IR Project Report

Saif Abushanab

202201838

1. Data Collection: Upload the data set file and read it as csv.
2. Preprocessing:
3. Cleaning: function to clean the data from special characters and twitter characters
4. Remove stop words: function to remove stop words (a , the , on , of …)
5. Stemming: convert the words to its original and lower case the words to make it easy to the machine to understand these words
6. Indexing : indexing the data
7. Query Processing:
8. enter any query to search for it on the collection and returns how many time this query occurred and in which documents.
9. Then define TF-IDF model to rank the retrieved documents with the highest occurrence.
10. Then rerank the TF-IDF model with the mono-bert model to check its ranking.
11. Query Expansion: used BM-25 model and RM-3 expander
12. User Interface: made a flask page to search with any query and will return the results and documents.
13. Evaluation: evaluate the quality of the search results I retrieved with the relevance documents to a query

A screenshot of a search engine

Description automatically generated