Package 'tipitaka.critical'

May 8, 2026

Type Package

Title Lemmatized Critical Edition of the Pali Canon

Version 1.0.0

Description A lemmatized critical edition of the complete Pali Canon (Tipitaka), the canonical scripture of Theravadin Buddhism. Based on a five-witness collation of the Pali Text Society (PTS) edition (via 'GRETIL'), 'SuttaCentral', the Vipassana Research Institute (VRI) Chattha Sangayana edition, the Buddha Jayanti Tipitaka (BJT), and the Thai Royal Edition. All text is lemmatized using the 'Digital Pali Dictionary', grouping inflected forms by dictionary headword. Covers all three pitakas (Sutta, Vinaya, Abhidhamma) with 5,777 individual text units. The companion package 'tipitaka' provides the original VRI edition data and Pali text tools. For background on the collation method, see Zigmond (2026) <https://github.com/dangerzig/tipitaka.critical>.

URL https://github.com/dangerzig/tipitaka.critical

BugReports https://github.com/dangerzig/tipitaka.critical/issues

License CC0

Encoding UTF-8

LazyData true

LazyDataCompression xz

RoxygenNote 7.3.3

Depends R (>= 3.5), Matrix

Suggests dplyr, knitr, rmarkdown, testthat (>= 3.0.0), tidyr

Config/testthat/edition 3

VignetteBuilder knitr

NeedsCompilation yes

Author Dan Zigmond [aut, cre]

Maintainer Dan Zigmond <djz@shmonk.com>

Repository CRAN

Date/Publication 2026-02-20 10:20:02 UTC

Contents

`dtm` *Document-Term Matrix (Sparse)*

Description

Sparse document-term matrix computed from `lemmas`. Each row is a text unit, each column is a lemma, and values are frequencies (proportions). Stored as a `dgCMatrix` from the Matrix package. Computed on first access.

Usage

```
dtm
```

Format

A sparse matrix of class `dgCMatrix` with text unit IDs as row names and lemma headwords as column names.

Examples

```
# Sparse document-term matrix
dim(dtm)

# Hierarchical clustering of text units
d <- dist(dtm[1:20, ])
plot(hclust(d))
```

lemmas *Lemma Frequency Table*

Description

Lemma frequency table computed from the lemmatized text. Tokenizes `texts$text_lemmatized` and counts word frequencies per text unit on first access (~5 seconds).

Usage

```
lemmas
```

Format

A data frame with the variables:

word Lemma (dictionary headword)

n Count of this lemma in this text unit

total Total lemma tokens in this text unit

freq Frequency (n/total)

id Text unit ID

collection Collection code

pitaka Pitaka name

Examples

```
# First access triggers computation (~5 seconds)
head(lemmas)

# Most frequent lemmas across the entire canon
totals <- tapply(lemmas$n, lemmas$word, sum)
head(sort(totals, decreasing = TRUE), 20)
```

search_lemma *Search for Lemma Occurrences*

Description

Finds all text units containing a specific lemma, sorted by frequency (most frequent first).

Usage

```
search_lemma(lemma)
```

Arguments

`lemma`	Character string of the lemma to search for.

Value

A data frame of occurrences with columns: word, n, total, freq, id, collection, pitaka. Returns an empty data frame if the lemma is not found.

Examples

```
# Find texts mentioning "nibbana"
nibbana <- search_lemma("nibbana")
head(nibbana)

# Find texts mentioning "dhamma"
dhamma <- search_lemma("dhamma")
head(dhamma[, c("id", "collection", "n", "freq")])
```

`texts` *Full Text of the Pali Canon (Critical Edition)*

Description

Surface-form and lemmatized text for every text unit in the Tipitaka. This is the only dataset shipped with the package; all other data is computed on demand from this text.

Usage

```
texts
```

Format

A data frame with 5,777 rows and 6 columns:

id Text unit ID (e.g., "dn1", "mn1", "sn1.1", "mahavagga")

collection Collection code (dn, mn, sn, an, kn, vinaya, abhidhamma)

pitaka Pitaka name (sutta, vinaya, abhidhamma)

title Pali title of the text

text Full surface-form Pali text

text_lemmatized Same text with each word replaced by its lemma headword

Source

Critical edition based on five-witness collation of PTS/GRETIL, SuttaCentral, VRI (Chattha Sangayana), Buddha Jayanti Tipitaka (BJT), and Thai Royal Edition. Lemmatization via the Digital Pali Dictionary.

Examples

```
# Number of text units per pitaka
table(texts$pitaka)

# Get text of the Brahmajala Sutta (DN 1)
dn1 <- texts[texts$id == "dn1", ]
cat(substr(dn1$text, 1, 200), "...\n")
```

`tipitaka.critical` *tipitaka.critical: Lemmatized Critical Edition of the Pali Canon*

Description

A lemmatized critical edition of the complete Pali Canon (Tipitaka) based on a five-witness collation with the Digital Pali Dictionary.

Details

This package ships the full text data (`texts`) and computes derived data on first access:

- `lemmas`: lemma frequency table
- `dtm`: sparse document-term matrix
- `search_lemma`: search for a lemma across all texts

For the original VRI edition and Pali text tools, see the companion package **tipitaka**.

Author(s)

Maintainer: Dan Zigmond <djz@shmonk.com>

See Also

Useful links:

- https://github.com/dangerzig/tipitaka.critical
- Report bugs at https://github.com/dangerzig/tipitaka.critical/issues

Index