**Project Data Mining Report**

1. **Data Collection:** I use ir\_datasets to download Vaswani documents, queries, and Qrels, then I iterate over each single document and add it in a dataframe using pandas to deal with it easily.I did the same for queries and Qrels.
2. **Preprocessing:**I have imported all important libraries used for the preprocessing stage to complete all processes(tokenization, removing stop words, steeming, and cleaning) and apply each function on the documents dataframe.
3. **Indexing:** I use the index method to build an inverted index for each term and its document id and its frequency over the whole collection.
4. **Query Preprocessing:**I have created a preprocess function to apply preprocessing steps on the (query) to retrieve and match the input query with its relevant documents , then I use the TF-IDF model to rank the relevant documents for the input query.
5. **Query Expansion:** I have applied relevance feedback for each query by analyzing top-ranked documents using the rm3\_ expander function,then I have printed all the analyzed documents and expanded queries.In addition I use Glove embeddings model and ELMO to check embeddings for synonym words and similarity between them in different texts.Additionally, I have used BERT model in my code for embeddings.
6. **Evaluation:** I use the Qrels dataframe to evaluate the results by calculating (Precision , MAP , Recall).
7. **User Interface:** I use a Flask web application with two routes:for the home page and search for processing search queries. When a query is submitted, the search results are generated and displayed on a separate HTML page.