Nome da ferramenta:
itmp

Identificação da equipe:
Duhan Caraciolo (dcms2)
Mateus Moury (mmfrb)

Descrição do funcionamento da ferramenta:
A ferramenta tem por objetivo fazer um índice completo utilizando um array de sufixos de um texto, para que depois possa ser usado na busca por padrões em tempo linear sem precisar pré-processar o texto. Para se evitar o uso excessivo de espaço ao armazenar o índice no HD, o mesmo é salvo em um arquivo em formato comprimido utilizando o algoritmo de compressão LZW, que é baseado no LZ78.
Para tal, a ferramenta consiste em dois modos:
→ index: dado um texto, cria o índice completo e salva o texto mais o índice em um arquivo com formato comprimido.
→ search: dado uma lista de padrões e vários arquivos de índice, inicialmente descomprime o índice e depois calcula as ocorrências de cada padrão em cada um dos arquivos.

A busca de um padrão P em um texto T é $O(|P| + \log |T|)$.
A compressão/descompressão é linear no tamanho do texto.

Detalhes de implementação relevantes:
→ Compressão/Descompressão:
Para se fazer a compressão é necessário saber qual o código de uma dada palavra em um dicionário, e também devemos inserir uma nova palavra (que na verdade é uma palavra que já existe no dicionário mais um novo caractere no final) no dicionário com um dado código.
Implementamos uma Trie adaptada que permite fazer as duas operações em tempo constante, ou seja, independente do tamanho da palavra a ser buscada ou inserida. Pois, da forma que o algoritmo funciona, basta mantermos em qual nó da trie estamos e para inserir a nova palavra basta criar um novo filho para esse nó; já para saber qual o código da palavra, basta olharmos o código que se encontra nesse nó.

Limitações de desempenho notáveis:
→ Busca de padrões:
A construção do array de sufixo de um texto T é $O(|T| * \log |T|)$, portanto um texto com tamanho 10^7 já nos dá um pouco de dor de cabeça.
É necessário também $O(|T| * \log |T| * 4)$ bytes de memória para passos intermediários da construção do índice.

Extras:
Implementamos três opções extras, a primeira é o número de ocorrências por padrão em cada um dos textos, a segunda é a soma das quantidades de ocorrências de todos os padrões em cada texto e a terceira é a opção da impressão das linhas ser ou não ordenada em relação ao texto, o valor padrão é não ordenar.
É possível também pesquisar pelo mesmo conjunto de padrões em vários arquivos de índice, utilizando wildcards ou listando de um por um, ou ambos. Infelizmente não conseguimos fazer o mesmo com vários arquivos de padrões, apenas um pode ser utilizado.