MOVIE RECOMMENDED SYSTEM

INTRODUCTION

Recommender systems used in a various form of areas together with movies, music, news, books, analysis articles, search queries, social tags, and merchandise normally. Recommendation System is a filtration program whose prime goal is to predict the movie to a user towards a domain-specific item.

In our case, this domain-specific item is a movie, so the most focus of our recommendation system is to filter and predict solely those movies that a user would favour given some information concerning the user him or herself. There are many alternative ways to create a movie recommendation system however we've selected the content based recommender system in order that users will simply get the foremost similar movies based on the user's interest. As our recommender system recommends the top high five movies as like movie that user is Selected.

Given below are the two types of Movie Recommendation Systems

I. Content-based sifting method

II. Collaborative sifting method

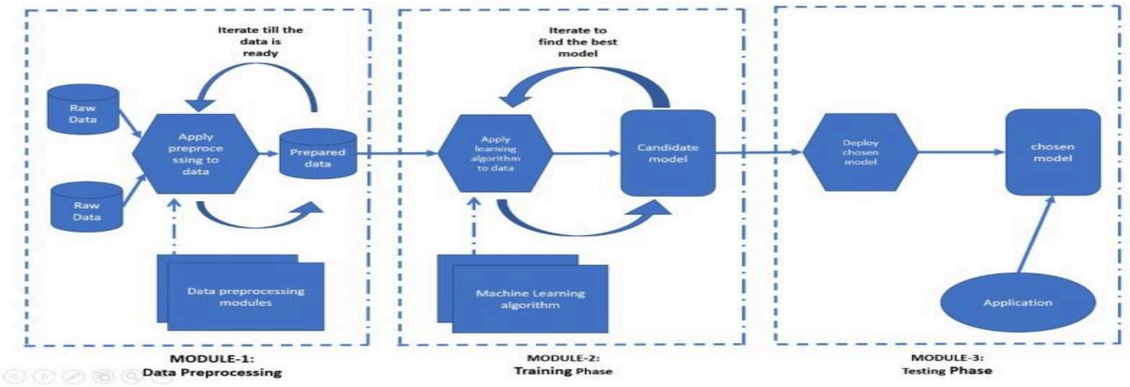
•Content-based filtering:

Content-based filtering is created based on keeping in mind the profile of the client's affinity and the initial database data. In this, to precisely predict the things we have castoff the ratings recorded by the clients to the movies or TVSeries and users favoured likes and dislikes to the shows. And by the end of the day the background of the software using Collaborative Filtering method and estimations endorse those things or like those things that were favoured before previously. It calculates and predicts the new shows and or earlier based predicted things and proposes best movies or shows based on his likes and dislikes items. It uses different strategies and projection methods on different areas of use. This method is mostly used in Hybrid Recommender Systems.An older calculations or the predictions of motion pictures or movies through MOVIEGEN datasets have different implementations, for ex, this demonstrates the movement of users' requests, with what had been searched in the past is also saved in the history or in the database. On learning these mistakes, we have developed Movie Recommendation System using Pearson Correlation Method, an advance method based structure that predicts and outputs movies to customers reliant for a data given by the customers in the past and the present examination, a customer is given the decision to pick his choices from a great deal of qualities based on No. Ratings and Rating to each movie, etc. We update the users choices in the database and compute a new set of results from the new data provided and based on the choices of the past visited history of customers. The software is developed using Python and a simple user interface.

DATA MINING PROCESS

Dataset Link: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

The basic architecture of the machine learning is as follows:


Iterate till the data is ready
Raw Data
Raw Data
Apply preproce ssing to data
Prepared data
Data preprocessing modules
MODULE-1: Data Preprocessing
Iterate to find the best model
Apply learning algorithm to data
Candidate model
Machine Learning algorithm
MODULE-2: Training Phase
Deploy chosen model
chosen model
Application
MODULE-3: Testing Phase

The methodology of the proposed system comprises of the following steps:

- Data-set Collection.
- Pre-Processing.
- Training and Validating.
- Prediction.
- Result.

Let's see some brief description on the above-mentioned topics:

A. Data-set collection

The most root part of the machine learning process is the dataset. So, here we have a valid data-set of 5000 Hollywood movies with different information about the movies.

B. Pre-processing

In machine learning, we use "pandas" and "numpy" libraries for pre-processing purposes.

NUMPY: NumPy stands for 'Numerical Python' or 'Numeric Python'. it's Associate in Nursing ASCII text file module of Python that provides quick mathematical computation on arrays and matrices. Numpy will be foreign into the notebook using:

```
>>>import numpy as np
```

PANDAS: Pandas is one in every of the foremost widely used python libraries in information science. It provides superior, straightforward to use structures and information analysis tools. Hence, with 2nd tables, pandas square measure capable of providing several further functionalities like making pivot tables, computing columns supporting different columns and plotting graphs. Pandas will be foreign into Python using:

```
>>>import pandas as pd
```

Choosing a model and strategy is an extremely vital method wherever we've victimised 2 libraries and machine learning techniques, specifically Scikit Learn, NLTK (Natural Language Toolkit), and victimisation formula referred to as circular function similarity.

SCI-KIT LEARN: Scikit-learn (Sklearn) is the most helpful and sturdy library for machine learning in Python. It provides a variety of economical tools for machine learning and applied maths modelling as well as classification, regression, clump and spatial property reduction via a consistent interface in Python. This library, that is basically written in Python, is made upon NumPy, SciPy and Matplotlib.Rather than that specialise in loading, manipulating and summarising information, Scikit-learn library is targeted on modelling the info. Stop words are simply a listing of words you don't wish to use as options. you'll be able to set the parameter stop words='english' to use an integral list. Alternatively, you'll be able to set stop words adequate to some custom list. This parameter defaults to none.

NLTK: NLTK (Natural Language Toolkit) Library could be a suite that contains libraries and programs for applied maths language processes. It's one in all the foremost powerful NLP libraries, that contains packages to form machines that perceive human language associated with an acceptable response. Stemming and Lemmatization in Python NLTK square measure text standardisation techniques for language process. These techniques square measure widely used for text preprocessing. The distinction between stemming and lemmatization is that stemming is quicker because it cuts words while not knowing the context, whereas

lemmatization is slower because it is aware of the context of words before the process. Stemming could be a

methodology of standardisation of words in language process. It is a way during which a collection of words in a very sentence square measure born-again into a sequence to shorten its search. During this methodology, the words having identical meanings however have some variations consistent with the context or sentence square measure normalised. Stemming and Lemmatization in Python NLTK area unit text normalisation techniques for language process.These techniques are unit wide used for text preprocessing. The distinction between stemming and lemmatization is that stemming is quicker because it cuts words while not knowing the context, whereas lemmatization is slower because it is aware of the context of words before the process. Stemming could be a methodology of normalisation of words in language process. It's a method within which a collection of words in an exceedingly sentence area unit is reborn into a sequence to shorten its operation. During this methodology, the words having an equivalent which means however have some variations in step with the context or sentence area unit normalised.

C. Training and Validation

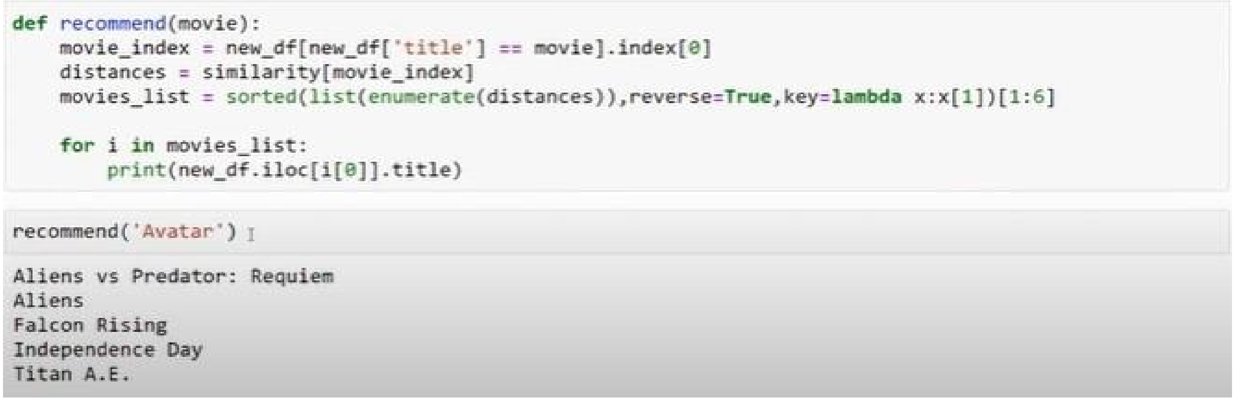
Now, the processed data is stored in the ".csv " file for further use. The processed data-set is divided into two parts :

- Training.(70 % of the data-set is used)
- Testing.(30 % of the data-set is used)

Now, comes the training part of the models. So, classification models are trained and tested to get the accuracy of the models. Once done with the accuracy part, we need to perform validation for further efficiency of the project.

D. Prediction

The presentation of the algorithm is based on accuracy and performance analysis and will provide a suggestion for the movies to the user whether movies are suggested or not upon the user's interest.



```
def recommend(movie):
    movie_index = new_df[new_df['title'] == movie].index[0]
    distances = similarity[movie_index]
    movies_list = sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[1:6]

    for i in movies_list:
        print(new_df.iloc[i[0]].title)
```

```
recommend('Avatar')
```

```
Aliens vs Predator: Requiem
Aliens
Falcon Rising
Independence Day
Titan A.E.
```

E. Result

The final result gives the recommendation of the movies.

ALGORITHMS

Some of the algorithms used in movie recommendation are COUNTVECTORIZER AND COSINE SIMILARITY.

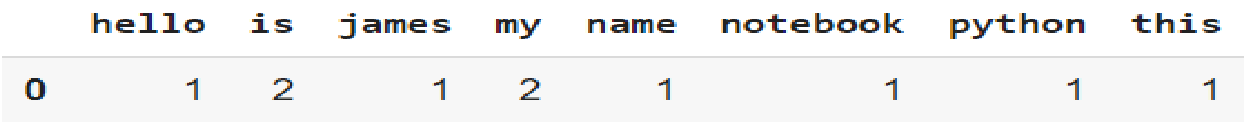
A. COUNT VECTORIZER

In order to use matter information for prophetic modelling, the text should be parsed to get rid of sure words. This method is termed tokenization. These words were then encoded as integers, or floating-point values, to be used as inputs in machine learning algorithms. This method is termed feature extraction (or vectorization).Scikit-learn's CountVectorizer is employed to convert a set of text documents to a vector of term/token counts.It conjointly permits the pre-processing of text information before generating the vector illustration. This practicality makes it an extremely versatile feature illustration module for text.

e.g

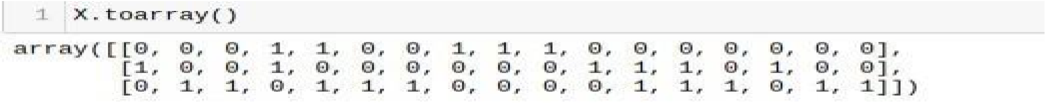
text = ['Hello my name is james, this is my jupyter notebook']

The text is transformed to a sparse matrix as shown below.



	hello	is	james	my	name	notebook	python	this
0	1	2	1	2	1	1	1	1

Count vectorizer makes it easy for text data to be used directly in machine learning and deep learning mode such as text class such as text classification.



```
X.toarray()
```

```
array([[0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0],
       [1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0],
       [0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1]])
```

Text Vectorization is the method of changing text into numerical illustration. Vectorization is jargon for a classic approach of changing a computer file from its raw format (i.e. text) into vectors of real numbers, that is the format that millilitre models support. Here we have a tendency to area unit mistreatment Bag of Words technique to convert text to vectors.

B. COSINE SIMILARITY

The circular function Similarity menstruation begins by finding the circular function of the 2 non-zero vectors.

The output can manufacture a worth starting from -1 to one, indicating similarity wherever -1 is non-similar,zero is orthogonal (perpendicular), and one represents total similarity If 2 vectors area unit diametrically opposed, that means they're familiarised in mere opposite directions, then the similarity menstruation is -1.circular function Similarity is employed in positive area, between the bounds zero and one circular function.

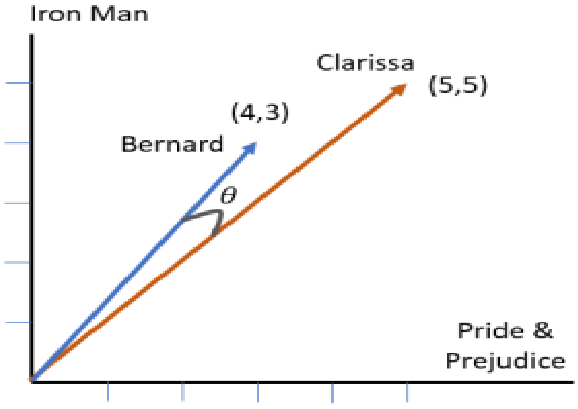
Similarity isn't involved, and doesn't live, variations is magnitude (length), and is simply a illustration of similarities in orientation. The library contains each procedure and functions to calculate similarity between sets of knowledge. The operation is best used once calculating the similarity between little numbers of sets. The procedures lay the computation and area unit thus additional acceptable for computing similarities

on large datasets.

$$similarity = \cos\theta = \frac{b.c}{||b||\,||c||}$$

$$b.c \Rightarrow Is\ the\ Dot\ product\ of\ the\ two\ vectors$$

$$||b||\,||c|| \Rightarrow Is\ the\ product\ of\ each\ vector's\ magnitude$$



$$Calculating:$$

$$b.c = \sum_{i=1}^{n} b_i c_i = (4 \times 5) + (3 \times 5) = 35$$

$$||b|| = \sqrt{4^2 + 3^2} = 5$$

$$||c|| = \sqrt{5^2 + 5^2} = 5\sqrt{2}$$

$$similarity = \frac{35}{5 \times 5\sqrt{2}} \sim 0.989$$

Theoretically, the perform circular function similarity is any range between -1 and +1 as a result of the image of the cos function, however during this case, there'll not be any negative picture show rating therefore the are going to be between zero° and 90° bounding the cos similarity between 0 and one. If the angleθ = 0° =>cosine similarity = one, if θ = 90° => cos similarity =0.

C. Cross Validation

In machine learning, we tend to not work the model on the coaching information and can't say that the model can work accurately for the important information. For this, we tend to assure that our model got the right patterns from the info, and it's not obtaining an excessive amount of noise. For this purpose, we tend to use the cross-validation technique.

Cross-validation may be a technique within which we tend to train our model to exploit the set of the data set and so value exploitation of the complementary set of the data-set.

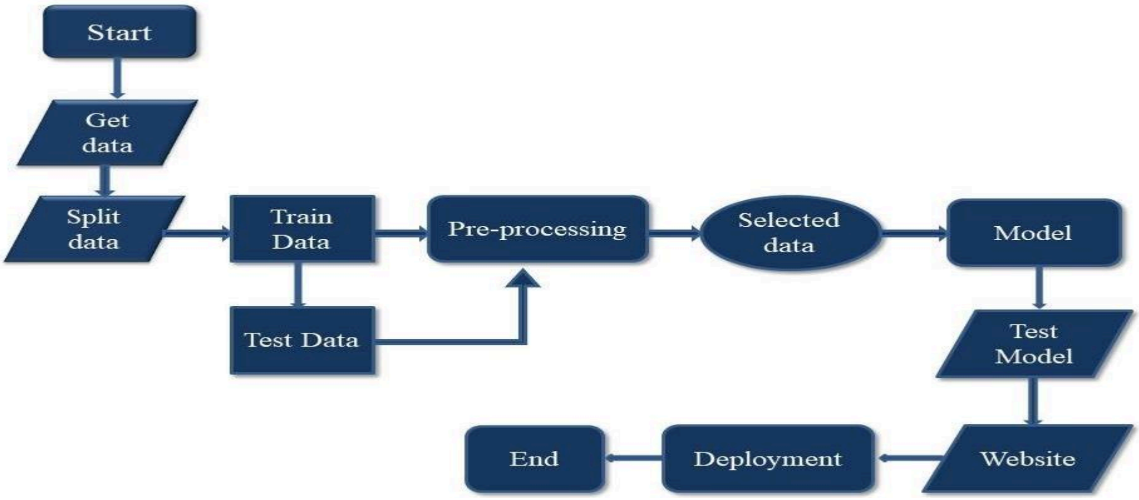
The 3 steps concerned in cross-validation square measure as follows :
Reserve some portion of sample data-set.

- Using the rest data-set, train the model.
- Test the model using the reserve portion of the data-set

PROPOSED METHODOLOGY

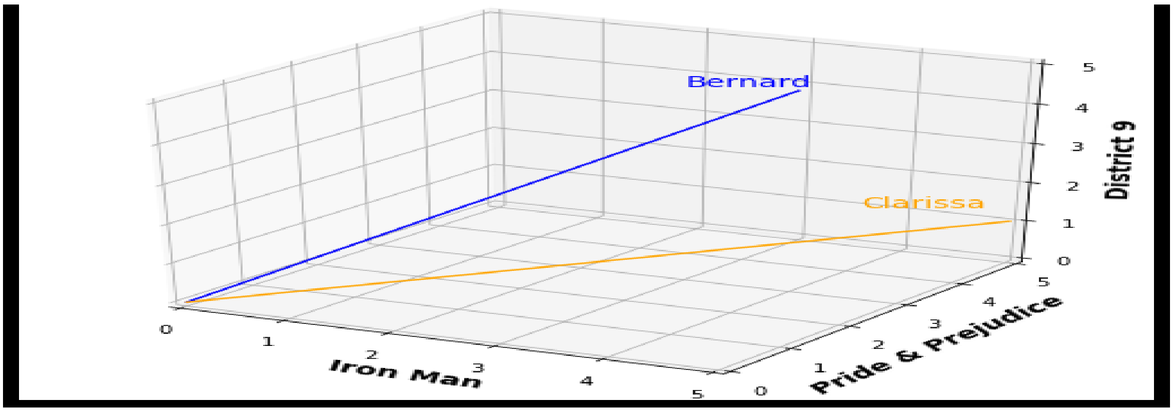
The methodology of the project is meant in six steps:

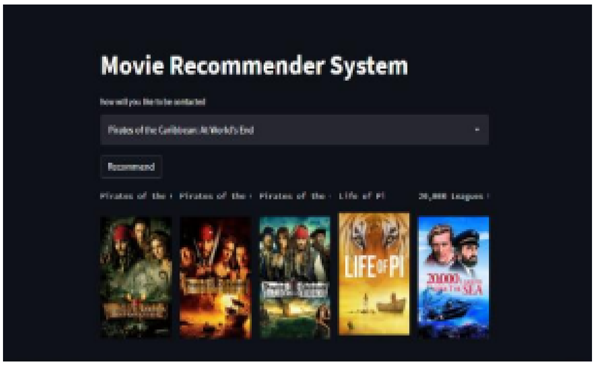
- Installing the Python and SciPy platform. we want to mount our ".ipynb" file on our google drive for more access.
- Loading the dataset. The dataset of picture show recommendation is required to be foreign in ".csv" format.
- Summarising the dataset. Sorting and improvement of knowledge is that the necessary method to extend the potency of the
- project. we are able to fill the missing information victimisation "imputer" perform.
- Visualising the dataset. We are able to visualise our "tmdb_5000_movies.csv" and "tmdb_5000_credits dataset through the Kaggle.com and so pre process method thereon.
- Evaluating some algorithms. when visualising the dataset, currently comes the coaching and testing part!!! Let's divide {the information|the info|the information} into 7:3 magnitude relation wherever seventieth data are trained and half-hour are tested. Now, let's choose the suitable models and train them to urge the accuracy of the prediction. We've got used two models: COUNT VECTORIZER AND cos SIMILARITY. When obtaining the accuracy of every model and scrutinising them, let's cross Making some predictions. Now , comes the last stage of the project, i.e., to form predictions. Here, the user will manually provide the input and acquire the advice of flick as per his/her interest.
- For content-based recommender systems specifically, we have a tendency to notice a brand new thanks to improve the accuracy of the representative of the flick and suggest 5 similar flicks to the user as per the interest of the movie. Now, to make the project additional easy, we've got to design a frontend as well!! The face consists of a web site with functions particularly recommendation and show. The face will be created victimisation flask module in python and preparation victimisation Heroku to link.



RESULT ANALYSIS

With the model trained, it must be tested to check if it'd operate well on planet objects. That's why a part of the info set created for analysis is employed to examine the model's proficiency. This puts the model during a state of affairs wherever it encounters things that weren't a district of its coaching. In short, for analysis purposes we tend to mistreat our tested knowledge and model to verify whether or not the model is functioning fine or not. Machine learning is a mistreatment of knowledge to answer queries. Therefore recommendation, or reasoning, is that the step wherever we tend to get to answer some queries. This is often the purpose of all this work, wherever the worth of machine learning is complete. We are able to finally use our model to predict whether or not the similar motion picture is usually recommended to the user or not as per his/her interest , supporting the similarity of the films.

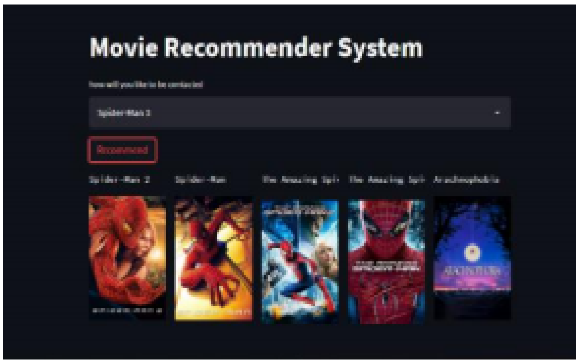




Movie Recommender System

How will you like to be contacted

Pirates of the Caribbean: At World's End



Movie Recommender System

How will you like to be contacted

Spider-Man 3

CONCLUSION

The main motivation of creating this project is to spice up every ., in order that we are able to perform our day-after-day of the movie, which is diverse and unique. We have successfully got the output of top five recommended movies as the user is selected by it’s choice. We develop the movie recommendation model using machine learning and algorithms.

Hence, our project “Movie recommendation system” is justified.