Package 'pdftools'

December 3, 2016

Type Package

Title Text Extraction and Rendering of PDF Documents

Version 1.0

Author Jeroen Ooms

Maintainer Jeroen Ooms <jeroen.ooms@stat.ucla.edu>

Description Utilities based on libpoppler for extracting text, fonts, attachments and metadata from a pdf file. Also implements rendering of PDF to bitmaps on supported platforms.

License MIT + file LICENSE

URL https://ropensci.org/blog/2016/03/01/pdftools-and-jeroen
https://github.com/ropensci/pdftools

BugReports https://github.com/ropensci/pdftools/issues

SystemRequirements Poppler C++ interface library and headers

Imports Rcpp (>= 0.12.0)

LinkingTo Rcpp

Suggests jpeg, png, webp

RoxygenNote 5.0.1.9000

NeedsCompilation yes

Repository CRAN

Date/Publication 2016-12-03 18:12:11

R topics documented:

pdf_info *PDF utilities*

Description

Utilities based on libpoppler for extracting text, fonts, attachments and metadata from a pdf file.

Usage

```
pdf_info(pdf, opw = "", upw = "")

pdf_text(pdf, opw = "", upw = "")

pdf_fonts(pdf, opw = "", upw = "")

pdf_attachments(pdf, opw = "", upw = "")

pdf_toc(pdf, opw = "", upw = "")
```

Arguments

pdf	file path or raw vector with pdf data
opw	string with owner password to open pdf
upw	string with user password to open pdf

Details

Poppler is pretty verbose when encountering minor errors in PDF files, in especially pdf_text. These messages are usually safe to ignore, use suppressMessages to hide them alltogether.

See Also

Other pdftools: pdf_render_page

Examples

```
# Just a random pdf file
pdf_file <- file.path(R.home("doc"), "NEWS.pdf")
info <- pdf_info(pdf_file)
text <- pdf_text(pdf_file)
fonts <- pdf_fonts(pdf_file)
files <- pdf_attachments(pdf_file)
```

pdf_render_page *Render PDF to bitmap*

Description

Renders a PDF page to a bitmap array which can be written to e.g. png, jpeg or webp using the respective R packages. This function is only available if libpoppler was compiled with cairo support.

Usage

```
pdf_render_page(pdf, page = 1, dpi = 72, numeric = FALSE, opw = "",
  upw = "")

poppler_config()
```

Arguments

pdf	file path or raw vector with pdf data
page	which page to render
dpi	resolution (dots per inch) to render
numeric	convert raw output to (0-1) real values
opw	owner password
upw	user password

See Also

Other pdftools: pdf_info

Examples

```
# Rendering not supported on Windows
if(poppler_config()$can_render){

file.copy(file.path(Sys.getenv("R_DOC_DIR"), "NEWS.pdf"), "news.pdf")
bitmap <- pdf_render_page("news.pdf")

# save to bitmap formats
png::writePNG(bitmap, "page.png")
jpeg::writeJPEG(bitmap, "page.jpeg")
webp::write_webp(bitmap, "page.webp")

# Higher quality
bitmap <- pdf_render_page("news.pdf", page = 1, dpi = 300)
png::writePNG(bitmap, "page.png")

# slightly more efficient
bitmap_raw <- pdf_render_page("news.pdf", numeric = FALSE)
```

```
webp::write_webp(bitmap_raw, "page.webp")
}
```

Index