BM25: Best Matching 25

BM25 (Best Matching 25) is a ranking function used by search engines to estimate the relevance of documents to a given search query. It is one of the most successful and widely-used ranking algorithms in information retrieval.

How BM25 Works

BM25 improves upon the classic TF-IDF approach by adding two key refinements:

1. Term Frequency Saturation: Unlike TF-IDF where more occurrences always mean higher scores, BM25 applies diminishing returns. The 10th occurrence of a word contributes less than the 2nd occurrence.

2. Document Length Normalization: Longer documents naturally contain more words, so BM25 adjusts scores based on how a document's length compares to the average corpus length.

The algorithm considers three main factors:
- Term Frequency (TF): How often does the query term appear in the document?
- Inverse Document Frequency (IDF): How rare is this term across all documents?
- Document Length: Is this document longer or shorter than average?

The BM25 Formula

The BM25 score for a document D given query Q is calculated as:

Score(D, Q) = sum of IDF(term) * (TF * (k1 + 1)) / (TF + k1 * (1 - b + b * docLen/avgDocLen))

Where:
- k1 controls term frequency saturation (typically 1.2 to 2.0)
- b controls document length normalization (typically 0.75)
- docLen is the document length
- avgDocLen is the average document length in the corpus

Strengths of BM25

BM25 excels at:
- Exact keyword matching: Finding documents with specific terms
- Rare term boosting: Giving higher weight to uncommon, specific words
- Speed: Very fast to compute, even on large document collections

- No training required: Works out-of-the-box without machine learning

BM25 is the default ranking algorithm in Elasticsearch, Apache Solr, and many other search systems.

Limitations

BM25 struggles with:
- Synonyms: "car" won't match "automobile"
- Semantic meaning: Cannot understand that "king" relates to "royalty"
- Typos: "serch" won't match "search"
- Context: Treats words independently, ignoring word order and context

This is why modern search systems often combine BM25 with semantic search methods like dense vector embeddings for hybrid search.