Lab Class IR

Exercise 1 : Formal Ranker

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

Exercise 2 : Probability Ranking Principle

What does the Probability Ranking Principle (PRP) state? What assumptions does it make? Give an example where the PRP fails because the assumptions are not fulfilled.

Exercise 3 : Different Ranking Approaches

Match the following terms from each group.

Type: Logical Algebraic Probabilistic Bayesian Information Theoretic

Formular: $-\log_2 P(\mathbf{d} \mid \mathrm{tf}(t_1), \ldots, \mathrm{tf}(t_{|\mathbf{q}|}))$ $\mathcal{I}(\mathbf{d} \to \mathbf{q})$ $P(\mathrm{rel}(d, q){=}1 \mid \mathbf{d}, \mathbf{q})$ $\varphi(\mathbf{d}, \mathbf{q})$ $P(\mathbf{d} \mid \mathbf{q})$

Models: BM25 monoT5 duoT5 TF-IDF Boolean Retrieval Language Models

Exercise 4 : BM25

(a) Where do tf-idf and BM25 differ?

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

(c) What are the strengths of BM25?

(d) What must be watched out for when comparing to BM25?