READ ME

Assignment.py, QueryModule.py and correlate\_files.py Python scripts implement a text processing and information retrieval system for document collection. It comprises three main components: document pre-processing(Assignment.py), query processing(QueryModule.py), and result visualization(correlate\_files.py).

All the document corpus and queries used can be found in the files folder.

**Document Pre-processing:** The Assignment.py script reads a collection of documents from a specified folder(Edit your own file paths for E.g.“/Users/rohinmehra/Desktop/Assignment1/files/IRW1.txt”), and pre-processes the text by removing special characters, converting to lowercase, eliminating stop words, and applying stemming. It then computes the TF-IDF matrix for the documents and builds an inverted index associating terms with document IDs.

**Query Processing:** For query processing, the QueryModule.py script reads queries from a file(/Users/rohinmehra/Desktop/Assignment1/files/IR1\_Queries.txt)( Edit your own file paths for these files in the code ), pre-processes them similarly to the documents, and computes TF-IDF vectors for the queries using a precomputed TF-IDF matrix for the document collection. The resulting query vectors are written to an output file.

**Result Visualization:** The correlate\_files.py script reads the inverted index(/Users/rohinmehra/Desktop/Assignment1/inverted\_index.txt) and query vectors(/Users/rohinmehra/Desktop/Assignment1/query\_vector.txt) from files, displaying relevant information for each query. This includes the query vector, the top terms and their TF-IDF weights in the query(/Users/rohinmehra/Desktop/Assignment1/tfidf\_matrix.txt), and documents related to the query based on the inverted index.

Please run files in the following sequence;

1. Open IDE (PyCharm preferred) and Open the 3 files (Assignment.py, QueryModule.py and correlate\_files.py)
2. Run Assignment.py status 0 will come on the terminal.
3. Run QueryModule.py status 0 will come on the terminal.
4. Run correlate\_files.py, this will show the data and the related document list with all five queries.

A screenshot of a computer

Description automatically generated