User Needs Assessment and Data Collection Report

Huaye Li

1. Introduction

The second week of the Personalized Movie Recommendation System project focused on understanding user needs and assembling the foundational dataset required to train and evaluate the recommendation model. A thoughtful and user-aligned data collection strategy is essential for designing a recommendation system that delivers meaningful results, aligns with user behavior, and performs reliably under real-world conditions. This report outlines the user needs analysis, data source selection, ethical considerations, and preprocessing steps undertaken in this phase.

2. User Needs Assessment

2.1 Objectives

The primary goals of the user needs assessment are:

- To determine how users prefer to find and engage with movies.
- To identify features (e.g., plot, genre, rating) that influence movie preferences.
- To anticipate the functional and non-functional requirements of the recommendation system.
- To ensure the chosen data supports both content-based and behavior-driven personalization.

2.2 Methodology

Rather than collecting new data through surveys or interviews, which would have introduced additional privacy and timeline complexities, this phase employed a **persona-based design approach** alongside analysis of a large real-world dataset.

Three fictional user personas were constructed to inform system functionality:

- **Persona A**: A casual moviegoer who watches popular genres like action or comedy. They tend to use keyword-based search terms (e.g., “superhero”, “funny”).
- **Persona B**: A cinephile who appreciates critically acclaimed, high-rated films. They often search by title or director.
- **Persona C**: An explorer who prefers niche topics such as “space travel” or “philosophical drama”, and often browses based on themes or concepts.

These personas helped identify the following functional needs:

- Natural language search functionality (by keyword, genre, or concept)
- Personalized movie suggestions based on historical user ratings
- Content-based recommendations leveraging textual similarity of movie descriptions

3. Data Collection

3.1 Data Source

Data was collected from the publicly available Kaggle dataset:
"The Movies Dataset" by Rounak Banik
Link: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Two key CSV files were used:

- **movies_metadata.csv**: Contains metadata for over 45,000 movies from The Movie Database (TMDb), including fields like id, imdb_id, original_title, overview, genres, release_date, and vote_average.
- **ratings.csv**: Contains over 26 million ratings by 270,000 users across thousands of movies. Fields include userId, movieId, rating, and timestamp.

3.2 Relevance to Project Objectives

These datasets were chosen due to:

- **Scale and diversity**: High volume of user interactions across a wide variety of films
- **Rich metadata**: Textual data (overview) suitable for NLP and TF-IDF vectorization
- **Real behavior**: User ratings represent genuine viewer preferences, not hypothetical scores

4. Data Handling and Preprocessing

4.1 Data Cleaning Steps

- Converted id in movies_metadata.csv to integers, removing rows with invalid or missing entries.
- Ensured movieId in ratings.csv was of consistent integer type.
- Removed rows in movies_metadata.csv with missing overviews, or filled missing overviews with an empty string to avoid vectorization errors.

4.2 Merging Datasets

The two datasets were joined using movieId (from ratings.csv) and id (from movies_metadata.csv). After merging, the resulting dataset contained:

- userId: User who rated the movie
- original_title: Movie title
- overview: Short plot description
- rating: User rating (0.5–5.0)
- imdb_id: Used to fetch posters from the OMDb API

4.3 Text Preprocessing for Overview Field

The overview field underwent light preprocessing for future use in TF-IDF:

- Converted to lowercase

- Removed non-alphanumeric punctuation via regular expressions
- Left advanced tokenization and stopword removal to be handled during vectorization

5. Ethical Considerations

Although the dataset is public and anonymized, the following ethical principles were observed:

- **Data anonymity**: No personal identifying information is included in the dataset.
- **Non-commercial use**: The dataset is used solely for academic research and learning purposes.
- **Transparency**: Data provenance is clear, and appropriate citations are provided.
- **Poster API usage**: OMDb API access complies with rate limits and terms of service. Poster retrieval is optional and does not affect core functionality.

6. Key Insights from Exploratory Data Analysis (EDA)

6.1 Ratings Distribution

- The majority of ratings are concentrated between 3.0 and 4.5.
- Ratings below 2.0 are infrequent, suggesting users tend to rate only content they've enjoyed.
- Some users have rated hundreds or even thousands of movies, offering strong personalization potential.

6.2 Overview Availability and Quality

- Approximately 80% of the movies have valid overview text suitable for NLP analysis.
- Overview lengths vary from ~10 to ~300 words, which is well-suited for TF-IDF.
- Descriptive quality is typically good, enabling accurate semantic similarity modeling.

6.3 Cold Start Observations

- Many movies have very few or no ratings. These items can benefit from content-based rather than collaborative methods.
- Popular movies (based on rating count) are useful for default suggestions when no user history exists.

7. Summary and Next Steps

This phase has successfully prepared the groundwork for the AI model. It has identified the key features for recommendation (overview text, user ratings), ensured data cleanliness and structural consistency, and validated the ethical integrity of the project's inputs.

Next steps include:

- Selecting the development framework and AI libraries (e.g., scikit-learn)
- Installing all technical dependencies
- Preparing the working environment (e.g., Jupyter Notebook, GitHub repo)
- Testing initial TF-IDF and cosine similarity pipelines

Appendix A: Sample Rows from Merged Dataset

userId	original_title	rating	overview_snippet
1	Toy Story	4.0	Led by Woody, Andy's toys live happily...
1	Jumanji	4.0	When two kids find and play a magical...
2	Grumpier Old Men	3.0	A family wedding reignites old flames...

Appendix B: Final Merged Dataset Schema

- **userId**: Anonymized user identifier
- **original_title**: Movie name as listed in metadata
- **overview**: Plot summary of the movie, used for NLP
- **rating**: User-assigned numerical score (0.5 to 5.0)
- **imdb_id**: Unique identifier used to fetch movie poster from OMDb API