NAME

pdftotext – Portable Document Format (PDF) to text converter (version 3.00)

SYNOPSIS

pdftotext [options] [*PDF-file* [*text-file*]]

DESCRIPTION

Pdftotext converts Portable Document Format (PDF) files to plain text.

Pdftotext reads the PDF file, *PDF-file*, and writes a text file, *text-file*. If *text-file* is not specified, pdftotext converts *file.pdf* to *file.txt*. If *text-file* is ´-', the text is sent to stdout.

CONFIGURATION FILE

Pdftotext reads a configuration file at startup. It first tries to find the user's private config file, ˜/.xpdfrc. If that doesn't exist, it looks for a system-wide config file, /etc/xpdf/xpdfrc. See the **xpdfrc**(5) man page for details.

OPTIONS

Many of the following options can be set with configuration file commands. These are listed in square brackets with the description of the corresponding command line option.

–f *number*
Specifies the first page to convert.

–l *number*
Specifies the last page to convert.

–layout
Maintain (as best as possible) the original physical layout of the text. The default is to ´undo' physical layout (columns, hyphenation, etc.) and output the text in reading order.

–raw Keep the text in content stream order. This is a hack which often "undoes" column formatting, etc. Use of raw mode is no longer recommended.

–htmlmeta
Generate a simple HTML file, including the meta information. This simply wraps the text in <pre> and </pre> and prepends the meta headers.

–enc *encoding-name*
Sets the encoding to use for text output. The *encoding−name* must be defined with the unicodeMap command (see **xpdfrc**(5)). The encoding name is case-sensitive. This defaults to "Latin1" (which is a built-in encoding). [config file: **textEncoding**]

–eol *unix* | *dos* | *mac*
Sets the end-of-line convention to use for text output. [config file: **textEOL**]

–nopgbrk
Don't insert page breaks (form feed characters) between pages. [config file: **textPageBreaks**]

–opw *password*
Specify the owner password for the PDF file. Providing this will bypass all security restrictions.

–upw *password*
Specify the user password for the PDF file.

–q Don't print any messages or errors. [config file: **errQuiet**]

–cfg *config-file*
Read *config-file* in place of ˜/.xpdfrc or the system-wide config file.

–v Print copyright and version information.

–h Print usage information. (**–help** and **––help** are equivalent.)

BUGS

Some PDF files contain fonts whose encodings have been mangled beyond recognition. There is no way (short of OCR) to extract text from these files.

EXIT CODES

The Xpdf tools use the following exit codes:

0 No error.

1 Error opening a PDF file.

2 Error opening an output file.

3 Error related to PDF permissions.

99 Other error.

AUTHOR

SEE ALSO

xpdf(1), **pdftops**(1), **pdfinfo**(1), **pdffonts**(1), **pdftoppm**(1), **pdfimages**(1), **xpdfrc**(5)
http://www.foolabs.com/xpdf/