ENHANCING CONTENT-BASED

Recommender System with Modern Sentence Embedding Methods

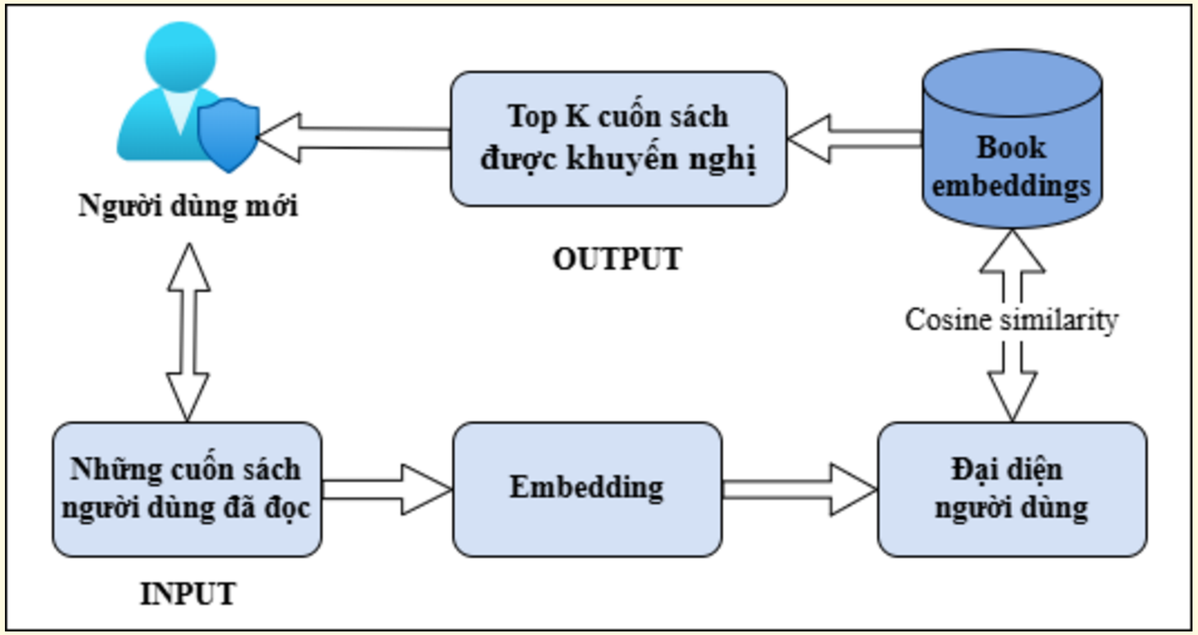
GVHD: Huỳnh Văn Tín

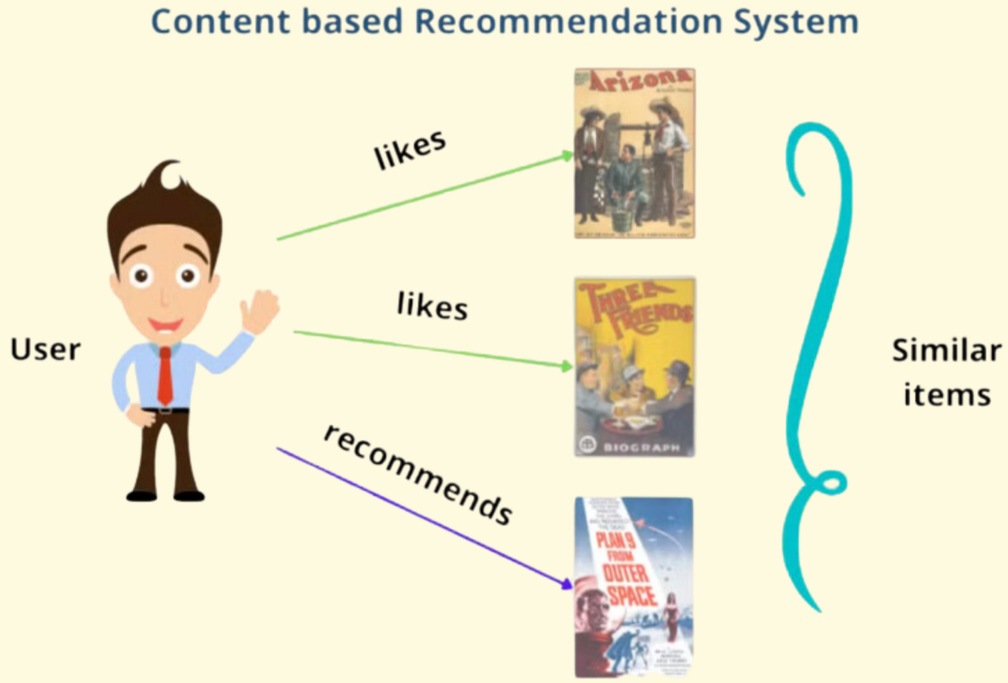
Team 05

CONTENT TABLE

INTRODUCTION

This study tackles the "cold start" problem in recommender systems by evaluating traditional and advanced embedding methods, including TFIDF, ST5, SBERT, and SGPT, on the Goodreads dataset.





RELATED WORD

- Cross-domain approach: Combining information across domains to alleviate data sparsity, as proposed in "Cross-Domain Collaborative Filtering".
- LinUCB algorithm: Uses user ratings as context to enhance recommendations \cite{article_v1}.
- **Sentence embedding**: More effectively captures context with models like BERT “**Sentence-BERT: Sentence embeddings using Siamese BERT-networks**”

Challenges: Cold start problem; Sparse and heterogeneous data; Limitations of word embeddings.

Current goals: Exploit sentence embeddings and deep learning to improve recommendation quality and address the cold start problem.

GOODREADS DATASET

Preprocessing
Lowercasing
Tokenization
Removing stopwords, extra whitespaces
Stemming & Lemmatization

Dataset	Feature of Dataset
MetaBooks (comic & graph genre only)	1. book_id: numeric 2. author_id: numeric 3. author_name: text 4. description: text 5. title: text
UsersHistory (book in comic & graph genre only)	1. user_id : numeric 2. book_id: numeric 3. ratings: numeric (1-5)

Dataset3 (Des + Title + Author)	
Total documents (D)	74,125
Vocabulary size (V)	302,159
Average number of words per document (Nd)	68

EXPERIMENTAL METHOD

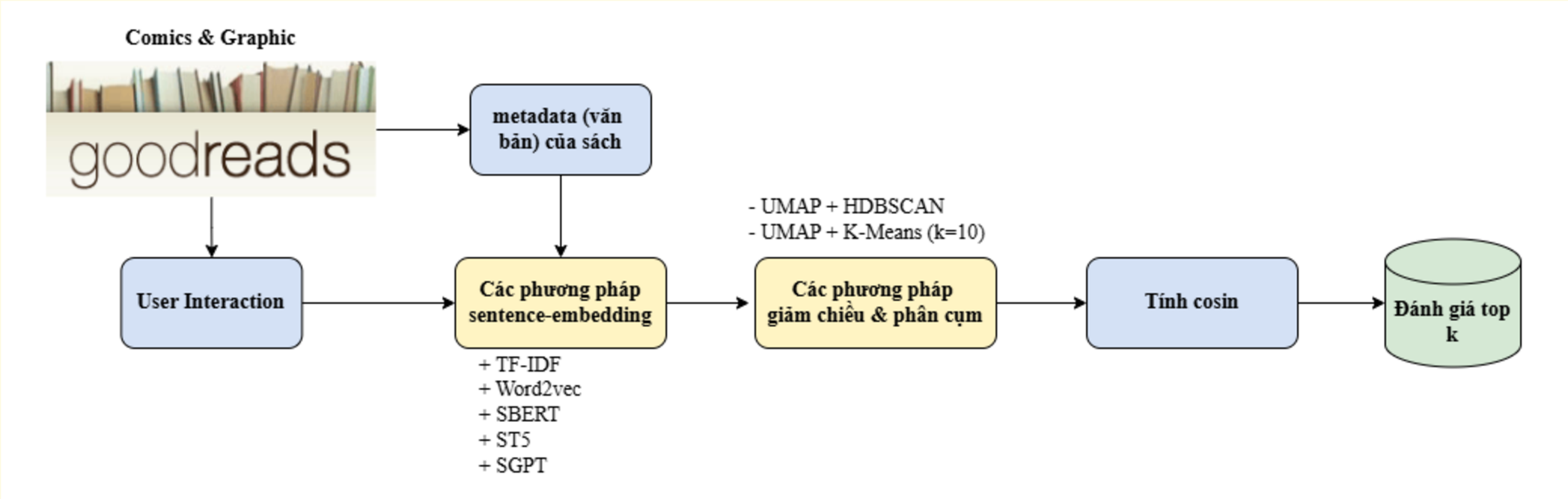


Figure 1: Embedding Phrase

EXPERIMENTAL METHOD

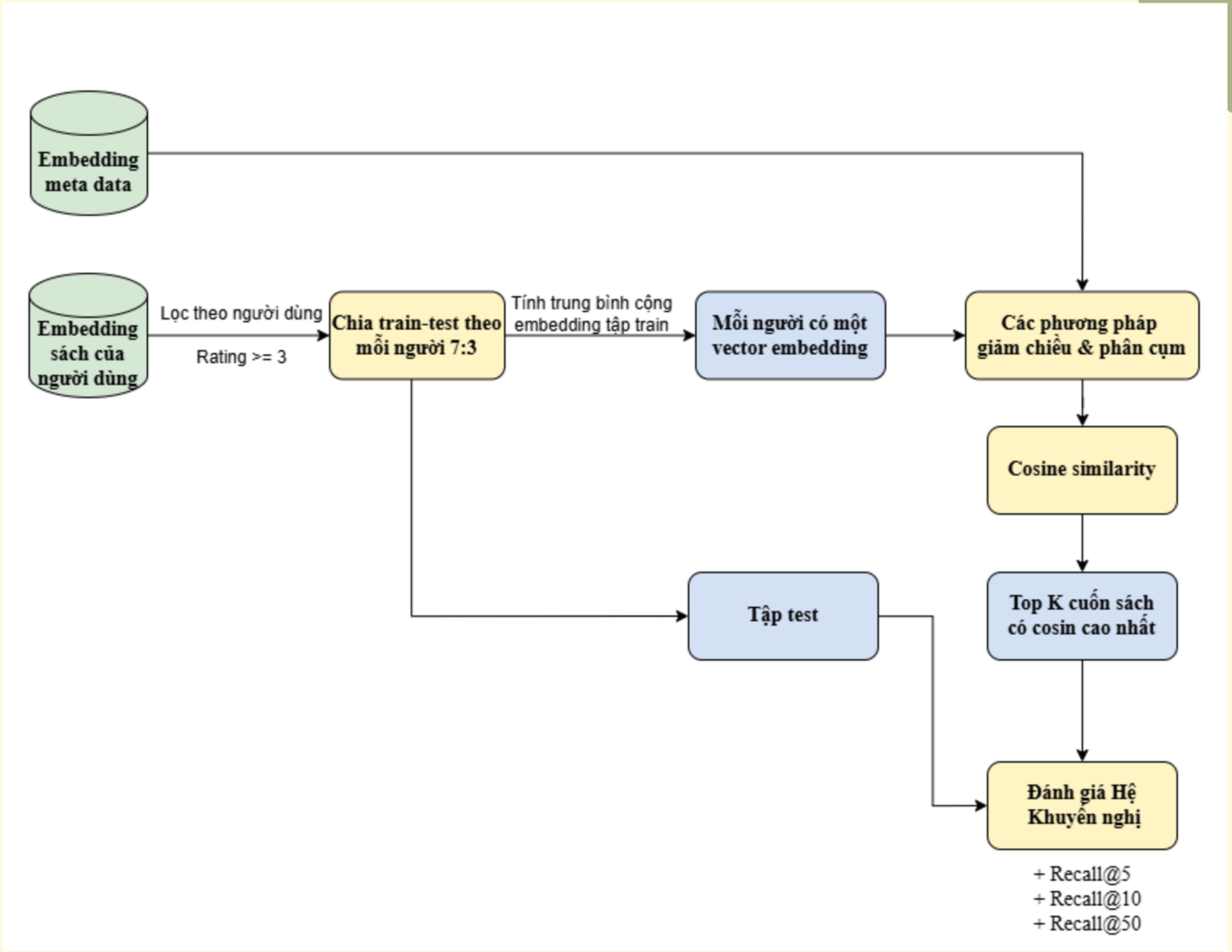


Figure 2: Evaluation Phrase

METHODOLOGY

Embedding	
TFIDF	A survey (Beel et al., 2015) showed that 83% of text-based recommender systems in digital libraries used TF-IDF.
ST5	A powerful word embedding model showcasing its efficacy across diverse Natural Language Processing tasks, including the embedding of textual features.
SBERT	Short for Sentence Embeddings using Siamese BERT-Networks, is a method for creating meaningful representations of sentences using the powerful BERT language model.
SGPT	Utilizes GPT's strengths to convert sentences into dense vectors, reflecting semantic meanings.

Topic modeling
UMAP
HDBCAN (for recommendation task)
K-Means (k=10) (for recommendation task)

Recommend
Recall@K

RESULT

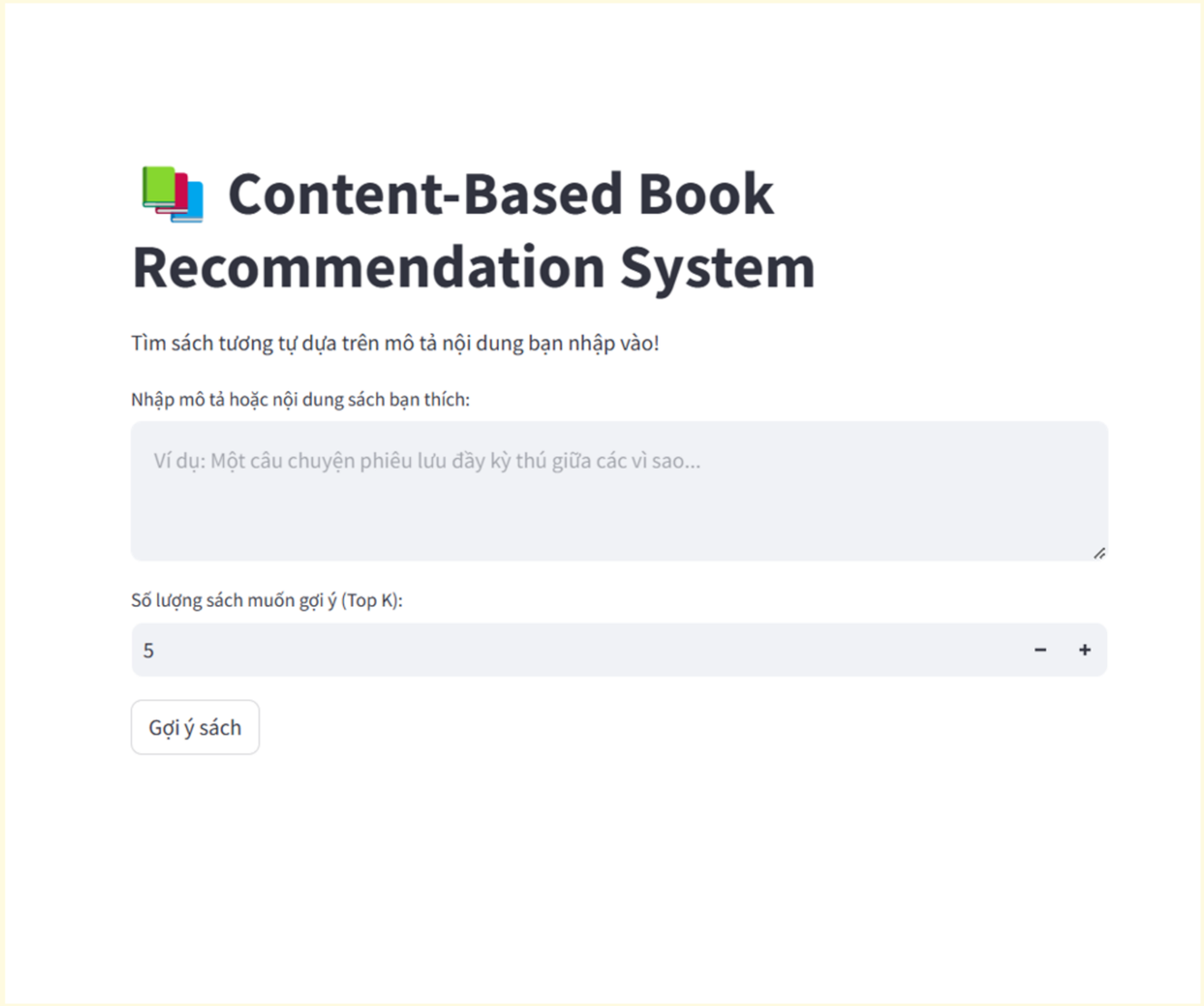
Embedding	Topic	Description only			Title + Description		
		Recall@5	Recall@10	Recall@50	Recall@5	Recall@10	Recall@50
TF-IDF	UMAP + HDBSCAN	0.1353	0.1624	0.2197	0.1081	0.1299	0.1719
ST5		0.2278	0.2305	0.2409	0.1910	0.2056	0.2236
SBERT		0.2357	0.2430	0.2549	0.2277	0.2324	0.2468
SGPT		0.2474	0.2569	0.2802	0.2394	0.2478	0.2698
TF-IDF	UMAP + K-means (k=10)	0.2756	0.3251	0.3889	0.2575	0.3173	0.3883
ST5		0.2296	0.2703	0.3905	0.2197	0.2694	0.3914
SBERT		0.2569	0.3224	0.4428	0.2595	0.3199	0.4346
SGPT		0.2682	0.3380	0.4757	0.2967	0.3519	0.4373

RESULT

Embedding	Topic	All		
		Recall@5	Recall@10	Recall@50
TF-IDF	UMAP + HDBSCAN	0.0846	0.1101	0.1549
ST5		0.2070	0.2171	0.2321
SBERT		0.2358	0.2387	0.2529
SGPT		0.2381	0.2494	0.2682
TF-IDF	UMAP + K-means (k=10)	0.2477	0.3057	0.3798
ST5		0.2315	0.2804	0.4058
SBERT		0.2586	0.3276	0.4531
SGPT		0.2853	0.3633	0.4768

DEMO

We develop a simple Streamlit web app for a content-based recommendation system using ST5 embeddings. Users can input keywords or book titles, adjust the number of recommendations (k), and receive the top k books with details like title, cover, description, and cosine similarity.



CONCLUSION

We built a book recommendation system using the Goodreads dataset, incorporating 8 configurations from 4 embedding methods (TFIDF, ST5, SBERT, SGPT) and 2 clustering techniques (HDBSCAN, K-means). The results demonstrate that modern sentence embeddings enhance recommendation accuracy and have the potential to replace traditional methods.

Limitations: Clustering instability, especially with HDBSCAN, may impact result consistency. Future work will focus on dimensionality reduction and improving stability.

THANK YOU

Nhóm 2