**Machine Learning Based Recommendation System**

# Thesis submitted in partial fulfilment of the requirements for the degree of

## B. Tech in Computer Science and Engineering

by

**Aditya Garg** (Roll No: 11901145)



Department of University College of Engineering

Computer Science and Engineering

Supervised by: - Dr. Gauravdeep Sir

# Declaration

This is to certify that the thesis entitle **” Machine Learning Based Recommendation System”**, submitted by **Aditya Garg, Roll No. 12901145** in partial fulfilment for the award of the degree of Bachler of Technology in Computer Science and Engineering, is a record of bona-fide work done by him, under our supervision, during the period of 2022.

This thesis, our opinion, is worthy of consideration for the award of degree of Bachler of Technology in accordance with the regulations of this institute.

# ACKNOWLEDGEMENTS

I extend my ineffable and deepest gratitude to my supervisor **Dr. Gauravdeep** for their esteem supervision, incessant inspiration, prolific encouragement, creative criticisms and also supplying the needs of runaway geniuses throughout the course of this thesis work. Their innovative ideas and invaluable suggestions have introduced me to a very interesting and dynamic field. They have been very much responsible for ensuring that my work progressed smoothly and surely in the right direction. I am grateful to them for the completion of my thesis work.

This research was supported in part by my family and friends with their constant moral support, love and encouragement throughout my endeavors.

## Aditya Garg

**Roll No. 11901145**

Department Of Computer Science and Engineering Punjabi University Patiala (PUP)

# Abstract

Recommendation system has been seen to be very useful for a user to select an item amongst many. Most existing recommendation systems rely either on a collaborative approach or a content based approach . We on the other hand have applied machine learning techniques to build recommender systems. We have taken two approaches in consideration . In the first approach a content based recommender system is built, which uses collaborative data, so as to get the effect of a hybrid approach to get better result of recommendation. Attributes used for content based recommendations are assigned weights depending on their importance to users. The weight values are estimated from a set of linear regression equations obtained from a social network graph which captures human judgment about similarity of items. In the second approach agent of call center have been recommended some procedure depending upon the current state of online call.

Overall, A combination of K-Means Algorithm and Hidden Markov model is used.

**Contents**

|  |  |
| --- | --- |
| **Title** | **Page no.** |
| Declaration | ii |
| Acknowledgements | iii |
| Abstract | iv |
| Table of Content | v |
| Introduction to Movie Recommendation System | 1 - 2 |
| Recommendation System | 3 - 5 |
| Technology of Project | 6 - 10 |
| Libraries | 11 - 14 |
| Project Description | 15 - 20 |
| Features and Drawback | 21 |
| Conclusion | 22 |

**1. Introduction to Movie Recommendation Systems**

Recommendation systems produce a ranked list of items on which a user might be interested, in the context of his current choice of an item. It helps the user to address the problem of information overload and finds the relevant objects for user. Recommendation system has been built for movies, book, communities, news, articles etc. There are two main approaches to build a recommendation system – collaborative filtering and content based [5]. Collaborative filtering recommenders use the opinions of other users to predict the value of items for each user in the community. Whereas content based recommendation systems recommend on basis of the content similarity between objects. There are several advantages and disadvantages of both which have been described in detail at Chapter 2. A hybrid recommender system also can be built by combining collaborative filtering and content based system.

Content based recommendation systems are very much useful when there is no or very less user data available. In that case depending on the similarity between the items the system recommends. The similarity of the content of the items is measured. To find this similarity various machine learning techniques (supervised or unsupervised) can be applied. The content also can be structured or unstructured. The later should be converted to structured content to make the similarity measure easier. Generally the structure of the content of two objects should be same to find the similarity. But in some situation it may possible that out of two objects, partial content is available for the one.

We have built recommender system for two application- Movie and Conversation Text. In movie recommendation when a user hits or selects one movie, or opens a page of a movie, the recommendation system recommends other movies which are similar to that selected movie. We used the structure content (feature set of director, writer, cast etc.) of movies. To find the similarities between two movies their feature sets are compared. But if we consider cosine similarity like measure that will give same weights to all the features. Though some features can be of more importance and some are of less. All the features may have different weights. These weights have been calculated in supervised technique, where the number of common reviewer has been used as a supervised data to the system. The same carries the collaborative information between two movies. Chapter 3 has the detail of the algorithm and experiments of this weight learning technique.

The second application we have considered conversational text data of call center, where the agents are recommended procedures that they should follow depending on the present status of the call. When a customer calls to the call center, while getting the query from the customer the agent access some knowledge base for the possible solution or answer to the customer’s query. Instead of this manual access to the knowledge base,

prompting the agent some recommendation of possible solution would be very effective. Depending on the current content of the call, it will produce a list of possible solution automatically.

This call data are text data which are unstructured. This unstructured data have been converted to structured data as term vector. We have used unsupervised machine learning technique to build this procedure recommendation. Chapter 4 has the detail of the algorithm and experiments of this agent prompting technique.

# Recommendation Systems

For many years recommendation systems had been a part of many online shopping systems. But in recent years it is evolving as a part of many other systems like portals, search engines, blogs, news, WebPages etc. We can put recommendation system on a top of another system, which have mainly two elements *Item* and *User*. To build the recommendation system one can use the Item data of the underlying system or both Item and User data. Examples of items are book, song, movie, news, blog, procedure etc.

There are mainly two approaches to build a recommendation system- Collaborative Filtering (CF) or Social Information Filtering (SF) and Content Based (CB).

1. *Collaborative Filtering:* Collaborative Filtering system maintains a database of many users’ ratings of a variety of items. For a given user, it finds other similar users whose ratings strongly correlate with the current user. It recommends items which are rated highly by these similar users, but not rated by the current user. Almost all existing commercial recommenders use this approach (e.g. Amazon). To build a Collaborative Filtering system one need to use both user and item data.
2. *Content Based:* Content Based system uses only the item data. It maintains a profile for each item. Considering the attributes or feature of the item it CB finds the similarity between items, and recommends the most similar item for an item. Content of an item can be structured or unstructured. For structured content extracting features is straightforward. If we consider the content of a movie as director, writer, cast etc., then each of these attribute can be considered as a feature.

General approach to that is building a term vector of text and considers the terms as features.

Collaborative Filtering can be categorized into three broad types.

1. Active CF*:* In Active CF, user actively participate to build the recommendation system by giving ratings or direct recommendation opinion on an item. Along with item profile data, user profile data also required to be maintained in Active CF.
2. Inactive CF*:* Inactive CF captures and analyses user activities like buying sequence or web access nature etc. Inactive CF does not use the user profile data.

1. Item Based CF*:* In item based CF a team of experts analyses the data of similar items and give direct recommendation. This type of recommendation can be found in various computer magazines.

In the other way, recommendation system can be categorized into two types- *item to item* and *user to user*. Item to item recommendation is mainly based on item data by comparing the item data, but user to user recommendation is mainly based on user data. Active CF is an example of user to user recommendation system. Content based recommendation is an example of item to item recommendation system. Some inactive CF also falls under the item to item recommendation category. If we apply association rule on the purchase history of the items and uses associated item as recommendation, it will be a type of inactive CF and also item to item recommendation. There is no importance of notion of sequence of items in a purchase session. But in some cases, the sequence of the item in a session is also important. For example if we consider the web page hit by a user in a single session, the sequence can be considered to make the recommendation.

It Is possible to make the recomme”dati’n systems on a collaborative filtering approach or on a content based approach. But there are some advantages and disadvantages of both. Collaborative filtering has these set of well known disadvantages.

1. *Cold Start:* There needs to be enough other users already in the system to find a match.
2. *Sparsity:*Most users do not rate most items and hence the user-item matrix is typically very sparse. If there are many items to be recommended, even if there are many users, because the user-ratings matrix is sparse, and it is hard to find users that have rated the same items.
3. *First Rater:*It is not possible to recommend an item that has not been previously rated. This problem comes for new items mostly. An obscure item also may face this problem.

Content based recommendation solves the above problems, but has its own constrains like-

1. CB is possible only for similar items like recommendation of movie. But itis not possible for recommend items on a super store. So this constraint is not a problem for movie recommendation.
2. It requires content that can be encoded as meaningful features. This is also not a constraint for movie recommendation as it has a nice set of features like director, writer, release date, genre etc. which can be easily encoded.

There are various attempts to combine content based system and collaborative filtering depending on the domain, data availability and requirement. A hybrid recommendation system is the combination of content based system and collaborative filtering.

# Technologies of Project

In this, I have use “Python” for building the code moreover “ML” is use for the objective of the Projects.

**3.1 INTRODUCTION TO PYTHON:**

Python is a simple, general purpose, high level, and object-oriented programming language. Python is an interpreted scripting language also. *Guido Van Rossum* is known as the founder of Python programming.

**3.1.1 What is Python?**

* **Python** is a general purpose, dynamic, [high-level,](https://www.javatpoint.com/classification-of-programming-languages) and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is simple and easy to learn and provides lots of high-level data structures.
* Python is *easy to learn* yet powerful and versatile scripting language, which makes it attractive for Application Development.
* Python’s syntax and *dynamic typing* with its interpreted nature make it an ideal language for scripting and rapid application development.
* Python supports *multiple programming pattern*, including object-oriented, imperative, and functional or procedural programming styles.
* Python is not intended to work in a particular area, such as web programming. That is why it is known as *multipurpose* programming language because it can be used with web, enterprise, 3D CAD, etc.
* We don’t need to use data types to declare variable because it is *dynamically typed* so we can write a=10 to assign an integer value in an integer variable.
* Python makes the development and debugging *fast* because there is no compilation step included in Python development, and edit-test-debug cycle is very fast.

**3.1.2 FEATURES OF PYTHON:**

Python provides many useful features which make it popular and valuable from the other programming languages. It supports object-oriented programming, procedural programming approaches and provides dynamic memory allocation. We have listed below a few essential features.

1. Easy to Learn and Use
2. Expressive Language
3. Interpreted Language
4. Cross-platform Language
5. Free and Open Source
6. Object-Oriented Language
7. Large Standard Library
8. GUI Programming Support
9. Integrated
10. Embeddable

* + 1. Easy to Learn and Use

Python is easy to learn as compared to other programming languages. Its syntax is straightforward and much the same as the English language. There is no use of the semicolon or curly-bracket, the indentation defines the code block. It is the recommended programming language for beginners.

* + 1. Expressive Language

Python can perform complex tasks using a few lines of code. A simple example, the hello world program you simply type **print(“Hello World”)**. It will take only one line to execute, while Java or C takes multiple lines.

* + 1. Interpreted Language

Python is an interpreted language; it means the Python program is executed one line at a time. The advantage of being interpreted language, it makes debugging easy and portable.

* + 1. Cross-platform Language

Python can run equally on different platforms such as Windows, Linux, UNIX, and Macintosh, etc. So, we can say that Python is a portable language. It enables programmers to develop the software for several competing platforms by writing a program only once.

* + 1. Free and Open Source

Python is freely available for everyone. It has a large community across the world that is dedicatedly working towards make new python modules and functions. Anyone can contribute to the Python community. The open-source means, “Anyone can download its source code without paying any penny.”

* + 1. Object-Oriented Language

Python supports object-oriented language and concepts of classes and objects come into existence. It supports inheritance, polymorphism, and encapsulation, etc. The object-oriented procedure helps to programmer to write reusable code and develop applications in less code.

* + 1. Extensible

It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our Python code. It converts the program into byte code, and any platform can use that byte code.

* + 1. Large Standard Library

It provides a vast range of libraries for the various fields such as machine learning, web developer, and also for the scripting. There are various machine learning libraries, such as Tensor flow, Pandas, Numpy, Keras, and Pytorch, etc. Django, flask, pyramids are the popular framework for Python web development.

* + 1. GUI Programming Support

Graphical User Interface is used for the developing Desktop application. PyQT5, Tkinter, Kivy are the libraries which are used for developing the web application.

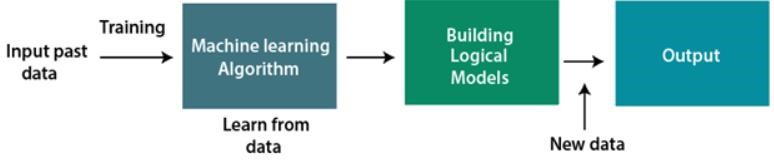
**3.2 Introduction to “ML”**

Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition**,** speech recognition**,** email filtering**,** Facebook auto-tagging**,** recommender system, and many more.

## 3.2.2 WORKING OF MACHINE LEARNING

A Machine Learning systemlearns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.



## 3.2.3 CLASSIFICATION OF MACHINE LEARNING

1. **Supervised learning**
2. **Unsupervised learning**
3. **Reinforcement learning**

1. Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample odellin data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using odellin data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Supervised learning can be grouped further in two categories of algorithms:

* + **Clustering**
  + **Regression**

1. Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been odellin, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don’t have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

* + **Clustering**
  + **Association**

1. Reinforcement Learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance.

# Libraries

## Pandas Library

Python Pandas is defined as an open-source library that provides high-performance data manipulation in Python. This tutorial is designed for both beginners and professionals.

Data analysis requires lots of processing, such as restructuring, cleaningor merging, etc. There are different tools are available for fast data processing, such as Numpy, Scipy, Cython**,** and Panda. But we prefer Pandas because working with Pandas is fast, simple and more expressive than other tools.

Key Features of Pandas

* It has a fast and efficient DataFrame object with the default and customized indexing.
* Used for reshaping and pivoting of the data sets.
* Group by data for aggregations and transformations.
* It is used for data alignment and integration of the missing data.
* Provide the functionality of Time Series.
* Process a variety of data sets in different formats like matrix data, tabular heterogeneous, time series.

## Numpy Library

NumPy stands for numeric python which is a python package for the computation and processing of the multidimensional and single dimensional array elements.

NumPy provides various powerful data structures, implementing multi-dimensional arrays and matrices. These data structures are used for the optimal computations regarding arrays and matrices.

NumPy provides a convenient and efficient way to handle the vast amount of data. NumPy is also very convenient with Matrix multiplication and data reshaping. NumPy is fast which makes it reasonable to work with a large set of data.

There are the following advantages of using NumPy for data analysis.

1. NumPy performs array-oriented computing.
2. It efficiently implements the multidimensional arrays.
3. It performs scientific computations.
4. It is capable of performing Fourier Transform and reshaping the data stored in multidimensional arrays.
5. NumPy provides the in-built functions for linear algebra and random number generation.

## Matplotlib

We can easily understand things when they are visualized. It is better to represent the data through the graph where we can analyze the data more efficiently and make the specific decision according to data analysis. Before learning the matplotlib, we need to understand data visualization and why data visualization is important.

There are five phases which are essential to make the decision for the organization**:**

* **Visualize:** We analyze the raw data, which means it makes complex data more accessible, understandable, and more usable. Tabular data representation is used where the user will look up a specific measurement, while the chart of several types is used to show patterns or relationships in the data for one or more variables.

* **Analysis:**  Data analysis is defined as cleaning, inspecting, transforming, and odelling data to derive useful information. Whenever we make a decision for the business or in daily life, is by past experience. **What will happen to choose a particular decision**, it is nothing but analyzing our past. That may be affected in the future, so the proper analysis is necessary for better decisions for any business or organization.
* **Document Insight:**  Document insight is the process where the useful data or information is organized in the document in the standard format.
* **Transform Data Set:**  Standard data is used to make the decision more effectively.

## Difflib

* Difflib is a built-in module in the Python programming language consisting of different simple functions and classes that allow users to compare data sets. The module offers the outputs of these sequence comparisons in a format that can be read by a human, using deltas to show the differences more efficiently.

* The difflib module is generally used to compare the sequence of the strings. But we can also use it to compare other data types as long as they are hash-able. We know that an object is hash-able if its hash value does not alter through the duration of its lifetime.

* The most commonly utilized classes in the Python difflib module are the Differ and the SequenceMatcher classes. There are also a few other helper classes and functions that can assist with more particular operations. Let us understand of some these functions in the following sections.

## Sklearn

An open-source Python package to implement machine learning models in Python is called Scikitlearn. This library supports modern algorithms like KNN, random forest, XGBoost, and SVC. It is constructed over NumPy. Both well-known software companies and the Kaggle competition frequently employ Scikit-learn. It aids in various processes of model building, like model selection, regression, classification, clustering, and dimensionality reduction (parameter selection).

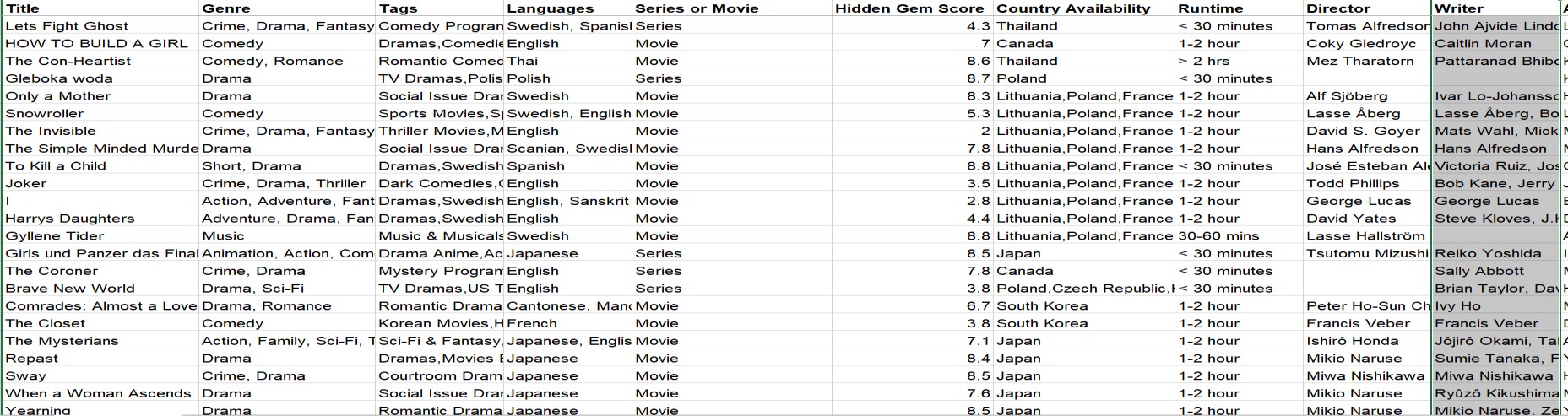
Scikit-learn is simple to work with and delivers successful performance. Scikit Learn, though, does not enable parallel processing. We can implement deep learning algorithms in sklearn, though it is not a wise choice, especially if using TensorFlow is an available option.

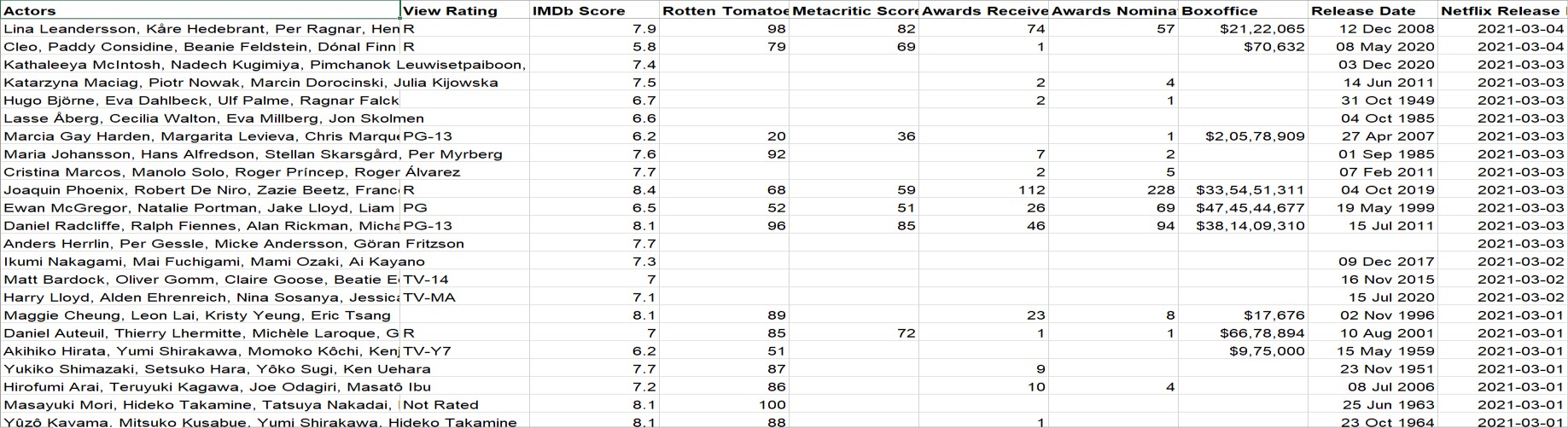
# Project Description

## Dataset

We are using the Netflix Dataset which has 27 features such as Title, Genre, Tags,

Language, etc. which can be used to find similarity between Movies and Series than the System can suggest the movie. The list of features are shown in images :-







It also provides the link of Movie and Serie’s Poster, Trailer, Netflix link, Poster, etc. By using this Dataset we can also check the rating of Movie and Series.

The data provided by the dataset is large (9426 Rows) and also has very old data, from 2015 to 2021 data. We can check Movie or Series is nominated for how many Award and how many id Awarded.

The data provide the information of its RunTime, Release Date , Boxoffice Collection which can be used to find success of it.

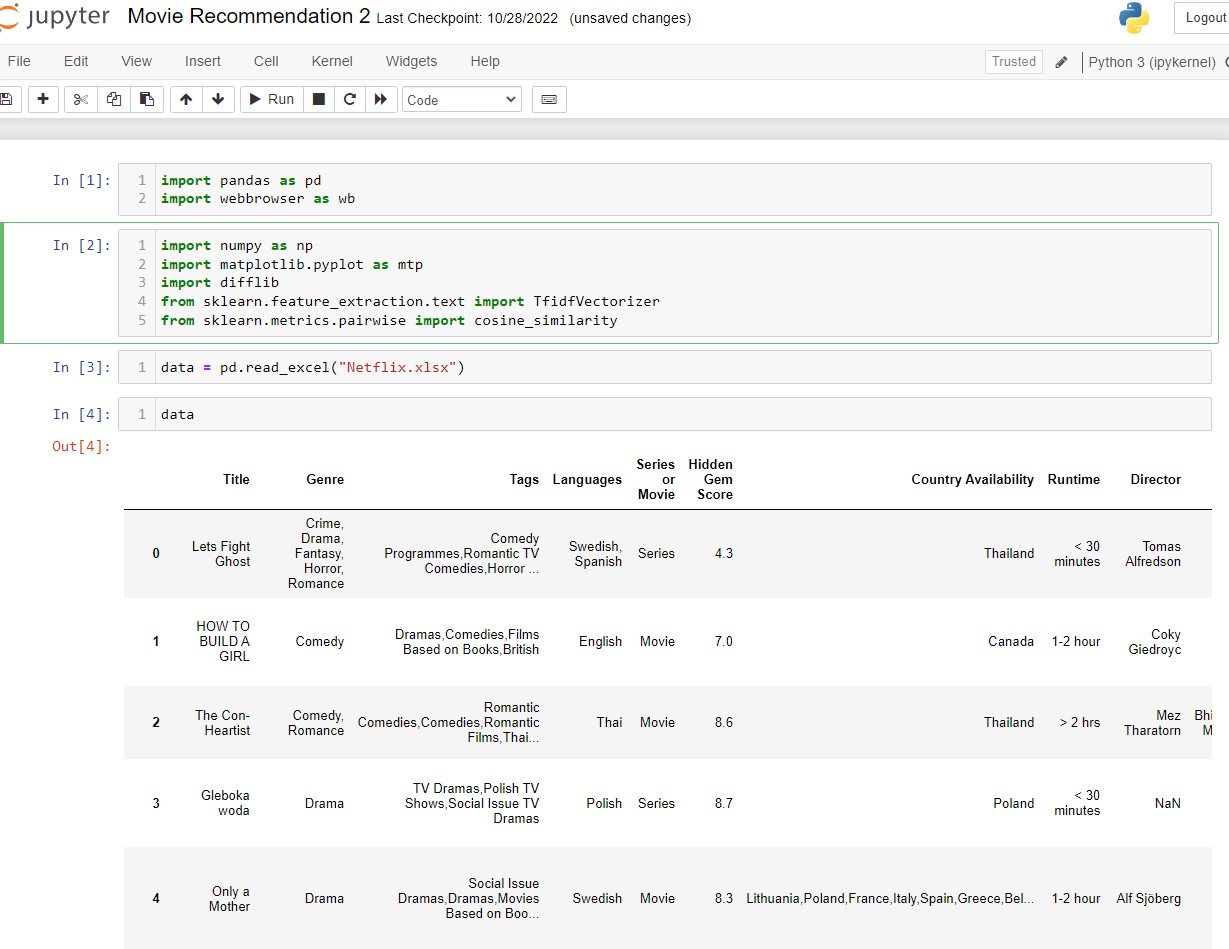
## Project Description

**Step 1:** - Loading/Importing the libraries.

1. Pandas: - used to make dataframe of the data present in the Dataset.
2. matplotlib: - used for visualization of data.
3. difflib: - used to correct the name of Movie/Series.
4. TfidVectorizer: - used for converting the Textual Data into Numeric Data.
5. cosine\_similarity: - used for find out the Similarity between Movies and Series.

**Step 2:** - Reading the dataset and display data.

read\_excel(): - used for reading the Dataset.

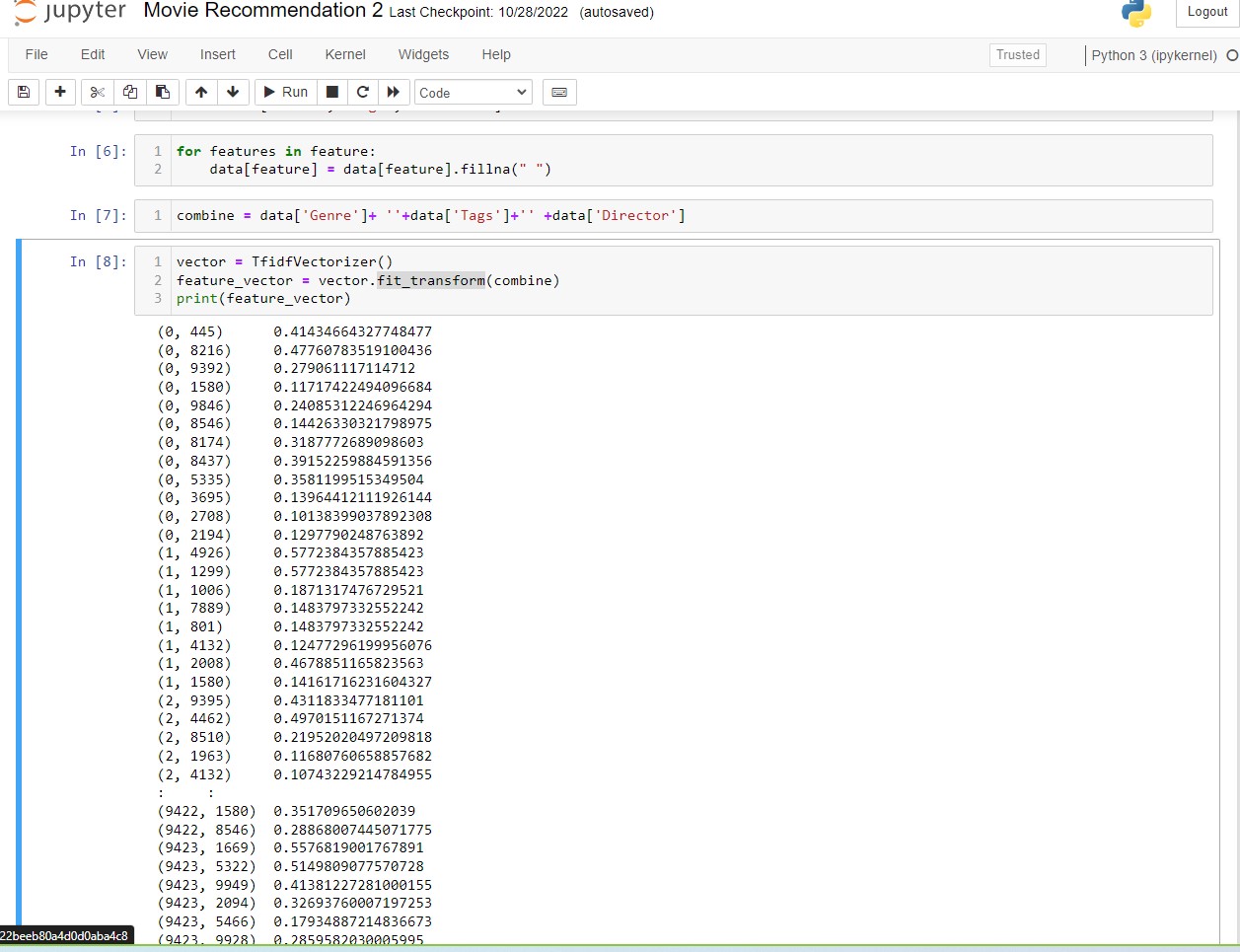


**Step 3:** - Select the Features which are useful for Recommend the Movie like “Tags”, “Director”, “Genre”, etc.

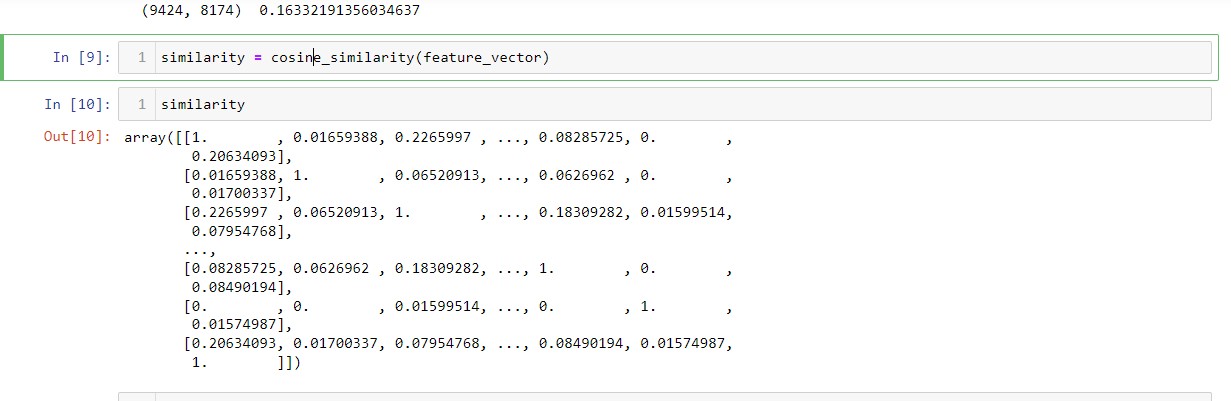
**Step 4:** - If any field of is empty in the selected column then we have to fill using null value.

**Step 5:** - TfidfVectorizer() is used to convert the covert the Textual data into Numerical data and fit\_transform() is used he transform method transforms all features using the corresponding means and variances.

It is very useful Library, as by using Textual values we can’t select the Movies/Series is very close match to the given value and moreover numeric value is easy to select the compare.



**Step 6:** - cosine\_similarity() is used to find out the similarity between the series/movies in the given dataset.



**Step 7:**

-

Take N

a

me

of Mo

vie/Serie for

recommend the movie.

Suppose

if we give input as

“

365

Days

”

.

After Selecting the mo

vie it will Display

index of mo

vie

/series an

d f

urt

her

goes for se

le

c

tion o

f similar mo

vies

usin

g th

e index of given Movie

/Serie

s

.

**Step 8**

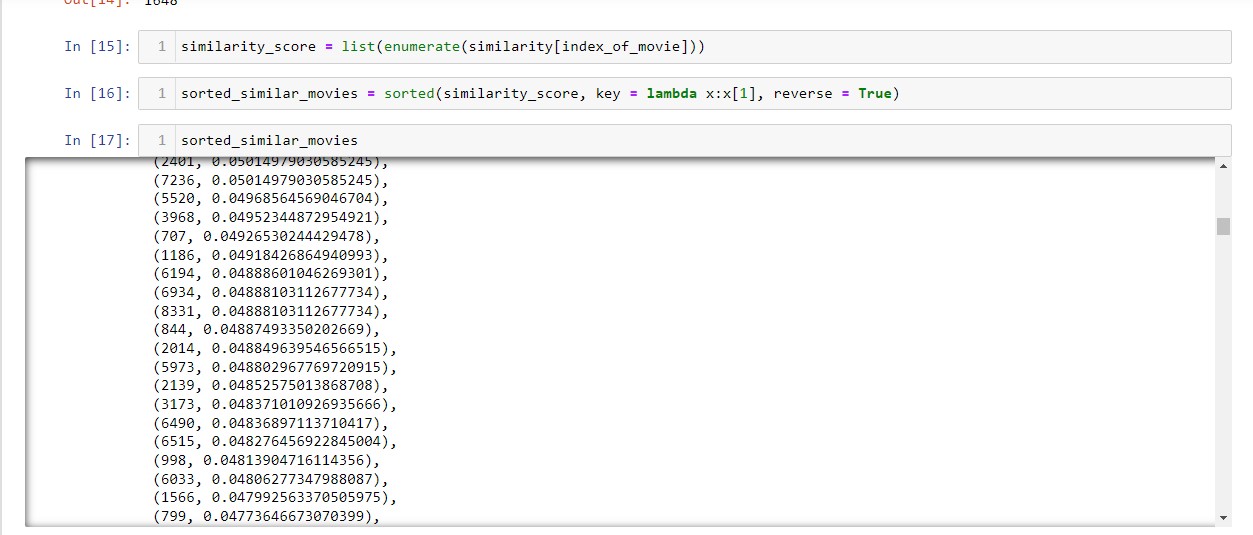
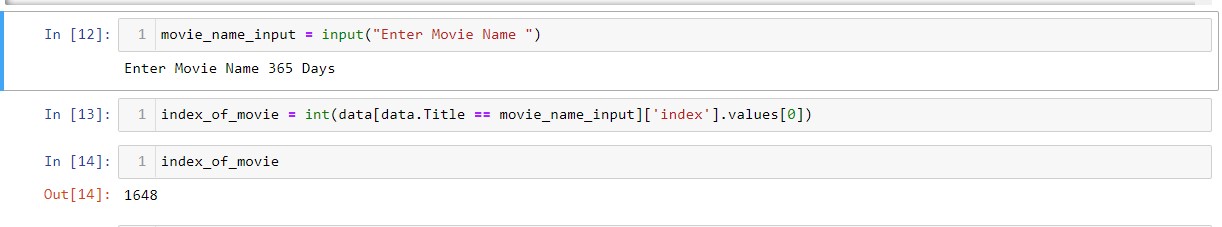
:

-

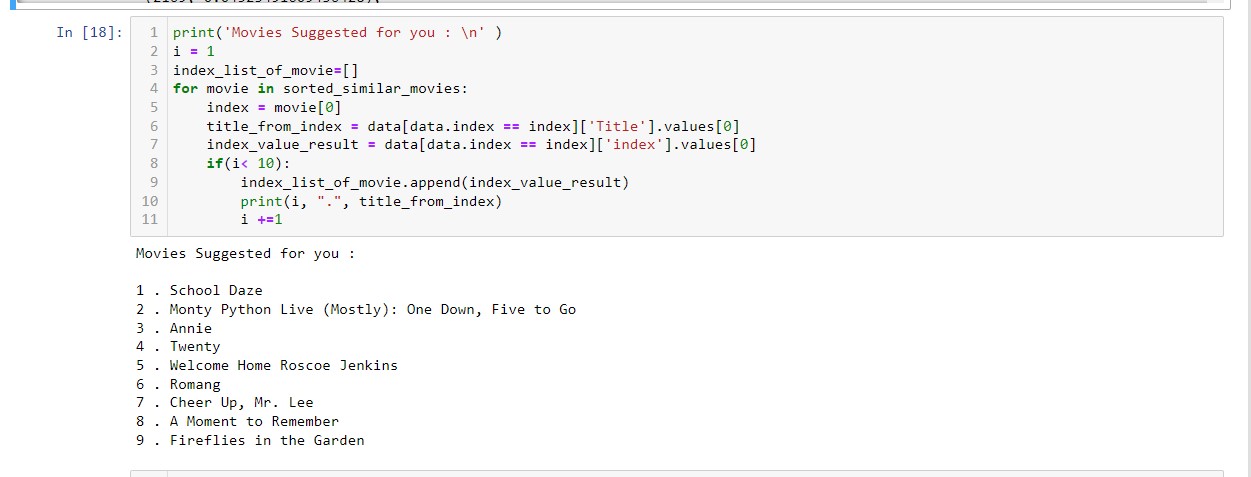
Now we create list of movie

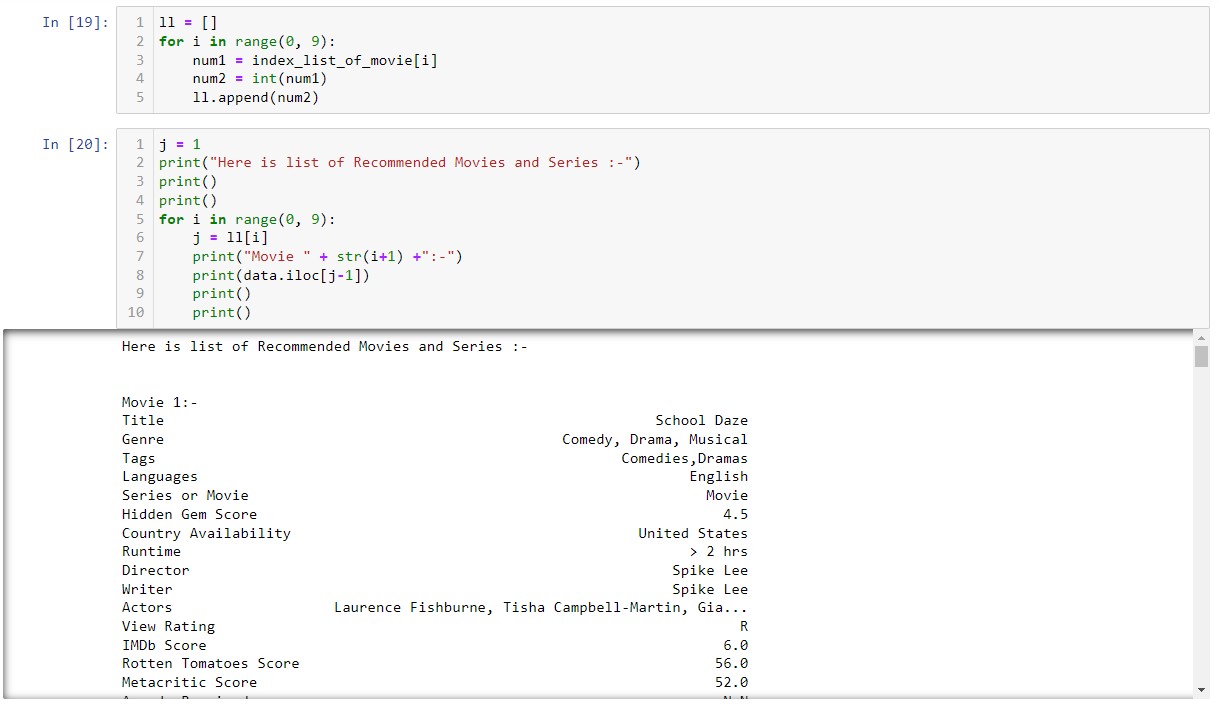
s similar to give

n movie and sort them according to there similarity.



**Step 9:** - Displaying the Name of the Movie/Series, including the information of them.





# Features & Drawback

## Features of Recommendation System

1. It has large dataset which include Movies/Series from 2015 to 2019.
2. It is easy to implement.
3. It can Recommend the Movies/series according the choices you prefer earlier.
4. It takes Every feature’s for Recommend the Movie/Series.

v. Reduce human affords to search the best Movie/Series.

vi. It provides Link for its Trailer, Poster, Netflix to open Directly from Dataset.

## Drawback of Recommendation System

1. Due to its large dataset it take long time to access dataset.
2. We can’t make Recommendation according to the IMDB Ratings.
3. We have to provide accurate name of movie/series according to the dataset.
4. It doesn’t have any proper Interface.
5. It doesn’t have any close match function features.
6. We can’t concatenate string and integer values together for feature extraction for Recommend the movie/series.

# Conclusion

In movie recommendation a hybridization of content based and collaborative filtering based recommendation is proposed. The weights of different attributes of an item are computed from the collaborative social network using regression analysis. Further studies to use this framework on other applications like web social network can give us more confidence on this concept. In agent prompting of call center we propose a method to extract procedural information from contact center transcripts. We define SPTS clusters and taken HMM based probabilistic approach to obtain better SPTS clusters. Procedures have been generated from HMM. We show that these procedures are useful in guiding an agent to the relevant procedure in an agent prompting application. This HMM based approach can be used to segmentation and classification of the call.