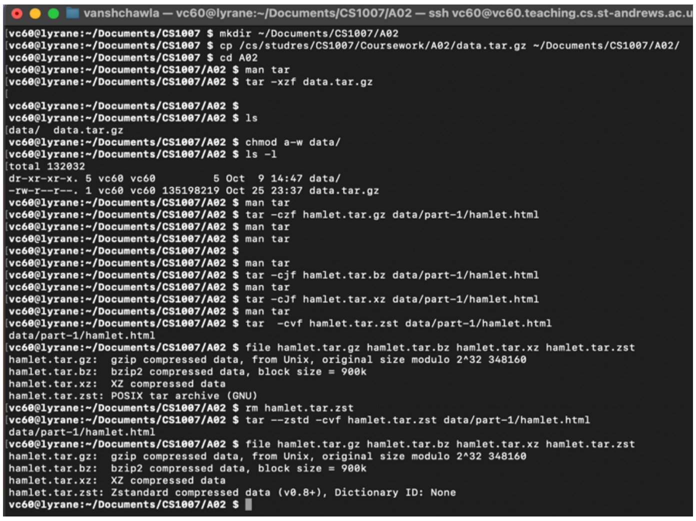
CS1007 A02 240021654 25 October 2024

Overview: In this practical we were asked to focus on data compression techniques.

Part 1:

```
mkdir ~/Documents/CS1007/A02
cp /cs/studres/CS1007/Coursework/A02/data.tar.gz
~/Documents/CS1007/A02/
cd ~/Documents/CS1007/A02
tar -xzf data.tar.gz
tar -czf hamlet.tar.gz data/part-1/hamlet.html
tar -cjf hamlet.tar.bz data/part-1/hamlet.html
tar -cJf hamlet.tar.xz data/part-1/hamlet.html
tar --zstd -cvf hamlet.tar.zst data/part-1/hamlet.html
file hamlet.tar.gz hamlet.tar.bz hamlet.tar.xz hamlet.tar.zst
```

Box 1: commands which were ran for part 1.



vanshchawla — vc60@lyrane:~/Documents/CS1007/A02 — ssh vc60@vc60.teaching.cs.st-andrews.ac.u

```
vc60@lyrane:~/Documents/CS1007 $ mkdir ~/Documents/CS1007/A02
vc60@lyrane:~/Documents/CS1007 $ cp /cs/studres/CS1007/Coursework/A02/data.tar.gz ~/Documents/CS1007/A02/
vc60@lyrane:~/Documents/CS1007 $ cd A02
vc60@lyrane:~/Documents/CS1007/A02 $ man tar
vc60@lyrane:~/Documents/CS1007/A02 $ tar -xzf data.tar.gz

vc60@lyrane:~/Documents/CS1007/A02 $
vc60@lyrane:~/Documents/CS1007/A02 $ ls
data/  data.tar.gz
vc60@lyrane:~/Documents/CS1007/A02 $ chmod a-w data/
vc60@lyrane:~/Documents/CS1007/A02 $ ls -l
total 132032
dr-xr-xr-x. 5 vc60 vc60         5 Oct  9 14:47 data/
-rw-r--r--. 1 vc60 vc60 135198219 Oct 25 23:37 data.tar.gz
vc60@lyrane:~/Documents/CS1007/A02 $ man tar
vc60@lyrane:~/Documents/CS1007/A02 $ tar -czf hamlet.tar.gz data/part-1/hamlet.html
vc60@lyrane:~/Documents/CS1007/A02 $ man tar
vc60@lyrane:~/Documents/CS1007/A02 $ man tar
vc60@lyrane:~/Documents/CS1007/A02 $
vc60@lyrane:~/Documents/CS1007/A02 $ man tar
vc60@lyrane:~/Documents/CS1007/A02 $ tar -cjf hamlet.tar.bz data/part-1/hamlet.html
vc60@lyrane:~/Documents/CS1007/A02 $ man tar
vc60@lyrane:~/Documents/CS1007/A02 $ tar -cJf hamlet.tar.xz data/part-1/hamlet.html
vc60@lyrane:~/Documents/CS1007/A02 $ man tar
vc60@lyrane:~/Documents/CS1007/A02 $ tar  -cvf hamlet.tar.zst data/part-1/hamlet.html
data/part-1/hamlet.html
vc60@lyrane:~/Documents/CS1007/A02 $ file hamlet.tar.gz hamlet.tar.bz hamlet.tar.xz hamlet.tar.zst
hamlet.tar.gz:  gzip compressed data, from Unix, original size modulo 2^32 348160
hamlet.tar.bz:  bzip2 compressed data, block size = 900k
hamlet.tar.xz:  XZ compressed data
hamlet.tar.zst: POSIX tar archive (GNU)
vc60@lyrane:~/Documents/CS1007/A02 $ rm hamlet.tar.zst
vc60@lyrane:~/Documents/CS1007/A02 $ tar --zstd -cvf hamlet.tar.zst data/part-1/hamlet.html
data/part-1/hamlet.html
vc60@lyrane:~/Documents/CS1007/A02 $ file hamlet.tar.gz hamlet.tar.bz hamlet.tar.xz hamlet.tar.zst
hamlet.tar.gz:  gzip compressed data, from Unix, original size modulo 2^32 348160
hamlet.tar.bz:  bzip2 compressed data, block size = 900k
hamlet.tar.xz:  XZ compressed data
hamlet.tar.zst: Zstandard compressed data (v0.8+), Dictionary ID: None
vc60@lyrane:~/Documents/CS1007/A02 $
```

Image 1: Screenshot of the results.

Part 4:

I used Microsoft excel for the analysis of my findings (AnalysisFinal.xlsx). First, I found out the ratios of all the 4 compression types, by dividing the compression size by original size and then I found their average (mean). Then I concluded that the means of the ratios are not enough to conclude for this analysis, so I made a different table which consisted of more statistics of the various compression ratios.

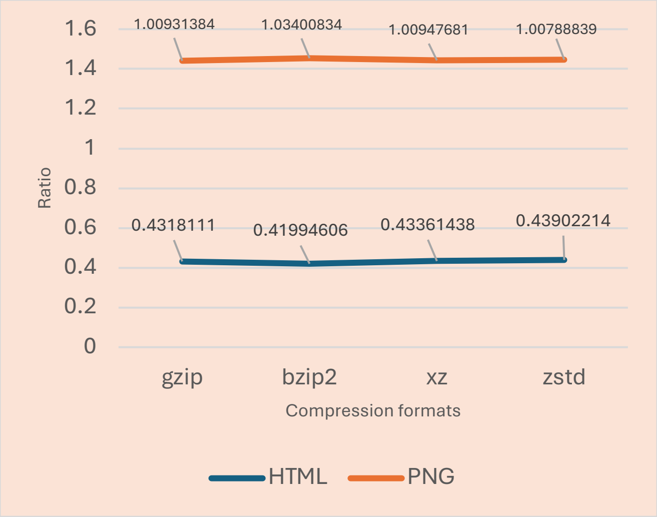
Compression	Mean ratio	Median ratio	Std Dev	Min ratio	Max ratio
gzip	0.92622486	1.00925996	0.21111591	0.25889566	1.03038005
bzip2	0.94565931	1.03125169	0.22630332	0.18576741	1.08653723
xz	0.92662384	1.0094496	0.21429856	0.22378194	1.02687001
zstd	0.926042	1.00801302	0.20742284	0.26772522	1.0240049

Mean and median compressions as per the various file types:

	html		png	
Compression	Mean	Median	Mean	Median
gzip	0.4318111	0.37848157	1.00931384	1.01052165
bzip2	0.41994606	0.35673287	1.03400834	1.03273697
xz	0.43361438	0.36740233	1.00947681	1.01085442
zstd	0.38878495	1.00788839	1.00788839	1.0087087

Table 1: A statistical analysis of the results.

Table 2: Mean and median compression of the various file formats.



Graph 1: A visual analysis of the mean for png and html files.

Lower mean and median compression ratios relate to the effectiveness of the algorithm in reducing file size. Besides, lower standard deviation values across the data sets imply the algorithm's effectiveness and its uniform performance across the HTML and png files due to low variance with respect to mean values. Based on my findings provided below, it is evident that the compression type with the lowest file size for html (bzip2) is the highest file size for a png file. Similarly, the smallest for a png file (zstd) is the highest for html. Hence, we can conclude that bzip2 is the preferred choice for compressing png files, while zstd is better for compressing HTML files. zstd would be the best format to export files if the directory has both html and png files due to its lowest mean compression ratio. Lastly, in the png comparison we can see that the median value is more than the mean value which implies that there is more variation in the compression of png files between the various formats. The reason behind the png files having a larger file size is that html's have more white spaces, taking up less storage, while pngs have more pixel usage. Both htmls and pngs don't consider semantics and encoding, but htmls removes extra whitespace, comments, and sometimes even inlines CSS/JS code which compressing, hence the less file size.

Box 2: Conclusion.