Personalized Movie Recommendation System (Group 12)

Team Members and Roles:

1. **Sanja Marinković:** Data Preprocessing, Implementation + Testing
2. **Anamarija Momić:** Model Adaptation, Implementation + Testing
3. **Amar Šehović:** Evaluation and Metrics, Implementation + Testing
4. **Dajana Tomašević:** Design Documentation, Implementation + Testing

Abstract

The MovieLens dataset provides a comprehensive collection of user and movie information, making it an excellent resource for developing advanced recommendation techniques. In this project, we aim to create a personalized movie recommendation system that predicts and suggests movies tailored to a user's preferences. This addresses the challenge of efficient content filtering in large-scale datasets. By utilizing advanced machine learning models, our goal is to enhance the accuracy of recommendations while ensuring scalability and user satisfaction. Additionally, we will develop a user-friendly graphical interface to offer intuitive and seamless access to movie recommendations, combining technical implementation with practical usability.

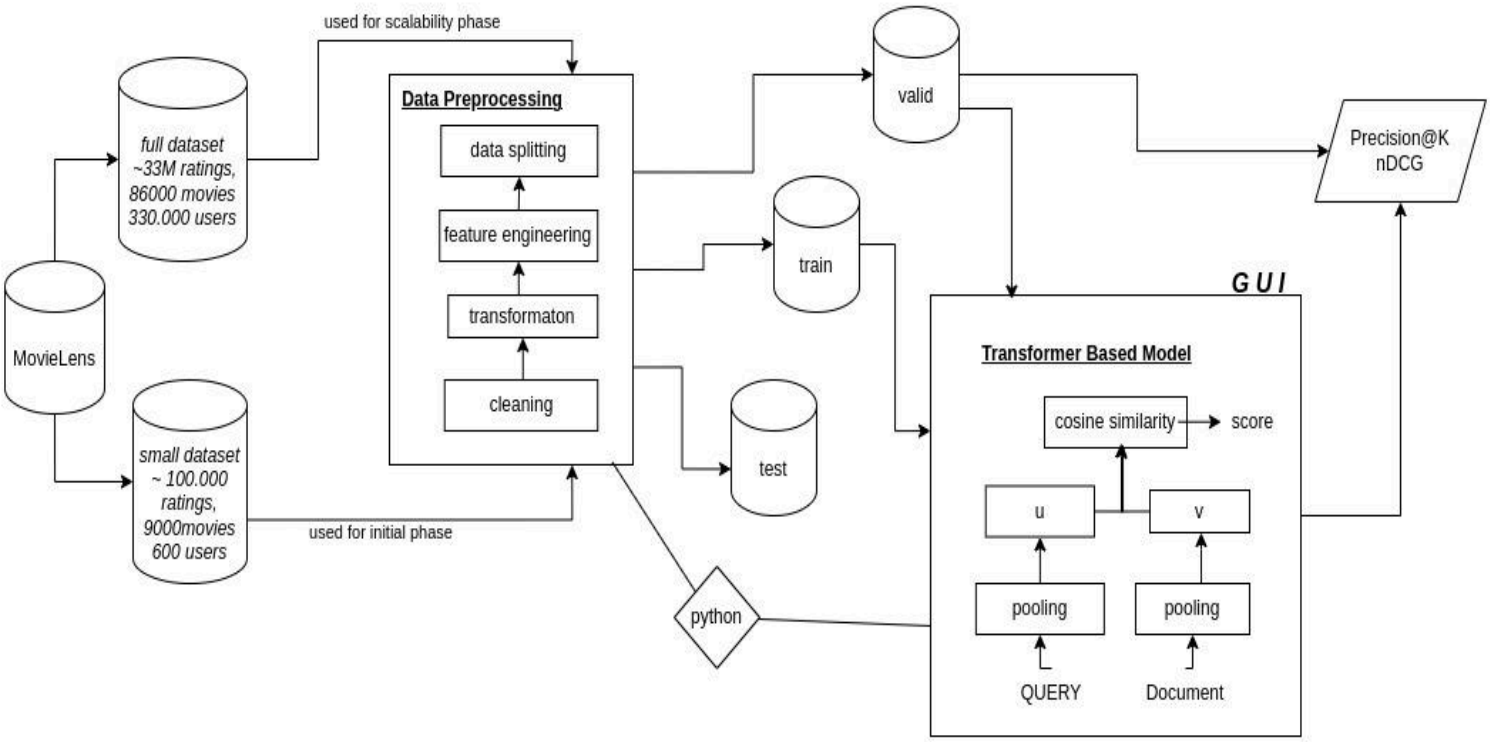

Figure 1: System utilizing the MovieLens dataset. The system begins with two datasets as shown. The data preprocessing phase involves cleaning, transformation, feature engineering, and data splitting into training, validation, and test sets. It processes user queries and movie documents through a pooling mechanism to generate embeddings. Embeddings are compared using cosine similarity to calculate scores, which rank movies for personalized recommendations. The system is evaluated using metrics like Precision@K and nDCG.

Idea

The main idea of our project is to create a personalized movie recommendation system that suggests movies based on a user's individual preferences. We plan to use advanced machine learning models, specifically transformer-based architectures to tackle challenges like handling large datasets and improving the accuracy of recommendations. Our focus is on making the recommendations as relevant and scalable as possible to ensure a great user experience and alongside the technical implementation we aim to develop a simple and user-friendly interface that allows users to easily access their movie recommendations in a smooth and intuitive way.

Main Task

The main focus of our project is to design and implement a movie recommendation system that is not only accurate but also provides an easy-to-use and visually appealing interface. We aim to use the MovieLens dataset to analyze user interaction data and predict their future preferences, but special attention will be given to the design of the **Graphical User Interface (GUI)**.

- **GUI Design**: We will create a simple and intuitive interface that displays personalized Top-5 movie recommendations. The design will focus on user experience, ensuring that the interface is clean, visual and easy to navigate. Features will include clear recommendation lists, movie details and options to explore similar movies.
- **Dataset Handling**: The MovieLens dataset will be processed to extract relevant features, but the focus will be on how this data is presented in the GUI to make it useful and understandable for the user.
- **Model Development**: Because we will use some advanced transformer-based models to generate recommendations our main goal is to ensure these recommendations are presented in a way that feels personalized and meaningful through the GUI.
- **Evaluation**: Metrics like Precision@K and nDCG will help us measure the quality of recommendations, but we will also gather user feedback to refine the interface and improve its usability.
- **Scalability Testing**: As we scale the system to handle the full dataset, we will keep the GUI responsive and functional even with larger amounts of data.

By focusing on the GUI design we aim to create a system that not only makes accurate recommendations but also enhances the overall user experience.

Dataset + Processing

Dataset Overview

For this project, we will utilize the MovieLens dataset. It is particularly suitable for this project because it provides a diverse range of features (ratings, tags, genres, and timestamps), enabling us to explore complex recommendation techniques. Specifically, we will work with the following versions:

a) MovieLens Latest Small Dataset:
Size: Contains ~100,000 user ratings and 3,600 tag applications.
Scope: Covers 9,000 movies rated by 600 users.
Usage: Suitable for quick prototyping, educational purposes, and rapid experimentation due to its manageable size.

b) MovieLens Latest Full Dataset:
Size: Contains ~33,000,000 user ratings and 2,000,000 tag applications.
Scope: Includes 86,000 movies rated by 330,975 users. Additionally, the dataset includes tag genome data with 14 million relevance scores across 1,100 tags.
Usage: Ideal for scalability tests and performance benchmarking in large-scale systems.

Firstly, we will focus on the Small Dataset to establish the pipeline and complete initial experiments. Once the foundational model is validated, we will scale to the Full Dataset to demonstrate the robustness of our approach.

Data Structure
The dataset includes the following components:
Ratings File: Contains user IDs, movie IDs, ratings (on a 1–5 scale), and timestamps.
Movies File: Provides movie titles and associated genres.
Tags File: Includes user-applied tags for movies and their relevance.
Links File: Maps MovieLens IDs to external references (e.g., IMDb, TMDb) for enriched metadata (available in the Full Dataset).

Data Preprocessing

To ensure the dataset is clean, consistent, and suitable for machine learning models, we will perform the following preprocessing steps:

1. *Data Cleaning*
   -Remove missing or invalid entries.
   -Handle duplicate records, if any.
   -Filter out users or movies with very few interactions to reduce noise.
2. *Data Transformation*
   -Encoding Features: Convert categorical variables (e.g., genres) into numerical representations using one-hot encoding or embedding layers.
   -Timestamps: Parse and convert timestamps into meaningful time features (e.g., year, month) for trend analysis.
3. *Feature Engineering :* User Profiles: Represent users based on their past ratings, preferred genres, and rating patterns.
4. *Data Splitting :*Divide the dataset into training, validation, and test subsets.

Models

In this project, we utilize Transformers to build a recommendation system that utilizes previous user interaction data (from MovieLens website) to deliver personalized movie suggestions.

Transformers are advanced models for sequential data, ideal for analyzing user movie-watching histories and generating personalized recommendations. By applying their ability to capture contextual relationships, we can predict movies suited to user preferences.

Transformers are highly effective at modeling long sequences, making them ideal for identifying patterns in users' historical preferences. Pre-trained models like BERT allow us to fine-tune them specifically on our MovieLens dataset. Additionally, Transformers are highly scalable and effective at handling large datasets and complex interaction patterns, making them a valuable choice for building recommendation systems.

In this project, movie information from a user's interaction history is first converted into numerical tokens using a pre-trained tokenizer. These tokens serve as input for the Transformer model, which analyzes the sequence of interactions to identify preferences and trends. A pre-trained Transformer model (e.g., BERT) is then fine-tuned on historical user-movie sequences, enabling it to predict future preferences and recommend i.e. Top 5 movies most relevant to the user. Finally, the recommendations generated by the model should be displayed in GUI, where users can view personalized movie suggestions based on their input.

Evaluation

To evaluate our movie recommendation system, we will use ranking-based metrics to measure the relevance of the recommendations.

1. Precision@K: Measures the accuracy of the Top-N (i.e. Top 5) recommendations by calculating the fraction of recommended movies that are relevant. This helps us evaluate how well the system identifies relevant movies within a limited number of recommendations.

2. nDCG (Normalized Discounted Cumulative Gain): Assesses both the relevance and ranking quality of the recommendations. By giving higher weight to relevant movies appearing in the ranked list, nDCG ensures that the system prioritizes the most relevant results.

For evaluation, the model will generate a ranked list of movie recommendations for each user in the test set. Precision will measure the accuracy of the recommendations, while nDCG will assess the quality of their ranking. Together, these metrics will provide a clear understanding of how effectively the system delivers personalized and relevant recommendations.
