Data Compression Project

1 Task

The project consists of experimenting with various data compression schemes that we discussed in the class, comparing their performance (compression ratio, compression/decompression times) on a given collection of datasets and analyzing the experimental results.

The project can be done either individually or in groups of size two. Let me know your preference (individual/group) soon.

You need to compress the given collection of datasets using the following set of algorithms (discussed in the class):

1. Static Huffman coding
2. Adaptive Huffman coding (FGK)
3. Golomb codes (choose an appropiate order to suit each dataset)
4. Tunstall codes
5. Arithmetic coding
6. LZ77
7. LZW

For each of the algorithms, you can (a) get the code online, (b) implement the algorithm yourself, or (c) get the code and modify it to suit your purposes (to improve the performance). You need to mention which option you chose for each algorithm (and a brief description of the modifications you did, if any), and also the sources from which you obtained the code (where applicable). In addition, compress the datasets with the compressors:

8. gzip
9. bzip2

2 Datasets

Different datasets will be given to each team (group/individual), in a couple of weeks.

3 Submission

You need to submit the following items:

- A tar file containing all your source code along with a Makefile that can be used to compile and test your code.
- A detailed report containing
    1. A brief description of the code used for each algorithm
    2. Methodology of the algorithms applied to the datesets
    3. Experimetal environment
    4. Instructions to run the codebase
    5. Tables/charts/graphs showing the comparison of compression ratio, compression time and decompression time
    6. Analysis of the experimental results (possible reasons why a particular algorithm(s) performed better or worse on a specific dataset etc.)
    7. References (if any)

[The experimental analysis (Item 6 above) is the crucial part in the report.]

4 Deadline

23:59 Tuesday, 1^{st} December, 2020. No late submissions allowed.