Simples Sistema de Recuperação de Informação com BM25, CISI e GPT

Building a Simple Information Retrieval System using BM25 and GPT-3 and evaluated in the CISI collection.

Valéria Banachi Barreto
Universidade Estadual de Campinas
Campinas, Brasil
valeriabanachi@gmail.com

Resumo — **Sistemas de Recuperação de Informação estão cada vez mais sendo utilizados na democratização da informação, principalmente no que diz respeito às mídias digitais. Nesse trabalho utiliza-se a biblioteca BM25 para por em prática o funcionamento de um sistema de Recuperação de Informação simples.**

Palavras Chave – Sistemas de Recuperação de Informação; BM25.

Abstract —

Information Retrieval Systems are increasingly being used in the democratization of information, especially with regard to digital media. In this work, the BM25 library is used to put into practice the operation of a simple Information Retrieval system.

Keywords – Information Retrieval Systems; BM25.

I. INTRODUÇÃO

Um Sistema de Recuperação de Informação é um sistema capaz de armazenar, recuperar e manter informações. As informações podem estar no formado texto, aúdio, imagens, vídeos e outros objetos multimídia. Esse tipo de programa tem como objetivo facilitar o encontro de informações [1].

Os primeiros Sistemas de Recuperação de Informações(IR) surgiram com a necessidade de organizar as informações em repositórios centrais, como bibliotecas. Desta forma, catálogos e indices foram criados para facilitar a recuperação.

Com o advento de bancos de dados o gerenciamento das informações passou a ser cada vez mais digital[1]. A partir de 1990 com a difusão da informação em formato eletrônico ascendeu a era da desintermediação, no qual o usuário passa a ter mais independência na busca de informações pelos meios tecnológicos[2].

Atualmente percebe-se grande utilização Sistemas de Recuperação de Informação principalmente nos motores de buscas da Web enraizados no dia a dia dos usuários.

Nessa vertente de Sistemas de Recuperação de Informação esse trabalho expõe a implementação de de um algoritmo que tem como objetivo indicar dentro de diversas opções de texto o que mais de relaciona com os termos pesquisados.

Para tanto, o algoritmo foi implementado em Python, utilizando a biblioteca BM25 na Plataforma Colab. As bases de textos tanto para pesquisa quanto para o conteúdo a ser pesquisado são da base CISI.

O desenvolvimento realizado, bibliotecas utilizadas e resultados alcançados serão apresentados nas próximas seções e por isso o texto deste trabalho está organizado da seguinte forma: Na seção 2 é apresentada o conceito de Sistemas de Recuperação de Informação. Na seção 3 apresenta-se a biblioteca BM25. A seção 4 descreveo algoritmo desenvolvido e expõe os resultados obtidos e a seção 5 a conclusão da execução desse trabalho.

II. SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO

Sistemas de Recuperação de Informação (IR, na sigla em inglês) são sistemas que permitem a recuperação eficiente e precisa de informações relevantes a partir de grandes coleções de dados. As informações podem estar no formado texto, aúdio, imagens, vídeos e outros objetos multimídia. Esse tipo de programa tem como objetivo facilitar o encontro de informações [1].

III. BM25

BM25 é um modelo matemático para pontuar a relevância de documentos em sistemas de recuperação de informação. Ele é frequentemente usado em motores de busca para ranquear documentos relevantes em resposta a uma consulta de pesquisa.

BM25 é uma abreviação para "Best Match 25", que é uma referência à sua formulação matemática original, que incluía 25 parâmetros. No entanto, ao longo do tempo, a formulação foi simplificada para uma fórmula mais simples e computacionalmente eficiente.

O modelo BM25 é baseado em uma função de pontuação que leva em consideração a frequência dos termos em um documento e em toda a coleção, além do comprimento do documento e do número total de documentos na coleção. A fórmula resultante é usada para atribuir uma pontuação de relevância a cada documento para uma dada consulta.

O BM25 é considerado um dos modelos de pontuação mais eficazes em sistemas de IR, e é usado em muitos sistemas de busca em grandes corpora de texto, incluindo motores de busca na web, bibliotecas digitais e sistemas de vigilância e segurança. Ele é particularmente eficaz em lidar com consultas de pesquisa complexas e ambíguas.

Na próxima seção será apresentado os métodos e procedimentos aplicados nesse trabalho.

IV. MÉTODOS E PROCEDIMENTOS

Apresentado uma visão geral de Sistemas de Recuperação de Informação, esta seção tem como objetivo apresentar os métodos aplicados em cada segmento da atividade:

A. Base de Dados

A base de dados utilizada nesse trabalho encontra-se em: http://ir.dcs.gla.ac.uk/resources/test_collections/cisi/cisi.tar.gz. O primeiro passo do algoritmo é realizar o download dos arquivos CISI e descompactá-los para posterior tratamento e leitura.

Os arquivos descompactados estão no diretório “resource” e utiliza-se o “CISI.ALL” e “CISI.QRY”. No qual o primeiro extrai-se os textos e no segundo as queries de pesquisas.

B. Processamento dos dados

Com os dados classificados entre textos e queries (algoritmo encontrado via chat GPT) aplica-se a biblioteca BM25. O modelo “BM25Okapi” é utilizado seguido pela execução de todas as queries em todos os documentos.

Os resultados mais relevantes encontrados pelo “score” do BM25 são exibidos na console do algoritmo no qual consta respectivamente a query utilizada na iteração, o maior score para ela e o texto qualificado no score.

V. CONCLUSÃO

Percebe-se que a utilização do Sistemas de Recuperação de Informações esta implicito no dia a dia dos usuários e que o mercado de informação vem crescendo constantimente. A utilização da biblioteca BM25 elucidou principalmente o funcionamento dos motores de buscas na Web e como com o advento da tecnologia esta cada vez democratizando o acesso as informações.

REFERÊNCIAS BIBLIOGRÁFICA

[1] G, Kowalski, “Information Retrieval Systems”, Available: https://books.google.com.br/books?hl=pt-BR&lr=&id=hfT6hFXNT4sC&oi=fnd&pg=PP10&dq=Information+Retrieval+System&ots=LaKUDPcIuN&sig=-DsWzs-0zyq8YDSjhMOW8amo5zk#v=onepage&q=Information%20Retrieval%20System&f=false

[2] R. Raieli, “Multimedia Information Retrieval”, Available: https://www.google.com.br/books/edition/Multimedia_Information_Retrieval/_v9DAgAAQBAJ?hl=pt-BR&gbpv=1&dq=Information+Retrieval+System&printsec=frontcover