Package: PDFR (via r-universe)

October 22, 2024

Type Package

Title Extract Text From PDFs In An R Friendly Way

Version 0.1.0

Maintainer Allan Cameron <Allan.Cameron@nhs.scot>

Description Extracts text from PDF into an R dataframe giving the content, size, position and font of any text elements. This information can then be manipulated in R.

License MIT + file LICENSE

URL https://github.com/AllanCameron/PDFR

BugReports https://github.com/AllanCameron/PDFR/issues

Depends R (>= 2.10)

Imports cli, grDevices, grid, Rcpp, rlang

Suggests ggplot2, testthat

LinkingTo Rcpp, testthat

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

StagedInstall no

SystemRequirements C++11

Repository https://elipousson.r-universe.dev

RemoteUrl https://github.com/AllanCameron/PDFR

RemoteRef HEAD

RemoteSha 955c122cd0efa0ed4c0924e1ec8ec6340207d151

Contents

`draw_glyph` *draw_glyph*

Description

Draws glyphs from a truetype font as grid grobs

Usage

```
draw_glyph(fontfile, glyph)
```

Arguments

`fontfile`	a raw vector representing a font file
`glyph`	the character to be drawn. Can be text or an integer

Value

no return

Examples

```
## Not run:
if(interactive()){
 # ttf <- "raw vector with font file"
 draw_glyph(ttf, "a")
 }

## End(Not run)
```

getglyphmap *Return map of glyphs from a page*

Description

Used mainly for debugging, this function returns an R dataframe, one row for each byte that may be used as a glyph. It shows the unicode number of each interpreted glyph, as well as its width in text space.

Usage

```
getglyphmap(pdf, page = 1)
```

Arguments

pdf	a valid pdf file location
page	the page number from which to extract glyphs

Value

a dataframe of all entries of font encoding tables with width mapping

Examples

```
getglyphmap(pdfr_paths$leeds, 1)
```

getpagestring *pagestring*

Description

Returns contents of a pdf page description program

Usage

```
getpagestring(pdf, page)
```

Arguments

pdf	a valid pdf file location
page	the page number to be extracted

Value

a single string containing the page description program

Examples

```
getpagestring(pdfr_paths$leeds, 1)
```

`get_object` *Get the contents of a pdf object*

Description

Returns a list consisting of a named vector representing key:value pairs in a specified object. It also contains any stream data associated with the object.

Usage

```
get_object(pdf, number)
```

Arguments

`pdf`	a valid pdf file location
`number`	the object number

Value

a named vector of the dictionary and stream of the pdf object

Examples

```
get_object(pdfr_paths$leeds, 1)
```

`get_xref` *Get a pdf's xref table as an R dataframe*

Description

Get a pdf's xref table as an R dataframe

Usage

```
get_xref(pdf)
```

Arguments

`pdf`	a valid pdf file location or raw data vector

Value

a data frame showing the bytewise positions of each object in the pdf

Examples

```
get_xref(pdfr_paths$leeds)
```

`pdfboxes` *pdfboxes*

Description

Plots the bounding boxes of text elements from a page as a ggplot.

Usage

```
pdfboxes(pdf, pagenum)
```

Arguments

`pdf`	a valid pdf file location
`pagenum`	the page number to be plotted

Value

a ggplot

Examples

```
pdfboxes(pdfr_paths$leeds, 1)
```

`pdfdoc` *pdfdoc*

Description

Returns contents of all pdf pages

Usage

```
pdfdoc(pdf)
```

Arguments

`pdf`	a valid pdf file location

Value

a data frame of all text elements in a document

Examples

```
pdfdoc(pdfr_paths$leeds)
```

pdfgraphics *pdfgraphics*

Description

Plots the graphical elements of a pdf page as a ggplot

Usage

```
pdfgraphics(file, pagenum, scale = 1)
```

Arguments

file	a valid pdf file location
pagenum	the page number to be plotted
scale	Scale used for linewidth and text size. Passed to ‘ggplot2::geom_text()‘ size parameter as scale * size/3

Value

a ggplot

Examples

```
pdfgraphics(pdfr_paths$leeds, 1)
```

pdfgrobs *pdfgrobs*

Description

Plots the graphical elements of a pdf page as grobs

Usage

```
pdfgrobs(file_name, pagenum, scale = dev.size()[2]/10, enc = "UTF-8")
```

Arguments

file_name	a valid pdf file location
pagenum	the page number to be plotted
scale	Document scale. Defaults to ‘dev.size()[2]/10‘
enc	Document encoding. Defaults to "UTF-8"

Value

invisibly returns grobs as well as drawing them

Examples

```
pdfgrobs(pdfr_paths$leeds, 1)
```

---

pdfpage *pdfpage*

Description

Returns contents of a pdf page

Usage

```
pdfpage(pdf, page = 1, atomic = FALSE, table_only = TRUE)
```

Arguments

pdf	a valid pdf file location
page	the page number to be extracted
atomic	a boolean - should each letter treated individually?
table_only	a boolean - return data frame alone, as opposed to list

Value

a list containing data frames

Examples

```
head(pdfpage(pdfr_paths$leeds, page = 1))

head(pdfpage(pdfr_paths$chestpain, page = c(1:2)))
```

pdfplot *pdfplot*

Description

Plots the text elements from a page as a ggplot. The aim is not a complete pdf rendering but to help identify elements of interest in the data frame of text elements to convert to data points.

Usage

```
pdfplot(pdf, page = 1, atomic = FALSE, boxes = FALSE, textsize = 1)
```

Arguments

pdf	a valid pdf file location
page	the page number to be plotted
atomic	a boolean - should each letter treated individually?
boxes	Show the calculated text bounding boxes
textsize	the scale of the text to be shown

Value

a ggplot

Examples

```
pdfplot(pdfr_paths$leeds, 1)
```

pdfr_paths *Paths to test pdfs*

Description

A list of paths to locally stored test pdfs

Usage

```
pdfr_paths
```

Format

A list of 9 pdf files

barcodes a pdf constructed in Rstudio

chestpain a flow-chart for chest pain management

pdfinfo information about the pdf format

adobe an official adobe document

leeds a table-rich local government document

sams a document based on svg

testreader a simple pdf test

tex a simple tex test

rcpp a CRAN package vignette

`run_testthat_tests` *A tool used for symbol registration*

Description

A registered native symbol used in testing

Usage

```
run_testthat_tests
```

Format

A list of 4 fields

name run_testthat_tests

address a pointer to this symbol

dll the compiled file where the symbol is contained

numParameters no parameters

Index