509496 - Information Retrieval and Recommender Systems

31 January 2025

PERSONALIZED INFORMATION RETRIEVAL

SE-PQA dataset

Barbieri Chiara - 517096
Sotgia Francesca - 513067

How could we improve a model's accuracy at predicting relevant answers for a user's question in a community Q&A system?

Personalized search engine

Goal: develop a search engine tailored for a community Question Answering dataset (SE-PQA).

Implement traditional methods with advanced techniques to improve the model's ability of producing user-tailored recommendations.

Expected results: personalization improves the model's accuracy at predicting relevant answers.

Intuition of the idea

Main idea

Traditional **probabilistic IR models** are good rankers but lack of deep understanding of relationships between items.

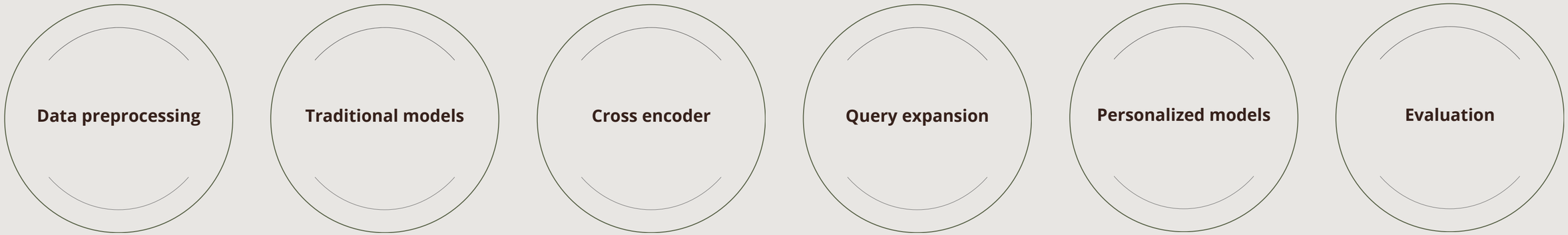
Implement:

- **Neural re-rankers** to implement contextual undestranding,
- **Personalization models** to retrieve user-tailored recommendations.

Expectations

After implementing said models, we expect improved performance of the model in retrieving relevant answers.

Methodology



Dataset

SE-PQA dataset consists of:

- User data
- Question
- Answers
- Metadata

The dataset is divided into train, validation and test subdirectories.

Preprocessing

We defined two preprocessing functions to have the text written in a consistent way.

- *preprocess_text_basics*: lowercasing, remove numbers, links and extra spaces.
- *preprocess_text_norm*: tokenization, stop word removal and stemming.

Convert the dataframes to be Retriv-friendly.

Model selection

We defined two SparseRetrievers to index models and compare performances.

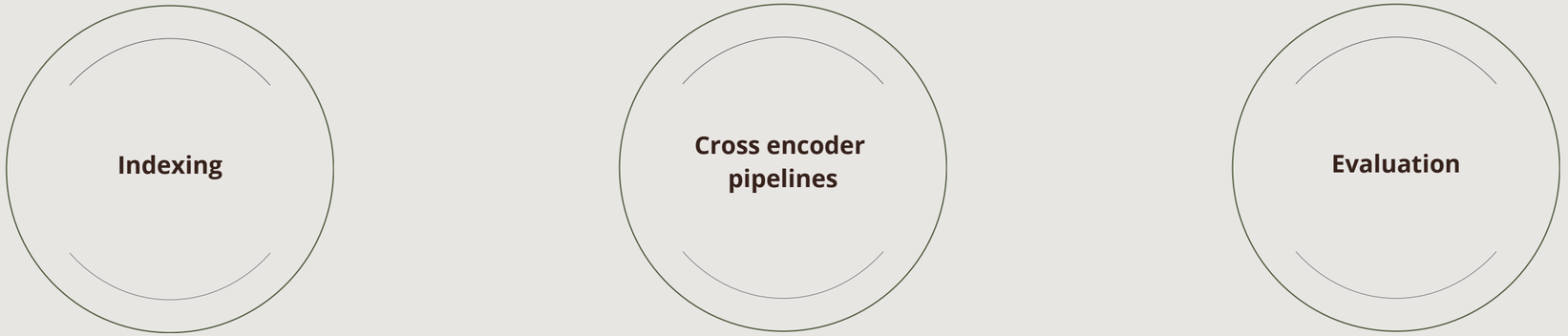
- **BM25 Retriever**: ranks documents based on their relevance to a given query.
- **TF-IDF Retriever**: ranks documents based on the frequency of their keywords.

BM25 returns better results and is thus the one to be applied.

Model	MAP@100
BM25	0.721
TF-IDF	0.380

Cross Encoder

We implement a Cross Encoder to verify question-answer pairs to improve the model's performance.



Advantages:

- Understand query-answer relationships.
- Contextual understanding.
- Better performance.

Query expansion with LLM

Goals: refine the queries and improve generalization and efficiency.

Expand queries

- Expand the queries using a pre-trained LLM and the users' data and context.

Dataset update

- Update the dataset, introducing a new column.
- Divide into train and test.
- Define qrels.

Evaluation

- Run an experiment.
- Worse results.

Personalization models

Goals: retrieve recommendations tailored to individual preferences and collective behavior.

Content based filtering

CBF suggests relevant answers for a user based on the characteristics of the answers and the user's context.

Collaborative filtering

User-based CF suggest answers based on what similar users interacted with.

Once we calculated the personalized score we evaluated the model's performance .

Results

Using the Cross Encoder, we obtained good results.
By implementing the query expansion and personalization the model’s performance at retrieving relevant documents decreases.

Model	MAP	nDCG
Baseline	0.8275	0.8568
Expanded queries	0.5793	0.6381
Personalized	0.4461	0.5306

Conclusion

In contrast with our expectations, the personalized model has a lower efficacy at retrieving relevant answers compared to the baseline model.
The outcome can be attributed to:

- Introduction of **noise** by the LLM during query expansion,
- Potential introduction of **biases** during the personalization.

Future implementations

Q&A

Barbieri Chiara - 517096

Sotgia Francesca - 513067

THANK YOU

Barbieri Chiara - 517096
Sotgia Francesca - 513067