Package 'pdfsearch'

December 16, 2016

Type Package

Version 0.1.1

License MIT + file LICENSE

Title Search Tools for PDF Files

Description Includes functions for keyword search of pdf files. There is also a wrapper that includes searching of all files within a single directory.

Depends R (>= 3.3.0), pdftools, tibble

Suggests shiny, testthat

Maintainer Brandon LeBeau <lebebr01+pdfsearch@gmail.com>

RoxygenNote 5.0.1

URL https://github.com/lebebr01/pdfsearch

BugReports https://github.com/lebebr01/pdfsearch/issues

NeedsCompilation no

Author Brandon LeBeau [aut, cre]

Repository CRAN

Date/Publication 2016-12-16 08:34:00

R topics documented:

heading_search *Function to locate sections of pdf*

Description

The ability to extract the location of the text and separate by sections. The function will return the headings with their location in the pdf.

Usage

```
heading_search(x, headings, path = FALSE, pdf_toc = FALSE,
  full_line = FALSE, ignore_case = FALSE, split_pdf = FALSE)
```

Arguments

x	Either the text of the pdf read in with the pdftools package or a path for the location of the pdf file.
headings	A character vector representing the headings to search for. Can be NULL if pdf_toc = TRUE.
path	An optional path designation for the location of the pdf to be converted to text. The pdftools package is used for this conversion.
pdf_toc	TRUE/FALSE whether the pdf_toc function should be used from the pdftools package. This is most useful if the pdf has the table of contents embedded within the pdf. Must specify path = TRUE if pdf_toc = TRUE.
full_line	TRUE/FALSE indicating whether the headings should reside on their own line. This can create problems with multiple column pdfs.
ignore_case	TRUE/FALSE/vector of TRUE/FALSE, indicating whether the case of the keyword matters. Default is FALSE meaning that case of the headings keywords are literal. If a vector, must be same length as the headings vector.
split_pdf	TRUE/FALSE indicating whether to split the pdf using white space. This would be most useful with multicolumn pdf files. The split_pdf function attempts to recreate the column layout of the text into a single column starting with the left column and proceeding to the right.

Examples

```
file <- system.file('pdf', '1501.00450.pdf', package = 'pdfsearch')

heading_search(file, headings = c('abstract', 'introduction'),
  path = TRUE)
```

keyword_directory *Wrapper for keyword search function*

Description

This will use the keyword_search function to loop over all pdf files in a directory. Includes the ability to include subdirectories as well.

Usage

```
keyword_directory(directory, keyword, split_pdf = FALSE,
  surround_lines = FALSE, ignore_case = FALSE, full_names = FALSE,
  recursive = FALSE, max_search = NULL)
```

Arguments

directory	The directory to perform the search for pdf files to search.
keyword	The keyword(s) to be used to search in the text. Multiple keywords can be specified with a character vector.
split_pdf	TRUE/FALSE indicating whether to split the pdf using white space. This would be most useful with multicolumn pdf files. The split_pdf function attempts to recreate the column layout of the text into a single column starting with the left column and proceeding to the right.
surround_lines	numeric/FALSE indicating whether the output should extract the surrouding lines of text in addition to the matching line. Default is FALSE, if not false, include a numeric number that indicates the additional number of surrounding lines that will be extracted.
ignore_case	TRUE/FALSE/vector of TRUE/FALSE, indicating whether the case of the keyword matters. Default is FALSE meaning that case of the keyword is literal. If a vector, must be same length as the keyword vector.
full_names	TRUE/FALSE indicating if the full file path should be used. Default is FALSE, see list.files for more details.
recursive	TRUE/FALSE indicating if subdirectories should be searched as well. Default is FALSE, see list.files for more details.
max_search	An optional numeric vector indicating the maximum number of pdfs to search. Will only search the first n cases.

Examples

```
# find directory
directory <- system.file('pdf', package = 'pdfsearch')

# do search over two files
keyword_directory(directory,
       keyword = c('repeated measures', 'measurement error'),
```

```
       surround_lines = 1, full_names = TRUE)

# can also split pdfs
keyword_directory(directory,
       keyword = c('repeated measures', 'measurement error'),
       split_pdf = TRUE,
       surround_lines = 1, full_names = TRUE)
```

---

keyword_search *Search a pdf file for keywords*

Description

This uses the pdf_text from the pdftools package to perform keyword searches. Keyword locations indicating the line of the text as well as the page number that the keyword is found are returned.

Usage

```
keyword_search(x, keyword, path = FALSE, split_pdf = FALSE,
  surround_lines = FALSE, ignore_case = FALSE, heading_search = FALSE,
  heading_args = NULL)
```

Arguments

x	Either the text of the pdf read in with the pdftools package or a path for the location of the pdf file.
keyword	The keyword(s) to be used to search in the text. Multiple keywords can be specified with a character vector.
path	An optional path designation for the location of the pdf to be converted to text. The pdftools package is used for this conversion.
split_pdf	TRUE/FALSE indicating whether to split the pdf using white space. This would be most useful with multicolumn pdf files. The split_pdf function attempts to recreate the column layout of the text into a single column starting with the left column and proceeding to the right.
surround_lines	numeric/FALSE indicating whether the output should extract the surrouding lines of text in addition to the matching line. Default is FALSE, if not false, include a numeric number that indicates the additional number of surrounding lines that will be extracted.
ignore_case	TRUE/FALSE/vector of TRUE/FALSE, indicating whether the case of the keyword matters. Default is FALSE meaning that case of the keyword is literal. If a vector, must be same length as the keyword vector.
heading_search	TRUE/FALSE indicating whether to search for headings in the pdf.
heading_args	A list of arguments to pass on to the heading_search function. See heading_search for more details on arguments needed.

Examples

```
file <- system.file('pdf', '1501.00450.pdf', package = 'pdfsearch')

keyword_search(file, keyword = c('repeated measures', 'mixed effects'),
  path = TRUE)

# Add surrounding text
keyword_search(file, keyword = c('variance', 'mixed effects'),
  path = TRUE, surround_lines = 1)

# split pdf
keyword_search(file, keyword = c('repeated measures', 'mixed effects'),
  path = TRUE, split_pdf = TRUE)
```

`run_shiny` *Run Shiny Application Demo*

Description

Function runs Shiny Application Demo

Usage

```
run_shiny()
```

Details

This function does not take any arguments and will run the Shiny Application. If running from RStudio, will open the application in the viewer, otherwise will use the default internet browser.

Examples

```
run_shiny()
```

Index