DOCUMENTATION OF ASSIGNMENT

Data Analysis

Objective

The purpose of the data analysis was to explore and understand the user interaction and metadata of the pratilipis, extract insights, and prepare the data for building a recommendation system.

1. Dataset Overview

Datasets

I worked with two datasets:

1. User Interaction Data (user_interaction.csv):
    - Contains user interactions with pratilipis.
    - Key columns:
        - user_id: Unique identifier for users.
        - pratilipi_id: Unique identifier for pratilipis.
        - read_percentage: Percentage of the pratilipi read by the user.
        - updated_at: Timestamp of the interaction.
2. Metadata (meta_data.csv):
    - Contains metadata about pratilipis.
    - Key columns:
        - author_id: Unique identifier for the author.
        - pratilipi_id: Unique identifier for pratilipis.
        - category_name: Category of the pratilipi (e.g., Romance, Mystery).
        - reading_time: Estimated reading time in seconds.
        - updated_at and published_at: Timestamps for the metadata update and publication.

2. Exploratory Data Analysis (EDA)

2.1 Basic Information

- I checked the structure and size of the datasets:
  - user_interaction.csv contained ~3.9M rows.
  - meta_data.csv contained metadata for ~2M pratilipis.
- Data Types:
  - user_id and pratilipi_id: Integer types.
  - read_percentage and reading_time: Numeric.
  - updated_at and published_at: Datetime.

2.2 Missing Values

- User Interaction Data:
  - No missing values were present.
- Metadata:
  - The columns category_name, reading_time, meta_data_updated, and published_at had 991,943 missing values.
  - Missing rows were removed during preprocessing, as their contribution to training the recommendation system would be negligible.

2.3 Insights from user_interaction.csv

1. Read Percentage Distribution:
   - A histogram showed that most pratilipis were partially read, with only a small percentage being fully read (100%).
   - Users often stopped reading before completing the pratilipi.
2. User Interactions Over Time:
   - Interaction trends were plotted over time by grouping interactions by month (resample('M')).
   - Interaction activity showed noticeable spikes around certain periods, indicating user engagement patterns.
3. Highly Active Users:
   - The top users (by the number of interactions) were identified.

    - A small group of users contributed to a large share of interactions, reflecting a power-law distribution.

2.4 Insights from meta_data.csv

1. Category Distribution:
    - The most common pratilipi categories (e.g., Romance, Mystery) were identified. These popular categories could drive more interactions.
2. Reading Time Distribution:
    - Most pratilipis had a relatively short reading time (e.g., less than 10 minutes).
    - A long tail of pratilipis had significantly longer reading times.

2.5 Merging Datasets

- Merged Dataset:
    - Combined user_interaction.csv and meta_data.csv using pratilipi_id as the key.
    - After merging:
        - Final dataset contained 3.9M rows.
        - Added metadata features: category_name, reading_time, published_at, and author_id.

3. Feature Engineering

To enhance the dataset for model training, the following steps were taken:

1. Derived Features:
    - Interaction Frequency: Count of interactions per user.
    - Category Encoding: One-hot encoded category_name to represent pratilipi genres numerically.
    - Normalized Reading Time: Standardized reading_time to improve compatibility with similarity computations.
2. User-item Interaction Matrix:
    - Created a sparse matrix where:
        - Rows = user_id
        - Columns = pratilipi_id

        - Values = read_percentage
3. This matrix served as the foundation for collaborative filtering models.

4. Key Insights

The analysis revealed the following:

1. Reading Completion Trends:
    - Only a small fraction of pratilipis were fully read (100% read_percentage).
2. Popular Categories:
    - Romance and Mystery were the most interacted categories, likely due to broader user interest.
3. User Behavior:
    - Users tended to interact more with shorter stories (lower reading_time).
4. Time-Based Interaction Trends:
    - Interaction spikes suggested the influence of events or promotions during specific months.

EDA Visualizations

1. Read Percentage Distribution:
    - A histogram showed the spread of read_percentage values.
2. Interactions Over Time:
    - A time series plot showed the frequency of interactions by month.
3. Category Popularity:
    - A bar chart highlighted the most common categories of pratilipis.

5. Preprocessing

1. **Train-Test Split:**
    - The dataset was split into train (75%) and test (25%) based on the updated_at timestamp to preserve chronological order.
2. **User-Item Interaction Matrix:**
    - A sparse matrix was created where:
        - Rows = user_id
        - Columns = pratilipi_id
        - Values = read_percentage
    - This matrix was used for SVD-based collaborative filtering.
3. **Content Similarity Matrix:**
    - One-hot encoded category_name and normalized reading_time.
    - Combined these features and computed the cosine similarity between pratilipis.

6. Model Selection

6.1 Collaborative Filtering

- **Approach:** Collaborative filtering recommends items based on user interaction patterns.
- **Technique:** Singular Value Decomposition (SVD)
    - Decomposes the user-item interaction matrix into latent features representing users and items.
    - Predicts the interaction score for unseen user-item pairs using the reconstructed matrix.

6.2 Content-Based Filtering

- **Approach:** Content-based filtering recommends items similar to what a user has previously interacted with based on pratilipi features.
- **Technique:** Cosine similarity
    - One-hot encoding of category_name and combining it with normalized reading_time created a feature vector for each pratilipi.
    - Cosine similarity calculated how similar two pratilipis are based on their feature vectors.

6.3 Hybrid Recommendation System

- Combined collaborative filtering and content-based filtering scores to generate predictions.
- **Rationale:** Leveraged both user behavior and pratilipi content for better personalization.

3. Evaluation

Metrics Used

1. **Precision@K:**
    - Proportion of recommended pratilipis that were actually interacted with by the user.
    - Precision@K=Relevant RecommendationsK\text{Precision@K} = \frac{\text{Relevant Recommendations}}{K}
2. **Recall@K:**
    - Proportion of relevant pratilipis recommended out of all relevant pratilipis.
    - Recall@K=Relevant RecommendationsTotal Relevant Items\text{Recall@K} = \frac{\text{Relevant Recommendations}}{\text{Total Relevant Items}}
3. **Mean Average Precision (MAP):**
    - Average precision across all users, considering ranking.