Assignment IT Entropy

Aidar Ahmetshin

September 2019

1 Report

Code you can get here: https://github.com/brooky56/IT
Any computer file is known to consist of bytes. The byte can take values from 0 to 255. Information entropy is a statistical parameter that indicates the probability of occurrence of certain bytes in a file.

Visually assess the degree of entropy can be using a plot like histogram, the probability distribution of repetitions of the same bytes in the file. By the entropy of the file, we can guess what type of file is in front of us, seeing only its histogram.

In our case we calculate bytes so for us the maximum entropy according to the Shannon formula:

$$H(x) = -\sum_{i=0}^{n} p(i) log_2(p(i)), \tag{1}$$

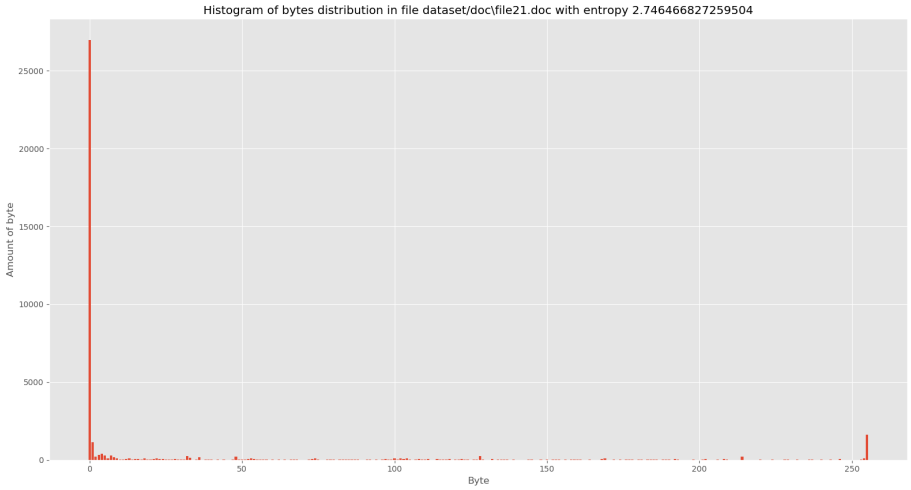
will be 8, why, because according to definition of the entropy we can assume that each byte from [0..255] occurs ones and then we calculate entropy for such situation
Entropy of independent random events x with 256 possible states (from 0 to n=255):

$$H(x) = -\sum_{i=0}^{n} \frac{1}{256} log_2(\frac{1}{256}) = 8.0 \tag{2}$$

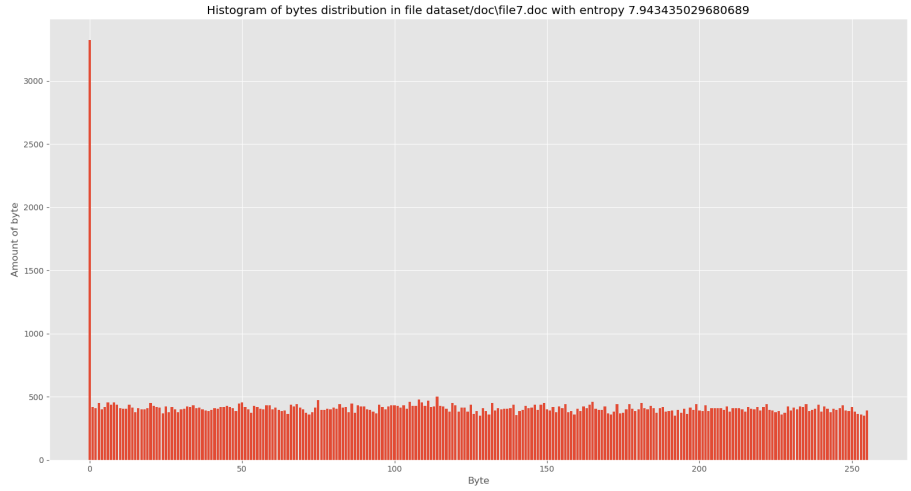
Now using our algorithm we can calculate entropies for each group of files in our dataset. We compute Min entropy, Max entropy and Average entropy for each file group:

```
Amount of files on this directory: 25
DOC:
Min entropy: 2.746466827259504
Max entropy: 7.943435029680689
Avg entropy: 4.953286428127747
Amount of files on this directory: 50
EXE:
Min entropy: 3.2219575413733796
Max entropy: 7.999110372457775
Avg entropy: 7.038762343238465
Amount of files on this directory: 25
JPG:
Min entropy: 7.787729962899609
Max entropy: 7.988069772317484
Avg entropy: 7.931090015975819
Amount of files on this directory: 24
PDF:
Min entropy: 6.402801741188575
Max entropy: 7.9873053013084965
Avg entropy: 7.764205337542925
Amount of files on this directory: 25
PNG:
Min entropy: 7.677151366353173
Max entropy: 7.997779200994514
Avg entropy: 7.9507714019713225
```

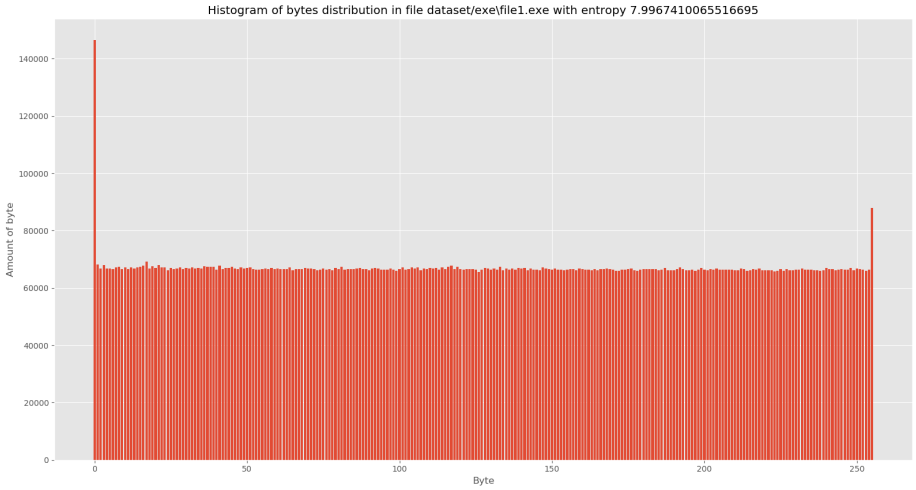
From this results we can do some conclusions: the minimum average entropy has .doc files, other groups has average score in very short limit [7.0 - 7.95]. Of course the average score gives us only general situation when we comparing files, because it highly depends on file it self and its content.
In the absence of information loss is numerically equal to the amount of information per symbol of the transmitted message, so the doc file shows us that it has problem with it. doc files has minimal value of it information per symbol, but for example pictures .png or .jpg has value close to the maximum, so the bring us the most value information. For more convenient visualization of entropy let plot histogram for file with smallest and maximum entropy. Doc file with the smallest entropy looks like:

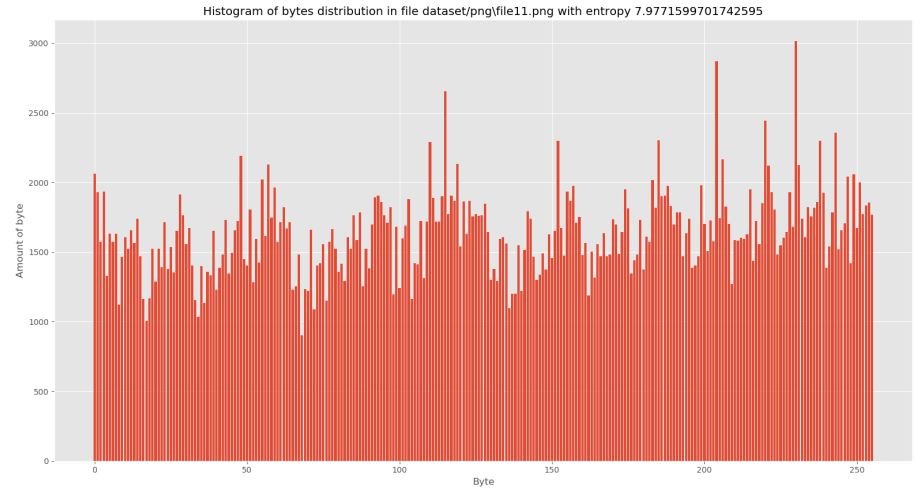


Doc file with high entropy looks like:



We see how differ it looks, also we see that a lot-of amount of 0 byte, it look likes spaces in file. But we can say that doc file with low entropy we can easily compress. Lets look for exe file histogram and png:





Png files have such distribution because it has already compressed by their formats, also pictures should have max entropy because they shows maximum information per symbol. Also zip file tends to 8 score.