International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Recognition of similar movies based on plot summaries

Sai Chaitanya.K, Sai Charan.R, Sai Harshitha.I, Sai Kamal.M, Sai Kiran.A., Prof. Kalyani

Artificial Intelligence and Machine Learning Department, Mallareddy University, Hyderabad, Telangana, India

ABSTRACT:

This project focuses on leveraging Natural Language Processing (NLP) techniques and clustering algorithms to identify similarities among movie plot summaries. The first step involves collecting a dataset of movie plot summaries, either from existing sources or from platforms IMDb or Wikipedia. Subsequently, the collected text data undergoes preprocessing. Feature extraction techniques are then employed to convert the textual content into numerical vectors. Common methods include tokenization, stemming, create Tfidf Vectorizer enabling the representation of the plot summaries in a format suitable for clustering analysis. The core of the project revolves around clustering algorithm K-Means clustering.

Introduction:

Natural Language Processing (NLP) is an exciting field of study for data scientists where they develop algorithms that can make sense out of conversational language used by humans. In this Project, we will use NLP to find the degree of similarity between movies based on their plots available on IMDb and Wikipedia.[1]

In the vast world of digital entertainment, finding the right movie can be overwhelming.
Traditional recommendation systems have their limitations, often missing the essence of a film's plot. Our project, "Movie Similarity Analysis Based on Plot Summaries, "aims to revolutionize movie recommendations by focusing on the core narrative.[2]

This project seeks to redefine how users discover movies, addressing challenges in semantic understanding, data representation, scalability, and ethical considerations.[3]

The goal is simple: providing users with smarter, more personalized movie suggestions in the ever-evolving landscape of digital entertainment.
The dataset contains the titles of the top 100 movies on IMDb as well as each movie's plot summary from both IMDb and Wikipedia.[4]

Limitations:

- **Subjectivity of Plot Summaries:** Plot summaries can vary in detail and interpretation. Different individuals might summarize a movie differently, focusing on different aspects of the plot. This subjectivity can lead to inconsistencies in similarity assessments.

- **Lack of Context:** Plot summaries often fail to capture the broader context of a film, including its visual style, cinematography, pacing, and atmosphere. These elements contribute significantly to the overall movie experience and can't be fully conveyed through a textual summary.
- **Limited Scope for Evolution and Change:** Movies evolve over time, and their significance, influence, and interpretation can change. Plot summaries provide only a snapshot of a film's content at a particular moment, failing to capture its evolving cultural impact or critical reception.

Proposed work:

Machine Learning Models:

Existing System: The existing system might not involve a machine learning model or may use a different architecture.
Proposed Method: Proposes experimenting with machine learning models, networks, or deep neural network.

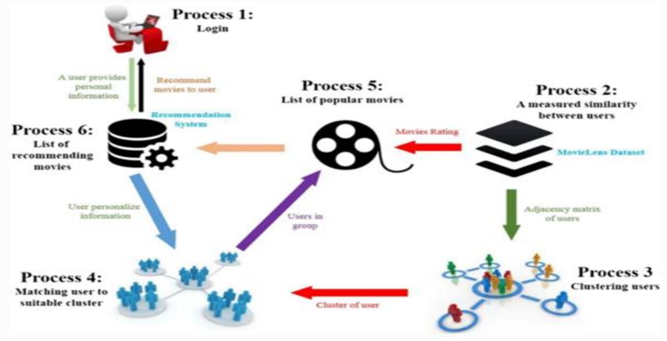
Flexibility and Adaptability:

Existing System: The existing system may lack details on adaptability and flexibility.
Proposed System: Proposes adapting the methodology based on specific project requirements and user feedback.

Database:

Creating a database for movie recognition based on plot summaries involves designing an efficient data storage system capable of storing movie information, similarity scores. Here's a high-level overview of how you might structure such a database:

Movie Table: This table stores information about each movie in your dataset. It could include attributes such as:
Movie ID (Primary Key) Title, Release Year, Genre, Director, Cast, Plot Summary Additional metadata (e.g., runtime, language, country of origin)



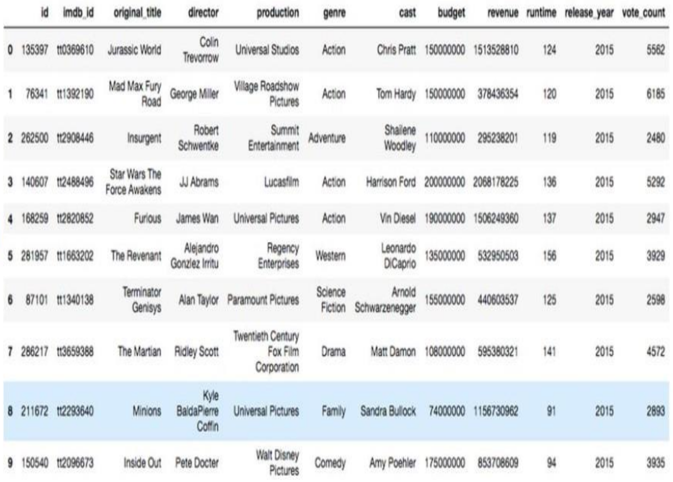
Data Set Descriptions:

To perform Finding similar movies based on plot summaries, you'll need a dataset that includes text of movies based on plots. Here's a description of the required dataset:
Dataset Format:
Characters: The dataset should consist of characters that is to be undergoing preprocessing, initially to be taken as input.
Vectors: Vectors are the later form that is formed after preprocessing and trained to machine in form of vectors by using vectorizing after tokenization and stemming.
Character Set: Define the set of movie plots as characters you want to recognize. This may include all the movies descriptions as characters.
Training and Testing Sets: Split the dataset into training and testing sets. A common split is to use a majority of the data for training (e.g., 70-80%) and the rest for testing to evaluate the model's performance.

Implementation Details: Choose a suitable implementation library, such as Scikit-learn in Python. Train the above using the training dataset and evaluate its performance on the testing dataset.

Dataset Source and Licensing:

We have collected data from IMDb and Wikipedia. They are a collection of movies which are based on plots. As the present study required a reasonable length of summary and we were also interested in the release year of movies, we filtered the IMDb dataset to remove records without a year and with a summary length less than 400 characters.



	id	imdb_id	original_title	director	production	genre	cast	budget	revenue	runtime	release_year	vote_count
0	135397	tt0369610	Jurassic World	Colin Trevorrow	Universal Studios	Action	Chris Pratt	150000000	1513528810	124	2015	5562
1	76341	tt1392190	Mad Max Fury Road	George Miller	Village Roadshow Pictures	Action	Tom Hardy	150000000	378436354	120	2015	6185
2	262500	tt2908446	Insurgent	Robert Schwentke	Summit Entertainment	Adventure	Shailene Woodley	110000000	295238201	119	2015	2480
3	140607	tt2488496	Star Wars The Force Awakens	JJ Abrams	Lucasfilm	Action	Harrison Ford	200000000	2068178225	136	2015	5292
4	168259	tt2820852	Furious	James Wan	Universal Pictures	Action	Vin Diesel	190000000	1506249360	137	2015	2947
5	281957	tt1663202	The Revenant	Alejandro Gonzlez Irritu	Regency Enterprises	Western	Leonardo DiCaprio	135000000	532950503	156	2015	3929
6	87101	tt1340138	Terminator Genisys	Alan Taylor	Paramount Pictures	Science Fiction	Arnold Schwarzenegger	155000000	440603537	125	2015	2598
7	286217	tt3659388	The Martian	Ridley Scott	Twentieth Century Fox Film Corporation	Drama	Matt Damon	108000000	595380321	141	2015	4572
8	211672	tt2293640	Minions	Kyle BaldaPierre Coffin	Universal Pictures	Family	Sandra Bullock	74000000	1156730962	91	2015	2893
9	150540	tt2096673	Inside Out	Pete Docter	Walt Disney Pictures	Comedy	Amy Poehler	175000000	853708609	94	2015	3935

A IMDb has multiple movies . A sample image is shown in Fig.

Research problem and contribution:

Enhancing Similarity Measures: Research can focus on improving the accuracy and effectiveness of similarity measures used compare plot summaries. This could take advanced natural language processing (NLP) techniques, such as deep learning semantic similar models incorporating domain-specific knowledge to better capture the nuances of movie plots.

Entertainment Industry: Improved movie recognition algorithms can benefit the entertainment industry by enabling better content recommendation systems for streaming platforms, movie databases, and recommendation engines. This can lead to increased user engagement satisfaction, and retention.

Cultural Studies: Research in this area can contribute to our understanding of cross-cultural story telling conventions, audience preferences and the globe impact of movies. By analyzing similarities and differences in plot summaries across cultures, researchers can gain insights into shared themes, motifs, and cultural influences.

Data Preprocessing Techniques:

Data preprocessing is a crucial step in building machine learning models for Finding movie similarity wise and based on plot summaries or any other predictive task.

It involves cleaning, transforming, enabling, foaming, organizing raw data to make it suitable for training and testing machine learning algorithms.

Character representation and Segmentation:

Initially, there is a requirement of segmenting the data to characters. The first step involves collecting a dataset of movie plot summaries, either from existing sources or from platforms IMDb or Wikipedia. Subsequently, the collected text data undergoes preprocessing. Feature extraction techniques are then employed to convert the textual content into numerical vectors.

K-means Clustering Analysis:

- It is a variant of clustering analysis.
- This algorithm group together movies with similar plot structures, allowing for the identification of thematic and narrative similarities.

Model Selection-Model Development:

Natural Language Processing (NLP) is an exciting field of study for data scientists where they develop algorithms that can make sense out of conversational language used by humans. In this Project, we will use NLP to find the degree of similarity between movies based on their plots available on IMDb and Wikipedia.

- **NLTK (Natural Language Toolkit)**: A comprehensive library for processing human language data, including modules for tokenization, stemming, and TF-IDF calculations.

- **SpaCy:** A library for advanced natural language processing tasks, providing pre-trained models for various languages and efficient tokenization.

TFID Transformers: A library that provides pre-trained transformer-based models, including BERT. You can use this library to easily obtain contextual embeddings for words and sentences.

Count Vectorizers: Count Vectorizer converts a collection of text documents into a matrix where the rows represent the documents, and the columns represent the tokens (words or n-grams).

Model Training and Testing:

One half for model training and also the other part. For model analysis or testing. During this study, the info set is separated into two part the first half is termed coaching knowledge and also the second called take a look at data, training data makes up for eighty percent of the whole data used, and the rest for test data. all of those models are trained with the training data part and so evaluated with the test data.

CONCLUSION:

By this project we can conclude that, finding movie similarity from plot summaries involves combining traditional and modern method to understand and measure how similar movie stories are. We start by cleaning and organizing the plot information, and then use different techniques to represent and also compare movies. This includes traditional methods like counting words and more advanced methods like using artificial intelligence to understand the meaning of words in context. The combination of these methods, including advanced neural networks and graph-based techniques, helps us create a holistic view of movie similarity.

REFERENCES:

1. Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? (and how to fix it using search-based software engineering). Information and Software Technology, 98(February), 74–88.
2. Bamman, D., O'Connor, B., & Smith, N. A. (2015). Learning latent personas of film characters. In Proceedings of the 51st annual meeting of the association for computational linguistics (pp. 352–361).
3. H. Yi, D. Rajan and L.-T. Chia, "Semantic video indexing and summarization using subtitles" in Advances in Multimedia Information Processing-PCM 2004 ser. LNCS, Springer, vol. 3331, pp. 634-641, 2004.
4. Bamman, David, Ted Underwood, and A. Noah Smith. 2019. A Bayesian Mixed Effects Model of Literary Character. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, MD, USA, pages 370–379.