Semantic Book Recommendation System Using a Large Language Model

This project aims to develop a content-based book recommendation system powered by a Large Language Model (LLM). The main goal is to suggest books that are semantically similar to a given title by analyzing their descriptive metadata. Unlike collaborative filtering methods that rely on user ratings or interactions, this system focuses purely on the textual content and metadata of the books themselves, which allows it to work effectively even for books with few or no ratings.

The dataset used for this project comes from Kaggle and includes metadata for over 7,000 books. Each entry contains useful information such as the book's title, authors, publication year, average rating, number of pages, categories, and a short description. To prepare this data for processing, a function is used to combine all relevant fields into a single text block for each book. This textual representation acts as a synthetic summary, which captures the most relevant aspects of the book in natural language format. These summaries are essential for generating meaningful embeddings.

To convert the book summaries into numerical vectors that a machine can understand, the project makes use of a pre-trained LLM, specifically the flan-t5-base model developed by Google. This model is a version of the T5 (Text-to-Text Transfer Transformer) architecture, fine-tuned to better understand instructions and general-purpose tasks. In this context, the encoder portion of the model is used to extract embeddings from the textual inputs. By tokenizing the text and passing it through the encoder, we obtain a high-dimensional representation of the book's content. The final embedding is computed by averaging the encoder's output across the token sequence, resulting in a single fixed-size vector that encapsulates the semantic meaning of the input text.

Once the embeddings are generated for all books in the dataset, they are stored in a matrix, where each row corresponds to a single book. These embeddings serve as the core feature set for measuring similarities between books. To compare books, cosine similarity is used—a metric that measures the angle between two vectors, rather than their absolute distance. This method is particularly useful for text data, as it focuses on the orientation (semantic similarity) rather than magnitude (length of text or number of tokens).

To generate recommendations, the user is asked to input the name of a book they are interested in. The system searches for matching titles in the dataset, and if multiple candidates are found, it prompts the user to choose the correct one. Once the desired book is identified, its embedding is compared to all others using cosine similarity. The system ranks the books by similarity score and selects the top results that exceed a

minimum threshold. This threshold ensures that only highly relevant recommendations are presented, filtering out weak or unrelated suggestions.

The final output consists of a list of recommended books, each accompanied by key details such as title, authors, publication year, page count, category, and a short excerpt of the description. Additionally, the similarity score is shown to give users an idea of how closely related each recommendation is to their original selection. This helps users discover new books that align well with their interests, even if the recommended titles are obscure or previously unknown to them.

In summary, this project leverages a powerful language model to build a content-based book recommendation engine that understands and compares books at a deep semantic level. By focusing on textual features and using advanced natural language processing techniques, the system offers meaningful and personalized recommendations that go beyond surface-level matching. This approach is particularly valuable for exploring large and diverse book collections, uncovering hidden literary gems, and supporting informed reading choices.