Task 1

ColBERT (Contextualized Late Interaction over BERT) is a new approach to information retrieval that uses BERT-based models. Unlike traditional methods that either compute a single vector representation for both queries and documents or compute query-document interaction for each pair, ColBERT encodes queries and documents separately into multiple vectors per token. ColBERT then uses "late interaction" where the relevance between a query and a document is computed by matching the token-level representations. The advantage of this approach is that it still retains the context depth provided by BERT while being able to pre-compute the document vectors, which speeds up the retrieval process as the expensive interaction phase is done only in the final matching stage.

In the "late interaction" phase, ColBERT uses a max-similarity operation to calculate the final relevance scores, which evaluates the highest similarity scores across token pairs between the query and the document. This offers a balance between efficiency and retrieval effectiveness. Thanks to these optimizations ColBERT can be used for real-time search in large-scale datasets.

ColBERT differs from both Bi-encoder and Cross-encoder models in the query-document interactions. Bi-encoders generate independent embeddings for queries and documents and compute relevance based on cosine similarity between these embeddings. This approach is fast but cannot capture deeper semantic relationships. Cross-encoders compute interactions by processing the query and document together through a single BERT model. This allows to capture the contextual relationship but at a higher computational cost. ColBERTcombines these methodologies by pre-computing deep context embeddings separately and then using them for interaction during the retrieval phase. This optimizes both performance and efficiency.