Trabalho Prático
Algoritmos 2
Manipulação de sequências

William Teles Dias
2020109977

Implementação:

Nesse trabalho prático, implementei o algoritmo de compressão de arquivos lz78. Para isso, utilizei uma trie tradicional para encontrar e guardar padrões que se repetiam no arquivo a ser comprimido.
A compressão se dá de acordo com o pseudocódigo presente no artigo da wikipédia: https://pt.wikipedia.org/wiki/LZ78
Faço esse algoritmo duas vezes, sendo que a primeira tem por objetivo apenas descobrir a quantidade de padrões que serão guardados, para saber quantos bytes são necessários para indexar todas as sequências e coloco esse número no início do arquivo de saída.

- A árvore começa com a string vazia associada ao valor 0.
- e guardo uma string
- Para cada caractere lido checamos se a string concatenada com o caractere já está na trie
    - Se estiver,
        - adiciono o caractere lido a nossa string
    - Senão,
        - Coloco o código associado à string lida até agora (Em forma de bytes) e o caractere que acabamos de ler na saída.
        - coloco a string + caractere na trie
        - string volta a ser vazia

A descompressão é feita usando a estrutura de dados de dicionário presente por padrão na linguagem Python, também conforme o artigo da wiki acima.

Taxas de Compressão:

Essas são as taxas de compressão dos arquivos, os números são a quantidade de bytes em cada arquivo, as razões são: arquivo.z78 / arquivo.txt.

the_iliad.txt: 777392 / 1135098 = 34.98%
moby_dick.txt: 817767 / 1253916 = 34.78%
romeo_and_juliet.txt: 99136 / 163623 = 39.41%
frankenstein.txt: 309579 / 440869 = 29.78%
alice.txt: 98139 / 170518 = 42.45%
dracula.txt: 566284 / 865819 = 35.60%
the_odyssey.txt: 463836 / 705354 = 34.24%
ulysses.txt: 1036723 / 1553117 = 34.24%
the_republic.txt: 734253 / 1219023 39.78%

peter_pan.txt: 154626 / 282113 = 45.19%
winnie_the_pooh.txt: 86677 / 147955 = 41.41%

O texto lorem ipsum parece ter uma taxa de compressão mais alta que o normal pois é bastante repetitivo
loremipsum.txt: 664887 / 2167737 = 69.33%

dom_casmurro.txt: 294587 / 409610 = 28.08%
os_lusiadas.txt : 195080 / 344538 = 43.38%
constituicao1988.txt: 353362 / 651790 = 45.79%