INTRODUÇÃO

Nesta atividade, trabalhamos para desenvolver um projeto utilizando o GitHub como plataforma de colaboração e versionamento. O objetivo foi criar um sistema de busca chamado "máquina de busca" usando um índice invertido. Durante o projeto, conseguimos aprimorar consideravelmente nossas habilidades de programação enquanto realizávamos o trabalho em equipe.

Para isso, implementamos classes e métodos para carregar e processar os documentos, criar e manter o índice invertido, e realizar consultas de busca. Tivemos que lidar com desafios, como normalizar termos, remover caracteres especiais e calcular frequências nos documentos.

Além disso, o uso do GitHub facilitou nossa colaboração, permitindo compartilhar e integrar alterações de forma eficiente. Também nos ajudou a revisar o código e acompanhar o progresso do projeto.

IMPLEMENTAÇÃO

Bibliotecas inclusas e suas funcionalidades:

iostream: Entrada e saída de dados.

vector: Contêiner dinâmico para armazenar elementos.

string: Manipulação e armazenamento de sequências de caracteres.

fstream: Operações de entrada e saída em arquivos.

sstream: Manipulação de strings como fluxos de entrada e saída.

unordered_map: Contêiner associativo de chave-valor implementado usando tabela hash.

unordered_set: Contêiner para armazenar elementos únicos, sem ordem garantida.

algorithm: Algoritmos genéricos para operações em contêineres.

cassert: Macros para verificação de assertiva durante a execução do programa.

dirent.h: Operações relacionadas a diretórios e arquivos.

cctype: Manipulação de caracteres, como testar tipos e conversão de maiúsculas/minúsculas.

locale: Funcionalidades para localização e internacionalização.

EXPLICANDO O CÓDIGO E SUAS CLASSES

Document (classe): Essa classe representa um documento e possui três membros de dados: id (identificador do documento), title (título do documento) e content (conteúdo do documento).

InvertedIndex (classe): Essa classe implementa um índice invertido para pesquisa de documentos. Ela possui os seguintes membros de dados:

Index: Um unordered_map que mapeia termos normalizados para um segundo unordered_map que mapeia IDs de documentos para a frequência de ocorrência desse termo no documento.

documents: Um unordered_map que mapeia IDs de documentos para os próprios documentos.

next_id: Um inteiro que representa o próximo ID disponível para atribuir a um documento.

normalize(const string& term): Essa função recebe um termo e retorna sua versão normalizada. A normalização envolve a remoção de caracteres especiais, conversão para letras minúsculas e remoção de números.

remove_accents(const string& term): Essa função recebe um termo e transforma o caractere acentuado na sua versão padrão, seguindo uma tabela de conversão.

load_documents(const string& folderPath): Essa função carrega os documentos de um diretório especificado pelo folderPath. Para cada arquivo no diretório, ele lê o título e o conteúdo do arquivo e adiciona o documento ao índice invertido por meio da função add_document.

get_file_names(const string& folderPath): Essa função retorna os nomes dos arquivos presentes em um diretório especificado pelo folderPath.
add_document(const Document& doc): Essa função adiciona um documento ao índice invertido. Ela adiciona o documento ao mapeamento documents e atualiza o índice invertido index com as palavras e suas frequências de ocorrência no documento.

tokenize(const string& str): Essa função recebe uma string e a divide em tokens (palavras) com base nos espaços como delimitadores. Ela retorna um vetor contendo os tokens resultantes.

search(const string& query): Essa função realiza uma busca no índice invertido com base em uma consulta (query). Ela percorre os termos da consulta, encontra os documentos relevantes e calcula uma pontuação para cada um com base nas frequências dos termos nos documentos. Os resultados são retornados em um vetor de pares (ID do documento, pontuação), ordenados pela pontuação.

print_titles(const vector<pair>& results, const string& query): Essa função imprime os títulos dos documentos encontrados na busca. Se não houver resultados, uma mensagem informando que nenhum documento foi encontrado é exibida.

main(): A função principal do programa. Ela cria uma instância do InvertedIndex, carrega os documentos do diretório especificado, e entra em um loop onde solicita ao usuário uma consulta de busca. A função realiza a busca no índice invertido e imprime os títulos dos documentos encontrados. O loop continua até que o usuário digite "exit" para sair do programa.

CONCLUSÃO

Participar deste projeto no GitHub foi uma experiência positiva de trabalho em equipe e colaboração. Aprendemos a importância da organização e comunicação para o sucesso do projeto. Melhoramos nossas habilidades de programação, especialmente na implementação de algoritmos e manipulação de dados. Durante a implementação do sistema de busca, percebemos como é crucial usar algoritmos eficientes e estruturas de dados adequadas para obter um desempenho satisfatório. A escolha correta dos algoritmos de normalização de termos, remoção de caracteres especiais e cálculo de frequência de termos foi fundamental para obter resultados precisos nas consultas. O uso do GitHub como plataforma de versionamento e colaboração facilitou a integração do trabalho de cada membro do grupo. Foi mais fácil revisar o código, compartilhar ideias e resolver conflitos. A ferramenta também permitiu acompanhar o progresso do projeto e manter um histórico completo das alterações realizadas. Resumindo, essa experiência não apenas resultou no desenvolvimento de um sistema funcional, mas também nos proporcionou aprendizado valioso em programação. Através desse projeto, conseguimos aplicar os conhecimentos, adquirir novas habilidades e trabalhar de forma eficaz em equipe utilizando ferramentas de colaboração como o GitHub.

INTEGRANTES

Bernardo Prosdocimi Lamounier Soares - 2021032250
João Henrique Voss Teixeira - 2022056080