Package 'pdfminer'

July 23, 2025

Type Package

Title Read Portable Document Format (PDF) Files

Version 1.0

Description Provides an interface to 'PDFMiner' <https://github.com/pdfminer/pdfminer.six>
a 'Python' package for extracting information from 'PDF'-files.
'PDFMiner' has the goal to get all information available in a 'PDF'-file,
position of the characters, font type, font size and informations about lines.
Which makes it the perfect starting point for extracting tables from 'PDF'-files.
More information can be found in the package 'README'-file.

License MIT + file LICENSE

Imports checkmate, jsonlite

Suggests PythonInR, RSQLite

SystemRequirements Python>=3.6, pdfminer.six>=20200402, pandas

RoxygenNote 7.1.0

NeedsCompilation no

Author Florian Schwendinger [aut, cre, cph],
Benjamin Schwendinger [aut, cph]

Maintainer Florian Schwendinger <FlorianSchwendinger@gmx.at>

Repository CRAN

Date/Publication 2020-06-22 09:20:02 UTC

Contents

is_pdfminer_installed *Check if* **pdfminer** *is Installed*

Description

The function

Usage

```
is_pdfminer_installed(
  method = c("csv", "sqlite", "PythonInR"),
  pyexe = "python3"
)
```

Arguments

method	a character string giving the data transfer method. Allowed values are "csv" (default), "sqlite" and "PythonInR".
pyexe	a character string giving the path to the python executable (default is "python3"). Only used when method is "csv" or "sqlite".

Value

Returns TRUE if **pdfminer** is installed.

Examples

```
is_pdfminer_installed()
```

layout_control *Read a* PDF *document.*

Description

Extract PDF document

Usage

```
layout_control(
  line_overlap = 0.5,
  char_margin = 2,
  line_margin = 0.5,
  word_margin = 0.1,
  boxes_flow = 0.5,
  detect_vertical = FALSE,
  all_texts = FALSE
)
```

Arguments

line_overlap	a double, if two characters have more overlap than this they are considered to be on the same line. The overlap is specified relative to the minimum height of both characters.
char_margin	a double, if two characters are closer together than this margin they are considered part of the same line. The margin is specified relative to the width of the character.
line_margin	a double, if two characters on the same line are further apart than this margin then they are considered to be two separate words, and an intermediate space will be added for readability. The margin is specified relative to the width of the character.
word_margin	a double, if two lines are are close together they are considered to be part of the same paragraph. The margin is specified relative to the height of a line.
boxes_flow	a double, Specifies how much a horizontal and vertical position of a text matters when determining the order of text boxes. The value should be within the range of `-1.0` (only horizontal position matters) to `+1.0` (only vertical position matters). You can also pass `NULL` to disable advanced layout analysis, and instead return text based on the position of the bottom left corner of the text box.
detect_vertical	a logical, If vertical text should be considered during layout analysis
all_texts	a logical, If layout analysis should be performed on text in figures.

Value

Returns a list with the layout control variables.

Examples

```
layout_control()
```

---

read.pdf *Read a* PDF *document.*

Description

Extract PDF document

Usage

```
read.pdf(
  file,
  pages = integer(),
  method = c("csv", "sqlite", "PythonInR"),
  laycntrl = layout_control(),
  encoding = "utf8",
```

```
  password = "",
  caching = TRUE,
  maxpages = Inf,
  rotation = 0L,
  image_dir = "",
  pyexe = "python3"
)
```

Arguments

file	a character string giving the name of the PDF-file the data are to be read from.
pages	an integer giving the pages which should be extracted (default is integer()).
method	a character string giving the data transfer method. Allowed values are "csv" (default), "sqlite" and "PythonInR" (recommended).
laycntrl	a list of layout options, created by the function layout_control.
encoding	a character string giving the encoding of the output (default is "utf8").
password	a character string giving the password necessary to access the PDF (default is "").
caching	a logical if TRUE (default) **pdfminer** is faster but uses more memory.
maxpages	an integer giving the maximum number of pages to be extracted (default is Inf).
rotation	an integer giving the rotation of the page, allowed values are c(0, 90, 180, 270).
image_dir	a character string giving the path to the folder, where the images should be stored (default is "").
pyexe	a character string giving the path to the python executable (default is "python3"). Only used when method is "csv" or "sqlite".

Value

Returns a object of class "pdf_document".

Examples

```
if (is_pdfminer_installed()) {
pdf_file <- system.file("pdfs/cars.pdf", package = "pdfminer")
read.pdf(pdf_file)
}
```

Index