TF-IDF & BM25

Information Retrieval

Information retrieval (IR) is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.

Automated information retrieval systems are used to reduce what has been called information overload. An IR system is a software system that provides access to books, journals and other documents; stores and manages those documents. Web search engines are the most visible IR applications.

Apache Lucene

Apache Lucene is a free and open-source search engine software library, originally written completely in Java by Doug Cutting. While suitable for any application that requires full text indexing and searching capability, Lucene is recognized for its utility in the implementation of Internet Search Engines and local, single-site searching.

TF-IDF

Term Frequency (TF) is used to describe the number of key word's appearance.

TF Score = The number of appearance of a word in document / the length of the document

Suppose the key word w appeared in n documents, then the larger n, the less weight of w. Inverse Document Frequency (IDF): IDF = log (N/n)

$$similarity = log(numDocs/(docFreq + 1)) \times sqrt(tf) \times (1/sqrt(length))$$

fieldNorms can be seen as the Normalization of the length of the document. So we can also have this formula:

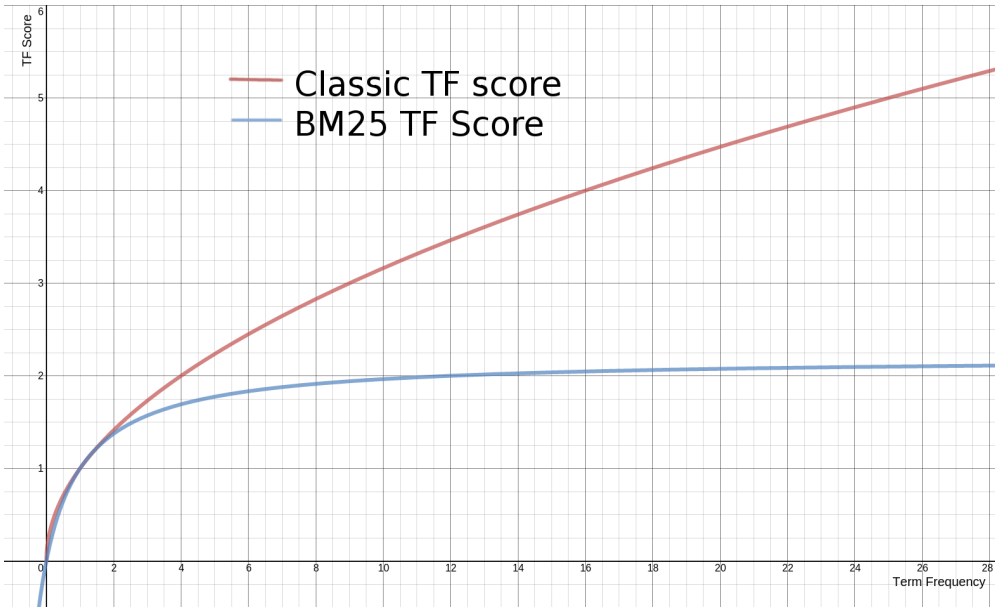
$similarity = IDF\ score \times TF\ score \times fieldNorms$

BM25

Traditional TF value can be infinitely big. What's different in BM25 is it added a regular number k, to restrict the boundary of the increase of TF value.

TF score in TF-IDF = $sqrt(tf)$

TF score in BM25 = $((k+1) \times tf)/(k+tf)$

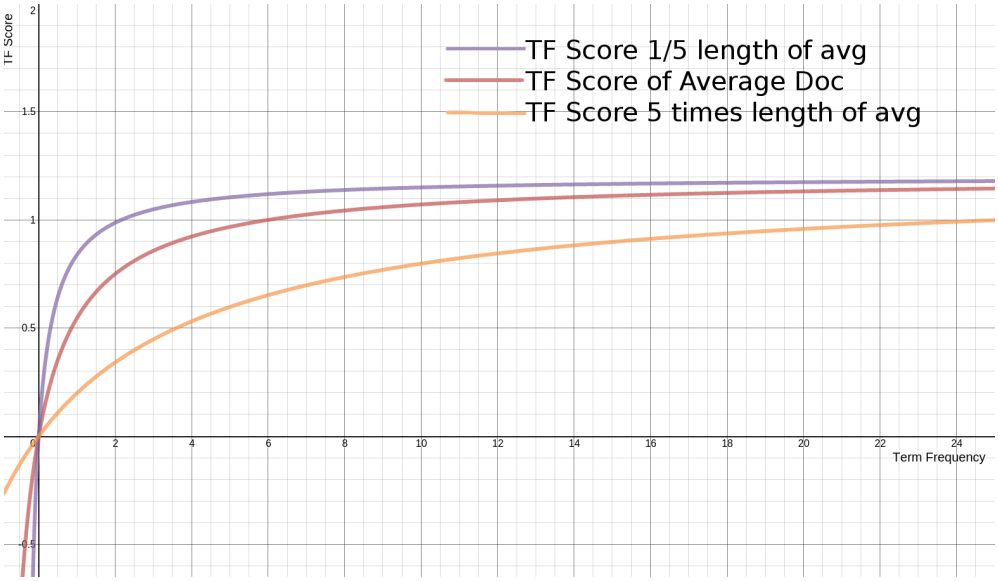


Also, BM25 introduced the concept of average document length. How the length of a single document affect the similarity is related to its relation with average length.

L = length of document / average length of documents

b: how effective L is to the score

TF score = $((k+1) \times tf)/(k \times (1.0 - b + b \times L) + tf)$



similarity = $IDF * ((k+1) \times tf)/(k \times (1.0 - b + b \times (|d|/avgDl)) + tf)$