**Methodology**

Once the model receives a query from a user, the query is fed from Rasa to a semantic search pipeline, which involves two retrieval models: BM25 and a Sentence Transformer. This is where the model decides on an appropriate response and sends it back to Rasa, before its delivered to the user. Potential responses consist of wiki summaries as seen in the collated raw\_outputs.csv, where each summary is given a fitness coefficient by the search pipeline. After the summaries with the highest coefficients are identified, given that a coefficient exceeds a threshold of 0.4, the top response is sent back to Rasa.

The dataset, stored as a CSV file (raw\_outputs.csv), contained multiple generated summaries per row in list format. These were parsed using ast.literal\_eval (Python [1]) to convert the strings into list objects. The entries that were fit for further processing were kept, with a few anomalous examples being redacted from the dataset.

Each summary was preprocessed with a tokenise function that removed punctuation, made all letters lowercase, and sorted by whitespace. This step prepared the data for Bag-of-Words (BoW) (Zhang et. Al [2]) tokenisation – a crucial part of the BM25 ranking algorithm used. BM25, and specifically, Okapi Best Matching 25 (Brown et. Al [3]), is a term-weighting mechanism commonly used in information retrieval (reference). It uses term frequency, inverse document frequency, and document length among other factors to rank documents based on how relevant they are to the user's query (reference).

Initial candidate summaries were retrieved using BM25, selecting the top five documents that carry the most relevance the user’s query. However, BM25 does not account for semantic similarity. Therefore, it cannot be used to compare results for different queries effectively, as coefficients are not normalised. BM25 turned out to be excellent for quickly gathering five relevant summaries though, which made it vital as a step in the pipeline. The model then reranks the top five using sentence embeddings from all-MiniLM-L6-v2: a lightweight (but deep compared to BM25) model from SentenceTransformers (Hugging Face [4]). All-MiniLM-L6-v2 uses semantic similarity, allowing the system to identify the most contextually relevant summary even when exact keywords differ – something BM25 is not capable of. The query and selected five candidate summaries were converted into a vector embedding, and reranked with cosine similarity, using the util.cos\_sim function to identify the most semantically relevant summary.

Through trial and error, a confidence threshold of 0.4 was found to be effective at filtering selected summaries that provide a reasonable answer to the user’s query. For example, asking the model, “where was Barack Obama born?”, results in an appropriate summary with a relevance coefficient of 0.65 to be selected, whereas asking the model, “How old is Alan Carr?” results in an irrelevant summary with a coefficient of 0.25. If the coefficient of the top summary falls below 0.4, the system returns “not confident” response. In this scenario, SBert is called by Rasa to generate a clarifying question.

The entire retrieval pipeline was modularised into the retrieve\_summary function and integrated into a command-line suitable form. Ultimately, this retrieval system finds a middle ground between lexical precision and semantic depth, and forms a basis for identifying the most contextually relevant response to all potential user queries.

[1] <https://docs.python.org/3/library/ast.html#ast.literal_eval>

[2] <https://link.springer.com/article/10.1007/s13042-010-0001-0>

[3] <https://zenodo.org/records/6106156>

[4] <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>