TECHNOLOGY,SINDRI is an authentic record of training work.

Signature of the students

**CERTIFICATE OF APPROVAL**

The project “**NETFLIX-MOVIE RECCOMMENDATION**” made UTKARSH RAJ is hereby approved as a creditable study for the Bachelor of Technology COMPUTER SCIENCE AND ENGINEERING and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned this project only for the purpose for which it is submitted.

MS. MOUSITA DHAR

(Project In-charge)

**ACKNOWLEDGEMENT**

WE would like to express our special thanks of gratitude to our trainer MS.MOUSITA DHAR as well as our HOD MR.CHANDAN BANERJEE who gave us the golden opportunity to do this wonderful project on the topic NETFLIX-MOVIE RECOMMENDATION BY MACHINE LEARNING USING PYTHON which also helped us in doing a lot of Research and we came to know about so many new things. We are really thankful to them.Finally, we would also like to thank our parents and friends who helped us a lot in finalizing this project within the limited time frame.

**1.INTRODUCTION**

**1.1 PYTHON**

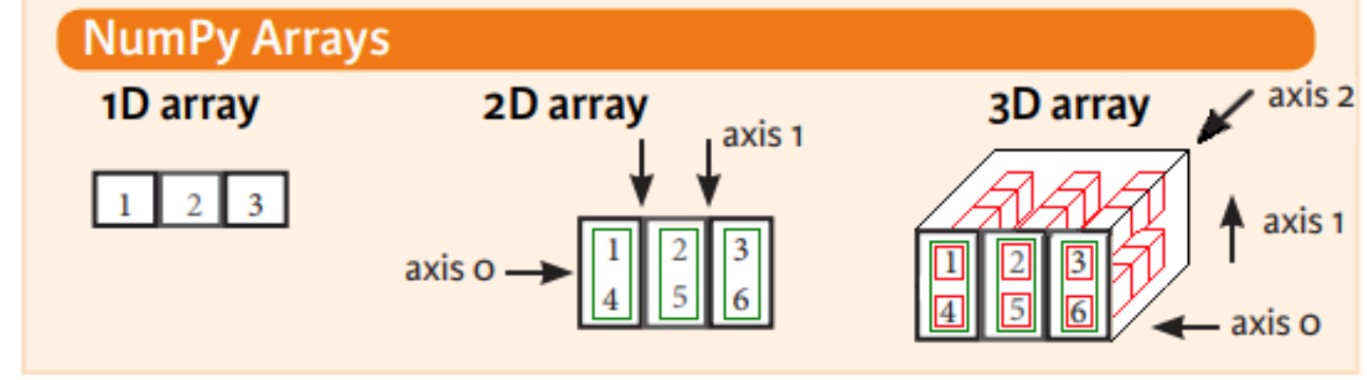
Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

* **Python is Interpreted** − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* **Python is Interactive** − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
* **Python is Object-Oriented** − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
* **Python is a Beginner's Language** − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.
* **PYTHON FEATURES**
* **Easy** to Learn and Use. Python is **easy** to learn and use
* Expressive Language
* Interpreted Language
* Cross-**platform** Language
* Free and Open Source
* **Object-Oriented** Language
* Extensible
* Large Standard Library
* **APPLICATIONS OF PYTHON**
* Web and internet develpoment
* Scientific and numeric computing
* Data Analysis
* Desktop GUIs
* Machine Learning
* Data visualization
* Game Deveploment
* Software Develpoment
* Business Application
  1. **ANACONDA**

**Anaconda** is a free and open distribution of Python programming languages for data science and machine learning related applications (large-scale data processing, predictive analytics, scientific computing), that aims to simplify package management and deployment. Package versions are managed by the package management system *conda* . Conda is an open source, cross platform, language-agnostic package manager and environment management system that installs, runs, and updates packages and their dependencies.The Anaconda distribution is used by over 6 million users, and it includes more than 250 popular data science packages suitable for Windows, Linux, and MacOS.

* 1. **PYTHON PACKAGES**
* **NUMPY**
* NumPy is the fundamental package for scientific computing with Python. It contains among other things:
* a powerful N-dimensional array object
* sophisticated (broadcasting) functions
* tools for integrating C/C++ and Fortran code
* useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

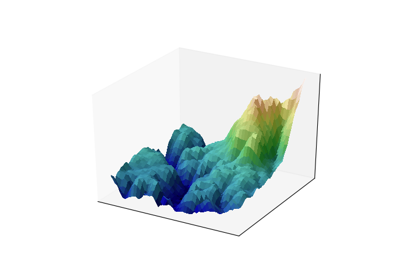


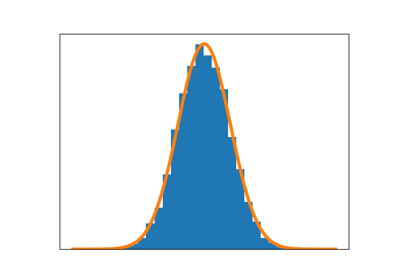
* **Matplotlib**

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when

combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.





* **Pandas**

*Pandas* is an open source, BSD-licensed library providing high-performance, easy- to-use data structures and data analysis tools for the *Python* programming language. Pandas library is well suited for data manipulation and analysis using python. In particular, it offers data structures and operations for manipulating numerical tables and time series.

* **Scikit-learn**

Scikit-learn provides machine learning libraries for python some of the features of Scikit- learn includes:

* Simple and efficient tools for data mining and data analysis
* Accessible to everybody, and reusable in various contexts
* Built on NumPy, SciPy, and matplotlib
* Open source, commercially usable - BSD license

**TRAINING WORK UNDERTAKEN**

* **COLLECTING DATA FROM KAGGLE**

**Kaggle** is a platform for predictive modelling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowd sourcing approach relies on the fact that there are countless strategies that can be applied to any predictive modelling task and it is impossible to know beforehand which technique or analyst will be

most effective.

On 8 March 2017, Google announced that they were acquiring Kaggle. They will join the Google Cloud team and continue to be a distinct brand. In January 2018, Booz Allen and Kaggle launched Data Science Bowl, a machine learning competition to analyse cell images and identify nuclei.

* **DATA SCIENCE**

**Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

When Harvard Business Review called it "The Sexiest Job of the 21st Century “the term became a buzzword, and is now often applied to business analytics, business intelligence, predictive modelling, or any arbitrary use of data, or used as a glamorized term for statistics. In many cases, earlier approaches and solutions are now simply rebranded as

"data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness." While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents. Because of the current popularity

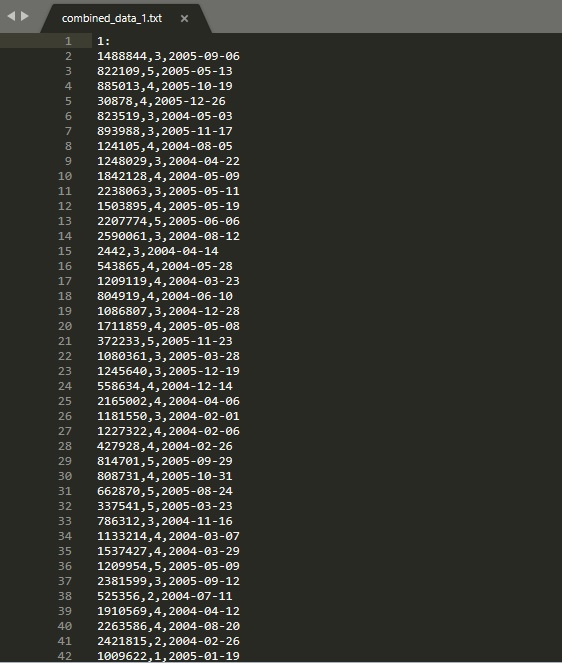
of this term, there are many "advocacy efforts" surrounding the field. To its discredit, however, many data science and big data projects fail to deliver useful results, often as a result of poor management and utilization of resources.

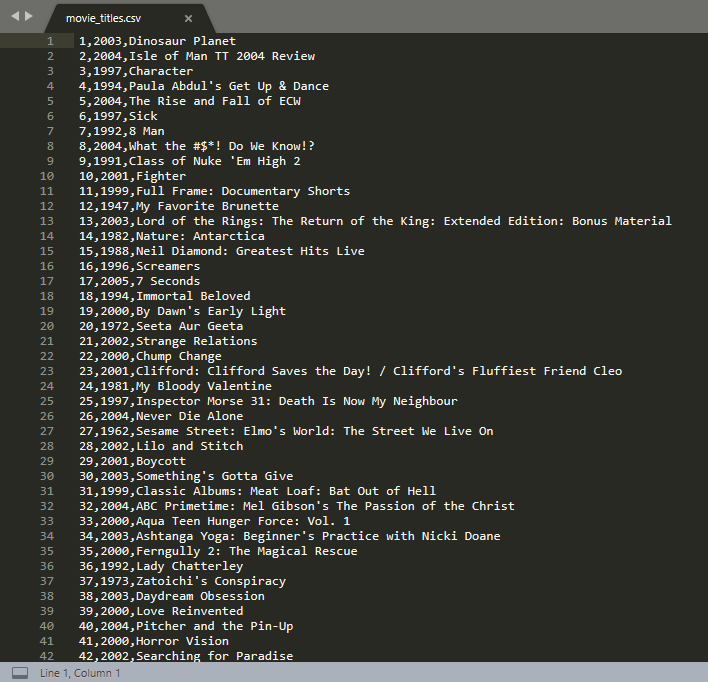
* **DATASET**

**This project contains 2 Dataset**

1st dataset contains 24058263 rows and 2 columns

2nd dataset contains 17770 rows and 3columns





* **SOURCE CODE & OUTPUT**

**1 Import Modules**

import numpy as np

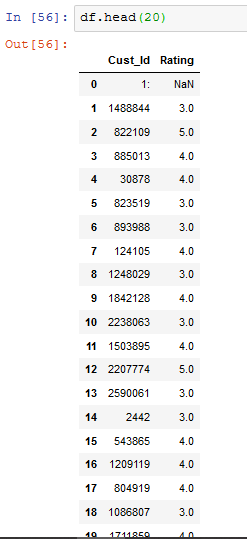
import pandas as pd

2 **Load the data**

dataset=pd.read\_csv(r"C:\Users\Raj\Desktop\NF Dataset\Machine-Learning\combined\_data\_1.txt",names = ['Cust\_Id', 'Rating'])

print(dataset.shape)

(24058263, 2)



* In this project, the User Id was under each movie id followed by NaN value.
* To use the Movie Id along with customer id and their respective rating to the respective movie the movie id was to be added with the dataset alongside with the customer rating of the respective user
* This was too much time consuming to run a loop for each user id

So it was handled by numpy approach. An array of as the same size in which to NaN value appears was created and appended with the dataset.

* The new column was created and the rows with NaN values were deleted.
* The missing values are removed by calculating the mean value.
* Then indexing is applies to all the columns.

1. **Algorithms used :**

* **SVD (Single Value Decomposition) or Collaborative Filtering**

**SVD** or **Single Value Decomposition** experimental conditions. The SVD theorem states:

**A*nxp*= U*nxn* S*nxp* VT*pxp***  Where,

**U**T**U** = **I**nxn

**V**T**V** = **I**pxp (i.e. U and V are orthogonal)

Where the columns of U are the left singular vectors (*gene coefficient vectors*); S (the same dimensions as *A*) has singular values and is diagonal (*mode amplitudes*); and VT has rows that are the right singular vectors (*expression level vectors*). The SVD represents an expansion of the original data in a coordinate system where the covariance matrix is diagonal.

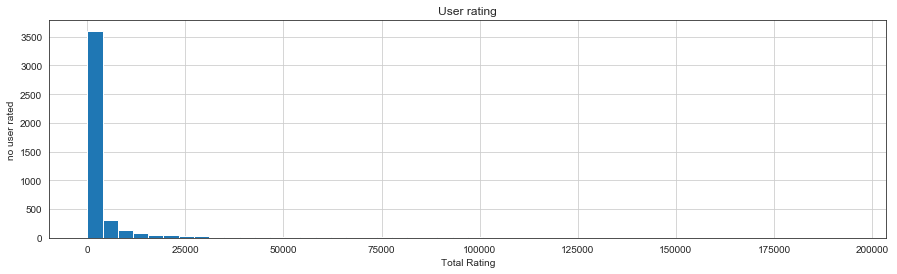
Calculating the SVD consists of finding the eigenvalues and eigenvectors of *AAT* and *ATA*. The eigenvectors of *ATA* make up the columns of *V* , the eigenvectors of *AAT* make up the columns of *U*. Also, the singular values in **S** are square roots of eigenvalues from *AAT* or *ATA*. The singular values are the diagonal entries of the *S* matrix and are arranged in descending order. The singular values are always real numbers. If the matrix *A* is a real matrix, then *U* and *V* are also real.

* **Pearsons' R correlation**

**Pearson’s correlation coefficient** is the test statistics that measures the statistical relationship, or association, between two continuous variables.  It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.  It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

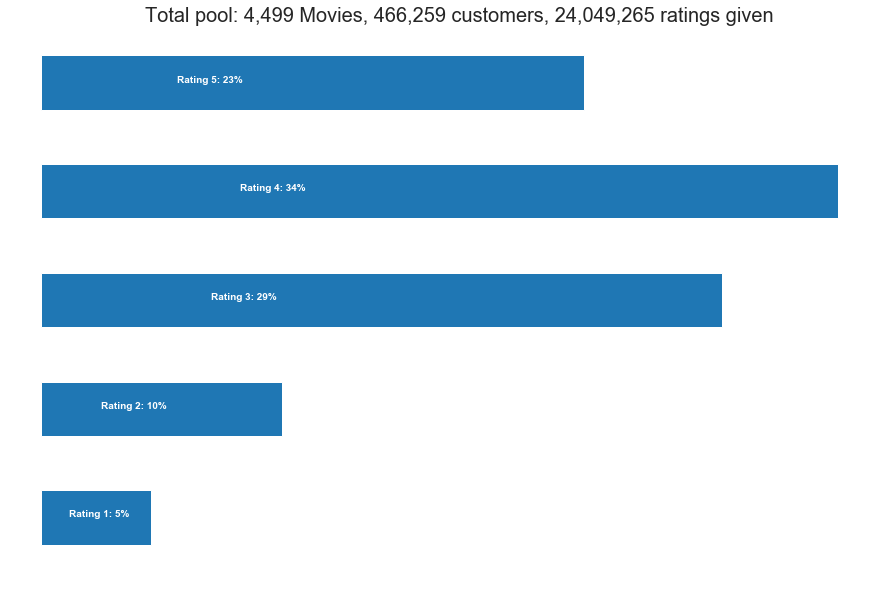
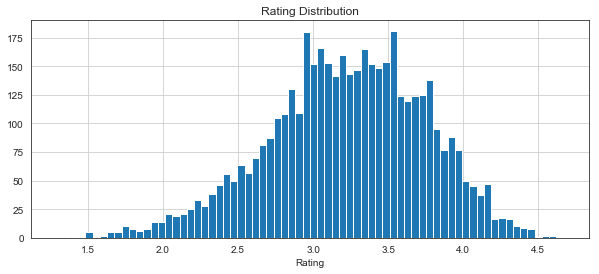
The way it works is we use Pearsons' R correlation to measure the linear correlation between review scores of all pairs of movies, then we provide the top 10 movies with highest correlations.

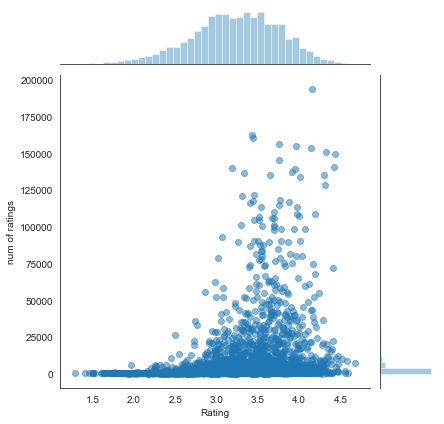
**Data Visualisation**



This graph shows that most user don’t tend to rate movies

This graph shows that most rating is between 3 to 4 This graph shows Rate ratio



****

This graph shows relation and between number of rating and rating value.

The obtained dataset was super huge. We can guess that the by applying following parameters we can significantly improve efficiency and reduce the dataset size with relatively important movies:

\* Remove movie with too less ratings (they are relatively not popular)

\* Remove customer who give too less ratings (they are relatively less active)

On experimenting a lot with the dataset we came across about the fact that the movies who got ratings less than 10K are relatively unpopular and can be removed from the working dataset.

* **RMSE and MAE**

RMSE stands for ROOT MEAN SQUARE ERROR and MAE stands for MEAN ABSOLUTE ERROR

Calculated RMSE and MAE

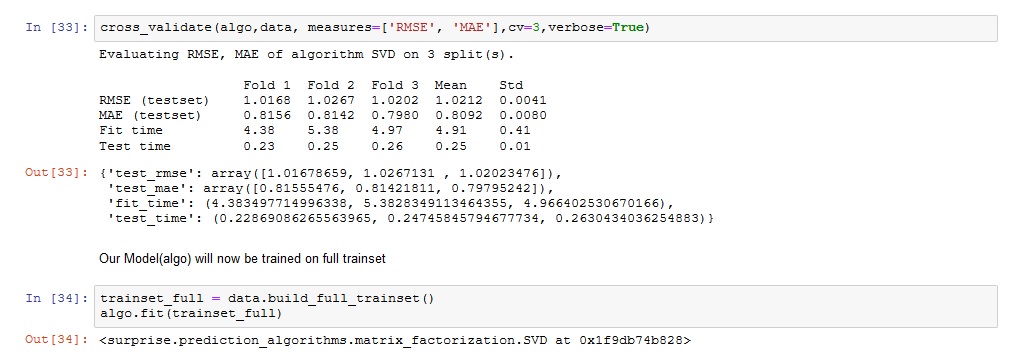
RMSE: 1.0255

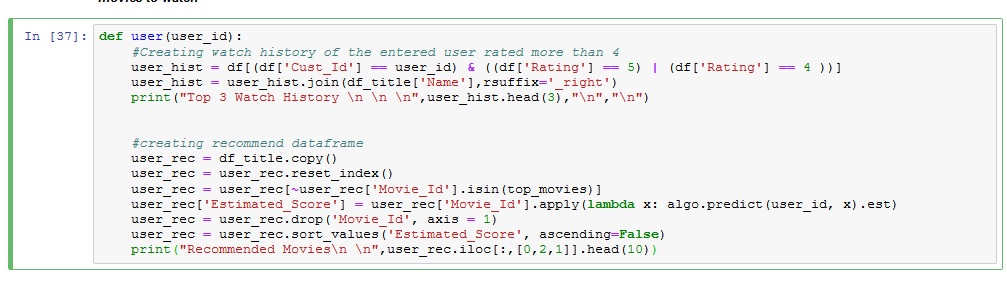
1.0254673813461568

MAE: 0.8245

0.8245087338717182

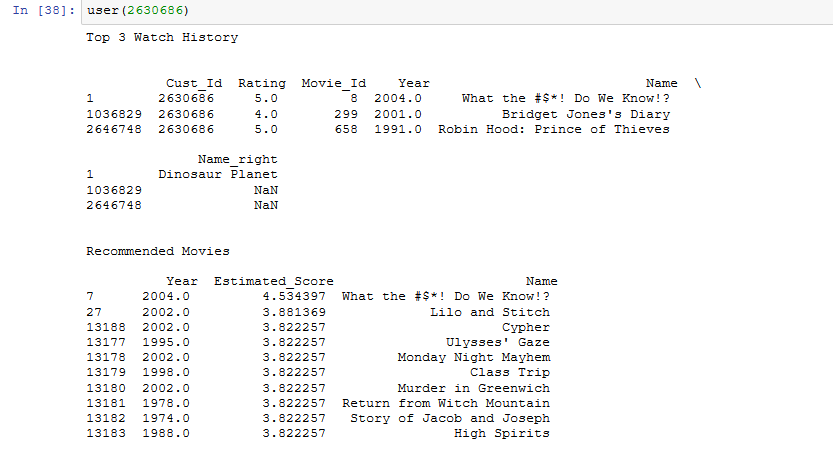
Since we cannot calculate the accuracy as the recommendation is unsupervised and hence we will choose the algorithm with lower RMSE and MAE value



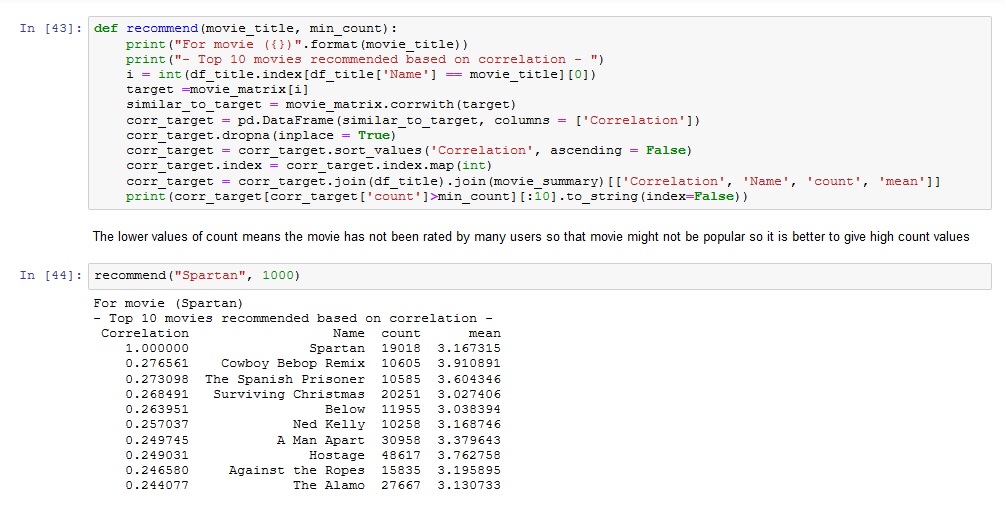


* **Prediction**

**COLLABORATIVE FILTERING (based on watch history of individual user)**: We define a function named user which will take user id as input and return the watch history and based on the watch history the model will predict the next movies to watch. Using SVD Algorithm

****

**Pearsons' R correlations (based on correlation with other Movie):** The way it works is we use Pearsons' R correlation to measure the linear correlation between review scores of all pairs of movies, then we provide the top 10 movies with highest correlations:

****

**RESULTS AND DISCUSSION**

* **RESULT**

By using SVD RMSE: 1.0255 & MAE: 0.824 error and Pearsons’ R correlation the chance of error is lower (as there is known way to check accuracy)

* **DISCUSSION**

Netflix is all about connecting people to the movies they love. To help customers find those movies. Recommending Movies is always a crucial task as one will like a movie or not is almost unpredictable but it can be guessed somehow as on the basis of the watch history and the rating given by the customer and comparing with other user who watched the same movie and liked other movie may bring us to the conclusion that this user may also like the same. Although a similar movie can be recommended by the users previously liked movies. Here comes our Pearsons’ R correlation method.

**CONCLUSION**

By using SVD and Pearsons’ R correlation we can recommend different movies on the basis of their watch history and recommend other similar movie.

We can say that our Movie recommendation in most cases will give a good movie to watch suggestion.

**REFERENCES**

https://www.python.org/

https://anaconda.org/anaconda/python

http://www.numpy.org/

https://matplotlib.org/

http://scikit-learn.org/

https://pandas.pydata.org/

https://pandas.pydata.org/

<https://ipython.org/>

<https://www.kaggle.com/netflix-inc/netflix-prize-data>