BM25 Algorithm

BestMatching25(BM25) is an algorithm which focuses on text-matching and is basically an extension of the Term frequency Inverse Document Frequency (Tf-IDF) algorithm which is a retireval algorithms focuses on importance of the keywords with respect to the document retrieved within a collection of documents.

The Tf-Idf algorithms calculates the importance of a keyword based on the following formula :

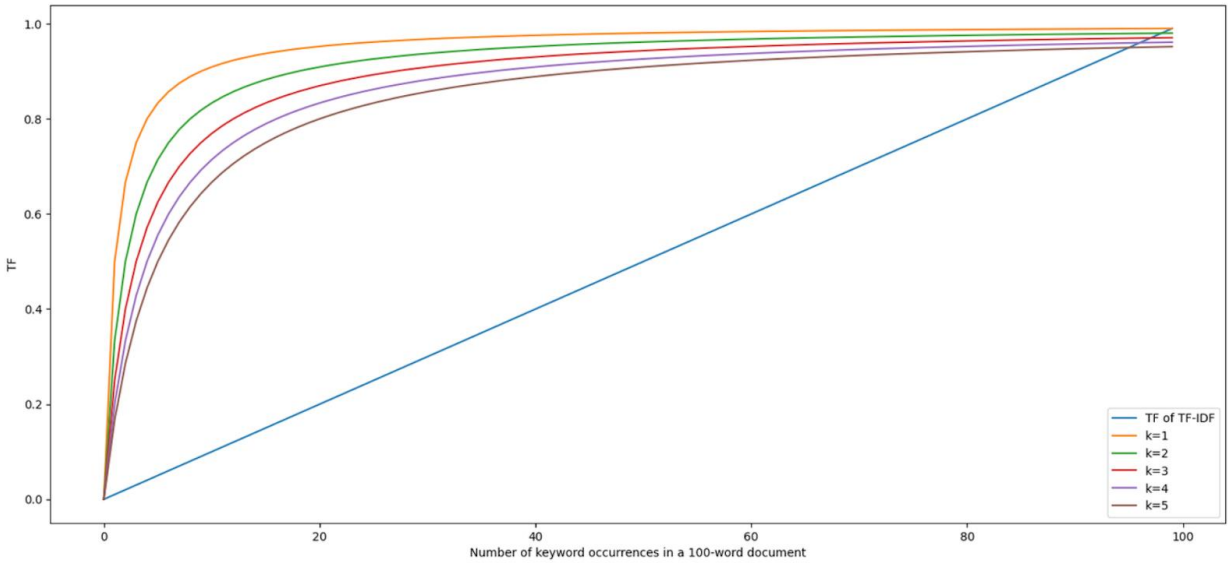
$$TF(x, y) = number\ of\ occurrences\ of\ keyword\ x\ within\ document\ y$$

$$IDF(x) = log(\frac{total\ number\ of\ documents\ in\ the\ corpus}{number\ of\ documents\ containing\ keyword\ x})$$

$$TF–IDF(x, y) = TF(x, y) * IDF(x)$$

Now based on this formula those documents will be retrieved whose score is highest with that of the importance of keyword in our query/text.

The difference between these two algorithms rise when considering that Tf-Idf does not consider the length of the document when calculating the relevance score since the longer the document the more chances of the keyword appearing in the document thus increasing the relevance, there may be a doubt regarding spamming the document completely with the single key word thus maximising the relevance but the case is that the saturation term in the BM25 algorithm addresses this issue by gradually diminishing the influence of a keyword's occurrence as its frequency increases. Furthermore, the saturation term value can be adjusted to control how rapidly the keyword occurrence reaches the saturation point, as you can see in the visualization below:



Thus enhancing the capabilities of tf-idf algorithm and deliver improved information retrieval results.

The BM25 is an informative retrieval algorithm offers predictable behaviour and easily predictable results and its vector containing few non-zero values compared to that of the splade giving higher efficiences with the retrieval of documents thus giving an easier time for searching in a large corpus of data.

BM25 also relies on string word matching and does require any fine-tuning to identify relevant documents but the vocabulary used in the query must be present in that of the documents that are to included in that of the BM25 corpus otherwise it will result in no matching documents.

Note that BM25 is a sparse embedding and thus meaning the number of unique entities present will be the result of the dimensions that are present in the document collection and non-zero entities in BM25 produced vector correspond to the dimensions where the keyword from the query is in and the values represent the keyword's relevancy score.

If we have a single-word keyword, and each entity represents a word, then only one element of the BM25 vector will have a non-zero value. This condition applies when the exact match of the keyword can be found in the document. Otherwise, all values in the vector will be zero.

This highlights a drawback of traditional sparse vectors. When the exact match of the query keyword cannot be found in a document, the sparse vector of BM25 fails

to capture the importance of the keyword, even though the document may discuss a similar topic.