Objective

Develop an Information Retrieval (IR) system for the CISI Dataset using the BM25 ranking algorithm with the support of chatGPT in finding solutions and coding.

BM25

BM25 is a ranking algorithm that uses term matching between a query and a set of documents. Its equation is presented below.

$$\mathrm{BM25}(D,q) = \underbrace{\frac{\mathrm{f(q,D)} * (\mathrm{k}+1)}{\mathrm{f(t,D)} + \mathrm{k} * \left(1 - \mathrm{b} + \mathrm{b} * \frac{\mathrm{D}}{d_{avg}}\right)}}_{\text{TF}} * \underbrace{log\left(\frac{N - N(q) + 0.5}{N(q) + 0.5} + 1\right)}_{\text{IDF}}$$

In fact, BM25 is an improvement of TF-IDF algorithm that considers document length and introduces tuning parameters *b* and *k*. Parameter *b* (~0.75) controls the impact of the document length and *k* (~1.25) is a constant that adjusts term frequency. Also, $\mathrm{f(q,D)}$ is the frequency of the query terms in a document and $\mathrm{f(t,D)}$ is the terms frequency in a document. Finally: D is the document length; d_{avg} is the average document length of the *corpus*; N the total of documents and $N(q)$ the total documents that contain the query.

Methodology

The BM25 algorithm used was from the Rank-BM25 library. Preprocessing was done with the spaCy library using lemmatization and stop words removal. The following evaluation metrics were used: Recall@K, Precison@K, Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP). Finally, chatGPT was used for better understanding of the CISI dataset, the BM25 equation and the evaluation metrics for IR.

Results

Table 1 presents the evaluation results of the BM25 in the CISI Dataset. From this it is possible to note that there is lot of room for improvement.

Table 1 – Evaluation Results

Metric	**Value**
Recall@10	0.103
Precision@10	0.261
MRR	0.541
MAP	0.056