Task 1

Does basic tokenization (based only on whitespace) and uses cosine similarity to rank the most relevant documents.

Task 2

Errors arise from word tokenization – words separated by dashes, for example, are treated as a single word. So I tried splitting tokens based on all non-word characters (instead of only whitespace) and the MRR score improved a lot.

Errors also arise from the same word capitalized in different ways not contributing to the cosine similarity. I made all words lowercase first before adding them to the vectors, which also improved the MRR score a lot.

Errors also arise from very common words that do not contribute to the meaning of the entire sentence in any meaningful way. These are called stop words. I removed them from the vectors and in doing so also improved the MRR score a lot.

The similarity function itself could be improved a lot. Instead of making the magnitude of vectors based on word frequency alone, we could make it based on tf-idf instead. Instead of cosine similarity, we could for example use BM25.

All in all, the changes I made improved the MMR score from 0.5167 to 0.7125, a vast improvement.