Book Recommendation System using NLP

Group 11: Dhruv Gorasiya, Henil Patel, Om Agarwal, Tisha Patel

1. Description

In this modern world, many people prefer to read books. There is a growing challenge among readers to find books that match their interests and help them to explore various topics. With thousands of books published every year finding relevant content can be difficult. By using Natural Language Processing we can analyze the content and provide meaningful recommendations. We want to make a system which can give valuable book recommendations to readers and students which can enhance their reading based on their preference.

2. Dataset

We will use the Book Recommendation Dataset from Kaggle, which contains approximately 1.1 million book ratings across three structured tables:

1. Books
- ISBN (unique identifier), title, author, publication year, publisher
- Optional image URLs (small/medium/large cover images)

2. Users
- User-ID (anonymized), location (free-form text), and age
- Note: Age may contain null values or outliers (e.g., ages > 110)

3. Ratings
- Explicit user-book interactions with User-ID, ISBN, and book rating (scale: 0–10)

Key Characteristics:
- 271,379 books, 278,858 users, and 1,149,780 ratings
- Requires integration of metadata (authors, publishers) with user behavior
- Primary challenge: Handling sparse user-item interactions and incomplete user demographic data

Preprocessing Tasks:
- Clean invalid publication years and age entries
- Merge relational tables into a unified interaction matrix
- Address implicit vs. explicit feedback signals for recommendation modeling

3. Methodology

We plan to build a content-based book recommendation system using NLP techniques, primarily BERT embeddings and TF-IDF. The first step will involve cleaning and preprocessing the dataset to retain only English-language books with meaningful descriptions. We will then use the KeyBERT library to extract contextual keywords from each book's description, capturing the semantic core of its content. These keywords will be converted into numeric vectors using TF-IDF to quantify their relevance and uniqueness.

Next, we will compute cosine similarity between these vector representations to identify books with similar themes. Given a book title, the system will recommend top-N similar books based solely on their descriptions. We intend to use tools such as scikit-learn, NLTK, and matplotlib for preprocessing, modeling, and visualization. The final goal is to develop an interpretable and efficient recommendation engine powered by both statistical and contextual text.

4. Timeline

Week 9: Cleaning and preprocessing the data

Week 10: Exploratory Data Analysis and modelling

Week 11: Extract keywords using BERT and implement the models

Week 12: Test the system, refine adjustments and visualization

Week 13: Write project report and make the final presentation

5. Responsibilities

1. Henil Patel – Data Preprocessing & Cleaning

Explore and clean the dataset, Handle missing values, duplicates, and irrelevant entries, Prepare the final dataset for keyword extraction

2. Dhruv Gorasiya – Keyword Extraction with BERT

Use KeyBERT to extract meaningful keywords from book descriptions, Optimize extraction parameters (e.g., number of keywords, diversity)

3. Om Agarwal – TF-IDF Vectorization & Similarity Computation

Apply TF-IDF to extracted keywords, Build cosine similarity matrix for book comparisons, Generate and evaluate top-N recommendations

4. Tisha Patel – Testing, Visualization

Test system outputs and refine recommendations, Create visualizations (e.g., similarity heatmaps, keyword importance)

All the project group members will contribute to prepare the final project report.