Package 'taxaudit'

November 25, 2026

Type Package

Title Taxonomic Audit, Character Ontology, and Pairwise Diagnosability

Version 0.1.0

Description Automates the extraction of morphological traits and taxonomic literature auditing from unstructured botanical treatments. Implements controlled ontology mapping, regular expression-based quantitative dimension parsing, benchmark corpus management, and comparative baseline evaluation protocols for biodiversity informatics.

License MIT + file LICENSE

Encoding UTF-8

Imports furrr, future, ggplot2, httr2, rvest, utils, xml2

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

VignetteBuilder knitr

Depends R (>= 4.1.0)

Config/roxygen2/version 8.1.1

NeedsCompilation no

Author Aamna Naveed [aut, cre]

Maintainer Aamna Naveed <naamna588@gmail.com>

Repository CRAN

Date/Publication 2026-11-25 20:20:02 UTC

Contents

calculate_DI 2
extract_description 3
ta_audit_agreement 3
ta_baseline_extract 4
ta_compare_baseline 4
ta_contrast_table 5

ta_extract_characters 5
ta_extract_traits 6
ta_fetch_treatment 6
ta_generate_report 7
ta_get_ontology 7
ta_load_corpus 8
ta_pairwise_di 8
ta_parse_treatment 9
ta_plot_heatmap 9
ta_resolve_names 10
ta_run_audit 10
ta_run_batch_audit 11
ta_split_evaluation 11

Index **12**

`calculate_DI` *Calculate Pairwise Diagnosability Index*

Description

Calculate Pairwise Diagnosability Index

Usage

```
calculate_DI(taxon_a, taxon_b)
```

Arguments

`taxon_a`	Character or numeric vector of character states for taxon A
`taxon_b`	Character or numeric vector of character states for taxon B

Value

A numeric value representing the Diagnosability Index (0 = identical, 1 = fully disjoint)

extract_description *Extract Description for a Given Taxon*

Description

Extract Description for a Given Taxon

Usage

```
extract_description(taxon)
```

Arguments

taxon	Character string specifying taxon name

Value

Character string containing extracted taxonomic description

ta_audit_agreement *Calculate Agreement and Overlap Between Two Annotation Sets*

Description

Calculate Agreement and Overlap Between Two Annotation Sets

Usage

```
ta_audit_agreement(annotations_a, annotations_b)
```

Arguments

annotations_a	Data frame from Annotator A (must include trait, span_start, span_end).
annotations_b	Data frame from Annotator B (must include trait, span_start, span_end).

Value

A list containing matching statistics and agreement score.

ta_baseline_extract *Baseline Keyword Trait Extraction*

Description

A naive regex-based baseline extractor for comparison against ontology-mapped extraction.

Usage

```
ta_baseline_extract(text)
```

Arguments

text	Character string of a botanical treatment.

Value

A data frame of naive extracted keyword matches.

ta_compare_baseline *Compare Baseline vs Ontology-Mapped Extraction*

Description

Compares naive baseline extraction performance against the taxaudit pipeline.

Usage

```
ta_compare_baseline(corpus)
```

Arguments

corpus	Loaded corpus data frame.

Value

A data frame containing comparative extraction metrics.

ta_contrast_table *Generate a contrast table comparing character states between two taxon treatments*

Description

Generate a contrast table comparing character states between two taxon treatments

Usage

```
ta_contrast_table(text1, text2, name1 = "Species A", name2 = "Species B")
```

Arguments

text1	Character string for species 1 description.
text2	Character string for species 2 description.
name1	Character. Name of species 1 (default 'Species A').
name2	Character. Name of species 2 (default 'Species B').

Value

A data frame contrast table.

ta_extract_characters *Extract characters from taxon description text with expanded ontology*

Description

Extract characters from taxon description text with expanded ontology

Usage

```
ta_extract_characters(text)
```

Arguments

text	Character string of the taxon description.

Value

A list of character categories and extracted states.

ta_extract_traits *Extract Traits with Evidence Spans and Ontology Terms*

Description

Extract Traits with Evidence Spans and Ontology Terms

Usage

```
ta_extract_traits(text)
```

Arguments

text	Character string containing treatment text.

Value

A data frame containing extracted traits, character spans, and ontology metadata.

ta_fetch_treatment *Fetch botanical treatment text for a given taxon*

Description

Fetch botanical treatment text for a given taxon

Usage

```
ta_fetch_treatment(taxon_name)
```

Arguments

taxon_name	Character string name of the taxon

Value

Character string of treatment prose

ta_generate_report *Generate Automated Taxonomic Audit Report*

Description

Generates automated taxonomic ambiguity reports for floristic datasets.

Usage

```
ta_generate_report(flora_list, output_dir = tempdir())
```

Arguments

flora_list	Named list of taxon descriptions.
output_dir	Directory path to save output reports and data.

Value

Invisibly returns a list containing audit summary statistics and results.

ta_get_ontology *Define standard botanical character ontology for Flora audits*

Description

Retrieves the controlled vocabulary and categorical trait definitions used for botanical literature auditing.

Usage

```
ta_get_ontology()
```

Value

A named list where each element represents a botanical trait category (e.g., leaf shape, indumentum) containing a character vector of valid state terms.

Examples

```
ontology <- ta_get_ontology()
names(ontology)
head(ontology$leaf_shape)
```

ta_load_corpus *Load the Expert-Annotated Benchmark Corpus*

Description

Retrieves the gold-standard benchmark corpus containing plant treatments, morphological traits, and evidence spans from package internal files.

Usage

```
ta_load_corpus()
```

Value

A `data.frame` containing gold-standard plant treatments, traits, and evidence spans.

Examples

```
corpus <- ta_load_corpus()
head(corpus)
```

ta_pairwise_di *Calculate Pairwise Diagnosability Index (DI) between two species character profiles*

Description

Calculate Pairwise Diagnosability Index (DI) between two species character profiles

Usage

```
ta_pairwise_di(chars1, chars2)
```

Arguments

`chars1`	Named list of character states for species 1.
`chars2`	Named list of character states for species 2.

Value

A list containing the DI score and differentiating characters.

ta_parse_treatment *Parse a cached taxonomic treatment HTML file*

Description

Parse a cached taxonomic treatment HTML file

Usage

```
ta_parse_treatment(html_path)
```

Arguments

html_path	Character. Path to the cached HTML file.

Value

A data frame containing extracted treatment text and status.

ta_plot_heatmap *Plot Pairwise Diagnosability Index Heatmap*

Description

Plot Pairwise Diagnosability Index Heatmap

Usage

```
ta_plot_heatmap(batch_results)
```

Arguments

batch_results	Data frame resulting from batch pairwise audits.

Value

A ggplot object.

ta_resolve_names *Resolve and Standardize Taxonomic Names*

Description

Resolve and Standardize Taxonomic Names

Usage

```
ta_resolve_names(taxon_names, synonym_db = NULL)
```

Arguments

taxon_names	Character vector of taxon names to resolve.
synonym_db	Optional data frame containing 'Synonym' and 'Accepted' columns. If NULL, uses built-in default mappings.

Value

A data frame containing original names, resolved accepted names, and resolution status.

ta_run_audit *Run a complete botanical trait audit for a taxon*

Description

Run a complete botanical trait audit for a taxon

Usage

```
ta_run_audit(taxon)
```

Arguments

taxon	Character string name of the taxon or path to treatment file

Value

A structured list containing audit results and extracted traits

ta_run_batch_audit *Run taxonomic audit across multiple treatment URLs*

Description

Run taxonomic audit across multiple treatment URLs

Usage

```
ta_run_batch_audit(urls)
```

Arguments

urls	Character vector of treatment URLs.

Value

A combined data frame containing audit results for all successfully processed taxa.

ta_split_evaluation *Split Corpus Data for Held-Out Evaluation*

Description

Split Corpus Data for Held-Out Evaluation

Usage

```
ta_split_evaluation(
  corpus_df,
  holdout_type = c("taxon", "source"),
  target_holdout
)
```

Arguments

corpus_df	Data frame containing the annotated corpus (must include columns like taxon and source).
holdout_type	Character string: either "taxon" or "source".
target_holdout	Value of the taxon or source to hold out (e.g., "Cedrus deodara" or "Flora of Pakistan Vol 2").

Value

A list with training/evaluation data splits.

Index

calculate_DI, 2

extract_description, 3

ta_audit_agreement, 3
ta_baseline_extract, 4
ta_compare_baseline, 4
ta_contrast_table, 5
ta_extract_characters, 5
ta_extract_traits, 6
ta_fetch_treatment, 6
ta_generate_report, 7
ta_get_ontology, 7
ta_load_corpus, 8
ta_pairwise_di, 8
ta_parse_treatment, 9
ta_plot_heatmap, 9
ta_resolve_names, 10
ta_run_audit, 10
ta_run_batch_audit, 11
ta_split_evaluation, 11