

Package ‘PDE’

November 27, 2020

Type Package

Title Extract Sentences and Tables from PDFs with User Interface

Version 1.1.2

Author Erik Stricker [aut, cre]

Maintainer Erik Stricker <erik.stricker@gmx.com>

Description The PDE (Pdf Data Extractor) allows the extraction of information and tables optionally based on search words from PDF (Portable Document Format) files and enables the visualization of the results, both by providing a convenient user-interface.

License GPL-3 | file LICENSE

Encoding UTF-8

LazyData true

Imports tcltk

Depends tcltk2 (>= 1.2.11), R (>= 3.5)

SystemRequirements XPDF
(>=4.0)(<https://www.xpdfreader.com/download.html>)

RoxygenNote 7.1.1

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2020-11-27 02:20:25 UTC

R topics documented:

| | |
|-----------------------------------|---|
| .PDE_extr_data_from_pdf | 2 |
| PDE | 5 |
| PDE-deprecated | 5 |
| PDE_analyzer | 6 |
| PDE_analyzer_i | 7 |

| | |
|--|----|
| PDE_check_Xpdf_install | 8 |
| PDE_extr_data_from_pdfs | 8 |
| PDE_install_XpdfReader4.02 | 12 |
| PDE_install_XpdfTools4.02 | 12 |
| PDE_path | 13 |
| PDE_pdfs2table | 14 |
| PDE_pdfs2table_searchandfilter | 15 |
| PDE_pdfs2txt_searchandfilter | 18 |
| PDE_reader_i | 20 |

| | |
|--------------|-----------|
| Index | 22 |
|--------------|-----------|

.PDE_extr_data_from_pdf

Extracting data from a PDF (Portable Document Format) file

Description

PDE_extr_data_from_pdf extracts sentences or tables from a single PDF file and writes output in the corresponding folder.

Usage

```
.PDE_extr_data_from_pdf(
  pdf,
  whattoextr,
  out = ".",
  filter.words = "",
  ignore.case.fw = FALSE,
  filter.word.times = 20,
  table.heading.words = "",
  ignore.case.th = FALSE,
  search.words,
  ignore.case.sw = FALSE,
  eval.abbrevs = TRUE,
  out.table.format = ".csv (WINDOWS-1252)",
  dev = 20,
  context = 0,
  write.table.locations = FALSE,
  exp.nondetc.tabs = FALSE,
  write.tab.doc.file = FALSE,
  write.txt.doc.file = FALSE,
  delete = TRUE,
  verbose = TRUE
)
```

Arguments

| | |
|-----------------------|--|
| pdf | String. Path to the PDF file to be analyzed. |
| whattoextr | String. Either <i>txt</i> , <i>tab</i> , or <i>tabandtxt</i> for PDFS2TXT (extract sentences from a PDF file) or PDFS2TABLE (table of a PDF file to a Microsoft Excel file) extraction. <i>tab</i> allows the extraction of tables with and without search words while <i>txt</i> and <i>tabandtxt</i> require search words. |
| out | String. Directory chosen to save analysis results in. Default: ".". |
| filter.words | List of strings. The list of filter words. If not NA or "" a hit will be counted every time a word from the list is detected in the article. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). Default: "". |
| ignore.case.fw | Logical. Are the filter words case-sensitive (does capitalization matter)? Default: FALSE. |
| filter.word.times | Numeric. The minimum number of hits described for <code>filter.words</code> for a paper to be further analyzed. Default: 20. |
| table.heading.words | List of strings. Different than standard (TABLE, TAB or table plus number) headings to be detected. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). Default = "". |
| ignore.case.th | Logical. Are the additional table headings (see <code>table.heading.words</code>) case-sensitive (does capitalization matter)? Default = FALSE. |
| search.words | List of strings. List of search words. To extract all tables from the PDF file leave <code>search.words = ""</code> . Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). |
| ignore.case.sw | Logical. Are the search words case-sensitive (does capitalization matter)? Default: FALSE. |
| eval.abbrevs | Logical. Should abbreviations for the search words be automatically detected and then replaced with the search word + "\$*"? Default: TRUE. |
| out.table.format | String. Output file format. Either comma separated file <code>.csv</code> or tab separated file <code>.tsv</code> . The encoding indicated in parentheses should be selected according to the operational system exported tables are opened in, i.e., Windows: "(WINDOWS-1252)"; Mac: (macintosh); Linux: (UTF-8). Default: ".csv" and encoding depending on the operational system. |
| dev | Numeric. For a table the size of indentation which would be considered the same column. Default: 20. |
| context | Numeric. Number of sentences extracted before and after the sentence with the detected search word. If 0 only the sentence with the search word is extracted. Default: 0. |
| write.table.locations | Logical. If TRUE, a separate file with the headings of all tables, their relative location in the generated html and txt files, as well as information if search words were found will be generated. Default: FALSE. |

| | |
|---------------------------------|--|
| <code>exp.nondetc.tabs</code> | Logical. If TRUE, if a table was detected in a PDF file but is an image or cannot be read, the page with the table will be exported as a png. Default: FALSE. |
| <code>write.tab.doc.file</code> | Logical. If TRUE, if search words are used for table detection and no search words were found in the tables of a PDF file, a no.table.w.search.words . Default: FALSE. |
| <code>write.txt.doc.file</code> | Logical. If TRUE, if no search words were found in the sentences of a PDF file, a file will be created with the PDF filename followed by no.txt.w.search.words . If the PDF file is empty, a file will be created with the PDF filename followed by no.content.detected . If the filter word threshold is not met, a file will be created with the PDF filename followed by no.txt.w.filter.words . Default: FALSE. |
| <code>delete</code> | Logical. If TRUE, the intermediate txt , keeplayouttxt and html copies of the PDF file will be deleted. Default: TRUE. |
| <code>verbose</code> | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |

Value

If tables were extracted from the PDF file the function returns a list of following tables/items: 1) **htmltablelines**, 2) **txttablelines**, 3) **keeplayouttxttablelines**, 4) **id**, 5) **out_msg**. The **tablelines** are tables that provide the heading and position of the detected tables. The **id** provide the name of the PDF file. The **out_msg** includes all messages printed to the console or the suppressed messages if `verbose=FALSE`.

See Also

[PDE_pdfs2table](#), [PDE_pdfs2table_searchandfilter](#), [PDE_pdfs2txt_searchandfilter](#)

Examples

```
## Running a simple analysis with filter and search words to extract sentences and tables
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- .PDE_extr_data_from_pdf(pdf = "examples/Methotrexate/29973177_!.pdf",
    whattoextr = "tabandt",
    out = paste0(PDE_path(), "examples/MTX_all_files+-0_test/"),
    filter.words = strsplit("cohort;case-control;group;study population;study participants", ";")[[1]],
    ignore.case.fw = TRUE,
    search.words = strsplit("(M|m)ethotrexate;(T|t)rexal;(R|r)heumatrex;(O|o)trexup", ";")[[1]],
    ignore.case.sw = FALSE)
}

## Running an advanced analysis with filter and search words to
## extract sentences and tables and obtain documentation files
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- .PDE_extr_data_from_pdf(pdf = paste0(PDE_path(),
    "examples/Methotrexate/29973177_!.pdf"),
    whattoextr = "tabandt",
```

```

out = paste0(PDE_path(),"examples/MTX_all_files+-1_test/"),
context = 1,
dev = 20,
filter.words = strsplit("cohort;case-control;group;study population;study participants", ";")[[1]],
ignore.case.fw = TRUE,
filter.word.times = 20,
table.heading.words = "",
ignore.case.th = FALSE,
search.words = strsplit("(M|m)ethotrexate;(T|t)rexal;(R|r)heumatrex;(O|o)trexup", ";")[[1]],
ignore.case.sw = FALSE,
eval.abbrevs = TRUE,
out.table.format = ".csv (WINDOWS-1252)",
write.table.locations = TRUE,
write.tab.doc.file = TRUE,
write.txt.doc.file = TRUE,
exp.nondetc.tabs = TRUE,
delete = TRUE)
}

```

PDE

*PDE: Extract Sentences and Tables from PDF Files.***Description**

The package includes two main components: 1) The PDE analyzer performs the sentence and table extraction while 2) the PDE reader allows the user-friendly visualization and quick-processing of the obtained results.

PDE functions

[PDE_analyzer](#), [PDE_analyzer_i](#), [PDE_extr_data_from_pdfs](#), [PDE_pdfs2table](#), [PDE_pdfs2table_searchandfilter](#), [PDE_reader_i](#), [PDE_path](#), [PDE_install_Xpdftools4.02](#), [PDE_check_Xpdf_install](#)

PDE-deprecated

*Deprecated functions in package 'PDE'***Description**

These functions are provided for compatibility with older versions of 'PDE' only, and will be defunct at the next release.

Details

The following functions are deprecated and will be made defunct; use the replacement indicated below:

- [PDE_install_XpdfReader4.02](#): [PDE_install_Xpdftools4.02](#)

PDE_analyzer *Extracting data from PDF (Portable Document Format) files*

Description

The PDE_analyzer allows the sentence and table extraction from multiple PDF files.

Usage

```
PDE_analyzer(PDE_parameters_file_path = NA, verbose = TRUE)
```

Arguments

| | |
|--------------------------|---|
| PDE_parameters_file_path | String. This file includes all parameters to run PDE_extr_data_from_pdfs on multiple PDF files. If PDE_parameters_file_path does not exist or is NA a dialog box is opened prompting the user to select the parameter file. |
| verbose | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |

Value

If tables were extracted from the PDF file the function returns a list of following tables/items: 1) **htmltablelines**, 2) **txttablelines**, 3) **keeplayouttxttablelines**, 4) **id**, 5) **out_msg**. The **tablelines** are tables that provide the heading and position of the detected tables. The **id** provide the name of the PDF file. The **out_msg** includes all messages printed to the console or the suppressed messages if verbose=FALSE.

Details

The parameter file (also referred to as .tsv file) can either manually or with the help of the [PDE_analyzer_i](#) interface be filled.

Note

A detailed description of the parameters in the TSV file can be found in the markdown file (README_PDE.md) and in the description of [PDE_extr_data_from_pdfs](#).

See Also

[PDE_extr_data_from_pdfs](#)

Examples

```
if(PDE_check_Xpdf_install() == TRUE){
  PDE_analyzer(paste0(PDE_path(),"examples/tsvs/PDE_parameters_v1.0_all_files+-0.tsv"))
}

## Not run:
## requires user file choice:
PDE_analyzer()

## End(Not run)
```

| | |
|----------------|---|
| PDE_analyzer_i | <i>Extracting data from PDF (Portable Document Format) files using a user interface</i> |
|----------------|---|

Description

The PDE_analyzer_i provides a user interface for the sentence and table extraction from multiple PDF files.

Usage

```
PDE_analyzer_i(verbose = TRUE)
```

Arguments

| | |
|---------|--|
| verbose | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |
|---------|--|

Note

A detailed description of the elements in the user interface can be found in the markdown file (README_PDE.md).

Examples

```
PDE_analyzer_i()
```

PDE_check_Xpdf_install

Check if the Xpdf tools are installed in the system path

Description

PDE_check_Xpdf_install runs a version test for pdftotext, pdftohtml and pdftopng.

Usage

```
PDE_check_Xpdf_install(sysname = NULL, verbose = TRUE)
```

Arguments

| | |
|---------|--|
| sysname | String. In case the function returns "Unknown OS" the sysname can be set manually. Allowed options are "Windows", "Linux", "SunOS" for Solaris, and "Darwin" for Mac. Default: NULL. |
| verbose | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |

Value

The function returns a Boolean for the installation status and a message in case the commands are not detected.

Examples

```
PDE_check_Xpdf_install()
```

PDE_extr_data_from_pdfs

Extracting data from PDF (Portable Document Format) files

Description

PDE_extr_data_from_pdfs extracts sentences or tables from a single PDF file and writes output in the corresponding folder.

Usage

```

PDE_extr_data_from_pdfs(
  pdfs,
  whattoextr,
  out = ".",
  filter.words = "",
  ignore.case.fw = FALSE,
  filter.word.times = 20,
  table.heading.words = "",
  ignore.case.th = FALSE,
  search.words,
  ignore.case.sw = FALSE,
  eval.abbrevs = TRUE,
  out.table.format = ".csv (WINDOWS-1252)",
  dev = 20,
  context = 0,
  write.table.locations = FALSE,
  exp.nondetc.tabs = FALSE,
  write.tab.doc.file = FALSE,
  write.txt.doc.file = FALSE,
  delete = TRUE,
  verbose = TRUE
)

```

Arguments

| | |
|----------------------------------|--|
| <code>pdfs</code> | String. A list of paths to the PDF files to be analyzed. |
| <code>whattoextr</code> | String. Either <i>txt</i> , <i>tab</i> , or <i>tabandtxt</i> for PDFS2TXT (extract sentences from a PDF file) or PDFS2TABLE (table of a PDF file to a Microsoft Excel file) extraction. <i>tab</i> allows the extraction of tables with and without search words while <i>txt</i> and <i>tabandtxt</i> require search words. |
| <code>out</code> | String. Directory chosen to save analysis results in. Default: ".". |
| <code>filter.words</code> | List of strings. The list of filter words. If not NA or "" a hit will be counted every time a word from the list is detected in the article. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). Default: "". |
| <code>ignore.case.fw</code> | Logical. Are the filter words case-sensitive (does capitalization matter)? Default: FALSE. |
| <code>filter.word.times</code> | Numeric. The minimum number of hits described for <code>filter.words</code> for a paper to be further analyzed. Default: 20. |
| <code>table.heading.words</code> | List of strings. Different than standard (TABLE, TAB or table plus number) headings to be detected. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). Default = "". |
| <code>ignore.case.th</code> | Logical. Are the additional table headings (see <code>table.heading.words</code>) case-sensitive (does capitalization matter)? Default = FALSE. |

| | |
|------------------------------------|--|
| <code>search.words</code> | List of strings. List of search words. To extract all tables from the PDF files leave <code>search.words = ""</code> . Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). |
| <code>ignore.case.sw</code> | Logical. Are the search words case-sensitive (does capitalization matter)? Default: FALSE. |
| <code>eval.abbrevs</code> | Logical. Should abbreviations for the search words be automatically detected and then replaced with the search word + "\$*"? Default: TRUE. |
| <code>out.table.format</code> | String. Output file format. Either comma separated file <code>.csv</code> or tab separated file <code>.tsv</code> . The encoding indicated in parantheses should be selected according to the operational system exported tables are opened in, i.e., Windows: "(WINDOWS-1252)"; Mac: (macintosh); Linux: (UTF-8). Default: ".csv" and encoding depending on the operational system. |
| <code>dev</code> | Numeric. For a table the size of indention which would be considered the same column. Default: 20. |
| <code>context</code> | Numeric. Number of sentences extracted before and after the sentence with the detected search word. If 0 only the sentence with the search word is extracted. Default: 0. |
| <code>write.table.locations</code> | Logical. If TRUE, a separate file with the headings of all tables, their relative location in the generated html and txt files, as well as information if search words were found will be generated. Default: FALSE. |
| <code>exp.nondetc.tabs</code> | Logical. If TRUE, if a table was detected in a PDF file but is an image or cannot be read, the page with the table with be exported as a png. Default: FALSE. |
| <code>write.tab.doc.file</code> | Logical. If TRUE, if search words are used for table detection and no search words were found in the tables of a PDF file, a no.table.w.search.words . Default: FALSE. |
| <code>write.txt.doc.file</code> | Logical. If TRUE, if no search words were found in the sentences of a PDF file, a file will be created with the PDF filename followed by no.txt.w.search.words . If the PDF file is empty, a file will be created with the PDF filename followed by no.content.detected . If the filter word threshold is not met, a file will be created with the PDF filename followed by no.txt.w.filter.words . Default: FALSE. |
| <code>delete</code> | Logical. If TRUE, the intermediate txt , keplayouttxt and html copies of the PDF files will be deleted. Default: TRUE. |
| <code>verbose</code> | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |

Value

If tables were extracted from the PDF file the function returns a list of following tables/items: 1) **htmltablelines**, 2) **txttablelines**, 3) **keplayouttxttablelines**, 4) **id**, 5) **out_msg**. The **tablelines** are tables that provide the heading and position of the detected tables. The **id** provide the name of the PDF file. The **out_msg** includes all messages printed to the console or the suppressed messages if `verbose=FALSE`.

See Also

[PDE_pdfs2table](#), [PDE_pdfs2table_searchandfilter](#), [PDE_pdfs2txt_searchandfilter](#)

Examples

```
## Running a simple analysis with filter and search words to extract sentences and tables
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- PDE_extr_data_from_pdfs(pdfs = c(paste0(PDE_path(),
    "examples/Methotrexate/29973177_!.pdf"),
    paste0(PDE_path(),
    "examples/Methotrexate/31083238_!.pdf")),
  whattoextr = "tabandtxt",
  out = paste0(PDE_path(),"examples/MTX_all_files+-0_test/"),
  filter.words = strsplit("cohort;case-control;group;study population;study participants", ";")[[1]],
  ignore.case.fw = TRUE,
  search.words = strsplit("(M|m)ethotrexate;(T|t)rexal;(R|r)heumatrex;(O|o)trexup", ";")[[1]],
  ignore.case.sw = FALSE)
}

## Running an advanced analysis with filter and search words to
## extract sentences and tables and obtain documentation files
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- PDE_extr_data_from_pdfs(pdfs = c(paste0(PDE_path(),
    "examples/Methotrexate/29973177_!.pdf"),
    paste0(PDE_path(),
    "examples/Methotrexate/31083238_!.pdf")),
  whattoextr = "tabandtxt",
  out = paste0(PDE_path(),"examples/MTX_all_files+-1_test/"),
  context = 1,
  dev = 20,
  filter.words = strsplit("cohort;case-control;group;study population;study participants", ";")[[1]],
  ignore.case.fw = TRUE,
  filter.word.times = 20,
  table.heading.words = "",
  ignore.case.th = FALSE,
  search.words = strsplit("(M|m)ethotrexate;(T|t)rexal;(R|r)heumatrex;(O|o)trexup", ";")[[1]],
  ignore.case.sw = FALSE,
  eval.abbrevs = TRUE,
  out.table.format = ".csv (WINDOWS-1252)",
  write.table.locations = TRUE,
  write.tab.doc.file = TRUE,
  write.txt.doc.file = TRUE,
  exp.nondetc.tabs = TRUE,
  delete = TRUE)
}
```

PDE_install_XpdfReader4.02

Install the XpdfReader 4.02

Description

PDE_download_XpdfReader4.02 is deprecated. Please run PDE_install_Xpdftools4.02() instead.

Usage

```
PDE_install_XpdfReader4.02(sysname = NULL, bin = NULL, verbose = TRUE)
```

Arguments

| | |
|---------|---|
| sysname | String. In case the function returns "Unknown OS" the sysname can be set manually. Allowed options are "Windows", "Linux", and "Darwin" for Mac. Default: NULL. |
| bin | String. In case the function returns "Unknown OS" the bin of the operational system can be set manually. Allowed options are "64", and "32". Default: NULL. |
| verbose | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |

Examples

```
## Not run:  
  
PDE_install_XpdfReader4.02()  
  
## End(Not run)
```

PDE_install_Xpdftools4.02

Install the Xpdf command line tools 4.02

Description

PDE_install_Xpdftools4.02 downloads and installs the XPDF command line tools 4.02.

Usage

```
PDE_install_Xpdftools4.02(sysname = NULL, bin = NULL, verbose = TRUE)
```

Arguments

| | |
|---------|--|
| sysname | String. In case the function returns "Unknown OS" the sysname can be set manually. Allowed options are "Windows", "Linux", "SunOS" for Solaris, and "Darwin" for Mac. Default: NULL. |
| bin | String. In case the function returns "Unknown OS" the bin of the operational system can be set manually. Allowed options are "64", and "32". Default: NULL. |
| verbose | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |

Value

The function returns a Boolean for the installation status and a message in case the commands are not installed.

Examples

```
## Not run:  
  
PDE_install_Xpdftools4.02()  
  
## End(Not run)
```

PDE_path

Export the installation path the PDE (PDF Data Extractor) package

Description

PDE_path returns the installation path of the PDE tool.

Usage

```
PDE_path()
```

Value

The function returns a potential path for the PDE package. If the PDE tool was not correctly installed it returns "".

PDE_pdfs2table

Extracting all tables from a PDF (Portable Document Format) file

Description

PDE_pdfs2table extracts all tables from a single PDF file and writes output in the corresponding folder.

Usage

```
PDE_pdfs2table(
  pdfs,
  out = ".",
  table.heading.words = "",
  ignore.case.th = FALSE,
  out.table.format = ".csv (WINDOWS-1252)",
  dev = 20,
  write.table.locations = FALSE,
  exp.nondetc.tabs = FALSE,
  delete = TRUE,
  verbose = TRUE
)
```

Arguments

| | |
|-----------------------|---|
| pdfs | String. A list of paths to the PDF files to be analyzed. |
| out | String. Directory chosen to save tables in. Default: ".". |
| table.heading.words | List of strings. Different than standard (TABLE, TAB or table plus number) headings to be detected. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). Default = "". |
| ignore.case.th | Logical. Are the additional table headings (see table.heading.words) case-sensitive (does capitalization matter)? Default = FALSE. |
| out.table.format | String. Output file format. Either comma separated file .csv or tab separated file .tsv. The encoding indicated in parantheses should be selected according to the operational system exported tables are opened in, i.e., Windows: "(WINDOWS-1252)"; Mac: (macintosh); Linux: (UTF-8). Default: ".csv" and encoding depending on the operational system. |
| dev | Numeric. For a table the size of indentation which would be considered the same column. Default: 20. |
| write.table.locations | Logical. If TRUE, a separate file with the headings of all tables, their relative location in the generated html and txt files, as well as information if search words were found will be generated. Default: FALSE. |

| | |
|------------------|---|
| exp.nondetc.tabs | Logical. If TRUE, if a table was detected in a PDF file but is an image or cannot be read, the page with the table will be exported as a png. Default: FALSE. |
| delete | Logical. If TRUE, the intermediate txt , keeplayouttxt and html copies of the PDF file will be deleted. Default: TRUE. |
| verbose | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |

See Also

[PDE_extr_data_from_pdfs,PDE_pdfs2table_searchandfilter](#)

Examples

```
## Running a simple table extraction
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- PDE_pdfs2table(pdf = paste0(PDE_path(),
      "examples/Methotrexate/29973177_!.pdf"),
    out = paste0(PDE_path(),"examples/29973177_tables/"))
}

## Running a the same table extraction as above with all paramaters shown
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- PDE_pdfs2table(pdf = paste0(PDE_path(),
      "examples/Methotrexate/29973177_!.pdf"),
    out = paste0(PDE_path(),"examples/29973177_tables/"),
    dev = 20,
    table.heading.words = "",
    ignore.case.th = FALSE,
    out.table.format = ".csv (WINDOWS-1252)",
    write.table.locations = FALSE,
    exp.nondetc.tabs = FALSE,
    delete = TRUE)
}
```

PDE_pdfs2table_searchandfilter

Extracting tables from a PDF (Portable Document Format) file

Description

PDE_pdfs2table_searchandfilter extracts tables from a single PDF file according to filter and search words and writes output in the corresponding folder.

Usage

```

PDE_pdfs2table_searchandfilter(
  pdfs,
  out = ".",
  filter.words = "",
  ignore.case.fw = FALSE,
  filter.word.times = 20,
  table.heading.words = "",
  ignore.case.th = FALSE,
  search.words,
  ignore.case.sw = FALSE,
  eval.abbrevs = TRUE,
  out.table.format = ".csv (WINDOWS-1252)",
  dev = 20,
  write.table.locations = FALSE,
  exp.nondetc.tabs = FALSE,
  write.tab.doc.file = FALSE,
  delete = TRUE,
  verbose = TRUE
)

```

Arguments

| | |
|----------------------------------|--|
| <code>pdfs</code> | String. A list of paths to the PDF files to be analyzed. |
| <code>out</code> | String. Directory chosen to save analysis results in. Default: ".". |
| <code>filter.words</code> | List of strings. The list of filter words. If not NA or "" a hit will be counted every time a word from the list is detected in the article. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). Default: "". |
| <code>ignore.case.fw</code> | Logical. Are the filter words case-sensitive (does capitalization matter)? Default: FALSE. |
| <code>filter.word.times</code> | Numeric. The minimum number of hits described for <code>filter.words</code> for a paper to be further analyzed. Default: 20. |
| <code>table.heading.words</code> | List of strings. Different than standard (TABLE, TAB or table plus number) headings to be detected. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). Default = "". |
| <code>ignore.case.th</code> | Logical. Are the additional table headings (see <code>table.heading.words</code>) case-sensitive (does capitalization matter)? Default = FALSE. |
| <code>search.words</code> | List of strings. List of search words. To extract all tables from the PDF file leave <code>search.words = ""</code> . Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). |
| <code>ignore.case.sw</code> | Logical. Are the search words case-sensitive (does capitalization matter)? Default: FALSE. |


```

search.words = strsplit("(M|m)ethotrexate;(T|t)rexal;(R|r)heumatrex;(O|o)trexup", ";")[[1]],
ignore.case.sw = FALSE
}

## Running an advanced analysis with filter and search words to
## extract tables and obtain documentation files
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- PDE_pdfs2table_searchandfilter(pdf = paste0(PDE_path(),
                                                             "examples/Methotrexate/29973177_!.pdf"),
  out = paste0(PDE_path(), "examples/29973177_tables/"),
  dev = 20,
  filter.words = strsplit("cohort;case-control;group;study population;study participants", ";")[[1]],
  ignore.case.fw = TRUE,
  filter.word.times = 20,
  table.heading.words = "",
  ignore.case.th = FALSE,
  search.words = strsplit("(M|m)ethotrexate;(T|t)rexal;(R|r)heumatrex;(O|o)trexup", ";")[[1]],
  ignore.case.sw = FALSE,
  eval.abbrevs = TRUE,
  out.table.format = ".csv (WINDOWS-1252)",
  write.table.locations = TRUE,
  write.tab.doc.file = TRUE,
  exp.nondetc.tabs = TRUE,
  delete = TRUE)
}

```

PDE_pdfs2txt_searchandfilter

Extracting sentences from a PDF (Portable Document Format) file

Description

PDE_pdfs2txt_searchandfilter extracts sentences from a single PDF file according to search and filter words and writes output in the corresponding folder.

Usage

```

PDE_pdfs2txt_searchandfilter(
  pdfs,
  out = ".",
  filter.words = "",
  ignore.case.fw = FALSE,
  filter.word.times = 20,
  search.words,
  ignore.case.sw = FALSE,
  eval.abbrevs = TRUE,
  out.table.format = ".csv (WINDOWS-1252)",
  context = 0,

```

```

write.txt.doc.file = FALSE,
delete = TRUE,
verbose = TRUE
)

```

Arguments

| | |
|---------------------------------|--|
| <code>pdfs</code> | String. A list of paths to the PDF files to be analyzed. |
| <code>out</code> | String. Directory chosen to save analysis results in. Default: ".". |
| <code>filter.words</code> | List of strings. The list of filter words. If not NA or "" a hit will be counted every time a word from the list is detected in the article. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). Default: "". |
| <code>ignore.case.fw</code> | Logical. Are the filter words case-sensitive (does capitalization matter)? Default: FALSE. |
| <code>filter.word.times</code> | Numeric. The minimum number of hits described for <code>filter.words</code> for a paper to be further analyzed. Default: 20. |
| <code>search.words</code> | List of strings. List of search words. Regex rules apply (see also https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf). |
| <code>ignore.case.sw</code> | Logical. Are the search words case-sensitive (does capitalization matter)? Default: FALSE. |
| <code>eval.abbrevs</code> | Logical. Should abbreviations for the search words be automatically detected and then replaced with the search word + "\$*"? Default: TRUE. |
| <code>out.table.format</code> | String. Output file format. Either comma separated file <code>.csv</code> or tab separated file <code>.tsv</code> . The encoding indicated in parantheses should be selected according to the operational system exported tables are opened in, i.e., Windows: "(WINDOWS-1252)"; Mac: (macintosh); Linux: (UTF-8). Default: ".csv" and encoding depending on the operational system. |
| <code>context</code> | Numeric. Number of sentences extracted before and after the sentence with the detected search word. If 0 only the sentence with the search word is extracted. Default: 0. |
| <code>write.txt.doc.file</code> | Logical. If TRUE, if no search words were found in the sentences of a PDF file, a file will be created with the PDF filename followed by no.txt.w.search.words . If the PDF file is empty, a file will be created with the PDF filename followed by no.content.detected . If the filter word threshold is not met, a file will be created with the PDF filename followed by no.txt.w.filter.words . Default: FALSE. |
| <code>delete</code> | Logical. If TRUE, the intermediate txt , keeplayouttxt and html copies of the PDF file will be deleted. Default: TRUE. |
| <code>verbose</code> | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |

See Also

[PDE_extr_data_from_pdfs](#)

Examples

```
## Running a simple analysis with filter and search words to extract sentences
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- PDE_pdfs2txt_searchandfilter(pdf = paste0(PDE_path(),
    "examples/Methotrexate/29973177_!.pdf"),
  out = paste0(PDE_path(), "examples/MTX_txt+-0/"),
  filter.words = strsplit("cohort;case-control;group;study population;study participants", ";")[[1]],
  ignore.case.fw = TRUE,
  search.words = strsplit("(M|m)ethotrexate;(T|t)rexal;(R|r)heumatrex;(O|o)trexup", ";")[[1]],
  ignore.case.sw = FALSE)
}

## Running an advanced analysis with filter and search words to
## extract sentences and obtain documentation files
if(PDE_check_Xpdf_install() == TRUE){
  outputtables <- PDE_pdfs2txt_searchandfilter(pdf = paste0(PDE_path(),
    "examples/Methotrexate/29973177_!.pdf"),
  out = paste0(PDE_path(), "/examples/MTX_txt+-1/"),
  context = 1,
  filter.words = strsplit("cohort;case-control;group;study population;study participants", ";")[[1]],
  ignore.case.fw = TRUE,
  filter.word.times = 20,
  search.words = strsplit("(M|m)ethotrexate;(T|t)rexal;(R|r)heumatrex;(O|o)trexup", ";")[[1]],
  ignore.case.sw = FALSE,
  eval.abbrevs = TRUE,
  out.table.format = ".csv (WINDOWS-1252)",
  write.txt.doc.file = TRUE,
  delete = TRUE)
}
```

PDE_reader_i

Browsing the PDE (PDF Data Extractor) analyzer results.

Description

The PDE_reader_i allows the user-friendly visualization and quick-processing of the obtained results.

Usage

```
PDE_reader_i(verbose = TRUE)
```

Arguments

| | |
|---------|--|
| verbose | Logical. Indicates whether messages will be printed in the console. Default: TRUE. |
|---------|--|

Note

A detailed description of the elements in the user interface can be found in the markdown file (README_PDE.md)

Examples

`PDE_reader_i()`

Index

.PDE_extr_data_from_pdf, [2](#)

PDE, [5](#)

PDE-deprecated, [5](#)

PDE_analyzer, [5](#), [6](#)

PDE_analyzer_i, [5](#), [6](#), [7](#)

PDE_check_Xpdf_install, [5](#), [8](#)

PDE_extr_data_from_pdfs, [5](#), [6](#), [8](#), [15](#), [17](#), [19](#)

PDE_install_XpdfReader4.02, [12](#)

PDE_install_XpdfTools4.02, [5](#), [12](#)

PDE_path, [5](#), [13](#)

PDE_pdfs2table, [4](#), [5](#), [11](#), [14](#), [17](#)

PDE_pdfs2table_searchandfilter, [4](#), [5](#), [11](#),
[15](#), [15](#)

PDE_pdfs2txt_searchandfilter, [4](#), [5](#), [11](#),
[18](#)

PDE_reader_i, [5](#), [20](#)