okapi bm25

by franklin liu

professional idiot

googly woogly

how to maek search engine
ranking function

- # of times term occurs
- length of document
- importance of a query term

okapi bm25

ranking function

$$\mathrm{BM25}(D,Q) = \sum_{i=1}^{n} IDF(q_i, D) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i) + k_1 \cdot (1 - b + b \cdot |D|/d_{avg}))}$$

where

- f(qi, D) is # of times qi appears in document D
- |D| is length of D in words, d_avg is average document length
- and k1 and b are tuned parameters

idf

inverse document frequency
how important a term is
penalizes common terms
'the doctor' vs 'doctor'

$$\text{IDF}(q_i, D) = \log \frac{N - N(q_i) + 0.5}{N(q_i) + 0.5}$$

denominator

words here

$$f(q_i) + k_1 \cdot (1 - b + b \cdot |D|/d_{avg}))$$

numerator

blah

$$f(q_i, D) \cdot (k_1 + 1)$$