PATENTSCOPE search scoring algorithm

Default scoring now uses Okapi BM25 by default. It ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity)

Some scoring factors include:

- The number of times a search term appears in the document field: more matches produce a higher score. The saturation function of BM25 asymptotically approaches a limit for high term frequencies (fig 1) and therefore high term frequency doesn't have an impact on the final score
- The size of the document field: longer fields produce a lower score, with the idea being that for a given number of term matches, shorter is better (more specific match)
- Smarter document length weighting: A search term occurring once in a short doc is more relevant than a single term occurring in a longer doc (a book). BM25 penalizes/rewards document length relative to a document's average document length, as opposed to just having a constant multiple based on document length. The average length of the field across the entire corpus (BM25 considers this, classic tf-idf does not)
- How common the query terms are across the entire corpus: the idea being that rarer terms carry more information. For example, if searched for "solenoid valve", patents about solenoid score higher than things about valve.

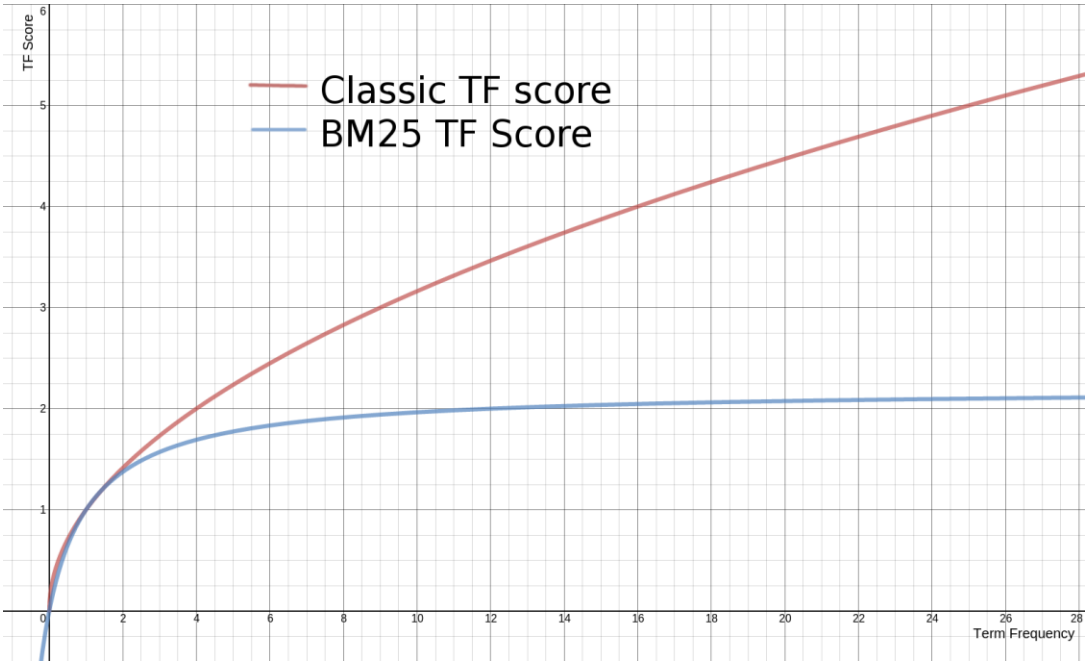

Fig 1