Package: inlpubs (via r-universe)

July 3, 2024

Title USGS INL Project Office Publications

Version 1.1.3

Description Contains bibliographic information for the U.S. Geological Survey (USGS) Idaho National Laboratory (INL) Project Office.

Depends R (>= 4.1)

Imports checkmate, stats, tm

Suggests chromote, connectapi, covr, graphics, htmltools, htmlwidgets, jsonlite, kableExtra, knitr, magick, markdown, pkgbuild, pkgdown, pkgload, pdftools, reactable, renv, rmarkdown, rsconnect, RWeka, stringi, tesseract, textutils, tinytest, utils, webshot2, wordcloud2

License CC0

URL https://rconnect.usgs.gov/INLPO/inlpubs-main/, https://code.usgs.gov/inl/inlpubs

BugReports https://code.usgs.gov/inl/inlpubs/-/issues

Copyright This software is in the public domain because it contains materials that originally came from the United States Geological Survey (USGS), an agency of the United States Department of Interior. For more information, see the official USGS copyright policy at https://www.usgs.gov/information-policies-and-instructions/copyrights-and-credits

Encoding UTF-8

SystemRequirements Complete functionality necessitates Amazon Corretto (win), and default-jre, pandoc, libxml2-dev, libpoppler-cpp-dev, libmagick++-dev, optipng, libtesseract-dev, libleptonica-dev, tesseract-ocr-eng (deb)

LazyData true

LazyDataCompression xz

RoxygenNote 7.3.1

NeedsCompilation no

Author Jason C. Fisher [aut, cre]
(<https://orcid.org/0000-0001-9032-8912>), Kerri C. Treinen
[aut] (<https://orcid.org/0000-0003-0645-6810>), Allison R.
Trcka [aut] (<https://orcid.org/0000-0001-8498-4737>)

Maintainer Jason C. Fisher <jfisher@usgs.gov>

Repository CRAN

Date/Publication 2024-07-02 07:00:02 UTC

Contents

`authors` *Contributing Authors to INLPO Publications*

Description

Authors who have contributed to the publications by the U.S. Geological Survey (USGS), Idaho Water Science Center, Idaho National Laboratory Project Office (INLPO).

Usage

```
authors
```

Format

An object of class 'author' that inherits behavior from the 'data.frame' class and includes the following columns:

`author_id` Unique identifier for the author.

`name` Name of author, surname first and initials or given name.

`person` Information about the person like email address and ORCiD identifier.

`pub_id` Identifier(s) of the publication(s) the author has contributed to, referes to the primry key of the pubs data table.

`total_pub` Total number of publications.

`single_authored` Number of single-authored publications.

`multi_authored` Number of multi-authored publications.

`first_authored` Number of multi-authored publications where the researcher appears as first author.

`first_year` First year author published.

`last_year` Last year author published.

Source

Curated by INLPO staff.

Examples

```
# Subset Jason Fisher's information and display structure:
author <- authors["jfisher", ]
str(author, max.level = 3, width = 75, strict.width = "cut")

# Print author's given name:
author$person |> format(include = "given")
```

`extract_pdf_image` *Extract Image from a PDF Document*

Description

Extract an image from any PDF document. Requires that the **pdftools** and **magick** packages are available.

Usage

```
extract_pdf_image(
  input,
  output = tempfile(fileext = ".jpg"),
  page = 1,
  width = 300,
  depth = 8,
  quality = 70
)
```

Arguments

`input`	'character' string. File path to PDF document.
`output`	'character' string. Location to write the JPEG image file.
`page`	'integer' number. Page number in the document. Defaults to page 1.
`width`	'integer' number. Image width in pixels.
`depth`	'integer' number. Image color depth (either 8 or 16). Defaults to 8.
`quality`	'integer' number. JPEG quality, a number between 0 and 100. Defaults to 70.

Value

Returns the path to the image file.

Author(s)

J.C. Fisher, U.S. Geological Survey, Idaho Water Science Center

See Also

`add_content` function to add cover images to the **inlpubs** package.

Examples

```
input <- system.file("extdata", "test.pdf", package = "inlpubs")
path <- extract_pdf_image(input)

unlink(path)
```

---

extract_pdf_text *Extract Text from a PDF Document*

Description

Extract text from any PDF document. Requires that the **pdftools** and **tesseract** packages are available.

Usage

```
extract_pdf_text(
  input,
  output = tempfile(fileext = ".txt"),
  dpi = 600,
  psm = 1
)
```

Arguments

`input`	'character' string. File path to PDF document.
`output`	'character' string. Location to write the text file.
`dpi`	'integer' number between 100 and 1200. Dots per inch (DPI). The resolution of an image, specifically the number of pixels per inch. For optimal optical character recognition (OCR) accuracy, 600 DPI (the default) is recommended.
`psm`	`integer` number between 0 and 13. Page Segmentation Mode (PSM). Describes the layout of the text you are trying to extract. For processing two columns of text you should use the page segmentation mode 1 (default). PSM 1 (default) is used to automatically segment the page into different text areas and also detect the orientation and script of the text.

Value

Returns the path to the text file. Each page from the PDF is transcribed as a separate line in the file.

Author(s)

J.C. Fisher, U.S. Geological Survey, Idaho Water Science Center

See Also

`add_content` function to add texts to the **inlpubs**-package corpus.

Examples

```
## Not run:
  input <- system.file("extdata", "test.pdf", package = "inlpubs")
  path <- extract_pdf_text(input)

  unlink(path)

## End(Not run)
```

`make_wordcloud` *Create Word Cloud*

Description

Create a word cloud from a frequency table of words, and save to a PNG file. Requires R-packages **htmltools**, **htmlwidgets**, **magick**, **webshot2**, and **wordcloud2** are available. System dependencies include the the following: ImageMagick for displaying the PNG image, OptiPNG for PNG file compression, and Chrome- or a Chromium-based browser with support for the Chrome DevTools protocol. Use `find_chromate` function to find the path to the Chrome browser.

Usage

```
make_wordcloud(
  x,
  max_terms = 200,
  size = 1,
  shape = "circle",
  ellipticity = 0.65,
  ...,
  width = 910,
  output = NULL,
  display = FALSE
)
```

Arguments

x	'data.frame'. A frequency table of terms that includes "term" and "freq" in each column.
max_terms	'integer' number. Maximum number of terms to include in the word cloud.
size	'numeric' number. Font size.
shape	'character' string. Shape of the "cloud" to draw. Possible shapes include a "circle", "cardioid", "diamond", "triangle-forward", "triangle", "pentagon", and "star".
ellipticity	'numeric' number. Degree of "flatness" of the shape to draw, a value between 0 and 1.
...	Additional arguments to be passed to the wordcloud2 function.
width	'integer' number. Desired image width in pixels.
output	'character' string. Path to the output file, by default the word cloud is copied to a temporary file.
display	'logical' flag. Whether to display the saved PNG file in a graphics window. Requires access to the **magick** package.

Value

File path to the word cloud plot in PNG format.

Author(s)

J.C. Fisher, U.S. Geological Survey, Idaho Water Science Center

See Also

mine_text function to perform a term frequency text analysis.

Examples

```
## Not run:
  d <- wordcloud2::demoFreq |> head(n = 10)
  colnames(d) <- c("term", "freq")
  file <- make_wordcloud(d, display = interactive())

  unlink(file)

## End(Not run)
```

mine_text *Mine Text*

Description

Performs a term frequency text analysis. A term is defined as a word or group of words.

Usage

```
mine_text(docs, ngmin = 1, ngmax = ngmin, sparse = NULL)
```

Arguments

docs	'list' or 'character' vector. Document text to analyze. Each list item contains the extracted text from a single document.
ngmin, ngmax	integer number. Splits strings into *n-grams* with given minimal and maximal numbers of grams. An n-gram is an ordered sequence of n words taken from the body of a text. Requires the **RWeka** package is available and that the environment variable JAVA_HOME points to where the Java software is located. Recommended for single text compoents only.
sparse	'numeric' number that is greater than 0 and less than 1. A threshold of relative document frequency for a term. It specifies the proportion of documents in which a term must appear to be retained. For example if you specify sparse equal to 0.99, it removes terms that are more sparse than 0.99. Conversely, at 0.01, only terms appearing in nearly every document will be retained.

Details

HTML entities are decoded when the **textutils** package is available.

Value

A term-frequency data table giving the number of times each word occurs in the text. A column in the table represents a single component in the docs argument, and each row provides frequency counts for a particular word (also known as a 'term').

Author(s)

J.C. Fisher, U.S. Geological Survey, Idaho Water Science Center

See Also

search_terms function to search for terms within the resulting term-frequency data table.

make_wordcloud function to create a word cloud.

Examples

```
d <- c(
  "The quick brown fox jumps over the lazy lazy dog.",
  "Pack my brown box.",
  "Jazz fly brown dog."
) |>
  mine_text()

d <- list(
  "A" = "The quick brown fox jumps over the lazy lazy dog.",
  "B" = c("Pack my brown box.", NA, "Jazz fly brown dog."),
  "C" = NA_character_
) |>
  mine_text()
```

---

pubs *Publications of the INLPO*

Description

Bibliographic information for reports, articles, maps, and theses related to scientific monitoring and research conducted by the U.S. Geological Survey (USGS), Idaho Water Science Center, Idaho National Laboratory Project Office (INLPO).

Usage

```
pubs
```

Format

An object of class 'pub' that inherits behavior from the 'data.frame' class and includes the following columns:

`pub_id` Unique identifier for the publication.

`institution` Name of the institution that published and/or sponsored the report.

`type` Type of publication.

`text_ref` Text reference (also known as the in-text citation) that excludes the year of publication.

`year` Year of publication.

`author_id` Identifier(s) of the author(s), referes to the primry key of the `authors` data table.

`title` Title of publication.

`bibentry` Bibliographic entry of class `bibentry`.

`abstract` Abstract of publication.

`annotation` Annotation of publication.

`annotation_src` Identifier for the annotation source publication (Knobel and others, 2005; Bartholomay, 2022).

`files` File names associated with the publication.

Source

Many of these publications are available through the USGS Publications Warehouse.

References

Bartholomay, R.C., 2022, Historical development of the U.S. Geological Survey hydrological monitoring and investigative programs at the Idaho National Laboratory, Idaho, 2002-2020: U.S. Geological Survey Open-File Report 2022-1027 (DOE/ID-22256), 54 p., doi:10.3133/ofr20221027.

Knobel, L.L., Bartholomay, R.C., and Rousseau, J.P., 2005, Historical development of the U.S. Geological Survey hydrologic monitoring and investigative programs at the Idaho National Engineering and Environmental Laboratory, Idaho, 1949 to 2001: U.S. Geological Survey Open-File Report 2005–1223 (DOE/ID–22195), 93 p., doi:10.3133/ofr20051223.

Examples

```
# Subset Fisher and others (2012) and display structure:
id <- "FisherOthers2012"
pub <- pubs[id, ]
str(pub, max.level = 3, width = 75, strict.width = "cut")

# Print suggested citation:
attr(unclass(pub$bibentry[[1]])[[1]], which = "textVersion")

# Print authors full name:
format(pub$bibentry[[1]]$author, include = c("given", "family"))

# Print abstract:
pub$abstract
```

`search_terms` *Search Terms*

Description

Pattern matches a search term within the term-frequency data table.

Usage

```
search_terms(
  x,
  data = inlpubs::terms,
  ignore.case = TRUE,
  ...,
  low_freq = 1,
  high_freq = Inf,
  simplify = TRUE
)
```

Arguments

x	'character' string. Term searched for in the term-frequency data table.
data	'term' and 'data.frame' class. Term-frequency data table. Defaults to using the term frequencies from the INLPO publications, see terms dataset for details.
ignore.case	'logical' flag. Whether to ignore character case during pattern matching.
...	Additional arguments passed to the grep function.
low_freq	'numeric' number. Lower frequency bound.
high_freq	'numeric' number. Upper frequency bound.
simplify	'logical' flag. Whether to return only the unique publication identifiers.

Value

A subset of the data table sorted by decreasing frequency.

Author(s)

J.C. Fisher, U.S. Geological Survey, Idaho Water Science Center

See Also

mine_text function to perform a term frequency text analysis.

Examples

```
search_terms("mlms")

out <- search_terms("mlms", simplify = FALSE)
head(out)
```

---

terms *Term Frequency from INLPO Publications*

Description

Term frequency from publications by the U.S. Geological Survey (USGS), Idaho Water Science Center, Idaho National Laboratory Project Office (INLPO).

Usage

```
terms
```

Format

An object of class 'term' that inherits behavior from the 'data.frame' class and includes the following columns:

`term` Term, a word or group of words, represented by an ASCII character string in lowercase.

`pub_id` Identifier for a publication, referes to the primry key of the `pubs` data table.

`freq` Frequency count from text analysis.

Source

The publication text was sourced from the original PDF documents using the `extract_pdf_text` function, and term frequencies were extracted from the text using the `mine_text` function.

Examples

```
str(terms, max.level = 3, width = 75, strict.width = "cut")
```

Index