BM25

- Popular and effective ranking algorithm based on binary independence model
  - adds document and query term weights

$$\sum_{i \in Q} \log \frac{(r_i+0.5)/(R-r_i+0.5)}{(n_i-r_i+0.5)/(N-n_i-R+r_i+0.5)} \cdot \frac{(k_1+1)f_i}{K+f_i} \cdot \frac{(k_2+1)qf_i}{k_2+qf_i}$$

  - k_1, k_2 and K are parameters whose values are set empirically
  - $K = k_1((1-b) + b \cdot \frac{dl}{avdl})$ *dl* is doc length
  - Typical TREC value for k_1 is 1.2, k_2 varies from 0 to 1000, b = 0.75

- r_i is the # of relevant documents containing term i
- (set to 0 if no relevancy info is known)
- n_i is the # of docs containing term i
- N is the total # of docs in the collection
- R is the number of relevant documents for this query
- (set to 0 if no relevancy info is known)
- f_i is the frequency of term i in the doc under consideration
- qf_i is the frequency of term i in the query
- k_1 determines how the tf component of the term weight changes as f_i increases. (if 0, then tf component is ignored.) Typical value for TREC is 1.2; so f_i is very non-linear (similar to the use of *log f* in term wts of the vector space model) --- after 3 or 4 occurrences of a term, additional occurrences will have little impact.
- k_2 has a similar role for the query term weights. Typical values (see slide) make the equation less sensitive to k_2 than k_1 because query term frequencies are much lower and less variable than doc term frequencies.
- K is more complicated. Its role is basically to normalize the tf component by document length.
- b regulates the impact of length normalization. (0 means none; 1 is full normalization.)

BM25 Example

- Query with two terms, "president lincoln", (*qf = 1)*
- No relevance information (*r and R are* zero)
- *N* = 500,000 documents
- *"president"* occurs in 40,000 documents (n_1 = 40,000)
- *"lincoln"* occurs in 300 documents (n_2 = 300)
- "president" occurs 15 times in doc (f_1 = 15)
- *"lincoln"* occurs 25 times (f_2 = 25)
- document length is 90% of the average length (*dl/avdl* = .9)
- k_1 = 1.2, *b* = 0.75, and k_2 = 100
- *K* = 1.2 · (0.25 + 0.75 · 0.9) = 1.11

BM25 Example

$$\begin{aligned} BM25(Q,D) \;=\; & \log \frac{(0+0.5)/(0-0+0.5)}{(40000-0+0.5)/(500000-40000-0+0+0.5)} \\ & \times \frac{(1.2+1)15}{1.11+15} \times \frac{(100+1)1}{100+1} \\ & + \log \frac{(0+0.5)/(0-0+0.5)}{(300-0+0.5)/(500000-300-0+0+0.5)} \\ & \times \frac{(1.2+1)25}{1.11+25} \times \frac{(100+1)1}{100+1} \end{aligned}$$

$$\begin{aligned} &= \log 460000.5/40000.5 \cdot 33/16.11 \cdot 101/101 \\ &\quad + \log 499700.5/300.5 \cdot 55/26.11 \cdot 101/101 \\ &= 2.44 \cdot 2.05 \cdot 1 + 7.42 \cdot 2.11 \cdot 1 \\ &= 5.00 + 15.66 = 20.66 \end{aligned}$$

BM25 Example

- Effect of term frequencies

Frequency of "president"	Frequency of "lincoln"	BM25 score
15	25	20.66
15	1	12.74
15	0	5.00
1	25	18.2
0	25	15.66