Package 'GermaParl'

September 23, 2020

Type Package

Title Download and Augment the Corpus of Plenary Protocols of the German Bundestag

Version 1.5.2

Date 2020-09-21

Depends R (>= 3.5.0)

Imports cwbtools (>= 0.3.0), tools, cli, zen4R

Suggests topicmodels, knitr, rmarkdown, testthat

LazyData yes

Description Data package to disseminate the 'GermaParl' corpus of parliamentary debates of the German Bundestag prepared in the 'PolMine Project'. The package includes a small subset of the corpus for demonstration and testing purposes. The package includes functionality to download the full corpus and supplementary data from the open science repository 'Zenodo'.

URL https://github.com/polmine/GermaParl

BugReports https://github.com/polmine/GermaParl/issues

License GPL-3

VignetteBuilder knitr

Encoding UTF-8

Collate 'GermaParl.R' 'download.R' 'lda.R'

RoxygenNote 7.1.1

NeedsCompilation no

Author Andreas Blaette [aut, cre], Christoph Leonhardt [ctb]

Maintainer Andreas Blaette <andreas.blaette@uni-due.de>

Repository CRAN

Date/Publication 2020-09-23 06:30:17 UTC

R topics documented:

`GermaParl-package` *GermaParl R Data Package.*

Description

GermaParl is a corpus of parliamentary debates in the German Bundestag. The package offers a convenient dissemination mechanism for the *GermaParl* corpus. The corpus has been linguistically annotated and indexed using the data format of the *Corpus Workbench* (CWB). To make full use if this data format, working with *GermaParl* in combination with the *polmineR* package is recommended.

Details

The GermaParl package initially only includes a subset of the GermaParl corpus which serves as a sample corpus ("GERMAPARLMINI"). To download the full corpus from the open science repository *Zenodo*, use the `germaparl_download_corpus` function.

The *GermaParl* R package and the *GermaParl* corpus are two different pieces of research data: The package offers a mechanism to ship, easily install and augment the data. The indexed corpus is the actual data. Package and corpus have different version numbers and should be quoted in combination in publications. We recommend to follow the instructions you see when calling `citation(package = "GermaParl")`. To ensure that the recommended citation fits the corpus you use, the citation for the corpus is available only when a version of *GermaParl* has been downloaded and installed.

Author(s)

Andreas Blaette <andreas.blaette@uni-due.de>

References

Blaette, Andreas (2018): "Using Data Packages to Ship Annotated Corpora of Parliamentary Protocols: The GermaParl R Package". ISBN 979-10-95546-02-3. Available online at http://lrec-conf.org/workshops/lrec2018/W2/pdf/15_W2.pdf.

See Also

Useful links:

- https://github.com/polmine/GermaParl
- Report bugs at https://github.com/polmine/GermaParl/issues

Examples

```
# This example uses the GERMAPARLSAMPLE corpus rather than the full GERMAPARL
# corpus in order to reduce the time required for testing the code. To apply
# everything on GERMAPARL rather than GERMAPARLSAMPLE, set variable 'samplemode'
# to FALSE, or simply omit argument 'sample'.

samplemode <- TRUE
corpus_id <- "GERMAPARLSAMPLE" # to get full corpus: corpus_id <- "GERMAPARL"

# This example assumes that the directories used by the CWB do not yet exist, so
# temporary directories are created.
cwb_dirs <- cwbtools::create_cwb_directories(prefix = tempdir(), ask = interactive())
registry_tmp <- cwb_dirs[["registry_dir"]]

# Download corpus from Zenodo
germaparl_download_corpus(
  registry_dir = registry_tmp,
  corpus_dir = cwb_dirs[["corpus_dir"]],
  verbose = FALSE,
  sample = samplemode
)

# Check availability of the corpus
germaparl_is_installed(sample = samplemode) # TRUE now
germaparl_get_version(sample = samplemode) # get version of indexed corpus
germaparl_get_doi(sample = samplemode) # get 'document object identifier' (DOI) of GERMAPARL corpus
```

---

germaparl_by_lp *Table with information on GermaParl by legislative period*

Description

A dataset with information on the corpus by legislative period is included in the package to be included in the data report of the package vignette.

Usage

```
germaparl_by_lp
```

Format

A `data.frame` with 5 rows and 6 variables with summary statistics on the GermaParl corpus on a year-by-year basis.

lp legislative period (`integer` value)

protocols total number of protocols included in the corpus for the respective legislative period (`integer` value)

first date of the first plenary protocol in the legislative period (`Date` class)

last date of the last plenary protocol in the legislative period (`Date` class)

size number of tokens in subcorpus for the respective legislative period (`integer` value)

unknown_total total number of words that cannot be lemmatized, resulting in #unknown# tag (`numeric` value)

unknown_share share of words that cannot be lemmatized, resulting in #unknown# tag (`numeric` value)

The table is based on v1.0.6 of the corpus. To prepare the table, the script available at data-raw/stats_for_vignette.R has been used.

Value

A `data.frame`.

`germaparl_by_year` *Table with information on GermaParl by year*

Description

A dataset with information on the corpus on a year-by-year basis is included in the package to be included in the data report of the package vignette.

Usage

```
germaparl_by_year
```

Format

A `data.frame` with 22 rows and 6 variables with summary statistics on the GermaParl corpus on a year-by-year basis.

year year reported on in the row (`integer` value)

protocols total number of protocols included in the corpus for the respective year (`integer` value)

txt number of protocols prepared based on plain text versions of the protocols (`integer` value)

pdf number of protocols prepared based on pdf versions of the protocols (`integer` value)

size number of tokens in subcorpus for the respective year (`integer` value)

unknown share of words that cannot be lemmatized, resulting in #unknown# tag (`numeric` value)

Details

The table is based on v1.0.6 of the corpus. The prepare the table, the script available at data-raw/stats_for_vignette.R has been used.

Value

A `data.frame`.

`germaparl_download_corpus` *Download full GermaParl corpus.*

Description

The GermaParl R package includes only a small subset of the GermaParl corpus (GERMAPARLMINI). The full corpus is deposited with Zenodo, an open science repository for research data. The `germaparl_download_corpus` function downloads a tarball with the indexed corpus from the Zenodo repository and moves the corpus data to the system corpus storage. If a corpus registry has not yet been created, an interactive dialogue will assist doing so. When calling the function, a stable internet connection is recommended. The size of the data to be downloaded is about 1 GB.

Usage

```
germaparl_download_corpus(
  doi = "https://doi.org/10.5281/zenodo.3742113",
  registry_dir = cwb_registry_dir(),
  corpus_dir = cwb_corpus_dir(registry_dir),
  verbose = interactive(),
  ask = interactive(),
  sample = FALSE
)
```

Arguments

`doi`	The DOI (Digital Object Identifier) of the GermaParl tarball at zenodo, presented as a hyperlink. Defaults to the latest version of GermaParl.
`registry_dir`	Path to the system registry directory. Defaults to value of `cwbtools::cwb_registry_dir()` to guess the registry directory. We recommend to state the registry directory explicitly.
`corpus_dir`	Directory where data directories of corpora are located. By default, the directory is guessed using `cwbtools::cwb_registry_dir`. We recommend to state the directory explicitly.
`verbose`	Whether to show messages, defaults to `TRUE`.
`ask`	A `logical` value, whether to ask for user input before replacing an existing corpus.
`sample`	A `logical` value, whether to download sample data (GERMAPARLSAMPLE) rather than full corpus (GERMAPARL) for testing purposes.

Details

After downloading and installing the tarball with the CWB indexed corpus, the registry file for the GERMAPARL corpus will be amended by the DOI and the corpus version. Afterwards, this information is available for a citation information fitting the corpus used that is provided when calling `citation(package = "GermaParl")`.

Value

Logical value. `TRUE` if the corpus has been installed successfully.

See Also

An example for using the `germaparl_download_corpus` function is part of the examples section of the overview documentation of the GermaParl package.

`germaparl_download_lda`

Use topicmodels prepared for GermaParl.

Description

A set of LDA topicmodels is part of the Zenodo release of GermaParl (k between 100 and 450). These topic models can be downloaded using `germaparl_download_lda` and loaded using `germaparl_load_lda`.

Usage

```
germaparl_download_lda(
  k = c(100L, 150L, 175L, 200L, 225L, 250L, 275L, 300L, 350L, 400L, 450L),
  doi = "10.5281/zenodo.3742113",
  data_dir,
  sample = FALSE,
  verbose = TRUE
)

germaparl_load_lda(
  k,
  registry_dir = cwbtools::cwb_registry_dir(),
  verbose = TRUE,
  sample = FALSE
)
```

Arguments

`k`	A `numeric` or `integer` vector, the number of topics of the topicmodel. Multiple values can be provided to download several topic models at once.
`doi`	The DOI of GermaParl at Zenodo.

data_dir	The data directory with the binary files of the GERMAPARL corpus. If missing, the directory will be guessed using the function cwb::cwb_corpus_dir
sample	A logical value, if TRUE, use GERMAPARLSAMPLE corpus rather than GERMAPARL.
verbose	logical
registry_dir	The registry directory where the registry file for GERMAPARL is located.

Details

The function germaparl_download_lda will download an rds-file that will be stored in the data directory of the GermaParl corpus.

germaparl_load_lda will load a topicmodel into memory. The function will return a LDA_Gibbs topicmodel, if the topicmodel for k is present; NULL if the topicmodel has not yet been downloaded.

Value

The function germaparl_download_lda will (invisibly) return TRUE if the operation has been succesful and FALSE if not.

Examples

```
# This example assumes that the directories used by the CWB do not yet exist, so
# temporary directories are created.
cwb_dirs <- cwbtools::create_cwb_directories(prefix = tempdir(), ask = FALSE)

samplemode <- TRUE
corpus_id <- "GERMAPARLSAMPLE" # for full corpus: corpus_id <- "GERMAPARL"

dir.create(file.path(cwb_dirs[["corpus_dir"]], tolower(corpus_id)))

# Download topic model
germaparl_download_lda(
  k = 30, # k = 250 recommended for full GERMAPARL corpus
  data_dir = file.path(cwb_dirs[["corpus_dir"]], tolower(corpus_id)),
  sample = samplemode
)
lda <- germaparl_load_lda(
  k = 30L, registry_dir = cwb_dirs[["registry_dir"]],
  sample = samplemode
)
lda_terms <- topicmodels::terms(lda, 10)
```

---

germaparl_get_doi *Get DOI of corpus*

Description

Auxiliary function that extracts the DOI (Document Object Identifier) from the registry file of the GERMAPARL corpus.

Usage

```
germaparl_get_doi(registry_dir = Sys.getenv("CORPUS_REGISTRY"), sample = FALSE)
```

Arguments

registry_dir	Path to the registry directory.
sample	A logical value, if FALSE (default), use GERMAPARL, if TRUE, use GERMAPARLSAMPLE.

Value

If the DOI is declared in the registry file, a length-one character vector with it is returned. If the corpus has not yet been installed, NULL is returned and a warning will be issued.

See Also

See the examples section of the overview documentation of the GermaParl package for an example.

---

germaparl_get_version *Get GERMAPARL version*

Description

germaparl_get_version is an auxiliary function that extracts the version of the GERMAPARL corpus from the registry.

Usage

```
germaparl_get_version(
  registry_dir = Sys.getenv("CORPUS_REGISTRY"),
  sample = FALSE
)
```

Arguments

registry_dir	Path to the registry directory.
sample	If TRUE, work with GERMAPARLSAMPLE corpus, if FALSE (default), use GERMAPARL corpus.

Value

The return value is the version of the corpus (class numeric_version). If the corpus has not yet been installed, NULL is returned, and a warning message is issued.

See Also

See the examples section of the overview documentation of the GermaParl package for an example.

germaparl_is_installed *Get installation status of GERMAPARL*

Description

Auxiliary function to detect whether GERMAPARL is installed or not.

Usage

```
germaparl_is_installed(
  registry_dir = Sys.getenv("CORPUS_REGISTRY"),
  sample = FALSE
)
```

Arguments

`registry_dir`	Path to the registry directory.
`sample`	A `logical` value. If `FALSE` (default), the GERMAPARL corpus will be used, if `TRUE`, the GERMAPARLSAMPLE corpus will be used.

Value

`TRUE` if the corpus has been installed, and `FALSE` if not.

See Also

See the examples section of the overview documentation of the GermaParl package for an example.

germaparl_lda_tuning *LDA Tuning Results*

Description

The R package ldatuning has been used to get guidance on the optimal number of topics when fitting an LDA topic model on the GermaParl corpus. Using around 250 topics is a good choice. The data object `germaparl_lda_tuning` reports the different metrics of the ldatuning package.

Usage

```
germaparl_lda_tuning
```

Format

An object of class `data.frame` with 10 rows and 5 columns.

Index