Universidade Federal de Minas Gerais

Programação e Desenvolvimento de Software II - (PDSII)

Alunos: André Luiz Alves Costa e Bruno Henrique Evangelista Pereira

Link do repositório: https://github.com/andrelac963/TP_maquina_busca

Trabalho Prático: Máquina de Busca

Introdução

O trabalho prático teve o objetivo de desenvolver uma máquina de busca feita a partir da linguagem de programação C++, onde em um banco de dados de arquivos txt, o usuário irá dizer o que deseja consultar e o programa desenvolvido retornará em quais arquivos as palavras dadas pelo usuário aparecem.

Ao início do projeto nos reunimos para decidir a estrutura da aplicação, a modelagem das classes que seriam necessárias e como seria implementado cada funcionalidade. Para isso foi necessário fazer uma abstração das funcionalidades aplicando os conhecimentos aprendidos na disciplina de PDS 2 e isso auxiliou na organização geral. E com as ideias bem organizadas podemos partir para a implementação do projeto de fato.

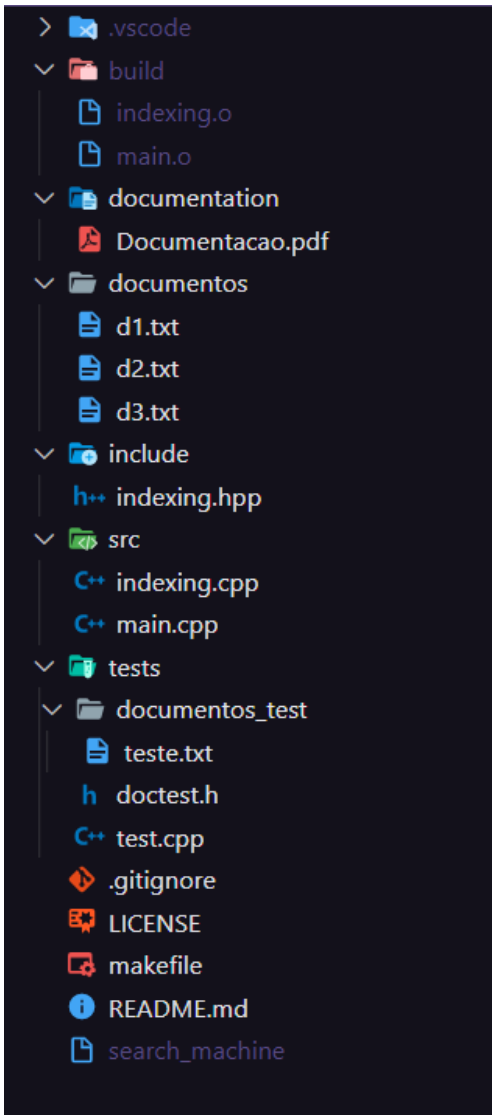
Implementação

Para versionamento do código foi utilizado o github, que funciona baseado em git, e também a utilização de pull requests para que os membros do projeto pudessem validar a implementação dos colegas de trabalho.

Para compilar o código foi utilizado o gcc versão 11.3.0 rodando no WLS com o Ubuntu 22.04.1 LTS. E para facilitar os testes foi feito um makefile com os comandos para compilar todos os arquivos do projeto.

A estrutura de pastas utilizadas foi da seguinte forma:

- build: para guardar os arquivos de compilação;
- documentos: para a base de dados lida pelo programa;
- include: para os arquivos de cabeçalho;
- src: para os arquivos de implementação.
- tests: para realização dos testes de unidade.



Decidimos que só seria necessário implementar uma classe indexing onde conteria as funcionalidades de indexação das palavras da base de dados e de recuperação delas.

```cpp
#include <map>
#include <set>
#include <vector>
#include <string>
#include <iostream>
#include <bits/stdc++.h>
#include <dirent.h>

using namespace std;

class indexing
{
private:
  map<string, set<string>> index;

public:
  indexing();
  ~indexing();

  set<string> read_directory(const string &);
  void read_files(const string &);

  void insert(string, string);
  map<string, set<string>> get_index();
  string normalize(string);

  vector<pair<string, int>> recovery(vector<string>);
  void print_ordered_files(vector<string>);
};
```

Para guardar as informações do index decidimos por utilizar um map responsável por manter a string dada pelo base de dados e o set de strings que são os nomes dos arquivos txt onde ela ocorre. Como há necessidade de que outros arquivos acessem os métodos da classe somente o container responsável por guardar os dados fique privado. Partindo para implementação do indexing, além do construtor e destrutor de instâncias foi preciso desenvolver métodos para a leitura dos diretórios em que os .txt se encontravam e para ler os arquivos em si .

```cpp
#include "indexing.hpp"

indexing::indexing()
{
}

indexing::~indexing()
{
}
```

```cpp
set<string> indexing::read_directory(const string &name)
{
  DIR *dirp = opendir(name.c_str());
  if (dirp == NULL)
  {
    return set<string>();
  }

  struct dirent *dp;
  set<string> files;

  while ((dp = readdir(dirp)) != NULL)
  {
    if (dp->d_name[0] != '.')
      files.insert(dp->d_name);
  }
  closedir(dirp);

  return files;
}
```

Como o nome dos arquivos dentro do diretório ./Documentos não é fixo, foi necessário pesquisar uma forma de saber o nome de todos os arquivos dentro de uma pasta, que é o que foi implementado na imagem acima. Utilizando a biblioteca dirent.h, foi possível percorrer o nome de cada arquivo e guardá-los em um container set.

```cpp
void indexing::read_files(const string &name)
{
  set<string> files = read_directory(name);
  fstream file;
  string word, filename;

  for (auto it = files.begin(); it != files.end(); it++)
  {
    filename = *it;
    file.open(name + filename);
    while (file >> word)
    {
      this->insert(word, filename);
    }
    file.close();
  }
}
```

Agora utilizando a biblioteca bits/stdc++, foi possível abrir cada arquivo com o nome salvo no set e inserir cada palavra com seu respectivo nome de arquivo com a função insert.

```cpp
void indexing::insert(string word, string filename)
{
  string normalized_word = this->normalize(word);
  if (normalized_word != "")
  {
    this->index[normalized_word].insert(filename);
  }
}

map<string, set<string>> indexing::get_index()
{
  return this->index;
}
```

A função insert, que recebe a palavra e o nome do arquivo onde ela foi encontrada, primeiro normaliza a palavra e depois a insere no index.

```
string indexing::normalize(string word)
{
  string normalized_word = "", aux;

  string comAcentos = "ÄÅÁÂÀÃäáâàãÉÊËÈéêëèÍÎÏÌíîïìÖÓÔÒÕöóôòõÜÚÛüúûùÇç";
  string semAcentos = "AAAAAAaaaaaEEEEeeeeIIIIiiiiOOOOOoooooUUUuuuuCc";

  for (int i = 0; i < word.length(); i++)
  {
    // Verificando se o caractere é maiúsculo
    aux = tolower(word[i]);

    // Verificando se o caractere tem acento
    for (int j = 0; j < comAcentos.length(); j += 2)
    {
      if (word[i] << word[i + 1] == comAcentos[j] << comAcentos[j + 1])
      {
        aux = semAcentos[j / 2];
        aux = tolower(aux[0]);
        break;
      }
    }

    // Verificando se o caractere é um caracter especial
    if (isalpha(aux[0]))
    {
      normalized_word += aux;
    }
  }

  return normalized_word;
}
```

Para permitir uma pesquisa mais precisa utilizamos o método normalize que recebe a palavra que será normalizada, e passando caractere por caractere dela, torna letras maiúsculas em minúsculas, através do tolower(), retira caracteres especiais, como pontuação, exclamação, vírgula, etc, com a função isalpha().

Fora isso temos que percorrer a string comAcentos comparando se ao caractere da palavra é acentuado, e se sim, substituí-lo por um caractere da string semAcentos. Uma dificuldade na implementação dessa função foi que mesmo que os caracteres acentuados pareçam visualmente ter o mesmo tamanho que um sem acentos, eles ocupam duas casas dentro de um array. Por isso foi necessário fazer a comparação de duas posições juntas da string word com a string comAcentos, e a posição de semAcentos dividido por dois, já que o laço estava sendo percorrido de dois em dois.

```cpp
vector<pair<string, int>> indexing::recovery(vector<string> query)
{
  set<string> normalized_query;
  string aux;

  for (int i = 0; i < query.size(); i++)
  {
    aux = this->normalize(query[i]);
    if (aux != "")
    {
      normalized_query.insert(aux);
    }
  }

  map<string, int> relevant_files;

  for (auto it = this->index.begin(); it != this->index.end(); it++)
  {
    if (normalized_query.find(it->first) != normalized_query.end())
    {
      for (auto it2 = it->second.begin(); it2 != it->second.end(); it2++)
      {
        relevant_files[*it2]++;
      }
    }
  }
```

Agora para a função de recovery, responsável por retornar os arquivos com a pesquisa feita pelo usuário, precisamos primeiramente normalizar os parâmetros de pesquisa recebidos.

Feito isso, podemos então criar um map responsável por contar quantas vezes a pesquisa do usuário está presente no index. Dentro de um loop que percorre as chaves do index, o programa compara se a palavra do index está presente no container normalized_query, se sim então ele irá percorrer o conteiner set do index, que guarda os arquivos onde a palavra ocorre, e colocar o arquivo no container relevant_files, onde cada vez que ele aparecer na pesquisa irá ser somado mais um ao seu valor.

```cpp
  vector<pair<string, int>> ordered_files;

  for (auto it = relevant_files.begin(); it != relevant_files.end(); it++)
  {
    ordered_files.push_back(make_pair(it->first, it->second));
  }

  sort(ordered_files.begin(), ordered_files.end(), conditional_sort);

  return ordered_files;
}
```

```cpp
bool conditional_sort(pair<string, int> &a, pair<string, int> &b)
{
  return a.second > b.second;
}
```

Para que fosse possível ordenar os arquivos por ordem de relevância, foi necessário transferi-los para um container vector, onde o valor de int, e utilizar a função sort() para ordenar os valores do container. A função sort recebe como parâmetro o início e fim do container que desejamos ordenar, e também a lógica de ordenação que foi implementada da funlçao conditional_sort. Com isso o vector ordered_files é ordenado tanto pela quantidade de ocorrências quanto pela ordem lexicográfica, e finalmente retorna o vector. E por fim, para imprimir o resultado no terminal do usuário foi implementado a função print_ordered_files, que recebe os o vetor ordenado e o percorre para imprimir em tela.

```cpp
void indexing::print_ordered_files(vector<string> query)
{
  vector<pair<string, int>> ordered_files = this->recovery(query);

  if (ordered_files.size() == 0)
  {
    cout << "\nNenhum arquivo encontrado!" << endl;
  }
  else
  {
    cout << "\nArquivos encontrados:" << endl;
    for (int i = 0; i < ordered_files.size(); i++)
    {
      cout << ordered_files[i].first << " - " << ordered_files[i].second << " ocorrências" << endl;
    }
  }
}
```

```
O que deseja buscar?
 casa quem apartamento quer

Arquivos encontrados:
d1.txt - 4 ocorrências
d2.txt - 3 ocorrências
d3.txt - 2 ocorrências


Deseja fazer outra pesquisa? (s/n)
```

```cpp
int main()
{

  string option, word, aux;
  indexing search;
  search.read_files("./documentos/");

  do
  {
    vector<string> query;

    cout << "O que deseja buscar?\n ";
    getline(cin, word);

    for (int i = 0; i <= word.length(); i++)
    {
      if (word[i] == ' ' || i == word.length())
      {
        query.push_back(aux);
        aux = "";
      }
      else
      {
        aux += word[i];
      }
    }

    search.print_ordered_files(query);
    do
    {
      cout << "\n\nDeseja fazer outra pesquisa? (s/n)" << endl;
      cin >> option;

      if (option != "s" && option != "n")
      {
        cout << "Opção inválida!" << endl;
      }

    } while (option != "s" && option != "n");

    if (option == "s")
    {
      cout << "\n-----------------------------------------\n"
           << endl;
    }

    cin.ignore();

  } while (option == "s");

  return 0;
}
```

Para que o usuário tenha uma interface onde utilizar o programa, no main foi implementado um menu com do/while. O usuário insere a string de pesquisa na variável através do getline, essa string é particionada palavra por palavra dentro de um vector para facilitar a manipulação, e é chamado a função de recovery passando este vector como parâmetro, retornando o resultado que foi mostrado mais acima. Além disso, o usuário tem a opção de sair do programa ou realizar uma nova consulta.

E com o objetivo de garantir o desempenho correto de todos os métodos do projeto, foi feito uma série de testes de unidade, via doctest.h, colocando diversas formas de entradas para corroboração da corretude do algoritmo desenvolvido.

```cpp
#define DOCTEST_CONFIG_IMPLEMENT_WITH_MAIN
#include "doctest.h"

#include "../include/indexing.hpp"

TEST_CASE("Testa Indexing::Indexing()")
{
  indexing test;
  CHECK(test.get_index().size() == 0);
}

TEST_CASE("Testa Indexing::read_directory()")
{
  indexing test;
  set<string> files = test.read_directory("./tests");
  CHECK(files.size() == 3);
  CHECK(files.find("test.cpp") != files.end());
  CHECK(files.find("doctest.h") != files.end());
  CHECK(files.find("documentos_test") != files.end());
}

TEST_CASE("Testa Indexing::read_files()")
{
  indexing test;
  test.read_files("./tests/documentos_test/");
  CHECK(test.get_index().size() == 3);
}

TEST_CASE("Testa Indexing::Insert()")
{
  indexing test;
  test.insert("teste", "teste.txt");
  CHECK(test.get_index().size() == 1);
}
```

Conclusão

Ao término da aplicação foi possível desenvolver não apenas conhecimentos a respeito da matéria e da linguagem C++, como também o desenvolvimento prático de um projeto.

A priori os fundamentos discutidos em aula foram fundamentais para o desenvolvimento do trabalho como boas práticas e modularização que permitiram uma boa compreensão e organização do código, bem como a programação orientada a objeto e controle de versão que tornaram possível o desenvolvimento do mesmo projeto em paralelo entre os integrantes.

Além dos conceitos discutidos em aula já citados, houve também a busca por informações mais técnicas da ferramenta como por exemplo as funções isalpha(), que verifica se o caracter é uma letra, tolower(), que transforma letras maiúsculas em minúsculas, além do getline(), que foi necessário utilizar a função cin.ignore() para que funcionasse mais de uma vez no menu do usuário.

Vale destacar que graças a esse TP descobrimos coisas como uma letra acentuada vale por 2 espaços numa string, por exemplo, a string comAcento tem o dobro do tamanho da semAcento, presentes no método normalize, e como um espaço sozinho gera um caractere inválido, necessitando percorrer esse vetor de duas em duas casas.

Concluindo, foi um projeto para treinar nossos conhecimentos básicos, mas também foi desafiador em certos pontos, o que agregou bastante para o aprendizado na disciplina.