Package 'samesies'

March 10, 2025

Title Compare Similarity Across Text, Factors, or Numbers

Version 0.1.0

Description Compare lists of texts, factors, or numerical values to measure their similarity. The motivating use case is evaluating the similarity of large language model responses across models, providers, or prompts. Approximate string matching is implemented using 'stringdist'.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.2

Imports cli, dplyr, ggbeeswarm, ggplot2, purrr, scales, stats, stringdist

Suggests testthat (>= 3.0.0), devtools

Config/testthat/edition 3

URL https://dylanpieper.github.io/samesies/

NeedsCompilation no

Author Dylan Pieper [aut, cre]

Maintainer Dylan Pieper <dylanpieper@gmail.com>

Repository CRAN

Date/Publication 2025-03-10 14:50:02 UTC

Contents

average_similarity 2
pair_averages 3
print.similar 3
print.similar_factor 4
print.similar_number 4
print.similar_text 5
print.summary.similar 5
print.summary.similar_factor 6
print.summary.similar_number 6

print.summary.similar_text 7
same_factor 7
same_number 8
same_text 9
similar 10
similar_factor 11
similar_number 12
similar_text 13
summary.similar 14
summary.similar_factor 15
summary.similar_number 15
summary.similar_text 16

Index **17**

`average_similarity` *Calculate Average Similarity Scores*

Description

Calculates and returns the average similarity score for each method used in the comparison.

Usage

```
average_similarity(x, ...)

average_similarity(x, ...)
```

Arguments

`x`	A similarity object
`...`	Additional arguments (not used)

Value

A named numeric vector of mean similarity scores for each method

A named numeric vector of mean similarity scores for each method

pair_averages *Calculate Average Similarity Scores By Pairs*

Description

Calculates and returns the average similarity scores for each pair of lists compared, broken down by method.

Usage

```
pair_averages(x, method = NULL, ...)

pair_averages(x, method = NULL, ...)
```

Arguments

x	A similarity object
method	Optional character vector of methods to include
...	Additional arguments (not used)

Value

A data frame containing:

method	The similarity method used
pair	The pair of lists compared
avg_score	Mean similarity score for the pair

A data frame containing pair-wise average scores

print.similar *Print a similarity object*

Description

Print a similarity object

Usage

```
## S3 method for class 'similar'
print(x, ...)
```

Arguments

x	A similarity object
...	Additional arguments (not used)

Value

The object invisibly

---

print.similar_factor *Print method for similar_factor objects*

Description

Print method for similar_factor objects

Usage

```
## S3 method for class 'similar_factor'
print(x, ...)
```

Arguments

x	A similar_factor object
...	Additional arguments (not used)

Value

The object invisibly

---

print.similar_number *Print method for similar_number objects*

Description

Print method for similar_number objects

Usage

```
## S3 method for class 'similar_number'
print(x, ...)
```

Arguments

x	A similar_number object
...	Additional arguments (not used)

Value

The object invisibly

print.similar_text *Print method for similar_text objects*

Description

Print method for similar_text objects

Usage

```
## S3 method for class 'similar_text'
print(x, ...)
```

Arguments

x	A similar_text object
...	Additional arguments (not used)

Value

The object invisibly

print.summary.similar *Print method for summary.similar objects*

Description

Print method for summary.similar objects

Usage

```
## S3 method for class 'summary.similar'
print(x, ...)
```

Arguments

x	A summary.similar object
...	Additional arguments (not used)

Value

The summary object invisibly

print.summary.similar_factor

Print method for summary.similar_factor objects

Description

Print method for summary.similar_factor objects

Usage

```
## S3 method for class 'summary.similar_factor'
print(x, ...)
```

Arguments

x	A summary.similar_factor object
...	Additional arguments (not used)

Value

The object invisibly

print.summary.similar_number

Print method for summary.similar_number objects

Description

Print method for summary.similar_number objects

Usage

```
## S3 method for class 'summary.similar_number'
print(x, ...)
```

Arguments

x	A summary.similar_number object
...	Additional arguments (not used)

Value

The object invisibly

print.summary.similar_text *Print method for summary.similar_text objects*

Description

Print method for summary.similar_text objects

Usage

```
## S3 method for class 'summary.similar_text'
print(x, ...)
```

Arguments

x	A summary.similar_text object
...	Additional arguments (not used)

Value

The object invisibly

same_factor *Compare Factor Similarity Across Lists*

Description

Compare Factor Similarity Across Lists

Usage

```
same_factor(
  ...,
  method = c("exact", "order"),
  levels,
  ordered = FALSE,
  digits = 3
)
```

Arguments

...	Lists of categorical values (character or factor) to compare
method	Character vector of similarity methods. Choose from: "exact", "order" (default: all)
levels	Character vector of all allowed levels for comparison
ordered	Logical. If TRUE, treat levels as ordered (ordinal). If FALSE, the "order" method is skipped.
digits	Number of digits to round results (default: 3)

Value

An S3 object of type "similar_factor" containing:

- scores: Numeric similarity scores by method and comparison
- summary: Summary statistics by method and comparison
- methods: Methods used for comparison
- list_names: Names of compared lists
- levels: Levels used for categorical comparison

---

same_number *Compare Numerical Similarity Across Lists*

Description

Computes similarity scores between two or more lists of numeric values using multiple comparison methods.

Usage

```
same_number(
  ...,
  method = c("exact", "raw", "exp", "percent", "normalized", "fuzzy"),
  epsilon = 0.05,
  epsilon_pct = 0.02,
  max_diff = NULL,
  digits = 3
)
```

Arguments

...	Two or more lists containing numeric values to compare
method	Character vector specifying similarity methods (default: all)
epsilon	Threshold for fuzzy matching (default: NULL for auto-calculation)
epsilon_pct	Relative epsilon percentile (default: 0.02 or 2%). Only used when method is "fuzzy"
max_diff	Maximum difference for normalization (default: NULL for auto-calculation)
digits	Number of digits to round results (default: 3)

Details

The available methods are:

- `exact`: Binary similarity (1 if equal, 0 otherwise)
- `percent`: Percentage difference relative to the larger value
- `normalized`: Absolute difference normalized by a maximum difference value
- `fuzzy`: Similarity based on an epsilon threshold
- `exp`: Exponential decay based on absolute difference (e^-diff)
- `raw`: Returns the raw absolute difference (|num1 - num2|) instead of a similarity score

Value

An S3 object containing:

- `scores`: A list of similarity scores for each method and list pair
- `summary`: A list of statistical summaries for each method and list pair
- `methods`: The similarity methods used
- `list_names`: Names of the input lists
- `raw_values`: The original input lists

Examples

```
nums1 <- list(1, 2, 3)
nums2 <- list(1, 2.1, 3.2)
result <- same_number(nums1, nums2)
```

`same_text` *Compare Text Similarity Across Lists*

Description

Compare Text Similarity Across Lists

Usage

```
same_text(
  ...,
  method = c("osa", "lv", "dl", "hamming", "lcs", "qgram", "cosine", "jaccard", "jw",
    "soundex"),
  q = 1,
  p = NULL,
  bt = 0,
  weight = c(d = 1, i = 1, s = 1, t = 1),
  digits = 3
)
```

Arguments

...	Lists of character strings to compare
method	Character vector of similarity methods from stringdist. Choose from: "osa", "lv", "dl", "hamming", "lcs", "qgram", "cosine", "jaccard", "jw", "soundex" (default: all)
q	Size of q-gram for q-gram based methods (default: 1)
p	Winkler scaling factor for "jw" method (default: 0.1)
bt	Booth matching threshold
weight	Vector of weights for operations: deletion (d), insertion (i), substitution (s), transposition (t)
digits	Number of digits to round results (default: 3)

Value

An S3 class object of type "similar_text" containing:

- scores: Numeric similarity scores by method and comparison
- summary: Summary statistics by method and comparison
- methods: Methods used for comparison
- list_names: Names of compared lists

Examples

```
list1 <- list("hello", "world")
list2 <- list("helo", "word")
result <- same_text(list1, list2)
```

---

similar *Abstract parent class for similarity comparison*

Description

similar is an S3 class for all similarity comparison objects. This class defines common properties shared among child classes like similar_text, similar_factor, and similar_number.

Usage

```
similar(scores, summary, methods, list_names, digits = 3)
```

Arguments

scores	List of similarity scores per method and comparison
summary	Summary statistics by method and comparison
methods	Character vector of methods used for comparison
list_names	Character vector of names for the compared lists
digits	Number of digits to round results (default: 3)

Details

This class provides the foundation for all similarity comparison classes. It includes common properties:

- scores: List of similarity scores per method and comparison
- summary: Summary statistics by method and comparison
- methods: Character vector of methods used for comparison
- list_names: Character vector of names for the compared lists
- digits: Number of digits to round results in output

Value

An object of class "similar" with the following components:

- scores: List of similarity scores per method and comparison
- summary: Summary statistics by method and comparison
- methods: Character vector of methods used for comparison
- list_names: Character vector of names for the compared lists
- digits: Number of digits to round results in output

The similarity scores are normalized values between 0 and 1, where 1 indicates perfect similarity and 0 indicates no similarity.

`similar_factor` *Factor similarity comparison class*

Description

`similar_factor` is an S3 class for categorical/factor similarity comparisons.

Usage

```
similar_factor(scores, summary, methods, list_names, levels, digits = 3)
```

Arguments

`scores`	List of similarity scores per method and comparison
`summary`	Summary statistics by method and comparison
`methods`	Character vector of methods used for comparison
`list_names`	Character vector of names for the compared lists
`levels`	Character vector of factor levels
`digits`	Number of digits to round results (default: 3)

Details

This class extends the `similar` class and implements categorical data-specific similarity comparison methods.

Value

An object of class "similar_factor" (which inherits from "similar") containing:

- scores: List of factor similarity scores per method and comparison
- summary: Summary statistics by method and comparison
- methods: Character vector of factor comparison methods used (exact, order)
- list_names: Character vector of names for the compared factor lists
- digits: Number of digits to round results in output
- levels: Character vector of factor levels used in the comparison

The factor similarity scores are normalized values between 0 and 1, where 1 indicates identical factors and 0 indicates completely different factors based on the specific method used.

`similar_number` *Numeric similarity comparison class*

Description

`similar_number` is an S3 class for numeric similarity comparisons.

Usage

```
similar_number(scores, summary, methods, list_names, raw_values, digits = 3)
```

Arguments

`scores`	List of similarity scores per method and comparison
`summary`	Summary statistics by method and comparison
`methods`	Character vector of methods used for comparison
`list_names`	Character vector of names for the compared lists
`raw_values`	List of raw numeric values being compared
`digits`	Number of digits to round results (default: 3)

Details

This class extends the `similar` class and implements numeric data-specific similarity comparison methods.

Value

An object of class "similar_number" (which inherits from "similar") containing:

- scores: List of numeric similarity scores per method and comparison
- summary: Summary statistics by method and comparison
- methods: Character vector of numeric comparison methods used (exact, percent, normalized, fuzzy, exp, raw)
- list_names: Character vector of names for the compared numeric lists
- digits: Number of digits to round results in output
- raw_values: List of raw numeric values that were compared

The numeric similarity scores are normalized values between 0 and 1, where 1 indicates identical numbers and 0 indicates maximally different numbers based on the specific method used. The exception is the "raw" method, which returns the absolute difference between values.

`similar_text` *Text similarity comparison class*

Description

`similar_text` is an S3 class for text similarity comparisons.

Usage

```
similar_text(scores, summary, methods, list_names, digits = 3)
```

Arguments

`scores`	List of similarity scores per method and comparison
`summary`	Summary statistics by method and comparison
`methods`	Character vector of methods used for comparison
`list_names`	Character vector of names for the compared lists
`digits`	Number of digits to round results (default: 3)

Details

This class extends the `similar` class and implements text-specific similarity comparison methods.

Value

An object of class "similar_text" (which inherits from "similar") containing:

- scores: List of text similarity scores per method and comparison
- summary: Summary statistics by method and comparison
- methods: Character vector of text similarity methods used (osa, lv, dl, etc.)
- list_names: Character vector of names for the compared text lists
- digits: Number of digits to round results in output

The text similarity scores are normalized values between 0 and 1, where 1 indicates identical text and 0 indicates completely different text based on the specific method used.

`summary.similar` *Summarize a similarity object*

Description

Summarize a similarity object

Usage

```
## S3 method for class 'similar'
summary(object, ...)
```

Arguments

`object`	A similarity object
`...`	Additional arguments (not used)

Value

A summary object

summary.similar_factor *Summary method for similar_factor objects*

Description

Summary method for similar_factor objects

Usage

```
## S3 method for class 'similar_factor'
summary(object, ...)
```

Arguments

object	A similar_factor object
...	Additional arguments (not used)

Value

A summary.similar_factor object

summary.similar_number *Summary method for similar_number objects*

Description

Summary method for similar_number objects

Usage

```
## S3 method for class 'similar_number'
summary(object, ...)
```

Arguments

object	A similar_number object
...	Additional arguments (not used)

Value

A summary.similar_number object

summary.similar_text *Summary method for similar_text objects*

Description

Summary method for similar_text objects

Usage

```
## S3 method for class 'similar_text'
summary(object, ...)
```

Arguments

object	A similar_text object
...	Additional arguments (not used)

Value

A summary.similar_text object

Index

average_similarity, 2

pair_averages, 3
print.similar, 3
print.similar_factor, 4
print.similar_number, 4
print.similar_text, 5
print.summary.similar, 5
print.summary.similar_factor, 6
print.summary.similar_number, 6
print.summary.similar_text, 7

same_factor, 7
same_number, 8
same_text, 9
similar, 10
similar_factor, 11
similar_number, 12
similar_text, 13
summary.similar, 14
summary.similar_factor, 15
summary.similar_number, 15
summary.similar_text, 16