Building a Movie Recommendation System with Large Language Models

Varun Bharahti Jayakumar 002751810

July 12, 2024

1 Introduction

This report details the approach taken to build a movie recommendation system using large language models leveraging Retrieval Augmented Generation

2 Approach

2.1 Data Collection and Preprocessing

- Collected a large dataset of movie titles, genres, descriptions, and user ratings.
- Preprocessed the data to remove duplicates, handle missing values, and normalize the text for analysis.

2.2 Model Selection

- Chose Chat-Gpt-3.5-turbo state-of-the-art large language model capable of understanding and generating natural language. The model's ability to capture semantic meaning and context in text was crucial for this task.

2.3 Embedding the data

- Used Jina AI's General Purpose text embedding model to embed text to vectors.

2.4 Uploading to Vector DB

- Upserted the data in pinecone (Vector DB with metadata of each movie)

2.5 Recommendation Algorithm

- Developed an algorithm to generate movie recommendations based on user input.
- Employed cosine similarity to match user preferences with movie plots.
- Employed metadata filtering from user prompt to provide accurate results.

3 Challenges Faced

3.1 Data Quality and Quantity

Challenge: Limited and inconsistent data quality, with missing values and duplicate entries.
Solution: Implemented robust data cleaning and augmentation techniques to improve data quality.

3.2 Model Complexity

Challenge: Managing the complexity and computational requirements of large language models.
Solution: Leveraged cloud computing resources and optimized model parameters to balance performance and resource usage.

4 Overcoming Challenges

- **Data Cleaning:** Developed automated scripts to clean and preprocess data efficiently, ensuring high-quality inputs for the model.
- **Resource Management:** Used cloud-based GPU resources for model training and inference, enabling efficient handling of large-scale computations.
- **Personalization:** Combined content-based filtering with collaborative filtering to create a hybrid recommendation system, enhancing personalization and accuracy.

5 Conclusion

The project successfully built a movie recommendation system using large language models, addressing various challenges through innovative solutions. This system demonstrates the potential of combining advanced NLP techniques with traditional recommendation algorithms to deliver personalized user experiences.

6 Future Work

- Enhance model training with larger and more diverse datasets.
- Integrate real-time user feedback for continuous improvement.
- Explore additional features such as user reviews and social media interactions to further refine recommendations.

This report provides a comprehensive overview of the methods and challenges encountered in developing a movie recommendation system with large language models.