Personalized Information Retrieval: Course Project Proposal

1. Goal

The goal of this project is to develop a search engine capable of retrieving relevant answers to user queries from a community Question Answering dataset. The engine will integrate personalization by tailoring search outcomes based on the user's historical activity and contextual features.

2. Project Description

This project emphasizes practical and theoretical aspects of **personalized information retrieval** (PIR). Students will leverage their knowledge of statistical retrieval methods (TF-IDF, BM25) and neural re-rankers (bi-encoders, cross-encoders). They will explore personalization using user-level features and recommender system scores. A specific focus will be integrating these components effectively. The project tasks include:

1. **Query Expansion:** Use LLMs to enrich user queries based on contextual or historical data.

and/or

2. **Personalization:** Incorporate user preferences, historical interactions, or tags from SE-PQA. with, **score Integration**, i.e, combining retrieval scores (e.g., BM25, neural rerankers) with personalization scores. The personalization model can be inspired from the recommender systems.

Evaluation: Benchmark models using precision and MAP and nDCG metrics. The choice of the metrics and the cutoff has to be justified in the final report.

Dataset:
A subset of the **SE-PQA dataset**, containing a manageable number of questions and answers, will be provided. This subset retains features like tags, reputation, and social and historical metadata to enable personalization.
https://drive.google.com/file/d/1HhgXzyEpsZNcenU9XhJuOYyDUKEzUse4/view?usp=drive_link
To download in google colab use the following command:
gdown 1HhgXzyEpsZNcenU9XhJuOYyDUKEzUse4

3. Participant Guidelines

- **Eligibility:** Open to all students enrolled in the course.
- **Team Formation:** Teams of 2-3 students must register by 28/12/2024, with one member designated as the communication lead.

4. Steps For completion

4.1 Timeline

1. **Team Registration:** Deadline - 28/12/2024, via email with the following subject: [Team Registration IR and RS]. List all the members of the team with enrollment number and Name and Surname in the following way.
    a. Team members:
        i. Mario Rossi **Communication Leader** - 828282
        ii. Pranav Kasela - 245632
        iii. Georgios Peikos - 659705

    If you have sent your team name previously, send it again with the correct subject.

2. **Phase I Development:** Implement retrieval and baseline neural models.
3. **Phase II Development:** Extend models with user features for personalized information retrieval (query expansion, recommender systems).
4. **Final Submission and Presentations:** Deadline - 20/01/2025

4.2 Development and Implementation: Phase I

1. **Baseline Retrieval:** Implement and analyze BM25 or TF-IDF models with SE-PQA data.
2. **Neural Reranking:** Incorporate cross-encoder or bi-encoder (re-)rankers for improved result ranking.
3. **Evaluation Metrics:** Evaluate using recall, precision, MAP and nDCG.

4.3 Development and Implementation: Phase II

1. **Query Expansion:** Use LLMs to suggest query refinements or expansions based on user data and context.
2. **Personalization Models:** Integrate user-level data, such as tags or historical queries, into the search pipeline.
3. **Recommender Systems:** Implement and train recommenders using SE-PQA data. Combine recommender scores with retrieval scores.
4. **Evaluation:** Reassess the models after adding advanced features, compare it to the baselines.

4.4 Project Report

The final project submission must contain the following.

- Pdf reporting what you did (Introduction, Intuition of your idea, methodology, experiments and results, conclusions, i.e., what did you learn). Max pages 4
- Source code (Colab notebook links are fine)
- Presentation in pdf to be presented on a date to be defined.

5. Assessment Criteria

Credits. The entire project carries a total of 3.5 credits. The initial part, comprising the proposal of your ideas and the development in Phase I, contributes to 2 credits. The subsequent Phase II is allocated the remaining 1.5 credits.

Projects will be graded on:

- **Functionality:** Quality and accuracy of search and personalization results.
- **Integration:** Effective combination of retrieval, query expansion, and recommender system features.
- **Technical Implementation:** Code quality and model design.
- **Team Collaboration:** Teamwork and communication.
- **Presentation:** Clarity and completeness of written and oral reports.

6. Additional Resources

- **SE-PQA Paper:** Details on dataset structure and baseline methods. https://arxiv.org/abs/2306.16261
- **IR Tools:** PyTerrier, HuggingFace Transformers, libraries for LLM integration.
- For any additional information or support, please send an email at: [redacted]