Exercise 1 : Binary Independence Model Ranking

The Binary Independence Model (BIM) is a probabilistic retrieval model. The relevance ranking function ρ_{BIM} under the BIM can be expressed as

$$\begin{aligned}\rho_{\text{BIM}}(\mathbf{d},\mathbf{q}) &= \mathsf{P}(r=1|\mathbf{d},\mathbf{q})\\ &\propto \sum_{t\in\mathbf{q}:t\in\mathbf{d}} \omega_{\mathbf{BIM}}(t,D)\\ &= \sum_{t\in\mathbf{q}:t\in\mathbf{d}} \log\frac{|D|-|D_t|+0.5}{|D_t|+0.5} && \texttt{IR:III-73}\\ &\overset{\text{rank}}{=} \sum_{t\in\mathbf{q}:t\in\mathbf{d}} \log\frac{1-s_t}{s_t}, && \texttt{IR:III-[48-63]}\end{aligned}$$

where $s_t = P(\mathbf{d}(t)=1|r=1,\mathbf{q}) = \frac{|D_t|+0.5}{|D|+1}$ (IR:III-64) denotes the probability of document d containing term t given d is not relevant for query q. D_t denotes the set of documents in D that contain term t.

A user issues the query $q=(a,c,h)$. The document collection D consists of the following six documents (term occurrences are indicated by listings in parentheses):

$$D=\{d_1=(a,b,c,b,d),d_2=(b,e,f,b),d_3=(b,g,c,d),d_4=(b,d,e),d_5=(a,b,e,g),d_6=(b,g,h)\}$$

(a) List some of the assumptions made by the Binary Independence Model.

(b) What does $\omega_{\text{BIM}}(t,D)$ approximate?

(c) Complete Table 1 with the missing values based on the document collection D.

Table 1: Term statistics for the Binary Independence Model

Term t	a	b	c	d	e	f	g	h
$\lvert D_t\rvert$								
s_t								

s_t values rounded to one decimal place.

(d) Complete Table 2 by calculating the relevance scores $\rho_{\text{BIM}}(\mathbf{d_i},\mathbf{q})$ for all documents $d_i\in D$ and ranking (i.e., correctly order them) the documents accordingly.

Table 2: Document ranking according to relevance score $\rho_{\text{BIM}}(\mathbf{d_i},\mathbf{q})$

Document i	$\rho_{\text{BIM}}(\mathbf{d_i},\mathbf{q})$	Ranking
d_1		
d_2		
d_3		
d_4		
d_5		
d_6		

$\rho_{\text{BIM}}(\mathbf{d_i},\mathbf{q})$ values rounded to four decimal places.

(e) Why is the first ranked document at the first postion?

Exercise 2 : Language Model Ranking

We denote the relevance function of language models by $\rho_{\text{LM}}(d, q)$. Recall the *query likelihood model* (IR:III-176) with *Jelinek–Mercer* (JM) and *Dirichlet smoothing* (IR:III-188):

$$\begin{aligned}
\rho_{\text{LM}}(d,q) = P(d|q) &\overset{\text{rank}}{=} P(d) \cdot P(q|d) && \texttt{IR:III-176} \\
&\overset{\text{rank}}{=} P(d) \cdot \prod_{i=1}^{|q|} P(t_i|d) && \texttt{IR:III-178} \\
&\overset{\text{rank}}{=} P(d) \cdot \sum_{i=1}^{|q|} \log P(t_i|d) && \texttt{IR:III-181} \\
&\overset{\text{rank}}{=} P(d) \cdot \sum_{i=1}^{|q|} \log(1-\lambda) \cdot P(t|d) + \lambda \cdot P(t|D) && \texttt{JM IR:III-188} \\
&\overset{\text{rank}}{=} P(d) \cdot \sum_{i=1}^{|q|} \log \frac{\mathsf{tf}(t,d) + \alpha \cdot P(t|D)}{|d| + \alpha}, && \texttt{Dirichlet IR:III-188}
\end{aligned}$$

where $P(t|D)$ denotes the probability of term t in the entire document collection D, i.e., $\frac{\sum_{d\in D} \mathsf{tf}(t,d)}{\sum_{d\in D} |d|}$ (IR:III-184), and λ and α are smoothing parameters.

The query is $q = \texttt{information retrieval}$, and the document collection D consists of documents $d_1, d_2, \ldots$ with the following term frequencies $\mathsf{tf}(t, d)$:

	information	retrieval	$\cdots$	$\sum_{t\in T} \mathsf{tf}(t, d_i)$
d_1	15	25	$\cdots$	$\lvert d_1\rvert = 1800$
d_2	15	1	$\cdots$	$\lvert d_2\rvert = 2000$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\sum_{d\in D} \mathsf{tf}(t, d)$	160 000	2400	$\cdots$	$\sum_{d\in D} \lvert d\rvert = 10^9$

For example, $\mathsf{tf}(\texttt{information}, d_1) = 15$ indicates that the term `information` occurs 15 times in document d_1.

(a) Given the following document collection D and query q, compute the *Dirichlet* smoothed query likelihood scores ρ_{LM} for documents d_1 and d_2 and rank them accordingly. Use a smoothing parameter of $\alpha = 2000$.

(b) What is the difference between *Jelinek–Mercer* smoothing and *Dirichlet* smoothing?

(c) Which smoothing method is more effective for (1) verbose, (2) keyword queries?