Documentation - Work Project 1

Class: Algorithms II 2023/1
Student: Gustavo Tavares Corrêa

Computer Science Department – Federal University of Minas Gerais (UFMG) – Belo Horizonte – MG – Brazil

1. Introduction

The project consisted in implementing a text compressing program through the use of the LZ78 algorithm implemented with a trie tree to represent the inherent dictionary of the algorithm. The texts the program should be able to compress can be encoded as UTF-8 characters, and so there is a variable width to the number of bytes used to represent a character which must be treated. Besides that, the overall number of nodes in the trie tree can also affect the number of bytes used to write the integers. The compression must result in a binary file in order to be efficient, and there must also be a way to decompress the files.

2. Implementation

I implement two different object classes. The TreeNode class, which has a value property to represent indicates which previous node in the dictionary may be a prefix to the text represented by that node, a prefix property which the last element of the prefix of the text represented by that node, and a sons property which is a list of all the nodes which are sons of that one. There is also a TrieTree class, which has a root initialized as an empty node and a num property to indicate the number of elements in the tree. It also has the functions printTree, to help visualize it; countNum, to count the number of nodes in it; compress, which compresses texts according to their encodings and the size of integers in binary; and decompress, which turns binary texts which were compressed to .z78 files back into .txt files.

There is also a main function, responsible for dealing with the command line arguments, creating the classes objects and calling the other functions. A problem the code has is dealing with oriental characters and emojis copied from the internet, which should be 3 and 4 bytes UTF-8 characters, and when printing their binary codification it does appear correct, but the encoding table warns of an error and is not able to convert them adequately. One of the possible reasons I've researched for this, but was unable to find a solution for, is the differing encodings of the internet from which the characters were copied without further analysis.

The program taxes of compression, as in size of the original .txt file / size of the .z78 generated file are as follows, for the 10 examples also present is this same repository:

- Brazilian Constitution of 1988 (constituição1988.txt): 637kb/346kb = 1.84
- Os Lusíadas (os_lusiadas.txt): 337kb/191kb = 1.76
- Dom Casmurro (dom_casmurro.txt): 401kb/288kb = 1.39
- Renato Vimieiro's Lattes curriculum (I hope you don't mind teacher, this is a small marketing of your work and I plan on replacing this later after you read it and hopefully have a laugh :)) (vimieiro.txt): 61kb/41kb = 1.49
- Bee Movie Script (bee.txt): 89kb/55kb = 1.62
- Angra albums and songs lyrics (angra.txt): 85kb/54kb = 1.57
- A text file of my name repeated until it got to the size I wanted (gustavo.txt): 1941kb/16kb = 122.56 (I do think this deserves a special mention considering how well it shows that the repetition of prefixes can be compressed extremely, as was the purpose of the project as application of the algorithm)
- Random ASCII characters (randomASCII.txt): 1084kb/1209kb = 0.90
- This very documentation written up to this point (doc.txt): 5kb/5kb = 1
- Shrek movie script (shrek.txt): 74kb/46kb = 1.61

3. Bibliography

- https://en.wikipedia.org/wiki/UTF-8
- https://pt.wikipedia.org/wiki/LZ78#Exemplo
- https://www.geeksforgeeks.org/how-to-convert-bytes-to-int-in-python/
- https://en.wikipedia.org/wiki/LZ77_and_LZ78
- https://stackoverflow.com/questions/1543613/how-does-utf-8-variable-width-encoding-work?noredirect=1&lq=1
- https://stackoverflow.com/questions/8815592/convert-bytes-to-bits-in-python
- https://stackoverflow.com/questions/606191/convert-bytes-to-a-string
- https://blog.gitnux.com/code/python-reading-binary-files/
- https://stackoverflow.com/questions/27238680/writing-integers-in-binary-to-file-in-python
- https://docs.python.org/3/library/stdtypes.html#str.encode
- https://stackoverflow.com/questions/2294608/how-to-write-integer-number-in-particular-no-of-bytes-in-python-file-writing
- https://stackoverflow.com/questions/34009653/convert-bytes-to-int
- https://stackoverflow.com/questions/14329794/get-size-in-bytes-needed-for-an-integer-in-python