Lab Class IR:III

Exercise 1 : Probability Ranking Principle

(a) What does the Probability Ranking Principle (PRP) state?

(b) What assumptions does it make?

(c) Give an example where the PRP fails because the assumptions are not fulfilled.

Exercise 2 : Boolean Retrieval

Figure 1 illustrates a query. (i, j, k) denotes whether term i, j, k is present. What query is illustrated?

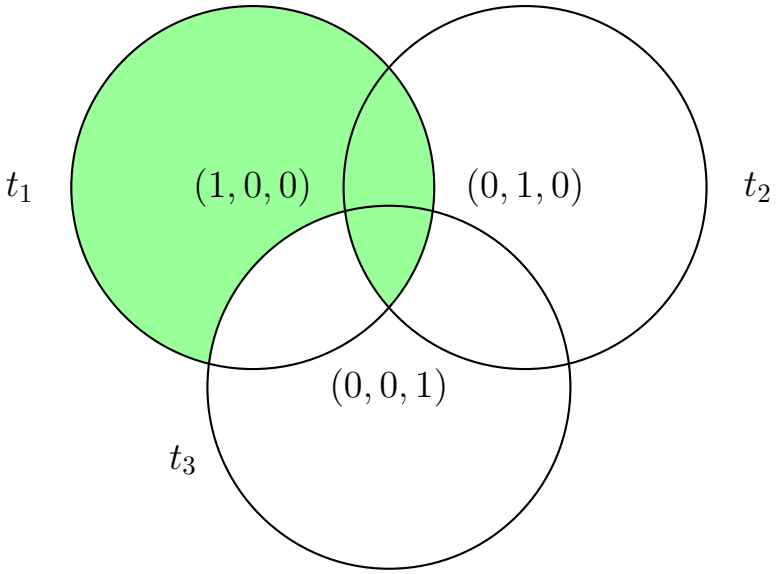

Figure 1: Boolean retrieval regions for t_1, t_2, and t_3.

Exercise 3 : Formal Ranker

In the lecture, we defined retrieval models for D, Q as a tuple $\mathcal{R} = \langle \mathbf{D}, \mathbf{Q}, \rho \rangle$. What does each of these symbols, $D, Q, \mathbf{D}, \mathbf{Q}$, and ρ, denote?

Exercise 4 : Okapi BM25

Okapi BM25 is a probabilistic ranking function derived from the Binary Independence Model [IR:III]. The following relevance ranking function ρ is used in the Okapi retrieval system (also known as Binary Independence Model [IR:III-47]):

$$\rho(\mathbf{d}, \mathbf{q}) = \sum_{t \in \mathbf{q}} \omega_{\mathsf{BM25}}(t, d, D),$$

where D denotes the document collection to be indexed [IR:III-93].

The BM25 term-weighting function is defined as

$$\omega_{\mathsf{BM25}}(t, d, D) = \omega_{\mathsf{dtf}}(t, d) \cdot \omega_{\mathsf{qtf}}(t, q) \cdot \omega_{\mathsf{BIM}}(t, D).$$

The individual components are given by

$$\omega_{\mathsf{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathsf{tf}(t,d)}{k_1\left((1-b)+b\cdot\frac{|d|}{|d|_{\mathsf{avg}}}\right)+\mathsf{tf}(t,d)},$$

$$\omega_{\mathsf{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathsf{tf}(t,q)}{k_2+\mathsf{tf}(t,q)},$$

,

$$\omega_{\mathsf{BIM}}(t,D) = \log\frac{|D|-\mathsf{df}(t,D)}{\mathsf{df}(t,D)}.$$

(a) Provide a high-level intuition for each component $\omega_{\mathsf{dtf}}(t,d)$, $\omega_{\mathsf{qtf}}(t,q)$, and $\omega_{\mathsf{BIM}}(t,D)$.

(b) Explain $\omega_{\mathsf{dtf}}(t,d)$ and its purpose.

(c) Explain $\omega_{\mathsf{qtf}}(t,q)$ and its purpose.

(d) Explain $\omega_{\mathsf{BIM}}(t,D)$ and its purpose.

(e) Explain the role of the parameters k_1, k_2, and b.

Exercise 5 : BM25 & TF–IDF

The following relevance ranking function ρ is used in the Okapi retrieval system:

$$\rho(\mathbf{d},\mathbf{q}) = \sum_{t\in\mathbf{q}} \omega_{\mathsf{BM25}}(t,d,D),$$

where D denotes the document collection to be indexed [IR:III-93].

The BM25 term-weighting function is defined as

$$\omega_{\mathsf{BM25}}(t,d,D) = \omega_{\mathsf{dtf}}(t,d)\cdot\omega_{\mathsf{qtf}}(t,q)\cdot\omega_{\mathsf{BIM}}(t,D).$$

The individual components are given by

$$\omega_{\mathsf{dtf}}(t,d) = \frac{(k_1+1)\cdot \mathsf{tf}(t,d)}{k_1\left((1-b)+b\cdot\frac{|d|}{|d|_{\mathsf{avg}}}\right)+\mathsf{tf}(t,d)},$$

$$\omega_{\mathsf{qtf}}(t,q) = \frac{(k_2+1)\cdot \mathsf{tf}(t,q)}{k_2+\mathsf{tf}(t,q)},$$

,

$$\omega_{\mathsf{BIM}}(t,D) = \log\frac{|D|-\mathsf{df}(t,D)}{\mathsf{df}(t,D)}.$$

For comparison, the following term-weighting function $\omega(t)$ is used, for example, in the Vector Space Model:

$$\omega(t) = \mathsf{tf}(t,d)\cdot\mathsf{idf}(t,D),$$

where $\mathsf{tf}(t,d)$ denotes the normalized term frequency of term t in document d [IR-III-35], and

$$\mathsf{idf}(t,D) = \log\frac{|D|}{\mathsf{df}(t,D)}$$

denotes the inverse document frequency of term t in the document collection D.

(a) In which aspects do TF–IDF and BM25 differ?

(b) Why is the term frequency $\mathsf{tf}(t, d)$ normalized in TF–IDF?

(c) How can the parameters k_1, k_2, and b be chosen for BM25?

(d) What are the strengths of BM25?

(e) What aspects should be considered when comparing other ranking functions to BM25?

Exercise 6 : Different Ranking Approaches

Match the following terms from each group.

Type	Formula	Model(s)
Logical		
Algebraic		
Probabilistic		
Bayesian		
Information Theoretic		

Formulae

(a) $-\log_2 P(\mathbf{d} \mid \mathsf{tf}(t_1), \ldots, \mathsf{tf}(t_{|\mathbf{q}|}))$

(b) $\mathcal{I}(\mathbf{d} \to \mathbf{q})$

(c) $P(\mathsf{rel}(d, q) = 1 \mid \mathbf{d}, \mathbf{q})$

(d) $\varphi(\mathbf{d}, \mathbf{q})$

(e) $P(\mathbf{d} \mid \mathbf{q})$

Models

(a) BM25

(b) monoT5

(c) duoT5

(d) TF–IDF

(e) Boolean Retrieval

(f) Language Models