KJ's Educational Institute
KJ COLLEGE OF ENGINEERING AND MANAGEMENT RESEARCH
Affiliated to Savitribai Phule Pune University (SPPU), Approved by Govt. of Maharashtra, Recognized by AICTE, New Delhi.

TE COMPUTER DEPARTMENT

YEAR 2023-24

Semester – VI

DSBDL
[310256]

PROJECT TITLE: Develop a movie recommendation model using the scikit-learn library in python.

Done By:

ROLL NO – 'A28'

ROLL NO – 'A36'

ROLL NO – 'A43'

KJ COLLEGE OF ENGINEERING AND MANAGEMENT RESEARCH

Department of Computer Engineering

CERTIFICATE

This is to certify that the project entitled

"Movie Recommendation Model "

submitted by

GHADAGE AMRUTA ANKUSH A28
HOLKAR TUSHAR VISHWANATH A36
JAGTAP MAHESH ROHIDAS A43

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Engineering) at KJ COLLEGE OF ENGINEERING AND MANAGEMENT RESEARCH, Pune under the University of Pune. This work will be done during the year 2023-2024, under our guidance.

Date: / /

Guide:
Prof. Tanvi. A. Ghodke

HOD:
Dr. Nikita Kulkarni

ABSTRACT

In this hustling world, entertainment is a necessity for each one of us to refresh our mood and energy. Entertainment regains our confidence for work and we can work more enthusiastically. For revitalizing ourselves, we can listen to our preferred music or can watch movies of our choice. For watching favourable movies online we can utilize movie recommendation systems, which are more reliable, since searching of preferred movies will require more and more time which one cannot afford to waste. In this paper, to improve the quality of a movie recommendation system, a Hybrid approach by combining content based filtering and collaborative filtering, using Support Vector Machine as a classifier and genetic algorithm is presented in the proposed methodology and comparative results have been shown which depicts that the proposed approach shows an improvement in the accuracy, quality and scalability of the movie recommendation system than the pure approaches in three different datasets. Hybrid approach helps to get the advantages from both the approaches as well as tries to eliminate the drawbacks of both methods.

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude and respect to KJ COLLEGE OF ENGINEERING AND MANAGEMENT RESEARCH, Pune for providing me a platform to pursue my studies and carry out my final year project

I would like to thank Dr. Nikita Kulkarni, Professor and Head, Department of Computer Engineering, KJ COLLEGE OF ENGINEERING AND MANAGEMENT RESEARCH, Pune who has been a constant support and encouragement throughout this project.

I consider it a privilege and honor to express my sincere gratitude to my guide Prof. Tanvi. A. Ghodke, Associative professor, Department of Computer Engineering, for the valuable guidance throughout the tenure of this review.

I also extend my thanks to all the faculty of Computer Engineering who directly or indirectly encouraged me.

INDEX

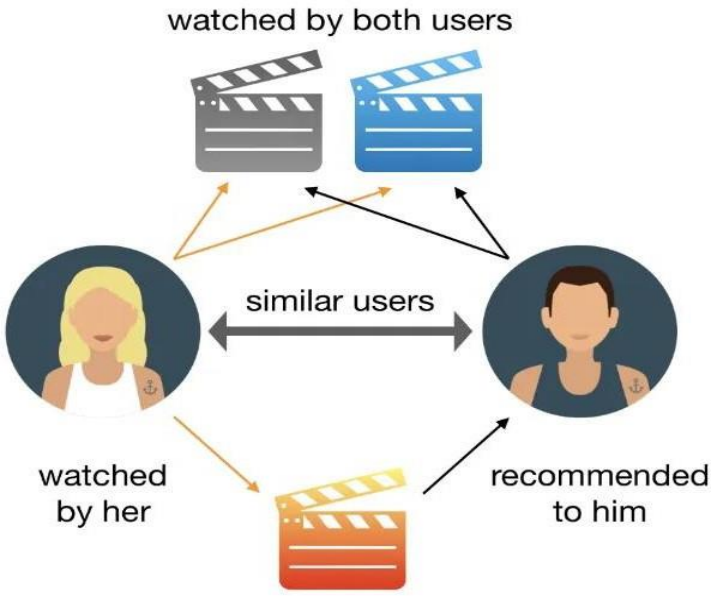
1.INTRODUCTION

1.1 Relevance of the Project

1. Personalized User Experience: With the abundance of content available on streaming platforms, users often struggle to find movies that match their preferences. A recommendation system addresses this issue by providing personalized suggestions based on individual tastes and viewing history, enhancing the overall user experience.
2. Enhanced Engagement and Retention: By offering relevant recommendations, platforms can increase user engagement and retention. When users discover content they enjoy, they are more likely to continue using the platform and exploring additional content, leading to increased satisfaction and loyalty.
3. Content Discovery and Diversity: Recommendation systems promote content discovery by introducing users to movies they may not have otherwise encountered. This facilitates the exploration of diverse genres, styles, and cultures, enriching the viewing experience and broadening horizons.
4. Efficient Content Curation: For streaming platforms and content providers, recommendation systems play a crucial role in efficiently curating and presenting content to users. By leveraging data-driven algorithms, these systems automate the process of content curation, saving time and resources while improving relevance.
5. Business Growth and Monetization: A well-functioning recommendation system can contribute to the growth and monetization of streaming platforms by increasing user engagement, retention, and satisfaction. Satisfied users are more likely to subscribe to premium services, purchase or rent movies, or engage with sponsored content and advertisements.
6. Technological Advancements: The development of recommendation systems involves cutting-edge technologies such as machine learning, deep learning, and big data analytics. Projects in this domain drive advancements in these fields and contribute to the evolution of recommendation algorithms and techniques.

1.2 Problem Statement

The goal of the project is to recommend a movie to the user. Providing related content out of relevant and irrelevant collection of items to users of online service providers.



1.3 Scope of the Project

- Data Collection and Preprocessing: Gathering data from diverse sources such as user ratings, viewing history, movie metadata, and contextual information. Preprocessing involves cleaning, transforming, and integrating the data to make it suitable for analysis.
- Model Training and Evaluation: Training recommendation models using historical user data and evaluating their performance using metrics such as accuracy, coverage, diversity, and serendipity. Iterative refinement of models based on feedback and validation results is also part of this phase.
- Scalability and Performance Optimization: Ensuring that the recommendation system can handle large volumes of data and concurrent user requests efficiently. Optimization techniques such as parallel processing, caching, and distributed computing may be employed to improve performance.
- Documentation and Maintenance: Documenting the system architecture, algorithms, data flows, and user interaction workflows. Regular maintenance and updates are also essential to ensure the continued effectiveness and relevance of the recommendation system.
- Testing and Quality Assurance: Conduct thorough testing to identify and address any bugs, usability issues, or performance bottlenecks. Quality assurance measures ensure that the recommendation system meets the desired standards of accuracy, reliability, and usability.

1.4 Movie Recommendation System

A recommendation system collect data about the user's preferences either implicitly or explicitly on different items like movies. An implicit acquisition in the development of movie recommendation system uses the user's behaviour while watching the movies. On the other hand, a explicit acquisition in the development of movie recommendation system uses the user's previous ratings or history. The other supporting technique that are used in the development of recommendation system is clustering. Clustering is a process to group a set of objects in such a way that objects in the same clusters are more similar to each other than to those in other clusters. KMeans Clustering along with K-Nearest Neighbour is implemented on the movie lens dataset in order to obtain the best-optimized result. In existing technique, the data is scattered which results in a high number of clusters while in the proposed technique data is gathered and results in a low number of clusters. The process of recommendation of a movie is optimized in the proposed scheme. The proposed recommender system predicts the user's preference of a movie on the basis of different parameters. The recommender system works on the concept that people are having common preference or choice. These users will influence on each other's opinions. This process optimizes the process and having lower RMSE.

2. SYSTEM REQUIREMENTS SPECIFICATION

This chapter involves both the hardware and software requirements needed for the project and detailed explanation of the specifications.

3.1 Hardware Requirements

- A PC with Windows/Linux OS
- Processor with 1.7-2.4gHz speed
- Minimum of 8gb RAM
- 2gb Graphic card

3.2 Software Specification

- Text Editor (VS-code/WebStorm)
- Anaconda distribution package (PyCharm Editor)
- Python libraries

Python Libraries:

For the computation and analysis we need certain python libraries which are used to perform analytics. Packages such as SKlearn, Numpy, pandas, Matplotlib, Flask framework, etc are needed.

SKlearn: It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

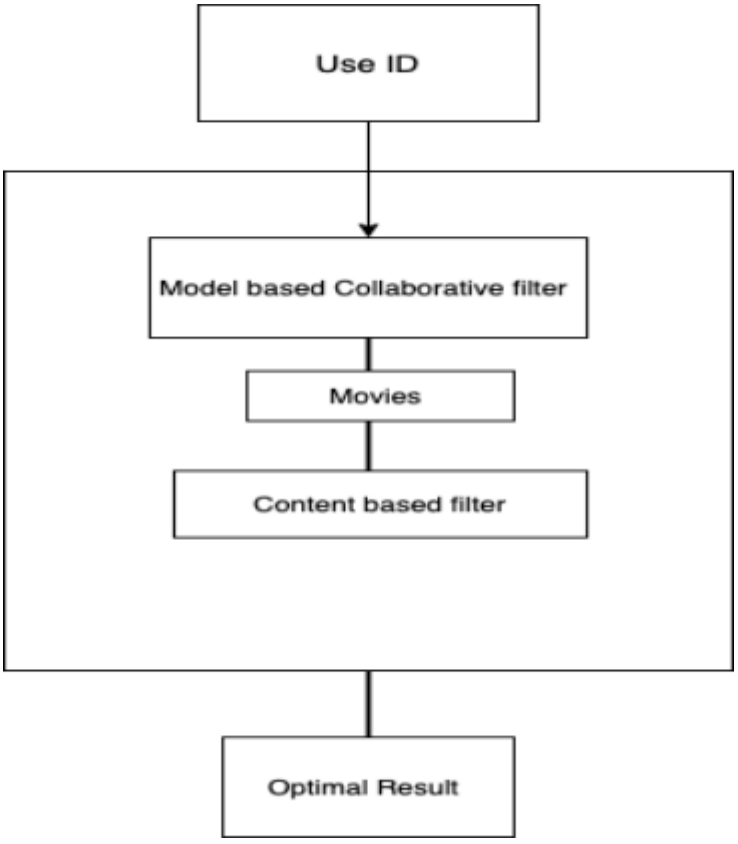
NumPy: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

Pandas: Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Data frame.

Flask: It is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It began as a simple wrapper around Werkzeug

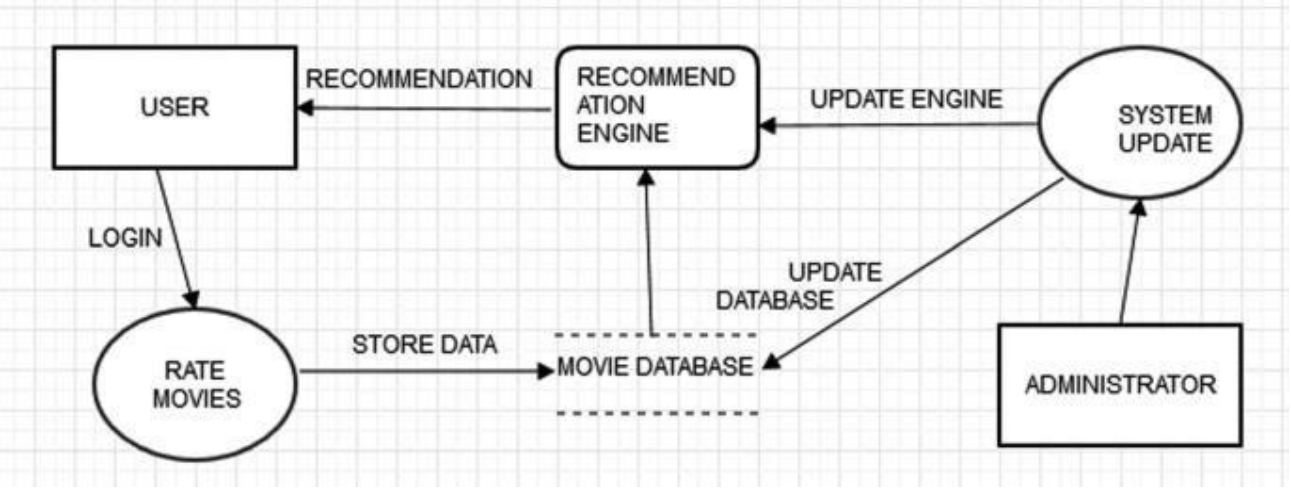
3. SYSTEM ANALYSIS AND DESIGN

4.1 System Architecture of Proposed System:



For each different individual use different list of movies are recommended ,as user login or enters the user id based on two different approaches used in the project each will recommend the set of movies to the particular user by combining the both the set of movie based on the user the hybrid model will recommend the single list of movie to the user.

4.2 Dataflow:



Initially load the data sets that are required to build a model the data set that are required in this project are movies.csv, ratinfg.csv, users.csv all the data sets are available in the Kaggle.com. Basically, two models are built in this project content based and collaborative filtering each produce a list of movies to a particular user by combining both based on the useid a single final list of movies are recommended to the particular user

4. IMPLEMENTATION

Develop a movie recommendation model using the scikit-learn library in python.

In [1]:
```
from sklearn.metrics.pairwise import cosine_similarity
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
```

In [2]:
```
df = pd.read_csv("https://raw.githubusercontent.com/rashida048/Some-NLP-Projects/master/movie_dataset.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	index	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	...	runtime	sp
0	0	237000000	Action Adventure Fantasy Science Fiction	http://www.avatarmovie.com/	19995	culture clash future space war space colony so...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	150.437577	...	162.0	[
1	1	300000000	Adventure Fantasy Action	http://disney.go.com/disneypictures/pirates/	285	ocean drug abuse exotic island east india trad...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	139.082615	...	169.0	[
2	2	245000000	Action Adventure Crime	http://www.sonypictures.com/movies/spectre/	206647	spy based on novel secret agent sequel mi6	en	Spectre	A cryptic message from Bond's past sends him o...	107.376788	...	148.0	

3	3	250000000	Action Crime Drama Thriller	http://www.thedarkknightrises.com/	49026	dc comics crime fighter terrorist secret ident...	en	The Dark Knight Rises	Following the death of District Attorney Harve...	112.312950	...	165.0	[
4	4	260000000	Action Adventure Science Fiction	http://movies.disney.com/john-carter	49529	based on novel mars medallion space travel pri...	en	John Carter	John Carter is a war-weary, former military ca...	43.926995	...	132.0	[

5 rows × 24 columns

In [4]:
```
features = ['keywords','cast','genres','director']
```

In [5]:
```
def combine_features(row):
    return row['keywords']+" "+row['cast']+" "+row['genres']+" "+row['director']
```

In [6]:
```
for feature in features:
    df[feature] = df[feature].fillna('')

df["combined_features"] = df.apply(combine_features,axis=1)
```

In [7]:
```
cv = CountVectorizer()
count_matrix = cv.fit_transform(df["combined_features"])
```

In [8]:
```
cosine_sim = cosine_similarity(count_matrix)
```

In [9]:
```
def get_title_from_index(index):
    return df[df.index == index]["title"].values[0]
def get_index_from_title(title):
    return df[df.title == title]["index"].values[0]
```

In [10]:
```
movie_user_likes = "Avatar"
movie_index = get_index_from_title(movie_user_likes)
similar_movies = list(enumerate(cosine_sim[movie_index]))
```

In [11]:
```
sorted_similar_movies = sorted(similar_movies,key=lambda x:x[1],reverse=True)[1:]
```

In [12]:
```
i=0
print("Top 5 similar movies to "+movie_user_likes+" are:\n")
for element in sorted_similar_movies:
    print(get_title_from_index(element[0]))
    i=i+1
    if i>5:
        break
```

```
Top 5 similar movies to Avatar are:

Guardians of the Galaxy
Aliens
Star Wars: Clone Wars: Volume 1
Star Trek Into Darkness
Star Trek Beyond
Alien
```

In [ ]:

5. RESULTS AND DISCUSSION

Since our project is movie recommendation system .one can develop a movie recommendation system by using either content based or collaborative filtering or combining both.

In our project we have developed a hybrid approach i.e combination of both content and collaborative filtering .Both the approaches have advantages and dis-advantages .in content based filtering the it based on the user ratings or user likes only such kind of movie will recommended to the user.

Advantages: it is easy to design and it takes less time to compute

Dis-advantages: the model can only make recommendations based on existing interests of the user. In other words, the model has limited ability to expand on the users' existing interests.

In Collaborative filtering the recommendation is comparison of similar users.

Advantages: No need domain knowledge because the embeddings are automatically learned. The model can help users discover new interests. In isolation, the ML system may not know the user is interested in a given item, but the model might still recommend it because similar users are interested in that item.

Dis-advantages: The prediction of the model for a given (user, item) pair is the dot product of the corresponding embeddings. So, if an item is not seen during training, the system can't create an embedding for it and can't query the model with this item. This issue is often called the cold-start problem.

The hybrid approach will resolves all these limitations by combining both content and collaborative filtering

6. REFERENCE

[1] Hirdesh Shivhare, Anshul Gupta and Shalki Sharma (2015), “Recommender system using fuzzy c-means clustering and genetic algorithm based weighted similarity measure”, IEEE International Conference on Computer, Communication and Control.

[2] Manoj Kumar, D.K. Yadav, Ankur Singh and Vijay Kr. Gupta (2015), “A Movie Recommender System: MOVREC”, International Journal of Computer Applications (0975 – 8887) Volume 124 – No.3.

[3] RyuRi Kim, Ye Jeong Kwak, HyeonJeong Mo, Mucheol Kim, Seungmin Rho,Ka Lok Man, Woon Kian Chong (2015),“Trustworthy Movie Recommender System with Correct Assessment and Emotion Evaluation”, Proceedings of the International MultiConference of Engineers and Computer Scientists Vol II.

[4] Zan Wang, Xue Yu*, Nan Feng, Zhenhua Wang (2014), “An Improved Collaborative Movie Recommendation System using Computational Intelligence”,Journal of Visual Languages & Computing,Volume 25, Issue 6.

[5] Debadrita Roy, Arnab Kundu, (2013), “Design of Movie Recommendation System by Means of Collaborative Filtering”, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 4.