Package ‘clinspacy’

March 20, 2021

Type Package

Title Clinical Natural Language Processing using ‘spaCy’, ‘scispaCy’, and ‘medspaCy’

Version 1.0.2


License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports reticulate (>= 1.16), data.table, assertthat, rappdirs, utils, magrittr

RoxygenNote 7.1.1

URL https://github.com/ML4LHS/clinspacy

BugReports https://github.com/ML4LHS/clinspacy/issues

Depends R (>= 2.10)

Suggests knitr, rmarkdown

NeedsCompilation no

Author Karandeep Singh [aut, cre], Benjamin Kompa [aut], Andrew Beam [aut], Allen Schmaltz [aut]

Maintainer Karandeep Singh <kdpsingh@umich.edu>

Repository CRAN

Date/Publication 2021-03-20 10:50:12 UTC
This function binds columns containing either the lemma of the entity or the UMLS concept unique identifier (CUI) with frequencies to a data frame. The resulting data frame can be used to train a machine learning model or for additional feature selection.

Usage

bind_clinspacy(
  clinspacy_output,
  df,
  cs_col = NULL,
  df_id = NULL,
  subset = "is_negated == FALSE"
)

Arguments

clinspacy_output
  A data.frame or file name containing the output from clinspacy.

df
  The data.frame to which you would like to bind the output of clinspacy.

cs_col
  Name of the column in the clinspacy_output that you would like to pivot. For example: "entity", "lemma", "cui", or "definition". Defaults to "lemma" if use_linker is set to FALSE and "cui" if use_linker is set to TRUE.

df_id
  The name of the id column in the data frame with which the clinspacy_id column in clinspacy_output will be joined. If you supplied a df_id in clinspacy, then you must also supply it here. If you did not supply it in clinspacy, then it will default to the row number (similar behavior to in clinspacy).
subset Logical criteria represented as a string by which the clinspacy_output will be subsetted prior to building the output data frame. Defaults to "is_negated == FALSE", which removes negated concepts prior to generating the output. Any column in clinspacy_output may be referenced here. To avoid any subsetting, set this to NULL.

Value
A data frame containing the original data frame as well as additional column names for each lemma or UMLS concept unique identifier found with values containing frequencies.

Examples
```r
## Not run:
mtsamples <- dataset_mtsamples()
mtsamples[1:5,] %>%
  clinspacy(df_col = 'description') %>%
  bind_clinspacy(mtsamples[1:5,])

## End(Not run)
```

bind_clinspacy_embeddings
This function binds columns containing entity or concept embeddings to a data frame. The entity embeddings are derived from the scispacy package, and the concept embeddings are derived from the dataset_cui2vec_embeddings dataset included with this package.

Description
The embeddings are derived from Andrew Beam’s cui2vec R package.

Usage
```r
bind_clinspacy_embeddings(
  clinspacy_output,
  df,
  type = "scispacy",
  df_id = NULL,
  subset = "is_negated == FALSE"
)
```

Arguments
- `clinspacy_output`: A data.frame or file name containing the output from clinspacy. In order for scispacy embeddings to be available to bind_clinspacy_embeddings, you must set return_scispacy_embeddings to TRUE when running clinspacy so that the embeddings are included within clinspacy_output.
bind_clinspacy_embeddings

- **df**: The data.frame to which you would like to bind the output of `clinspacy`.
- **type**: The type of embeddings to return. One of scispacy and cui2vec. Whereas cui2vec embeddings require the UMLS linker to be enabled, the scispacy embeddings do not. Defaults to scispacy.
- **df_id**: The name of the id column in the data frame with which the id column in `clinspacy_output` will be joined. If you supplied a df_id in `clinspacy`, then you must also supply it here. If you did not supply it in `clinspacy`, then it will default to the row number (similar behavior to in `clinspacy`).
- **subset**: Logical criteria represented as a string by which the `clinspacy_output` will be subsetted prior to building the output data frame. Defaults to "is_negated == FALSE", which removes negated concepts prior to generating the output. Any column in `clinspacy_output` may be referenced here. To avoid any subsetting, set this to NULL.

**Details**


- **License**: The cui2vec data is made available under a CC BY 4.0 license. The only change made to the original dataset is the renaming of columns.

**Value**

A data frame containing the original data frame as well as the concept embeddings. For scispacy embeddings, this returns 200 columns of embeddings. For cui2vec embeddings, this returns 500 columns of embeddings. The resulting data frame can be used to train a machine learning model.

**Examples**

```r
## Not run:
mtsamples <- dataset_mtsamples()
mtsamples[1:5,] %>%
  clinspacy(df_col = 'description', return_scispacy_embeddings = TRUE) %>%
  bind_clinspacy_embeddings(mtsamples[1:5,])

## End(Not run)
```
clinspacy

This is the primary function for processing both data frames and character vectors in the clinspacy package.

Description

This is the primary function for processing both data frames and character vectors in the clinspacy package.

Usage

clinspacy(
  x,
  df_col = NULL,
  df_id = NULL,
  threshold = 0.99,
  semantic_types = c(NA, "Acquired Abnormality", "Activity", "Age Group",
                    "Amino Acid Sequence", "Amino Acid, Peptide, or Protein", "Amphibian",
                    "Archaeon", "Bacterium", "Behavior", "Biologic Function", "Biologically Active Substance",
                    "Biomedical Occupation or Discipline", "Biomedical or Dental Material", "Bird",
                    "Body Location or Region", "Body Part, Organ, or Organ Component",
                    "Body Space or Junction", "Body Substance", "Bovine", "Carbohydrate Sequence",
                    "Cell", "Cell Component", "Cell Function", "Cell or Molecular Dysfunction",
                    "Chemical", "Chemical Viewed Functionally", "Chemical Viewed Structurally",
                    "Classification", "Clinical Attribute", "Clinical Drug", "Conceptual Entity",
                    "Congenital Abnormality", "Daily or Recreational Activity", "Diagnostic Procedure",
                    "Disease or Syndrome", "Drug Delivery Device", "Educational Activity",
                    "Element, Ion, or Isotope", "Embryonic Structure", "Entity",
                    "Experimental Model of Disease", "Family Group", "Finding", "Fish", "Food",
                    "Fully Formed Anatomical Structure", "Functional Concept", "Fungus",
                    "Gene or Genome", "Genetic Function", "Geographic Area",
                    "Governmental or Regulatory Activity", "Group", "Group Attribute",
                    "Hazardous or Poisonous Substance", "Health Care Activity",
                    "Health Care Related Organization", "Hormone", "Human",
                    "Human-caused Phenomenon or Process", "Idea or Concept", "Immunologic Factor",
                    "Indicator, Reagent, or Diagnostic Aid", "Individual Behavior",
                    "Injury or Poisoning", "Inorganic Chemical", "Intellectual Product",
                    "Laboratory or Test Result", "Laboratory Procedure", "Language", "Machine Activity",
                    "Mammal", "Manufactured Object", "Medical Device",
                    "Mental or Behavioral Dysfunction", "Mental Process",
                    "Molecular Biology Research Technique", "Molecular Function", "Molecular Sequence",
                    "Natural Phenomenon or Process", "Neoplastic Process",
                    "Nucleic Acid, Nucleoside, or Nucleotide", "Nucleotide Sequence",
                    "Occupation or Discipline", "Occupational Activity", "Organ or Tissue Function",
                    "Organic Chemical", "Organism", "Organism Attribute", "Organism Function"),
"Organization", "Pathologic Function", "Patient or Disabled Group",
"Pharmacologic Substance", "Phenomenon or Process", "Physical Object",
"Physiologic Function", "Plant", "Population Group",
"Professional or Occupational Group", "Professional Society", "Qualitative Concept",
"Quantitative Concept", "Receptor", "Regulation or Law", "Reptile",
"Research Activity", "Research Device", "Self-help or Relief Organization",
"Sign or Symptom", "Social Behavior", "Spatial Concept", "Substance",
"Temporal Concept", "Therapeutic or Preventive Procedure", "Tissue", "Vertebrate",
"Virus", "Vitamin"),
return_scispacy_embeddings = FALSE,
verbose = TRUE,
output_file = NULL,
overwrite = FALSE
)

Arguments

x Either a data.frame or a character vector
df_col If x is a data.frame then you must specify the name of the column containing
text as a string.
df_id If x is a data.frame then you may *optionally* specify an id column to help
match up each row of text in the original data frame with the resulting output.
If you do not specify an id, the resulting will contain the row number from the
original data.frame.
threshold Defaults to 0.99. The confidence threshold value used by clinspacy (can be
higher than the linker_threshold from clinspacy_init). Note that whereas
the linker_threshold can only be set once per session, this threshold can be up-
dated during the R session.
semantic_types Character vector containing any combination of the following: c(NA, "Acquired
Abnormality", "Activity", "Age Group", "Amino Acid Sequence", "Amino Acid,
Peptide, or Protein", "Amphibian", "Anatomical Abnormality", "Anatomical
Structure", "Animal", "Antibiotic", "Archaeon", "Bacterium", "Behavior", "Bi-
ologic Function", "Biologically Active Substance", "Biomedical Occupation or
Discipline", "Biomedical or Dental Material", "Bird", "Body Location or Re-
gion", "Body Part, Organ, or Organ Component", "Body Space or Junction",
"Body Substance", "Body System", "Carbohydrate Sequence", "Cell", "Cell
Component", "Cell Function", "Cell or Molecular Dysfunction", "Chemical",
"Chemical Viewed Functionally", "Chemical Viewed Structurally", "Classifica-
Abnormality", "Daily or Recreational Activity", "Diagnostic Procedure", "Dis-
case or Syndrome", "Drug Delivery Device", "Educational Activity", "Element,
Ion, or Isotope", "Embryonic Structure", "Entity", "Environmental Effect of
Humans", "Enzyme", "Eukaryote", "Event", "Experimental Model of Disease",
"Family Group", "Finding", "Fish", "Food", "Fully Formed Anatomical
Structure", "Functional Concept", "Fungus", "Gene or Genome", "Genetic Func-
tion", "Geographic Area", "Governmental or Regulatory Activity", "Group",
"Group Attribute", "Hazardous or Poisonous Substance", "Health Care Activity",
"Health Care Related Organization", "Hormone", "Human", "Human-caused

**return_scispacy_embeddings**

Defaults to FALSE. This is primarily intended for use by the `bind_clinspacy_embeddings` function to obtain scispacy embeddings. In order for scispacy embeddings to be available to `bind_clinspacy_embeddings`, you must set this to TRUE.

**verbose**

Defaults to TRUE.

**output_file**

Defaults to NULL. This is an optional argument that writes the output to a comma-separated value (CSV) file.

**overwrite**

Defaults to FALSE. If `output_file` already exists and `overwrite` is set to FALSE, then you will be prompted to confirm whether you would like to overwrite the file. If set to TRUE, then `output_file` will automatically be overwritten.

**Value**

If `output_file` is NULL (the default), then this function returns a data frame containing the UMLS concept unique identifiers (cui), entities, lemmatized entities, CyContext negation status (TRUE means negated, FALSE means *not* negated), other CyContext contexts, and section title from the clinical sectionizer. If `output_file` points to a file name, then the name of the created file will be returned.

**Examples**

```r
## Not run:
clinspacy('This patient has diabetes and CKD stage 3 but no HTN.')

clinspacy(c('This pt has CKD and HTN', 'Pt only has CKD but no HTN'))

data.frame(text = c('This pt has CKD and HTN', 'Diabetes is present'),
            stringsAsFactors = FALSE) %>%
  clinspacy(df_col = 'text')

if (!dir.exists(rappdirs::user_data_dir('clinspacy'))) {
```
clinspacy_init

**Initializes clinspacy.** This function is optional to run but gives you more control over the parameters used by scispacy at initiation. If you do not run this function, it will be run with default parameters the first time that any of the package functions are run.

### Description

Initializes clinspacy. This function is optional to run but gives you more control over the parameters used by scispacy at initiation. If you do not run this function, it will be run with default parameters the first time that any of the package functions are run.

### Usage

```r
clinspacy_init(
  miniconda = TRUE,
  use_linker = FALSE,
  linker_threshold = 0.99,
  ...
)
```

### Arguments

- `miniconda` Defaults to TRUE, which results in miniconda being installed (~400 MB) and configured with the "clinspacy" conda environment. If you want to override this behavior, set miniconda to FALSE and specify an alternative environment using use_python() or use_conda().

- `use_linker` Defaults to FALSE. To turn on the UMLS linker, set this to TRUE.

- `linker_threshold` Defaults to 0.99. This argument is only relevant if use_linker is set to TRUE. It refers to the confidence threshold value used by the scispacy UMLS entity linker. Note: This can be lower than the threshold from `clinspacy_init()`. The linker_threshold can only be set once per session.

### Value

No return value.
Description
This dataset contains definitions for the Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs). These come from Andrew Beam’s cui2vec R package.

Usage
dataset_cui2vec_definitions()

Format
A data frame with 3053795 rows and 3 variables:
cui A Unified Medical Language System (UMLS) Concept Unique Identifier (CUI)
semantic_type Semantic type of the CUI
definition Definition of the CUI

Details
License
This data is made available under a MIT license. The data is copyrighted in 2019 by Benjamin Kompa, Andrew Beam, and Allen Schmaltz. The only change made to the original dataset is the renaming of columns.

Value
Returns the cui2vec UMLS definitions as a data frame.

Source
https://github.com/beamandrew/cui2vec
dataset_cui2vec_embeddings

Cui2vec concept embeddings

Description

This dataset contains Unified Medical Language System (UMLS) concept embeddings from Andrew Beam’s cui2vec R package. There are 500 embeddings included for each concept.

Usage

dataset_cui2vec_embeddings()

Format

A data frame with 109053 rows and 501 variables:

- **cui** A Unified Medical Language System (UMLS) Concept Unique Identifier (CUI)
- **emb_001** Concept embedding vector #1
- **emb_002** Concept embedding vector #2
- ... and so on...
- **emb_500** Concept embedding vector #500

Details

This dataset is not viewable until it has been downloaded, which will occur the very first time you run clinspacy_init() after installing this package.

Citation


License

This data is made available under a [CC BY 4.0 license](https://figshare.com/s/00d69861786cd0156d81). The only change made to the original dataset is the renaming of columns.

Value

Returns the cui2vec UMLS embeddings as a data frame.

Source

[https://figshare.com/s/00d69861786cd0156d81](https://figshare.com/s/00d69861786cd0156d81)
**Description**

This dataset contains sample medical transcriptions for various medical specialties.

**Usage**

```r
dataset_mtsamples()
```

**Format**

A data frame with 4999 rows and 6 variables:

- **note_id** A unique identifier for each note
- **description** A description or chief concern
- **medical_specialty** Medical specialty of the note
- **sample_name** mtsamples.com note name
- **transcription** Transcription of note text
- **keywords** Keywords

**Details**

Acknowledgements

This data was scraped from https://mtsamples.com by Tara Boyle.

License This data is made available under a CC0: Public Domain license.

**Value**

Returns the mtsamples dataset as a data frame.

**Source**

https://www.kaggle.com/tboyle10/medicaltranscriptions/data
Index

bind_clinspacy, 2
bind_clinspacy_embeddings, 3, 3, 7

clinspacy, 2–4, 5
clinspacy_init, 6, 8, 8

dataset_cui2vec_definitions, 9
dataset_cui2vec_embeddings, 3, 10
dataset_mtsamples, 11