# Package 'highlightr'

October 19, 2025

**Title** Highlight Conserved Edits Across Versions of a Document

**Version** 1.2.0

**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** 7.3.2

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

**Depends** R (>= 2.10)

**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** 2025-10-19 02:10:03 UTC

# Contents



---

`collocate_comments` *Collocation of Comments*

---

## Description

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

## Usage

```
collocate_comments(transcript_token, note_token, collocate_length = 5)
```

## Arguments

| | |
|---|---|
| `transcript_token` | transcript token to act as baseline for notes, resulting from `token_transcript()` |
| `note_token` | tokenized document of notes, resulting from `token_comments()` |
| `collocate_length` | the length of the collocation. Default is 5 |

## 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. Matches between the two documents must be exact

## Value

data frame of the transcript and corresponding note frequency

## Examples

```
# Rename relevant column to page_notes in the derivative document
comment_example_rename <- dplyr::rename(comment_example, page_notes=Notes)
# Tokenize the derivative document
toks_comment <- token_comments(comment_example_rename[1:100,])
# Rename relevant column in the source document to text
transcript_example_rename <- dplyr::rename(transcript_example, text=Text)
# Tokenize source document
toks_transcript <- token_transcript(transcript_example_rename)
# Compute collocation frequencies
collocation_object <- collocate_comments(toks_transcript, toks_comment)
```

---

# collocate_comments_fuzzy

*Collocate Comments Fuzzy*

---

## Description

This function provides the frequency of collocations in comments that correspond to the provided transcript, using fuzzy matching.

## Usage

```
collocate_comments_fuzzy(
  transcript_token,
  note_token,
  collocate_length = 5,
  n_bands = 50,
  threshold = 0.7,
  n_gram_width = 4
)
```

## Arguments

| | |
|---|---|
| transcript_token | transcript token to act as baseline for notes, resulting from `token_transcript()` |
| note_token | tokenized document of notes, resulting from `token_comments()` |
| collocate_length | the length of the collocation. Default is 5 |
| 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 |

## 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. Due to fuzzy matching, 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

data frame of the transcript and corresponding note frequency

## Examples

```
# Rename relevant column to page_notes in the derivative document
comment_example_rename <- dplyr::rename(comment_example[1:10,], page_notes=Notes)
# Tokenize the derivative document
toks_comment <- token_comments(comment_example_rename)
# Rename relevant column in the source document to text
transcript_example_rename <- dplyr::rename(transcript_example, text=Text)
# Tokenize source document
toks_transcript <- token_transcript(transcript_example_rename)
# Compute collocation frequencies using fuzzy (or indirect) matching
fuzzy_object <- collocate_comments_fuzzy(toks_transcript, toks_comment)
```

---

# 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,
  n_scenario = 1,
  colors = c("#f251fc", "#f8ff1b")
)
```

## Arguments

| | |
|---|---|
| frequency_doc | document of frequencies (returned from transcript_frequency()) |
| n_scenario | number of scenarios for which this transcript appeared. Defualt is 1 |
| colors | list for color specification for the gradient. Default is c("#f251fc","#f8ff1b") |

## Value

list of plot, plot object, and frequency

## Examples

```
# Rename relevant column to page_notes in the derivative document
comment_example_rename <- dplyr::rename(comment_example, page_notes=Notes)
# Tokenize the derivative document
toks_comment <- token_comments(comment_example_rename)
# Rename relevant column in the source document to text
transcript_example_rename <- dplyr::rename(transcript_example, text=Text)
# Tokenize source document
toks_transcript <- token_transcript(transcript_example_rename)
# Compute collocation frequencies
collocation_object <- collocate_comments(toks_transcript, toks_comment)
# Merge frequencies with source document to provide averages by word and correct formatting
merged_frequency <- transcript_frequency(transcript_example_rename, collocation_object)
# Create a plot object to assign colors based on frequency
freq_plot <- collocation_plot(merged_frequency)
```

---

# comment_example *Comment Example Dataset*

---

## Description

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

## Usage

```
comment_example
```

## Format

`comment_example`**:**

A data frame with 125 rows and 2 columns:

**ID** Participant Identifier

**Notes** Participant notes

## Source

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

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

```
# Rename relevant column to page_notes in the derivative document
comment_example_rename <- dplyr::rename(comment_example, page_notes=Notes)
# Tokenize the derivative document
toks_comment <- token_comments(comment_example_rename)
# Rename relevant column in the source document to text
transcript_example_rename <- dplyr::rename(transcript_example, text=Text)
# Tokenize source document
toks_transcript <- token_transcript(transcript_example_rename)
# Compute collocation frequencies
collocation_object <- collocate_comments(toks_transcript, toks_comment)
# Merge frequencies with source document to provide averages by word and correct formatting
merged_frequency <- transcript_frequency(transcript_example_rename, collocation_object)
# 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)
```

## token_comments *Tokenize comments*

### Description

This function tokenizes comments that are to be used in `collocate_comments_fuzzy()` or `collocate_comments()`

### Usage

```
token_comments(comment_document)
```

### Arguments

`comment_document`
: document containing notes by individual, where the column containing the notes is named page_notes

### Value

tokenized comments

### Examples

```
# Rename relevant column to page_notes in the derivative document
comment_example_rename <- dplyr::rename(comment_example, page_notes=Notes)
# Tokenize the derivative document
toks_comment <- token_comments(comment_example_rename)
```

## token_transcript *Tokenize Transcript*

### Description

This function tokenizes a transcript document that is to be used in `collocate_comments_fuzzy()` or `collocate_comments()`

### Usage

```
token_transcript(transcript_file)
```

### Arguments

`transcript_file`
: data frame of the transcript, where the transcript text is in a column named text.

## Value

a tokenized object

## Examples

```
# Rename relevant column in the source document to text
transcript_example_rename <- dplyr::rename(transcript_example, text=Text)
# Tokenize source document
toks_transcript <- token_transcript(transcript_example_rename)
```

---

# transcript_example *Transcript Example*

---

## Description

Text corresponding to participant comments

## Usage

```
transcript_example
```

## Format

`transcript_example`**:**

A data frame with 1 row and 1 column:

**Text** Transcript text corresponding to the jury perception study

## Source

Jury Perception Study (see Rogers (2024) https://digitalcommons.unl.edu/dissertations/AAI31240449/ and Garrett et. al. (2020) doi:10.1037/lhb0000423)

---

# transcript_frequency *Mapping Collocation Frequency to Transcript Document*

---

## Description

This function connects the collocation frequency calculated in `collocate_comments_fuzzy()` to the base transcript.

## Usage

```
transcript_frequency(transcript, collocate_object)
```

### Arguments

| | |
|---|---|
| transcript | transcript document |
| collocate_object | collocation object (returned from collocate_comments_fuzzy() or collocate_comments()) |

### Value

a dataframe of the transcript document with collocation values by word

### Examples

```
# Rename relevant column to page_notes in the derivative document
comment_example_rename <- dplyr::rename(comment_example, page_notes=Notes)
# Tokenize the derivative document
toks_comment <- token_comments(comment_example_rename)
# Rename relevant column in the source document to text
transcript_example_rename <- dplyr::rename(transcript_example, text=Text)
# Tokenize source document
toks_transcript <- token_transcript(transcript_example_rename)
# Compute collocation frequencies
collocation_object <- collocate_comments(toks_transcript, toks_comment)
# Merge frequencies with source document to provide averages by word and correct formatting
merged_frequency <- transcript_frequency(transcript_example_rename, collocation_object)
```

---

## 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