5. Indexing + Document Retrieval

Task

- Download an existing IR dataset - Cranfield collection
    - Or you can use already preprocessed data cranfield.zip (d-documents, q-queries, r-relevances (a set of relevant document ids for each query id))
- Represent each document and query using the Vector Space Model with all following weightings:
    - Use Binary representation
    - Use pure Term Frequency
    - Use TF-IDF
    - Try existing model for text/document embeddings (or text/sentence similarity model from Hugging Face)
- Compute relevance scores for each combination of query, document
    - Use Euclidean distance
    - Use Cosine similarity measure
- Evaluate quality and difference of both scores and different weighting schemas
    - Compute Precision, Recall, F-measure (you can limit to top N relevant documents for each query)