

1. Does the compression ratio depends on the file type? If yes, then how?

Ratios are calculated using this formula:

$$\text{Compression Ratio} = \frac{\text{Uncompressed Size}}{\text{Compressed Size}}$$

These are the average compression ratio for each group.

pdf 0.9883181655136718

doc 1.239070437696408

png 0.8059341415379446

jpg 0.7986324350669967

exe 0.9907281567724946

As we can see, on average only files with doc type were compressed (reduced in size).

png, jpg, exe, pdf files contain more kinds of bytes, in other words their bytes distribution is more uniform compared to files of doc type and they have a higher entropy (explanation was presented in the previous report).

So, the number of possible unique strings (by unique strings I mean the entries in the dictionary created by the LZ78 algorithm) for the same size must be bigger for png, jpg, exe, pdf files compared to doc type. That is why the compression for doc files works better.

2. Does the compression ratio depends on file size? (i.e. large text file vs small one, just text *.doc vs *.doc with with additional formatting). Explain your answers

Pearson correlation coefficients between file size and compression according to type :

pdf 0.3114149614079196

doc -0.40285739634980855

png -0.01294664522405109

jpg 0.01720832728238147

exe -0.025594861666820772

Says that it depends on the file type. However, judging by plots, where x axis is size, y axis is compression ratio, we can see that we have not enough files with larger sizes, so results might be inaccurate.



