NAME

bzz - DjVu general purpose compression utility.

SYNOPSIS

Encoding:

bzz -e[blocksize] inputfile outputfile

Decoding:

bzz -d inputfile outputfile

DESCRIPTION

The first form of the command line (option -e) compresses the data from file *inputfile* and writes the compressed data into *outputfile*. The second form of the command line (option -d) decompressed file *inputfile* and writes the output to *outputfile*.

OPTIONS

-d Decoding mode.

-e[blocksize]

Encoding mode. The optional argument *blocksize* specifies the size of the input file blocks processed by the Burrows-Wheeler transform expressed in kilobytes. The default block sizes is 2048 KB. The maximal block size is 4096 KB. Specifying a larger block size usually produces higher compression ratios and increases the memory requirements of both the encoder and decoder. It is useless to specify a block size that is larger than the input file.

ALGORITHMS

The Burrows-Wheeler transform is performed using a combination of the Karp-Miller-Rosenberg and the Bentley-Sedgewick algorithms. This is comparable to (Sadakane, DCC 98) with a slightly more flexible ranking scheme. Symbols are then ordered according to a running estimate of their occurrence frequencies. The symbol ranks are then coded using a simple fixed tree and the ZP binary adaptive coder (Bottou, DCC 98).

The Burrows-Wheeler transform is also used in the well known compressor **bzip2**. The originality of **bzz** is the use of the ZP adaptive coder. The adaptation noise can cost up to 5 percent in file size, but this penalty is usually offset by the benefits of adaptation.

PERFORMANCE

The following table shows comparative results (in bits per character) on the Canterbury Corpus (http://corpus.canterbury.ac.nz). The very good bzz performance on the spreadsheet file excl puts the weighted average ahead of much more sophisticated compressors such as fsmx.

	Compression		I	
performance			·	
	text fax csrc excl sprc techpoemhtml lisp man play	Weight	edAverage	
+=====================================	ss 3.270.973.562.414.213.06 3.38 3.683.904.433.51	2.55	3.31	
gzip	2.850.822.241.632.672.71 3.23 2.592.653.313.12	2.08	2.53	
-9				
bzip2	2.270.782.181.012.702.02 2.42 2.482.793.332.53	1.54	2.23	
-9				
ppmd	2.310.992.111.082.682.19 2.48 2.382.433.002.53	1.65	2.20	
fsmx	2.10 0.79 1.89 1.48 2.521.84 2.21 2.242.292.912.35	1.63	2.06	
bzz	2.25 0.76 2.13 0.78 2.672.00 2.40 2.522.603.192.52	1.44	2.16	

Note that DjVu contributors have several entries in this table. Program compress was written some time ago by Joe Orost. Program ppmd is an improvement of the PPM-C method invented by Paul Howard.

CREDITS

Program bzz was written by Leon Bottou <leonb@users.sourceforge.net> and was then improved by Andrei Erofeev <andrew_erofeev@yahoo.com>, Bill Riemers <docbill@sourceforge.net> and many others.

SEE ALSO

djvu(1), compress(1), gzip(1), bzip2(1)