Information Theory
Assignment 1: Entropy

Results

The output of a program is:

Average entropy (all files) 7.108441217549163
Maximal entropy for that range of values: 8.0
pdf :
  entropy: 7.764205337542926
  variance: 0.1224598026279547
doc :
  entropy: 4.953286428127747
  variance: 1.850152394051591
png :
  entropy: 7.9507714019713225
  variance: 0.006312809554504041
jpg :
  entropy: 7.931090015975822
  variance: 0.002990988690714317
exe :
  entropy: 7.038762343238471
  variance: 1.0887946486561948

Sorted by entropy:
doc 4.953286428127747
exe 7.038762343238471
pdf 7.764205337542926
jpg 7.931090015975822
png 7.9507714019713225

Sorted by variance:
jpg 0.002990988690714317
png 0.006312809554504041
pdf 0.1224598026279547
exe 1.0887946486561948
doc 1.850152394051591

Explanation of the results

Average entropy (all files) 7.108441217549163

Sorted by entropy

1) doc 4.953286428127747
2) exe 7.038762343238471
3) pdf 7.764205337542926
4) jpg 7.931090015975822
5) png 7.9507714019713225

1) **Text files** have low entropy since *ordinary people's language contains less characters than number of possible bytes*. For example, English has about 52 symbols + some special like !,.? etc. against 256 values of bytes.

2) **Binary files**: binary files *use opcodes* which can be repeated again and again within the same binary file (that is why some bytes can be met more often that others). Also binary files *have .data section* and there can be some text or other data with low entropy

3, 4, 5) Next come **compressed types of files**, since compressed data *requires fewer bits to store the same information*, entropy is increased.
The entropy of those is close to the maximal one (8.0) because of the compression.

Pdf have less entropy among other compressed types probably because PDF files can choose to *use compression only for some sections* of their content such as images and leave other sections unchanged.

Here (looking at the variancy) we can see that *entropy differs for the files of the same type*:

1) jpg 0.002990988690714317
2) png 0.006312809554504041
3) pdf 0.1224598026279547
4) exe 1.0887946486561948
5) doc 1.850152394051591

Explanation:
1)jpg has low variance in entropy because of the compression
2)png has low variance in entropy because of the compression
3)pdf entropy vary more than png and jpg, probably because of the uncompressed parts
4)binaries' entropy can vary, because of the .data section (which contains global or static variables which have a pre-defined value)
5)text files' entropy can vary because one text file can be much more gibberish than another one (example: "War and peace" vs some private key)