



Linux Ubuntu 26.04 Manual Pages on command 'pdftotext.1'

\$ man pdftotext.1

pdftotext(1) General Commands Manual pdftotext(1)

NAME

pdftotext - Portable Document Format (PDF) to text converter (version 3.03)

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. If PDF-file is ?-', it reads the PDF file from stdin.

OPTIONS

-f number

Specifies the first page to convert.

-l number

Specifies the last page to convert.

-r number

Specifies the resolution, in DPI. The default is 72 DPI.

-x number

Specifies the x-coordinate of the crop area top left corner

-y number

Specifies the y-coordinate of the crop area top left corner

-W number

Specifies the width of crop area in pixels (default is 0)

-H number

Specifies the height of crop area in pixels (default is 0)

-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.

-fixed number

Assume fixed-pitch (or tabular) text, with the specified character width (in points).

This forces physical layout mode.

-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.

-nodiag

Discard diagonal text (i.e., text that is not close to one of the 0, 90, 180, or 270 degree axes). This is useful for skipping watermarks drawn on body text.

-htmlmeta

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

-bbox Generate an XHTML file containing bounding box information for each word in the file.

-bbox-layout

Generate an XHTML file containing bounding box information for each block, line, and word in the file.

-tsv Generate a TSV file containing the bounding box information for each block, line, and word in the file.

-cropbox

Use the crop box rather than the media box with -bbox and -bbox-layout.

-colspacing number

Specifies how much spacing we allow after a word before considering adjacent text to be a new column, measured as a fraction of the font size. Current default is 0.7, old releases had a 0.3 default.

-enc encoding-name

Sets the encoding to use for text output. This defaults to "UTF-8".

-listenc

Lists the available encodings

-eol unix | dos | mac

Sets the end-of-line convention to use for text output.

-nopgbrk

Don't insert page breaks (form feed characters) between pages.

-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.

-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

The pdftotext software and documentation are copyright 1996-2011 Glyph & Cog, LLC.

SEE ALSO

pdfdetach(1), pdffonts(1), pdfimages(1), pdfinfo(1), pdftocairo(1), pdftohtml(1),
pdftoppm(1), pdftops(1), pdfseparate(1), pdfsig(1), pdfunite(1)

15 August 2011

pdftotext(1)