

Benefit of Rewriting

- Better understanding of the ranking function
 - Smoothing with $p(w|C)$ → TF-IDF weighting + length norm.

$$\log p(q|d) = \sum_{\substack{w_i \in d \\ w_i \in q}} c(w, q) \left[\log \frac{p_{\text{Seen}}(w_i|d)}{\alpha_d p(w_i|C)} \right] + n \log \alpha_d + \sum_{i=1}^n \log p(w_i|C)$$

TF weighting (points to $p_{\text{Seen}}(w_i|d)$)

Doc length normalization (points to $n \log \alpha_d$)

matched query terms (points to $w_i \in q$)

IDF weighting (points to $p(w_i|C)$)

Ignore for ranking (points to the boxed sum $\sum_{i=1}^n \log p(w_i|C)$)

- Enable efficient computation