

Report - Desenvolvimento sistema de Busca com coleção CISI

Neste trabalho foi desenvolvido um sistema simples de ranqueamento de documentos usando o algoritmo BM25, implementado via Pyserini.

A primeira etapa foi baixar a coleção de teste do CISI, extrair os documentos, queries e Qrels. A partir das informações mapeadas, foi indexado os documentos com o BM25 (LuceneSearcher) e logo depois foi avaliado a recuperação dos documentos.

Para a avaliação foi iterado cada query sobre a indexação feita na etapa anterior, para essa etapa também foi usado o pyserini, mais especificamente a função "LuceneSearcher" e atribuído um k, esse valor de k consiste na quantidade de documentos retornados de resposta por query, por padrão foi definido 100.

Com a inferência feita em todas as queries foi então computada a avaliação com as métricas MRR e nDCG. Como resultado vemos que em MRR foi obtido 61% enquanto de nDCG@10 igual a 0.36, os resultados apontam que é necessário melhoria no ranqueamento, a métrica MRR diz que a probabilidade do documento relevante estar próximo da primeira posição é de 61%.

Uso do ChatGPT

Até o momento desse projeto o ChatGPT foi usado para entender melhor como o algoritmo do BM25 funciona; Gerar explicações sobre as métricas MRR e nDCG e por fim foi tentado resumir alguns documentos, porém, não foi satisfatória a saída via chatGPT.

Output ChatGPT

Query: Explain didactically how the BM25 works

ChatGPT:

The BM25 algorithm is a ranking function that is commonly used in information retrieval systems to retrieve relevant documents in response to a user's query. Here's how it works:

Preprocessing: The first step is to preprocess the query and the document collection. This typically involves tokenization (breaking the text into individual words or terms), stopwords removal (eliminating common words such as "the" and "and"), and stemming (reducing words to their base form, such as "running" to "run").

Term frequency: For each term in the query and document collection, the algorithm calculates the number of times the term appears in each document. This is known as the term frequency (TF).

Inverse document frequency: The algorithm also calculates the inverse document frequency (IDF) for each term. The IDF is a measure of how important a term is in the document collection. Terms that appear frequently in many documents (such as "the" or "and") have a low IDF, while terms that appear rarely in the document collection have a high IDF.

BM25 score: The BM25 score is calculated for each document in the collection based on its similarity to the query. The score is based on the TF and IDF values for each term in the query and the document. The formula for the BM25 score is:

$$\text{score}(D,Q) = \sum(w(q_i,D) * \text{IDF}(q_i) * ((k_1 + 1) * \text{TF}(q_i,D)) / (k_1 * (1 - b + b * (|D| / \text{avgdl}))) + \text{TF}(q_i,D)))$$

$\text{score}(D,Q)$ is the BM25 score for document D given query Q $w(q_i,D)$ is a weighting factor that determines the importance of each query term $\text{IDF}(q_i)$ is the IDF for query term q_i $\text{TF}(q_i,D)$ is the term frequency for term q_i in document D k_1 and b are tuning parameters that control the impact of term frequency and document length on the score $|D|$ is the length of document D in terms avgdl is the average document length in terms The BM25 algorithm ranks documents based on their BM25 scores, with higher scores indicating a greater similarity to the query.

In summary, the BM25 algorithm calculates a score for each document in the collection based on the similarity of its terms to those in the query, taking into account the importance of each term and the length of the document. This allows it to effectively retrieve relevant documents in response to a user's query.

Query: Explain in a didactic way the MRR and nDCG metrics for information retrieval

ChatGPT:

MRR (Mean Reciprocal Rank) and nDCG (Normalized Discounted Cumulative Gain) are evaluation metrics commonly used in information retrieval to measure the effectiveness of a system in returning relevant results to a user's query.

MRR is a measure of how quickly the relevant results are returned by the system. It is calculated as the average of the reciprocal ranks of the first relevant document for each

query. In other words, for each query, the reciprocal rank of the first relevant document is calculated (i.e., $1/\text{rank}$) and then the average of these reciprocal ranks is taken across all queries. The higher the MRR, the more quickly the relevant results are returned by the system.

nDCG is a measure of the quality of the results returned by the system. It takes into account both the relevance of the returned results and their rank in the list of results. It is calculated by first assigning a relevance score (usually a binary value) to each document based on how relevant it is to the query, and then calculating a cumulative gain score for the top k documents in the result list. The cumulative gain score is a measure of how much relevant information is contained in the top k documents, with greater weight given to documents that are higher in the result list. The cumulative gain score is then normalized by a discounted gain score, which gives greater weight to documents that are more relevant and appear higher in the result list. The nDCG score is the average of the normalized discounted cumulative gain scores across all queries. The higher the nDCG, the better the quality of the results returned by the system.

In summary, MRR measures how quickly the relevant results are returned by the system, while nDCG measures the quality of the results returned by the system. Both metrics are important for evaluating the effectiveness of an information retrieval system in returning relevant results to a user's query.

Query: Generate a summary

Measuring the Reliability of Subject Classification by Men and Machines Borko, Harold
Procedures for measuring the consistency of document classification are described..
Three subject specialists classified 997 abstracts of psychological reports into one of
eleven categories.. These abstracts were also mechanically classified by a computer
program using a factor score computational procedure.. Each abstract was scored for all
categories and assigned to the one with the highest score.. The three manual
classifications were compared with each other and with the mechanical classifications,
and a series of contingency coefficients was computed.. The average reliability of manual
classification procedures was equal to .870.. The correlation between automatic and
manual classification was .766..

ChatGPT:

In this study, the reliability of document classification was measured using both manual and automated methods. Three subject specialists manually classified 997 abstracts of psychological reports into one of eleven categories, while a computer program used a factor score computational procedure to mechanically classify the same abstracts. Each abstract was scored for all categories and assigned to the one with the highest score. The manual classifications were compared with each other and with the mechanical classifications, and contingency coefficients were computed.

The results showed that the average reliability of manual classification was .870, while the correlation between automatic and manual classification was .766