**Observation report**

The BM25 output tables for each query is almost similar to the ones obtained through VSM and Lucene.

The BM25 model gives scores to documents based on the factors query term, its frequency in the document and the length of the document.

As seen in the output tables of bm25 algorithm, we encounter some negative value of bm25 score for some documents. If the query term is a stopword, that occurs in many documents then the score for that term for the document can be negative as we are taking log (N-ni+0.5/ni + 0.5 will be less than 1).

Another reason for negative scores of documents can be because the documents may not contain more than half of the query terms and is a longer document than the average document length.

As observed from some documents that have negative scores, these documents do not contain many of the query terms and are also long or moderately long documents.

The BM25 algorithm gives negative values sometimes because of the frequency of a query term in the corpus. If a term appears in more than half the documents in the corpus, then the pi component will become less than 1. The resulting effect of this is that the log will give a negative value as the log of a number less than 1 but greater than 0 is negative.

The document length also plays a role in the occurrence of negative bm25 scores. If the frequency of a term in the document is 0 and the document length is more than the average document length, the component

Will give value lesser than 1.

Longer documents are suppressed by the BM25 algorithm, they receive a penalty and shorter documents are rewarded with respect to bm25 scoring.

We can avoid these negative values by using a stopwords list. Creating a list of frequent words that appear in many documents in the corpus, so that the tf component of these frequent words will not facilitate a negative value due to log.

Alternatively, we can add 1 to the entire value before taking log as the ranking will remain the same irrespectively.

The IDF can turn out to be very close to 0 which will also give a small value less than 1, whose log will be negative. Hence, the IDF component can be floored to give a value more than or equal to 1 by taking care of absent query terms and longer documents.