Huffman Empirical Analysis

Part A:

Which compresses more, binary files or text files?

--Text!

Calgary

bib from 111261 to 73791 in 0.165
book1 from 768771 to 439405 in 0.968
book1.unhf from 768771 to 439405 in 0.909
book2 from 610856 to 369331 in 0.751
geo from 102400 to 73588 in 0.172
news from 377109 to 247424 in 0.502
obj1 from 21504 to 17081 in 0.039
obj2 from 246814 to 195127 in 0.417
paper1 from 53161 to 34367 in 0.070
paper2 from 82199 to 48645 in 0.102
paper3 from 46526 to 28305 in 0.056
paper4 from 13286 to 8890 in 0.019
paper5 from 11954 to 8461 in 0.018
paper6 from 38105 to 25053 in 0.049
pic from 513216 to 107582 in 0.229
progc from 39611 to 26944 in 0.053
progl from 71646 to 44013 in 0.096
progp from 49379 to 31244 in 0.069
trans from 93695 to 66248 in 0.137
---
total bytes read: 4032556
total compressed bytes 2288306
total percent compression 43.254
compression time: 4.901

Waterloo

clegg.tif from 2149096 to 2034591 in 4.290
frymire.tif from 3706306 to 2188589 in 4.565
lena.tif from 786568 to 766142 in 1.613
monarch.tif from 1179784 to 1109969 in 2.284
peppers.tif from 786568 to 756964 in 1.550
sail.tif from 1179784 to 1085497 in 2.298
serrano.tif from 1498414 to 1127641 in 2.384
tulips.tif from 1179784 to 1135857 in 2.568
---
total bytes read: 12466304
total compressed bytes 10205250
total percent compression 18.137
compression time: 21.552

The results listed above are collected from running Huffmark on the Calgary and Waterloo folders: they contain text files and binary files respectively.

As indicated by the figure, the program can compress 43.254 percent of the original size of text files, which Waterloo's folder only get 18/137 compressed. Thus, it is obvious that text files can be compressed more than binary files. Also, the compression time of binary files are significantly longer that that for the text files: this is mainly due to the fact that image files are much larger than text files.

Part b

Second Time Compression
We modified the line to be:
if (!f.getName().endsWith(SUFFIX)) return;
so the program only compresses files that have been compressed once. ☺_

Calgary_ Second Time Compression

```
bib.hf from     73791 to  73743 in  0.181
book1.hf from   439405 to      434976 in      0.984
book1.unhf.hf from   439405 to      434976 in      0.920
book1comp.hf from    439405 to      434976 in      0.951
book2.hf from   369331 to      368055 in      0.786
geo.hf from     73588 to  74218 in  0.166
news.hf from    247424 to      247024 in      0.515
obj1.hf from    17081 to  17690 in  0.045
obj2.hf from    195127 to      194735 in      0.455
paper1.hf from  34367 to  34971 in  0.078
paper2.hf from  48645 to  49039 in  0.118
paper3.hf from  28305 to  28877 in  0.076
paper4.hf from  8890 to   9436 in   0.022
paper5.hf from  8461 to   9008 in   0.021
paper6.hf from  25053 to  25692 in  0.068
pic.hf from     107582 to      72678 in  0.149
progc.hf from   26944 to  27461 in  0.057
progl.hf from   44013 to  43824 in  0.096
progp.hf from   31244 to  31530 in  0.094
trans.hf from   66248 to  66377 in  0.199
--------
total bytes read: 2727711
total compressed bytes 2682176
total percent compression 1.669
compression time: 6.010
```

Waterloo (second time compression)

```
clegg.tif.hf from    2034591 to     2028478 in     4.613
frymire.tif.hf from  2188589 to     2053952 in     4.643
lena.tif.hf from     766142 to      767432 in      1.841
monarch.tif.hf from  1109969 to     1111378 in     2.475
peppers.tif.hf from  756964 to      758106 in      1.576
sail.tif.hf from     1085497 to     1086651 in     2.404
serrano.tif.hf from  1127641 to     1120053 in     2.403
tulips.tif.hf from   1135857 to     1137295 in     2.349
--------
total bytes read: 10205250
total compressed bytes 10063345
total percent compression 1.391
compression time: 22.304
```

As indicated by the data, second compression on Calgary yielded a rate of 1.669 percent. And second compression on Waterloo provides additional 1.391 percent of further compression. It is obvious that there is an eventual limit to the compressibility of files. Noticeably, compressing the files further would have a negligible effect on both directories. If a file is intentionally built to be compressed a lot, it might not be worthwhile to further compress after its compressibility is utilized. As shown in the empirical analysis, after the second compression is performed it is not very worthwhile to further compress. (This also attests to the effectiveness of Huffman program as it leaves little space for further compression possibility.)