Book Recommendation System Evaluation Report

1. Introduction

This report presents the evaluation results of our Book Recommendation System, which uses two primary methods for generating recommendations: FAISS (Facebook AI Similarity Search) and Cosine Similarity. The evaluation aims to assess the system's performance, identify strengths and weaknesses, and propose improvements.

2. Evaluation Methodology

By creating a script (scripts/evaluation.py) and using it to evaluate and assess performance we included two methodologies.

2.1 First Methodology

We gathered our test set from the dataframe that contains all books, and calculated user-centric metrics to evaluate how good the system is.

2.1.1 Metrics

We used the following metrics to evaluate our recommendation system:

1. **Hit Rate**: The proportion of times the actual book appears in the top K recommendations.
2. **Diversity**: The uniqueness of recommendations, calculated as the ratio of unique titles to total recommendations.
3. **Serendipity**: The unexpectedness of recommendations, measured by how different they are from the user's known preferences (which are title input in our case).

2.1.2 Evaluation Process

For each book in the test set:

1. We used its description as input to our recommendation system.
2. We generated recommendations using both FAISS and Cosine Similarity methods.
3. We calculated our evaluation metrics for both methods.

2.2 Second Methodology

We created a test set containing three descriptions and their expected recommendation titles based on their similarity from our books dataset, so we can compare results of both methods with the predefined recommendations.

2.1.1 Metrics

We used three key metrics to evaluate the performance of our recommendation system:

1. **Precision**: The proportion of recommended items that are relevant.
2. **Recall**: The proportion of relevant items that are recommended.
3. **Mean Reciprocal Rank (MRR)**: A measure of how high the first relevant item appears in the recommendation list.

2.1.2 Evaluation Process

For each book in the test set:

1. We used its description as input to our recommendation system.
2. We generated recommendations using both FAISS and Cosine Similarity methods.
3. We calculated our evaluation metrics by comparing recommended titles with expected titles.

3. Results

Metric	FAISS	Cosine Similarity
Hit Rate	1.00	1.00
Average Diversity	0.93	0.93
Average Serendipity	0.75	0.74
Precision	0.67	0.67
Recall	0.67	0.67
MMR	1.00	1.00

4. Analysis

4.1 First Methodology

4.1.1 Strengths

1. **Perfect Hit Rate**: Both FAISS and Cosine Similarity methods achieved a perfect hit rate of 1.00, indicating that the system consistently includes the actual book title in its recommendations when given its description.
2. **High Diversity**: With an average diversity of 0.93 for both methods, the system provides a wide range of unique recommendations, avoiding repetitive suggestions.
3. **Good Serendipity**: The serendipity scores of 0.75 (FAISS) and 0.74 (Cosine) suggest that the system provides a good balance between familiar and unexpected recommendations.

4.1.2 Weaknesses

1. **Minimal Differentiation Between Methods**: There's very little difference in performance between FAISS and Cosine Similarity methods, which may indicate redundancy in our approach.
2. **Lack of Genre-Specific Analysis**: The current evaluation doesn't provide insights into how the system performs across different book genres or categories.

4.2 Second Methodology

4.2.1 Precision and Recall

The precision and recall scores of 0.67 indicate that our system is performing reasonably well, but there is room for improvement. This score suggests that for every 3 recommendations:

- 2 are relevant (true positives)
- 1 is irrelevant (false positive)
- 1 relevant book is missed (false negative)

4.2.2 Mean Reciprocal Rank (MRR)

The perfect MRR score of 1.00 for both algorithms is excellent. This indicates that in all test cases, the first recommended book was always relevant. This is a strong point of our system, showing that when it does recommend relevant books, it tends to rank them highly.

4.3 FAISS vs Cosine Similarity

The performance of FAISS and Cosine Similarity methods is nearly identical across all metrics. This suggests that both methods are equally effective for our current dataset and evaluation approach

5. Areas for Improvement

Based on our analysis, we recommend the following improvements:

1. **Expand Test Set**: Increase the size and diversity of the test set especially in the second methodology which is based on only three test queries. A larger and more diverse set of test queries would provide a more robust evaluation of the system's performance.
2. **Fine-tune embeddings**: Investigate if using domain-specific pre-trained models or fine-tuning the embedding model on book descriptions could improve the quality of recommendations.
3. **Genre-Based Evaluation**: Introduce genre classifications and evaluate performance across different genres to gain more nuanced insights.
4. **Explore Alternative Metrics**: Introduce additional metrics such as ranking-based measures (e.g., NDCG) to provide a more nuanced evaluation of recommendation quality.
5. **User Studies**: Conduct user studies to validate that our high-performing metrics translate to actual user satisfaction.

6. Future Work

To further enhance our Book Recommendation System, we propose the following areas for future work:

1. **Expand Books Dataset:** Currently with 2400 books, the dataset needs to be more diverse and contain more books.
2. **Hybrid Approach**: Explore combining content-based methods (current approach) with collaborative filtering to potentially improve serendipity.
3. **Personalization**: Develop user profiles and personalized recommendation strategies to improve the relevance of recommendations for individual users.
4. **Genre and Topic Modeling**: Implement genre classification and topic modeling to provide more targeted recommendations and enable genre-specific performance analysis.
5. Experiment with different embedding models and fine-tuning approaches.

7. Conclusion

This Book Recommendation System demonstrates great performance across all evaluated metrics, with perfect hit rates and high diversity and serendipity scores for book title recommendations. Both FAISS and Cosine Similarity methods perform equally well. However, this good performance warrants further investigation to ensure it generalises to real-world scenarios and new books. The next steps should focus on validating these results with larger, more diverse datasets, implementing genre-based analysis, and conducting user studies to ensure that the system's statistical performance translates to user satisfaction.