Post Recommendation System

Introduction

This project focuses on developing a content-based post recommendation system. The objective is to recommend relevant posts to users based on their expressed interests and the available post metadata. Personalized recommendation systems are crucial in enhancing user engagement, improving content discovery, and increasing satisfaction in digital platforms.

The dataset consists of three key files:

- **Users.csv**: containing user information, including their top 3 interests.

- **Posts.csv**: containing post metadata such as tags.

- **Engagements.csv**: recording user–post interactions such as likes, comments, or shares.

The goal of the system is to connect users with posts most aligned with their interests by leveraging natural language processing and similarity-based ranking.

Approach

The recommendation pipeline is built using the following steps:

1. **Data Preparation**

   - User profiles were extracted from `Users.csv`, focusing on the `top_3_interests` field.

   - Post metadata was obtained from `Posts.csv`, particularly the `tags` field.

2. **Feature Representation**

   - A **TF-IDF** (Term Frequency–Inverse Document Frequency) Vectorizer was applied to capture the importance of terms in user interests and post tags.

   - This representation allows the system to compare textual information between users and posts effectively.

3. **Similarity Computation**

- Pairwise cosine similarity was computed between user-interest vectors and post-tag vectors.

- This similarity score serves as the basis for ranking posts for each user.

4. **Recommendation Generation**

- For each user, posts were ranked by descending similarity scores.

- The top 3 posts per user were selected as recommendations.

- Results were saved in CSV format for further analysis.

Metrics

The current implementation primarily computes similarity scores without explicit evaluation against ground-truth user interactions. While it demonstrates the feasibility of content-based matching, a rigorous assessment requires formal metrics. Suggested evaluation methods include:

- **Precision**: measures the fraction of recommended posts among the top K that are relevant.
- **Recall**: measures the fraction of relevant posts that are successfully recommended within the top K.
- **Mean Average Precision (MAP)**: provides a holistic view of ranking quality across users.
- **Normalized Discounted Cumulative Gain (NDCG)**: accounts for ranking order and relevance levels of recommended posts.

By comparing recommendations against actual user engagements from `Engagements.csv`, these metrics can validate system performance.

Possible Extensions

Several enhancements can be introduced to improve the recommendation quality:

1. **Hybrid Recommendations**

- Combine content-based filtering with collaborative filtering based on engagement patterns.

- This would help capture similarities among users and exploit implicit feedback.

2. **Advanced Embeddings**

   - Replace TF-IDF with pretrained language models such as Sentence-BERT to capture semantic similarity beyond keyword overlap.

3. **Incorporating Engagement Weights**

   - Weight similarity scores with engagement statistics (likes, comments, shares) to prioritize posts with higher relevance and quality.

4. **Temporal Dynamics**

   - Incorporate recency by applying decay functions so that newer posts are prioritized in recommendations.

5. **Feedback Loop**

   - Enable users to upvote/downvote recommendations, thereby refining the model iteratively using reinforcement learning techniques.

Conclusion

The current system demonstrates a simple but effective content-based approach using TF-IDF and cosine similarity. While it produces reasonable recommendations, its true potential lies in integrating engagement-based signals, advanced embeddings, and feedback-driven improvements. Introducing robust evaluation metrics will be key in quantifying and benchmarking its effectiveness.