Trabalho Prático 1

Samuel Gonçalves Leite

1 Introdução

O objetivo deste trabalho é implementar um compressor de texto seguindo o algorritmo LZ78. Para isso, usamos de conceitos e estruturas estudadas na seção de manipulação de sequências em sala de aula.

2 Implementação

O programa foi desenvolvido na linguagem C++.

O algoritmo consiste na criação de um dicionário, tanto no processo de compressão quanto de descompressão dos arquivos. Para armazenar este dicionário se fez uso de uma árvore de prefixos. Além de amenizar a complexidade de espaço gasta para $O(n)$, em comparação a $O(n^2)$ em um dicionário tradicional, as inserções e buscas se tornam mais rápidas.

A raiz da árvore não guarda nenhum caracter e tem índice 0, enquanto os nós subsequentes guardam um caracter cada e seus indices seguem ordem crescente de inserção. Foi feito o uso da biblioteca *map* para guardar os filhos de cada nó, assim não se gasta memória alocando desnecessariamente ponteiros nulos para cada possível letra a partir de um certo prefixo.

A saída é formada por blocos (índice, caracter), que no arquivo são representados por 3 bytes para o índice, que foram suficientes nos exemplos testados, e um byte para o caracter.

3 Resultados

Segue uma tabela com alguns exemplos de textos e suas taxas de compressão.

Exemplos			
Arquivo	Tamanho original	Tamanho comprimido	Diminuição
Constituição 1988	637 KB	348 KB	$45,4\%$
Os lusíadas - Luís de Camões	337 KB	255KB	$24,3\%$
Dom Casmurro - Machado de Assis	401 KB	290 KB	$27,7\%$
Ulysses - James Joyce	1525 KB	993 KB	$34,9\%$
os 1 milhão primeiros dígitos de π	977 KB	716 KB	$26,7\%$

Exemplos			
Arquivo	Tamanho original	Tamanho comprimido	Diminuição
Divina Comédia - Dante Alighieri	583 KB	395 KB	$32,2\%$
Moby Dick - Herman Melville	1247 KB	808KB	$35,2\%$
A Metamorfose - Franz Kafka	139 KB	109 KB	$21,6\%$
Os Irmãos Karamazov - Fiódor Dostoiévski	1994 KB	1161 KB	$41,1\%$
Triste Fim de Policarpo Quaresma - Lima Barreto	425 KB	310 KB	$27,1\%$
Total	8265 KB	5385 KB	$34,8\%$

4 Conclusão

Enquanto a média ponderada das compressões foi de $34,8\%$, a média simples foi de $31,6\%$, isso reflete como o algoritmo comprime melhor textos mais longos, nos quais é provável que haja mais repetição.

5 Referências

https://www.gutenberg.org/
https://assets.angio.net/pi1000000.txt