IA368DD - Relatório Exercício de Seleção
William Sueyoshi do Carmo - RA120282

A proposta do exercício foi implementar um buscador utilizando o algoritmo BM25, avaliado na CISI Collection e com o apoio do ChatGPT.

Inicialmente, foi estudado o que era a CISI Collection, a qual consiste em um conjunto de documentos e queries que, por incluir o retorno esperado de cada query, pode ser utilizado para avaliar sistemas de busca. Os arquivos presentes no pacote são:

CISI.ALL: Contém os documentos a serem pesquisados e retornados para as queries. Cada documento, em geral (nem todos os documentos contêm todos os campos), contém os campos ".I" (ID), ".T" (título), ".A" (autores), ".B" (referência, pode ser o ano de publicação ou o volume de um conjunto, por exemplo), ".W" (texto, ou resumo) e ".X" (referência cruzada, relaciona o documento a outros documentos).

CISI.QRY: Contém as queries a serem executadas. Cada query pode conter os campos ID, título, autores, referência e texto, seguindo os mesmos conceitos do arquivo CISI.ALL.

CISI.REL: Contém a prova real, ou seja, o retorno esperado para cada query. Relaciona IDs de queries aos IDs dos documentos correspondentes.

O arquivo CISI.BLN e o campo de referência cruzada do arquivo CISI.ALL não foram utilizados nessa implementação. Outro ponto referente aos arquivos é que, por estarem em texto corrido, apenas com "pontos + letra maiúscula" indicando cada campo de cada documento (p. ex.: ".I" seguido de um número para indicar um ID, ou ".A" seguido de caracteres para autores, etc.) e quebras de linha ("/n") espalhadas ao longo do texto, foi necessário realizar tratamentos no código. Para fins de clareza e organização, optou-se por armazenar cada documento em um dicionário, com as descrições e conteúdos dos campos nas chaves/valores; apesar disso, as buscas finais apenas concatenaram os campos relevantes.

A segunda parte foi entender o que é o BM25, algoritmo de busca que atribui, para uma dada query, uma nota a cada documento do corpus. Tal nota indica o quão relevante é o documento para a query e é baseada numa equação que leva em consideração os termos da query, dos documentos que compõem o corpus, a frequência de tais termos e parâmetros que permitem controlar a importância da frequência de termos e do comprimento dos documentos. Nesse contexto, a nota isolada de um documento (em termos absolutos) possui menos importância do que a comparação entre as notas dos diferentes documentos – é esta comparação que indica a ordem de relevância de cada um como retorno à query.

Entendidos os conceitos da CISI Collection e do BM25, partiu-se para a implementação do código. Como dito anteriormente, os arquivos da CISI Collection foram tratados e armazenados em dicionários antes de alimentarem o buscador. Este recebe o corpus completo, o conjunto de queries e a quantidade de documentos a serem retornados por query (função "executa_buscas" do código Python). Tal função, internamente (chamando

outras funções), separa os textos de input (tanto os documentos do corpus quanto as queries) em palavras, retira as stopwords, extrai o radical, calcula o TF/IDF e outros índices necessários para a nota do BM25, e retorna os IDs dos documentos com maior relevância por query. Boa parte do código do buscador foi consultada via ChatGPT, com ajustes sendo feitos para adequar o formato de input ao criado inicialmente no código. Os parâmetros internos k1 e b da equação do BM25 foram mantidos seguindo a sugestão feita pelo ChatGPT, assim como a quantidade de documentos retornados ser 10 para cada query.

Executadas as buscas e retornados os documentos mais relevantes, comparou-se o resultado à prova real, presente no arquivo CISI.REL. A métrica de avaliação foi a precisão (medida de verdadeiros positivos sobre total de positivos acusados), calculada para cada query e, por fim, feita a média de todas. O valor final de precisão obtido foi de 23,6%.

Um ponto importante em relação ao cálculo final da precisão é que nem todas as queries estão presentes na prova real, ou seja, existem queries que (segundo a CISI Collection) não deveriam retornar nenhum documento – do total de 112 queries, apenas 76 aparecem no arquivo CISI.REL. Para evitar qualquer tipo de viés na avaliação do buscador, uma vez que se pode argumentar que a prova real deveria ser usada apenas para indicar se cada documento é relevante ou não a cada query, e devido ao buscador sempre retornar 10 resultados (escolha feita arbitrariamente), a precisão para as queries que não aparecem no CISI.REL foram zeradas e contadas como zero na média final. Se tais casos fossem retirados da média, porém, a precisão subiria para próximo dos 35%.

Ainda sobre a avaliação do desempenho do buscador, vale dizer que o número de documentos relacionados a cada query na prova real é diferente para cada uma. Como o buscador foi configurado para retornar um número fixo de resultados, esse parâmetro modifica a quantidade de verdadeiros/falsos positivos/negativos, que por consequência modifica as métricas de precisão, recall, F1, etc. Foi escolhida a precisão para que se desse maior importância a saber, do conjunto de retorno, quantos eram verdadeiros positivos, porém pode ser interessante fazer a avaliação por outras métricas. Uma alternativa seria estabelecer um threshold baseado nas notas do BM25, o que faria com que o número de documentos retornados para cada query fosse variável; porém, como exposto anteriormente, a comparação entre notas traz mais informação do que a nota em si, o que de certa forma inviabiliza este método.

Em relação ao ChatGPT, sua utilização no exercício facilitou e, principalmente, acelerou a compreensão de conceitos como Information Retrieval System, BM25 e CISI Collection, além de prover exemplos de código que ajudaram na implementação do buscador. A forma com que se aprende utilizando o mesmo difere muito do método “tradicional” (de pesquisa em mecanismos de busca através de palavras-chave e com avaliação dos resultados apresentados sendo feita a critério do utilizador); “conversar” com o ChatGPT torna a experiência de pesquisa muito mais natural e permite que se diminua o trabalho de avaliar a relevância dos resultados por parte de quem pesquisa, já que é possível utilizar a própria máquina (de forma simples, como falar com uma pessoa) para direcionar os resultados ao que se está buscando.