# **COGS 118B - Project Proposal**

# **Names**

* Pranjli Khanna
* Vishnudeep Tyagi
* Hongjo Yoon
* Arturo Torres

# **Abstract**

This project endeavors to develop a movie recommendation system utilizing an open-source dataset from Kaggle, which includes detailed movie metadata but excludes personal user information, adhering to privacy standards. By applying unsupervised learning and natural language processing (NLP) techniques, such as TF-IDF and word embeddings, we aim to extract meaningful features from movie descriptions and metadata. We will test the accuracy of our findings with the rating that the users who watched the movies gave, under the column “user rating”. This foundation allows us to cluster movies based on similarities and generate personalized recommendations. This approach not only seeks to enhance user experience through personalized content discovery but also underscores the importance of ethical practices in the development of machine learning systems.

# **Background**

In an era dominated by digital streaming, the landscape of television and movie consumption has dramatically transformed, leading to an exponential increase in the volume of content available to viewers. This transformation has brought to the forefront the challenge of content discovery, making it increasingly difficult for viewers to navigate through endless options and select what to watch next. The inception of this project is motivated by the desire to tackle the growing dilemma of choice overload, aiming to develop a sophisticated movie recommendation system that harnesses the power of machine learning to offer personalized content suggestions.

The traditional approach to watching television, characterized by fixed schedules and limited selections, has been upended by the advent of streaming services, offering a boundless array of movies and TV shows at the fingertips of consumers. This shift has not only redefined viewer habits but has also introduced the problem of decision fatigue, where the abundance of choices paradoxically hampers the viewer's ability to make a selection. The project seeks to address this issue by creating a recommendation system that simplifies the decision-making process, enhancing the overall viewing experience.

According to a report highlighted in the context of streaming media consumption, the average American spends approximately 45 hours a year deciding what to watch next[[1]](https://github.com/COGS118B/group_template/blob/46e0ac6e2b28920d3e4f6a52d933fb8bba2888c0//#lorenznote). This statistic underscores the significant amount of time that could be better spent engaging with content rather than sifting through it. The project's relevance is thus underscored by this finding, as it aims to reduce the time spent in content selection, leveraging algorithms to streamline the process and make it more efficient.

Further emphasizing the shift in viewer preferences, a report also points to the rise of on-demand and mobile viewing, indicating a clear departure from traditional TV screen viewing[[2]](https://github.com/COGS118B/group_template/blob/46e0ac6e2b28920d3e4f6a52d933fb8bba2888c0//#lorenznote). This trend not only reflects changing technological landscapes but also a broader desire for flexibility and personalization in media consumption. By aligning the project's objectives with these trends, the proposed recommendation system seeks to cater to the modern viewer's needs, offering tailored suggestions that resonate with individual preferences and viewing habits.

The demand for solutions capable of addressing these challenges is also reflected in the job market, with a notable increase in listings for recommender systems positions[[3]](https://github.com/COGS118B/group_template/blob/46e0ac6e2b28920d3e4f6a52d933fb8bba2888c0//#lorenznote). This demand signals a growing recognition of the value and necessity of advanced recommendation systems within the industry. By tapping into this need, the project not only aligns with current technological and market trends but also contributes to an area of significant interest and investment.

In conclusion, the project is poised at the nexus of technological innovation and evolving viewer habits, seeking to address the pressing challenge of content discovery in the streaming era. By drawing on insights from consumer behavior, industry trends, and the evident demand for skilled professionals in recommender systems, the project underscores the urgency and relevance of developing a solution that enhances content accessibility and personalization. Through the development of an advanced movie recommendation system, this project aims to redefine how viewers interact with streaming content, making the process more intuitive, enjoyable, and aligned with individual preferences.

# **Problem Statement**

The problem that we are trying to solve through this project is how certain genres can underperform depending on the era that they are released in and how when people are trying to find a movie to watch they can spend upwards to an hour of trying to find the perfect movie wasting days of their lives in the long run.

# **Data**

Our [dataset](https://www.kaggle.com/datasets/ashpalsingh1525/imdb-movies-dataset) is sourced from the IMDB movies database. Contained within is detailed movie properties such as: their names, user ratings, genres, cast and crew members. Though these are not all the fields (there are 12 variables and over 10,000 (10,178!) movies are represented in the dataset) we will primarily focus on these in our journey to reduce indecision when choosing your next film.

While we would love to create a completely autonomous model, we believe that user input is important and as such will be breaking down the recommendations by genre to ensure a good user experience.

The data is pretty much in the form that will be useful to us, as the scores (ratings) are normalized by being in the same scale, however, if they weren’t we could have turned to logarithmic transformations to allow us to better maneuver it.

# **Proposed Solution**

The solution to our problem would be to use algorithms such as K-means and GMM in order to see what movies in our list are comparable with each other in order to gain an idea of what movie someone should be recommended based on a movie (within the list) that they like or what genre that they like. Our solution might work due to the fact that K-means and GMM would be able to cluster similar movies together based on things such as similar keywords in the synopsys and and similar ratings within the genres, which would try to recommend similar movies as to the one the person picked, which in theory they would enjoy since it is similar to the movies that they enjoy.

# **Evaluation Metrics**

The success of the recommendation system will be evaluated based on its ability to predict or match the user ratings of movies within the dataset. This approach involves comparing the system-generated recommendations with actual user ratings to assess the accuracy of the system's predictions. By leveraging the user rating column in the dataset, we can quantitatively measure how well the recommendation system understands and predicts user preferences.

### Metrics for Success

* Root Mean Square Error (RMSE) and Mean Absolute Error (MAE): These metrics will be used to quantify the average error between the predicted ratings by the recommendation system and the actual user ratings. Lower values of RMSE and MAE indicate higher accuracy in predictions.
* Precision and Recall: Precision measures the proportion of recommended movies that are relevant (i.e., highly rated by users), while recall measures the proportion of relevant movies that are recommended by the system. High precision and recall values signify that the system is effective in identifying movies that users will rate highly.
* F1 Score: The F1 Score combines precision and recall into a single metric by taking their harmonic mean. It balances the trade-off between precision and recall, providing a comprehensive measure of the system's overall performance.

By adopting this evaluation model, the project can systematically assess the effectiveness of the recommendation system in predicting user ratings and, consequently, its success in enhancing the content discovery experience for users. This quantitative approach allows for the identification of specific areas for improvement, guiding further development and optimization of the system.

# **Ethics & Privacy**

In developing a movie recommendation system, it's crucial to navigate the potential ethical and privacy concerns that accompany the collection, analysis, and application of user data.

In our project to develop a movie recommendation system, we utilize a dataset sourced from Kaggle, an open-source platform known for providing a wealth of datasets for various machine learning projects. This particular dataset comprises detailed information about movies, including genres, ratings, crew details, and overviews, but notably does not contain any direct user information. The open-source nature of the data and the absence of personal user data alleviate privacy concerns.

Some ethical concerns that we may have is the impact of our recommendation algorithm on future movies, if this project ends up being very successful and popular in recommending movies, movie companies may use this as a basis to try to make the most profitable movie, instead of a movie that they put their heart and soul into that ended up being very popular. This would end up being a detriment to society as the quality of media would be lowered and become something that panders to the most profitable group instead of being a medium to express artistic vision.

# **Team Expectations**

* *All team members are expected to communicate with each other on the group chat and all team members commit to regular updates and open communication in the group chat. This includes sharing progress, raising concerns, and providing feedback.*
* *All team members are expected to provide responses within a reasonable timeframe, acknowledging that everyone has different schedules.*
* *Decisions, especially those affecting the project's direction, will be made collectively, with each member having equal say.*
* *Each team member is expected to contribute equally to the workload, with tasks and responsibilities distributed based on skills, interests, and learning goals. Members are responsible for completing their tasks by the agreed deadlines.*
* *Deadlines are set for a reason and must be respected. All team members agree to deliver their work on time. If a delay is inevitable, the team should be notified as soon as possible, so adjustments can be made.*

# **Project Timeline Proposal**.

| **Meeting Date** | **Meeting Time** | **Completed Before Meeting** | **Discuss at Meeting** |
| --- | --- | --- | --- |
| 2/23 | 6 - 9 PM | Brainstorm topics/questions (all) Submit the project proposal. | Establish communication and decide on the final project topic. Find a problem from data to solve as a final project. Submit final project proposal. |
| 2/26 | 12 - 2 PM | Data verification and cleaning (Prune variables) | Begin to trim unused variables to create a lightweight dataset for efficient performance. |
| 2/29 | 6 -8 PM | Begin programming model. | Start deciding on model architecture and programming language. |
| 3/4 | 11 AM | Edge case testing.(Bias reduction) | Continue building model and unit testing to prevent errors or unintended performance. |
| 3/7 | 12 PM | Quality & performance check | Ensure that our model is outputting recommendations in our desired and expected form |
| 3/13 | 5 PM | Complete model and analysis. (Fine tuning) | Ensure that the model presents our intended recommendations with good precision and accuracy. |
| 3/19 | Before 11:59 PM | Final Look | Turn in Final Project |

# **Footnotes**

1.[^](https://github.com/COGS118B/group_template/blob/46e0ac6e2b28920d3e4f6a52d933fb8bba2888c0//#lorenz): Marcy.franklin. (2019, November 20). *How much time do you spend picking out what to watch next?*. Vox. https://www.vox.com/ad/20974139/streaming-content-movies-tv-shows-algorithm-human-choice 2.[^](https://github.com/COGS118B/group_template/blob/46e0ac6e2b28920d3e4f6a52d933fb8bba2888c0//#admonish): Consumerlab Report on TV and Media 2017 - Ericsson. (n.d.). https://www.ericsson.com/en/reports-and-papers/consumerlab/reports/tv-and-media-2017   
3.[^](https://github.com/COGS118B/group_template/blob/46e0ac6e2b28920d3e4f6a52d933fb8bba2888c0//#sota):$71K-$197K recommender systems jobs (now hiring) ... (n.d.-a). https://www.ziprecruiter.com/Jobs/Recommender-Systems