Package 'highlightr'

May 8, 2026

Title Highlight Conserved Edits Across Versions of a Document

Version 2.0.0

Description Input multiple versions of a source document,
and receive HTML code for a highlighted version of the source document
indicating the frequency of occurrence of phrases in the different versions.
This method is described in Chapter 3 of
Rogers (2024) <https://digitalcommons.unl.edu/dissertations/AAI31240449/>.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.3

Imports dplyr, ggplot2, magrittr, purrr, quanteda, quanteda.textstats,
stringi, stringr, tibble, tidyr, tm, zoomerjoin

Depends R (>= 3.5)

LazyData true

URL https://rachelesrogers.github.io/highlightr/,
https://github.com/rachelesrogers/highlightr

Suggests knitr, rmarkdown, testthat (>= 3.0.0), xml2

VignetteBuilder knitr

Config/testthat/edition 3

BugReports https://github.com/rachelesrogers/highlightr/issues

NeedsCompilation no

Author Center for Statistics and Applications in Forensic Evidence [aut, cph,
fnd],
Rachel Rogers [aut, cre] (ORCID:
<https://orcid.org/0000-0002-4145-9630>),
Susan VanderPlas [aut] (ORCID: <https://orcid.org/0000-0002-3803-0972>)

Maintainer Rachel Rogers <rrogers.rpackages@gmail.com>

Repository CRAN

Date/Publication 2026-04-10 23:10:02 UTC

Contents

collocation_frequency 2
collocation_plot 3
highlighted_text 4
notepad_example 5
wiki_pages 5

Index **6**

`collocation_frequency` *Mapping Collocation Frequency to Source Document*

Description

This function provides the frequency of collocations in comments that correspond to the provided source document.

Usage

```
collocation_frequency(
  tbl,
  source_row,
  text_column,
  collocate_length = 5,
  fuzzy = FALSE,
  n_bands = 50,
  threshold = 0.7,
  n_gram_width = 4,
  band_width = 8
)
```

Arguments

`tbl`	data frame containing documents, where each row represents a document
`source_row`	row containing text to be treated as source
`text_column`	string indicating the name of the column containing derivative text
`collocate_length`	the length of the collocation. Default is 5
`fuzzy`	whether or not to use fuzzy matching in collocation calculations
`n_bands`	number of bands used in MinHash algorithm passed to `zoomerjoin::jaccard_right_join()`. Default is 50
`threshold`	Jaccard distance threshold to be considered a match passed to `zoomerjoin::jaccard_right_join()`. Default is 0.7
`n_gram_width`	width of n-grams used in Jaccard distance calculation passed to `zoomerjoin::jaccard_right_join()`. Default is 4
`band_width`	width of band used in MinHash algorithm passed to `zoomerjoin::jaccard_right_join()`. Default is 8

Details

Collocations are sequences of words present in the source document. For example, the phrase "the blue bird flies" contains one collocation of length 4 ("the blue bird flies"), two collocations of length 3 ("the blue bird" and "blue bird flies"), and three collocations of length 2 ("the blue", "blue bird", and "bird flies"). This function counts the number of corresponding phrases in the 'notes', or the derivative documents. This count is divided by the number of times the phrase occurs in the source document. When fuzzy matching is included, indirect matches are included with a weight of (n*d)/m, where n is the frequency of the fuzzy collocation, d is the Jaccard similarity between the transcript and note collocation, and m is the number of closest matches for the note collocation.

Value

a dataframe of the transcript document with collocation values by word

Examples

```
src_row <- which(notepad_example$ID=="source")
merged_frequency <- collocation_frequency(notepad_example, src_row, "Text")
```

---

collocation_plot *Map collocation to ggplot object*

Description

This assigns colors based on frequency to the words in the transcript.

Usage

```
collocation_plot(
  frequency_doc,
  colors = c("#f251fc", "#f8ff1b"),
  values = "Freq",
  order = "word_num",
  text = "words"
)
```

Arguments

frequency_doc	document of frequencies (returned from collocation_frequency())
colors	list for color specification for the gradient. Default is c("#f251fc","#f8ff1b")
values	column name of values to use in gradient calculation. Default is "Freq", corresponding to document returned from collocation_frequency()
order	column name corresponding to the the word order of the text. Default is "word_num", corresponding to the document returned from collocation_frequency()
text	column name corresponding to text to map the gradient to. Default is "words", corresponding to the document returned from collocation_frequency()

Value

list of plot, plot object, and frequency

Examples

```
# Identify Source Row
src_row <- which(notepad_example$ID=="source")
merged_frequency <- collocation_frequency(notepad_example, src_row, "Text")
# Create a plot object to assign colors based on frequency
freq_plot <- collocation_plot(merged_frequency)
```

`highlighted_text` *Create Highlighted Testimony*

Description

Adds html tags to create a highlighted testimony corresponding to word frequency. To render correctly, the object produced from `highlighted_text()` can be added outside of a code chunk in an .Rmd document in the `` `r highlighted_text()` `` format. Alternatively, the html output can be saved by using the `xml2` package as follows: `xml2::write_html(xml2::read_html(highlighted_text(), "filepath.html"))`

Usage

```
highlighted_text(plot_object, labels = c("", ""))
```

Arguments

`plot_object`	plot object resulting from `collocation_plot()`
`labels`	lower and upper labels for the gradient scale

Value

html code for highlighted text

Examples

```
# Identify Source Row
src_row <- which(notepad_example$ID=="source")
# Calculate Frequency
merged_frequency <- collocation_frequency(notepad_example, src_row, "Text")
# Create a plot object to assign colors based on frequency
freq_plot <- collocation_plot(merged_frequency)
# Add html tags to create a highlighted version of the source document
page_highlight <- highlighted_text(freq_plot, merged_frequency)
```

notepad_example *Comment Example Dataset*

Description

Participant comments for the initial description used in the jury perception study

Usage

```
notepad_example
```

Format

`notepad_example`**:**
A data frame with 126 rows and 2 columns:

ID Participant Identifier, as well as source document identifier

Text Participant notes, as well as source transcript

Source

Jury Perception Study (see Rogers (2024) https://digitalcommons.unl.edu/dissertations/AAI31240449/)

wiki_pages *Wikipedia Edit History for "Highlighter"*

Description

Text corresponding to versions of the Wikipedia article for Highlighter

Usage

```
wiki_pages
```

Format

`wiki_pages`**:**
A data frame with 300 rows and 1 column:

page_notes text of the Wikipedia page for Highlighter

Source

Wikipedia: https://en.wikipedia.org/w/index.php?title=Highlighter&action=history

Index

∗ **datasets**
notepad_example, 5
wiki_pages, 5

collocation_frequency, 2
collocation_frequency(), *3*
collocation_plot, 3
collocation_plot(), *4*

highlighted_text, 4

notepad_example, 5

wiki_pages, 5