Universidade Federal de Pernambuco – UFPE

Centro de Informática – CIn

# Processamento de Cadeias de Caracteres

# Buscador ipmt

Recife – 2018.2

# Processamento de Cadeias de Caracteres

# Buscador ipmt

Relatório do Projeto

Escrito por:

Alexsandro Vítor Serafim de Carvalho (avsc@cin.ufpe.br).

Entrega do relatório: 21/10/2018

# Sumário

# 1. Introdução

O projeto que este relatório apresenta trata-se de um indexador / buscador de textos em índice feito com o objetivo de aplicar os conhecimentos adquiridos na cadeira de Processamento de Cadeias de Caracteres.
Este relatório descreverá seu funcionamento.

# 2. Implementação

Ao realizar a busca, o programa pode ser configurado para retornar os matches (match = local no texto onde o padrão é encontrado) encontrados ou apenas a quantidade deles, através das opções equivalentes -c ou --count.
Cada match é retornado da seguinte forma:

[path]:[linha]:[coluna]:Lorem ipsum **match** dolor sit amet

Onde [path] é o caminho do arquivo onde houve o match, [linha] e [coluna] definem as coordenadas no texto do primeiro caractere do match e são seguidas por uma transcrição da linha onde ocorreu o match com o mesmo destacado.

Mais detalhes sobre as opções são dados no arquivo README.md e usando as opções -h ou --help na aplicação.

## 2.1. Algoritmos

Os algoritmos implementados neste trabalho foram o LZ-77 para compressão e o array de sufixos para indexação, com busca através de busca binária.

## 2.2. Limitações

A especificação do C++17 conta com uma biblioteca de string views. Ela não foi usada no projeto porque o meu compilador não a possuía. Por isso tive que usar o método std::basic_string::insert, que possui custo O(n), para formatar a saída no lugar de usar views que poderiam ser cortadas em tempo constante.

Os algoritmos de busca são aplicados linha por linha. Por isso, não é possível buscar por strings com quebras de linha no meio delas. Além disso, uma string composta por mais de uma palavra só pode ser encontrada se as duas palavras estiverem na mesma linha do texto.

O caractere '\n' foi usado como separador para diferentes partes do índice. Por causa disso, ele não pode ser usado no alfabeto.

Os parâmetros ls e ll podem variar apenas de 0 a 255, pois eles são representados pelos 2 primeiros caracteres do arquivo de índice.

## 2.3. Bugs conhecidos

A formatação de cores só funciona no Linux ou no Git Bash do Windows, pois o Prompt de Comando do Windows não dá suporte à formatação usada no Linux.

Ao testar o programa com o arquivo sources50MB, do Pizza&Chili, houve um segmentation fault.