Package ‘textdata’

June 12, 2019

Title Download and Load Various Text Datasets

Version 0.1.0

Description Provides a framework to download, parse, and store text datasets on the disk and load them when needed. Includes various sentiment lexicons and labeled text data sets for classification and analysis.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports fs, readr, tibble, rappdirs

RoxygenNote 6.1.1

Collate 'dataset_sentence_polarity.R' 'lexicon_bing.R' 'lexicon_loughran.R' 'lexicon_afinn.R' 'download_functions.R' 'info.R' 'load_dataset.R' 'printer.R' 'process_functions.R'

Suggests knitr, rmarkdown, testthat (>= 2.1.0)

VignetteBuilder knitr

URL https://github.com/EmilHvitfeldt/textdata

BugReports https://github.com/EmilHvitfeldt/textdata/issues

NeedsCompilation no

Author Emil Hvitfeldt [aut, cre] (<https://orcid.org/0000-0002-0679-1945>), Julia Silge [ctb] (<https://orcid.org/0000-0002-3671-836X>)

Maintainer Emil Hvitfeldt <emilhhvitfeldt@gmail.com>

Repository CRAN

Date/Publication 2019-06-12 12:20:03 UTC

R topics documented:

`catalogue` *Catalogue of all available data sources*

Description

Catalogue of all available data sources

Usage

```
catalogue
```

Format

An object of class `data.frame` with 4 rows and 6 columns.

`dataset_sentence_polarity` *v1.0 sentence polarity dataset*

Description

5331 positive and 5331 negative processed sentences / snippets. Introduced in Pang/Lee ACL 2005. Released July 2005.

Usage

```
dataset_sentence_polarity(dir = NULL, delete = FALSE,
  return_path = FALSE)
```

Arguments

`dir`	Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.
`delete`	Logical, set `TRUE` to delete dataset.
`return_path`	Logical, set `TRUE` to return the path of the dataset.

Details

Citation info:

This data was first used in Bo Pang and Lillian Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.”, Proceedings of the ACL, 2005.

InProceedings{pang05,
author = {Bo Pang and Lillian Lee},
title = {Seeing stars: Exploiting class relationships for sentiment
categorization with respect to rating scales},
booktitle = {Proceedings of the ACL},
year = 2005
}

Value

A tibble with 10,662 rows and 2 variables:

text Sentences or snippets

sentiment Indicator for sentiment, "neg" for negative and "pos" for positive

Source

https://www.cs.cornell.edu/people/pabo/movie-review-data/

Examples

```
dataset_sentence_polarity()

# Custom directory
dataset_sentence_polarity(dir = "data/")

# Deleting dataset
dataset_sentence_polarity(delete = TRUE)

# Returning filepath of data
dataset_sentence_polarity(return_path = TRUE)
```

`lexicon_afinn` *AFINN-111 dataset*

Description

AFINN is a lexicon of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011.

Usage

```
lexicon_afinn(dir = NULL, delete = FALSE, return_path = FALSE)
```

Arguments

dir	Character, path to directory where data will be stored. If NULL, user_cache_dir will be used to determine path.
delete	Logical, set TRUE to delete dataset.
return_path	Logical, set TRUE to return the path of the dataset.

Details

This dataset is the newest version with 2477 words and phrases.

Citation info:

This dataset was published in Finn Ärup Nielsen (2011), "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages (2011) 93-98.

article{nielsen11,
author = {Finn Äruprup Nielsen},
title = {A new ANEW: Evaluation of a word list for sentiment analysis in microblogs},
journal = {CoRR},
volume = {abs/1103.2903},
year = {2011},
url = {http://arxiv.org/abs/1103.2903},
archivePrefix = {arXiv},
eprint = {1103.2903},
biburl = {https://dblp.org/rec/bib/journals/corr/abs-1103-2903},
bibsource = {dblp computer science bibliography, https://dblp.org}
}

Value

A tibble with 2,477 rows and 2 variables:

word An English word

sentiment Indicator for sentiment: integer between -5 and +5

Examples

```
lexicon_afinn()

# Custom directory
lexicon_afinn(dir = "data/")

# Deleting dataset
lexicon_afinn(delete = TRUE)
```

```
# Returning filepath of data
lexicon_afinn(return_path = TRUE)
```

`lexicon_bing` *Bing sentiment lexicon*

Description

General purpose English sentiment lexicon that categorizes words in a binary fashion, either positive or negative

Usage

```
lexicon_bing(dir = NULL, delete = FALSE, return_path = FALSE)
```

Arguments

`dir`	Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.
`delete`	Logical, set `TRUE` to delete dataset.
`return_path`	Logical, set `TRUE` to return the path of the dataset.

Details

Citation info:

This dataset was first published in Minqing Hu and Bing Liu, "Mining and summarizing customer reviews.", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), 2004.

inproceedings{Hu04,
author = {Hu, Minqing and Liu, Bing},
title = {Mining and Summarizing Customer Reviews},
booktitle = {Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining},
series = {KDD '04},
year = {2004},
isbn = {1-58113-888-1},
location = {Seattle, WA, USA},
pages = {168–177},
numpages = {10},
url = {http://doi.acm.org/10.1145/1014052.1014073},
doi = {10.1145/1014052.1014073},
acmid = {1014073},
publisher = {ACM},
address = {New York, NY, USA},
keywords = {reviews, sentiment classification, summarization, text mining},
}

Value

A tibble with 6,787 rows and 2 variables:

word An English word

sentiment Indicator for sentiment: "negative" or "positive"

Source

https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Examples

```
lexicon_bing()

# Custom directory
lexicon_bing(dir = "data/")

# Deleting dataset
lexicon_bing(delete = TRUE)

# Returning filepath of data
lexicon_bing(return_path = TRUE)
```

---

lexicon_loughran *Loughran-McDonald sentiment lexicon*

Description

English sentiment lexicon created for use with financial documents. This lexicon labels words with six possible sentiments important in financial contexts: "negative", "positive", "litigious", "uncertainty", "constraining", or "superfluous".

Usage

```
lexicon_loughran(dir = NULL, delete = FALSE, return_path = FALSE)
```

Arguments

`dir`	Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.
`delete`	Logical, set `TRUE` to delete dataset.
`return_path`	Logical, set `TRUE` to return the path of the dataset.

Details

Citation info:

This dataset was published in Loughran, T. and McDonald, B. (2011), “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” The Journal of Finance, 66: 35-65.

article{loughran11,
author = {Loughran, Tim and McDonald, Bill},
title = {When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks},
journal = {The Journal of Finance},
volume = {66},
number = {1},
pages = {35-65},
doi = {10.1111/j.1540-6261.2010.01625.x},
url = {https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x},
eprint = {https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2010.01625.x},
year = {2011}
}

Value

A tibble with 4,150 rows and 2 variables:

word An English word

sentiment Indicator for sentiment: "negative", "positive", "litigious", "uncertainty", "constraining", or "superfluous"

Source

https://sraf.nd.edu/textual-analysis/resources/

Examples

```
lexicon_loughran()

# Custom directory
lexicon_loughran(dir = "data/")

# Deleting dataset
lexicon_loughran(delete = TRUE)

# Returning filepath of data
lexicon_loughran(return_path = TRUE)
```

Index