**EE 541 – Computational Introduction to Deep Learning**

Project Proposal

Nissanth Neelakandan Abirami, Sam Devavaram Jebaraj

# November 15, 2023

**Project Title:** "CineLit Recommender: Elevating Entertainment Discovery"

# Topic summary

We are proposing the development of a recommendation system, "CineLit," designed to revolutionize content discovery. By seamlessly blending the realms of cinema and literacy, “Cinelit” aims to assist both cinephiles and bibliophiles. This innovative system will deliver personalized movie suggestions by integrating cutting-edge techniques such as reinforcement learning, content-based filtering, and collaborative filtering. Additionally, CineLit will extend its recommendations to include books based on users' preferred movies (Figure 1).



Figure 1: CineLit recommending the first Harry Potter book after the completion of the first movie. Additionally, the second movie is recommended following the order of the movie series.

# Introduction

This project's recommendation engine directly tackles this issue by providing a solution for tailored content discovery. Users will benefit from a more simplified and pleasurable content discovery experience, in which movies and books are more closely aligned with their personal tastes and preferences.

The importance of the topic addressed by this study resides in the current difficulties that users confront when traversing the enormous and ever-expanding world of entertainment content. Users are confronted with a dizzying array of possibilities as the volume of available movies and books continues to expand, leading to the problem of information overload.

This improved user experience is equally important for online streaming platforms and digital libraries, where user engagement is a key commercial aim. For example, Amazon’s Prime video content streams numerous movies and this recommendation system can open new impressions towards their Kindle webpage for obtaining books.

The proposed recommendation system has a multifaceted impact on the industry, influencing both user experience and the strategic positioning of entertainment platforms. Here are some significant characteristics of its possible industry impact:

*Enhanced User Engagement:*

By delivering individualized and relevant suggestions, the recommendation system has the ability to greatly increase user engagement. Users who discover material that matches their preferences are more likely to stay on the site for extended periods of time, resulting in higher user retention.

*Competitive Advantage:*

In a competitive entertainment industry, platforms that offer superior content recommendations gain a distinct edge. The use of advanced techniques such as content-based filtering, collaborative filtering, and reinforcement learning distinguishes the platform as technologically advanced and user-centric, distinguishing it from competitors.

*Increased Subscriber Base:*

A recommendation system that regularly provides accurate and interesting content suggestions attracts new users while also retaining existing members. Positive user experiences result in word-of-mouth referrals, which contribute to organic subscription growth.

*Dynamic Adaptability:*

The use of reinforcement learning and feed-forward networks enables the recommendation system to adapt dynamically to evolving user preferences. This elasticity guarantees that the platform remains relevant over time, addressing shifting user preferences and industry content trends.

This project involves a comprehensive examination of collaborative filtering, matrix factorization perspectives, feed-forward networks, EASE (Explicit Alternating Least Squares), sequential recommendation systems, the interplay between correlation and causality, propensity correction mechanisms, and the integration of reinforcement learning within recommendation systems.

**Theoretical Background:**

The theoretical foundation of this project rests on several key concepts and methodologies that shape the development of the recommendation system. Here are concise definitions for the keywords used:

*Content-Based Filtering:*

Content-based filtering involves recommending items to users based on the characteristics and features of the items themselves. In the context of movies, this would mean recommending films to users based on attributes such as genre, cast, and plot.

*Collaborative Filtering:*

Collaborative filtering recommends items to users based on the preferences and behaviors of similar users. It assumes that users who agreed in the past tend to agree again in the future. User-based collaborative filtering relies on similarities between users, while item-based collaborative filtering focuses on similarities between items.

*Reinforcement Learning:*

Reinforcement learning is when an agent learns to make decisions by interacting with its environment. In the context of the recommendation system, reinforcement learning may be employed to adapt and improve recommendations over time based on user feedback and interactions. This feedback can be taken as a penalty to punish the model for poor recommendations.

*Feed-Forward Network:*

A feed-forward neural network is a type of artificial neural network where information moves in one direction—from the input layer through hidden layers to the output layer. It's a fundamental architecture used in various machine learning tasks, including recommendation systems.

*Correlation vs. Causality:*

Correlation refers to a statistical relationship between two variables, whereas causality implies that one variable causes the other. Understanding the nuances of correlation vs. causality is essential in interpreting relationships in recommendation systems and discerning whether observed patterns are coincidental or causal.

*Sequential Recommendation System:*

A sequential recommendation system considers the order and context of user interactions. It takes into account the sequence of items a user has interacted with to make more contextually relevant recommendations, particularly important for content like movies and books where order can impact user preferences.

*Matrix Factorization View:*

Matrix factorization is a technique used in collaborative filtering where a large user-item interaction matrix is decomposed into the product of two lower-dimensional matrices. This view helps capture latent factors influencing user-item relationships, improving recommendation accuracy.

*EASE (Explicit Alternating Least Squares):*

EASE is an algorithm used in collaborative filtering that employs alternating least squares optimization to factorize the user-item interaction matrix. It helps uncover latent features and patterns in the data for improved recommendation accuracy.

# Related Work

[1] Previous research in the field of university library book recommendation systems has extensively investigated collaborative filtering, content-based filtering, and hybrid models. According to the findings of these studies, combining content-based filtering with Course Syllabus information, in addition to collaborative filtering, is effective in overcoming the limits of individual recommendation algorithms. Evaluation measures such as RSME and K-Fold Cross Validation have proved critical in measuring the performance of hybrid systems. The current research follows suit by providing a hybrid recommendation system distinguished by a novel weighting mechanism for increased resilience. The emphasis on tackling the cold start problem and boosting accuracy demonstrates the field's consistency of aims, emphasizing a consistent commitment to developing recommendation system approaches.

[2] This study addresses the problem of improving movie recommendations by combining Content-based and Collaborative-based filtering algorithms. Content-based techniques focus on previous data and preferences of users, whereas collaborative approaches take insights from comparable users' behaviors. The study compares the effectiveness of User-based, Item-based, SVD, and SVD++ algorithms in Collaborative filtering. Notable is the introduction of a revolutionary hybrid recommendation engine that combines Content-based and SVD models for maximum performance. In contrast to previous studies, this research coincides with current issues in recommendation systems but distinguishes itself by applying specialized algorithms, such as SVD and SVD++. The unique hybrid approach, which combines SVD and Content-based filtering, is a pioneering effort to integrate the capabilities of both techniques in order to overcome restrictions and demonstrate possible gains in both performance and efficiency.

[3] This study describes a unique unsupervised learning-based hybrid recommender system that incorporates Collaborative Filtering, a Content-Based Approach, and a Self-Organizing Map. The system surpasses state-of-the-art approaches in Collaborative Filtering when tested on a subset of the Movies Database, displaying higher accuracy, precision, and efficiency. The hybrid model combines three recommendation methodologies, combining collaborative filtering with a self-organizing map and age demographic attribute consideration. Despite a lengthier recommendation time, the system outperforms in terms of precision and performance improvement. The combination of Self-Organizing Map with collaborative filtering decreases RMSE in clusters, outperforming K-means clustering. Based on real user data, the Precision-Recall assessment highlights the system's capacity to improve movie suggestions.

# Dataset description

Movie Dataset : <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data>

The extensive Movies Metadata dataset from Kaggle forms the basis of the CineLit recommendation algorithm. This dataset includes a comprehensive set of metadata for more than 45,000 films, including a wide variety of data that is essential for developing and improving recommendation algorithms. This is a thorough analysis of the dataset:

*movies\_metadata.csv:*

The metadata file contains crucial information about 45,000 films in the Full MovieLens dataset, serving as a comprehensive resource for content-based filtering.

*keywords.csv:*

The movie plot keywords, essential for understanding the themes and characteristics of MovieLens films, are encapsulated in a JSON object. This object is stringified, offering a structured representation of keywords that proves valuable for identifying specific elements in the storyline.

*credits.csv:*

The movie plot keywords, essential for understanding the themes and characteristics of MovieLens films, are encapsulated in a JSON object. This object is stringified, offering a structured representation of keywords that proves valuable for identifying specific elements in the storyline.

*links.csv:*

The dataset includes essential identifiers for films, encompassing both IMDB IDs and TMDB (The Movie Database) IDs for every entry in the Full MovieLens dataset. This provision allows for data augmentation, facilitating connectivity between the MovieLens dataset and external datasets. These identifiers serve as pivotal links, enabling a seamless integration of information and enhancing the dataset's versatility for broader analyses and collaborative efforts.

*links\_small.csv:*

The TMDB and IMDB ID subset comprises the identifiers for 9,000 films, thoughtfully selected from the comprehensive Full Dataset. This subset, specifically curated to optimize computing resources, serves as a strategic resource for testing and early-stage model development.

ratings\_small.csv:

The Ratings in Subset consist of 100,000 user ratings specifically allocated to 9,000 films, forming a condensed yet comprehensive subset. This streamlined set of ratings serves as a valuable resource for preliminary testing and assessing model performance.

Books Dataset: <https://www.kaggle.com/datasets/zygmunt/goodbooks-10k/data>

This dataset encompasses ratings for ten thousand widely-read books, sourced from the internet. Each book has been reviewed approximately 100 times, though some have fewer ratings. The rating scale spans from one to five. Both book IDs and user IDs follow a consecutive sequence, ranging from 1 to 10000 for books and 1 to 53424 for users. Every user has submitted a minimum of two ratings, with the median number of ratings per user standing at 8. Additionally, users have indicated books they plan to read, and the dataset includes book metadata such as author, year, and tags.

*ratings.csv:*

Contains ratings for 10,000 popular books. Book IDs range from 1 to 10,000, and user IDs range from 1 to 53,424. Each book typically has 100 reviews, with ratings ranging from one to five. All users have made at least two ratings, and the median number of ratings per user is 8.

*to\_read.csv:*

Provides information on books marked "to read" by users. Lists user\_id, book\_id pairs for books users plan to read.

*books.csv:*

Holds metadata for each book, including goodreads IDs, authors, titles, and average ratings. Goodreads\_book\_id and best\_book\_id often point to the most popular edition. Goodreads\_work\_id refers to the book in an abstract sense. Description is obtained through Goodreads API using the book\_id tag.

*tags.csv:*

Translates tag IDs to names, providing a reference for the tags in book\_tags.csv. Contains user-assigned tags, shelves, and genres for books. Tags are represented by their IDs.

# Architecture Investigation Plan

The architectural investigation plan offers a targeted method for effectively designing, implementing, and optimizing a recommendation system. It consists of numerous major objectives and related tasks. The aim for content-based filtering is to extract important features from movies\_metadata.csv, create a feed-forward neural network, and evaluate using accuracy and recall metrics. Collaborative filtering strategies are investigated and reviewed, including user-based and item-based algorithms, as well as matrix factorization (SVD, SVD++).

The goal of reinforcement learning integration is to dynamically adjust suggestions based on user input over time, while also applying incentive systems for good interactions. The strategy includes expanding recommendations to include books based on movies, incorporating content-based filtering for books, and investigating matrix factorization perspectives. Propensity correction strategies are used to overcome biases and are evaluated for their influence on recommendation fairness.The development of a sequential recommendation system involves capturing sequential patterns and considering temporal aspects in user preferences.

Scalability and resource efficiency are ensured by optimizing algorithms and taking parallelization into account. To continually enhance the system, user input methods are created and their influence on overall performance is evaluated. This thorough approach ensures rigorous research of multiple recommendation algorithms, their integration, and system optimization for practical deployment, with an emphasis on flexibility and continual improvement via user input.

# Estimated Compute Needs

Compute resources will be strategically employed from platforms like Kaggle and Google Colab, harnessing GPU acceleration for efficient model training. Personal workstations, equipped with GPUs, CPUs, and substantial RAM, will be utilized for development and testing phases. The project's complexity, including reinforcement learning, may demand significant computational power, necessitating careful resource allocation for optimal performance.

# Likely Outcome and Expected Results

The CineLit recommendation system intends to have a significant influence on user pleasure and content discovery. The expected effects include greatly enhanced accuracy and relevance in movie and book suggestions, which will lead to higher user engagement and retention. The expansion of movie suggestions to books is anticipated to promote cross-domain synergy, providing consumers with an integrated entertainment experience. With the combination of reinforcement learning and feed-forward networks, the system will be able to react to user preferences in real time, with the system constantly improving recommendations over time.

# Primary References and Codebase

We propose to study and build on the approach gathered from the following:

* Pitiwat Arunruviwat, Veera Muangsin, “A Hybrid Book Recommendation System for University Library” 2022 26th International Computer Science and Engineering Conference (ICSEC) | IEEE | DOI: 10.1109/ICSEC56337.2022.10049318.
* Sakina Salmani, Sarvesh Kulkarni , “ Hybrid Movie Recommendation System Using Machine Learning”, 2021 International Conference on Communication information and Computing Technology (ICCICT) | IEEE | DOI: 10.1109/ICCICT50803.2021.9510058
* Yassine Afoudi, Mohamed Lazaar , Mohammed Al Achhab, “Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network,” Elsevier , [Simulation Modelling Practice and Theory](https://www.sciencedirect.com/journal/simulation-modelling-practice-and-theory) , [Volume 113](https://www.sciencedirect.com/journal/simulation-modelling-practice-and-theory/vol/113/suppl/C), December 2021, doi : 10.1016/ 102375.
* N. Ifada, T. F. Rahman and M. K. Sophan, ”Comparing Collaborative Filtering and Hybrid based Approaches for Movie Recommendation,” 2020 6th Information Technology International Seminar (ITIS), 2020, pp. 219-223, doi: 10.1109/ITIS50118.2020.9321014.
* Y. Xiong and H. Li, ”Collaborative Filtering Algorithm in Pictures Recommendation Based on SVD,” 2018 International Conference on Robots and Intelligent System (ICRIS), 2018, pp. 262-265, doi: 10.1109/ICRIS.2018.00074.
* S. Agrawal and P. Jain, ”An improved approach for movie recommendation system,” 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2017, pp. 336-342, doi: 10.1109/ISMAC. 2017.8058367.
* Blog Post, “https://medium.com/analytics-vidhya/building-a-movie-recommendation-engine-in-python-53fb47547ace”