NAME

utfcheck − Check a file to verify that it is valid UTF-8 or ASCII

SYNOPSIS

utfcheck [−a] [−q] [−−expurgated] [−i *input_file.beta*] [−o *output_file.utf8*]

DESCRIPTION

utfcheck(1) reads an input file and prints messages about contents that might be unexpected (even if legal Unicode) in a UTF-8 or ASCII file, such as embedded control characters or Unicode "noncharacters". No diagnostic messages are printed for the control characters horizontal tab, vertical tab, line feed, or form feed. A final summary will indicate if null, carriage return, or escape characters were read.

utfcheck will detect a UTF-16 big-endian or little-endian Byte Order Mark at the beginning of a file and quit if it sees one. Support for UTF-16 is not implemented.

OPTIONS

−a Test for a pure ASCII file. ASCII control characters are allowed, but **utfcheck** will fail if it encounters a byte with value greater than hexadecimal 7F (the delete control character).

−i Specify the input file. The default is STDIN.

−o Specify the output file. The default is STDOUT.

−q Quiet mode. Do not print any output unless an illegal byte sequence is detected.

−−expurgated
Check a UTF-8 file against the "expurgated" version of the Unicode Standard, the one without the Byte Order Mark, after Monty Python's "Bookshop" skit with the "expurgated" version of *Olsen's Standard Book of British Birds*, the one without the gannet—because the customer didn't like them. (But they've all got the Byte Order Mark. It's a standard part of the Unicode Standard, the Byte Order Mark. It's in all the books.) This option is not abbreviated, to keep the user mindful of the questionable nature of testing for the lack of something even though it is a legitimate part of the Unicode Standard. **utfcheck** will fail if this option is selected and the UTF-8 Byte Order Mark (officially the zero width no-break space) is detected anywhere in the file.

Sample usage:

utfcheck −i *my_input_file.txt* −o *my_output_file.log*

EXIT STATUS

utfcheck will exit with a status of EXIT_SUCCESS if all went well, or with a status of EXIT_FAILURE if undesired input was read.

FILES

ASCII or UTF-8 text files.

AUTHOR

utfcheck was written by Paul Hardy.

LICENSE

BUGS

No known bugs exist.