Program Report: Information Retrieval for Legal Documents

Murat Sahin (`m.sahin@bilkent.edu.tr`)
Mehmet Can Sakiroglu (`can.sakiroglu@bilkent.edu.tr`)

December 19, 2023

Project Structure

/Infer_Eval

- Evaluation.ipynb
- Inference.ipynb

/Preprocessing

- 1_Json_to_CSV.ipynb
- 2_CSV_to_Pickle.ipynb
- 3_Extract_Query_Document_Pairs.ipynb
- 4_Extract_BERTurk_Embeddings.ipynb
- 5_Preprocess_for_BM25.ipynb
- 6_Extract_BM25_Indexes.ipynb

/UI

- Inference.py
- UI.py
- templates/index.html

Infer_Eval

Contains code for final evaluation. `Evaluation.ipynb` loads precomputed embeddings and indexes, then evaluates for all test queries. `Inference.ipynb` does the same but computes for a given new query.

Preprocessing

Contains code for all preprocessing steps in order. Includes initial preprocessing of the dataset, query-document pair extraction, embedding, and index computations.

UI

Contains code for the user interface. `Inference.py` is a class version of `Inference.ipynb`. `UI.py` creates the user interface and uses an `Inference` instance in the backend.

Note

The dataset is available on Kaggle:
`https://kaggle.com/datasets/59eae2354b71cc2e3474ed78d789bc7fab457dc06fe8456368b2d7b3d1efe606`