Package ‘textdata’

May 2, 2022

Title Download and Load Various Text Datasets

Version 0.4.2

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

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

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

Imports fs, rappdirs, readr, tibble

Suggests covr, knitr, rmarkdown, testthat (>= 2.1.0)

VignetteBuilder knitr

Encoding UTF-8

RoxygenNote 7.1.2.9000

Collate 'cache_info.R' 'dataset_ag_news.R' 'dataset_dbpedia.R' 'dataset_imdb.R' 'dataset_sentence_polarity.R' 'dataset_trec.R' 'embedding_glove.R' 'lexicon_nrc_vad.R' 'lexicon_nrc_eil.R' 'lexicon_nrc.R' 'lexicon_bing.R' 'lexicon_loughran.R' 'lexicon_afinn.R' 'download_functions.R' 'info.R' 'load_dataset.R' 'printer.R' 'process_functions.R' 'textdata-package.R'

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 2022-05-02 20:30:02 UTC

R topics documented:

`cache_info` *List folders and their sizes in cache*

Description

This function will return a tibble with the name and sizes of all folder in specified directory. Will default to textdata's default cache.

Usage

```
cache_info(dir = NULL)
```

Arguments

`dir`	Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.

Value

A tibble with 2 variables:

name Name of the folder

size Size of the folder

Examples

```
## Not run:
cache_info()

## End(Not run)
```

`catalogue` *Catalogue of all available data sources*

Description

Catalogue of all available data sources

Usage

```
catalogue
```

Format

An object of class `data.frame` with 15 rows and 8 columns.

`dataset_ag_news` *AG's News Topic Classification Dataset*

Description

The AG's news topic classification dataset is constructed by choosing 4 largest classes from the original corpus. Each class contains 30,000 training samples and 1,900 testing samples. The total number of training samples is 120,000 and testing 7,600. Version 3, Updated 09/09/2015

Usage

```
dataset_ag_news(
  dir = NULL,
  split = c("train", "test"),
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = FALSE
)
```

Arguments

`dir`	Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.
`split`	Character. Return training ("train") data or testing ("test") data. Defaults to "train".
`delete`	Logical, set `TRUE` to delete dataset.
`return_path`	Logical, set `TRUE` to return the path of the dataset.
`clean`	Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.

`manual_download`
Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

Details

The classes in this dataset are

- World
- Sports
- Business
- Sci/Tech

Value

A tibble with 120,000 or 30,000 rows for "train" and "test" respectively and 3 variables:

class Character, denoting new class

title Character, title of article

description Character, description of article

Source

http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

https://github.com/srhrshr/torchDatasets/raw/master/dbpedia_csv.tar.gz

See Also

Other topic: `dataset_dbpedia()`, `dataset_trec()`

Examples

```
## Not run:
dataset_ag_news()

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

# Deleting dataset
dataset_ag_news(delete = TRUE)

# Returning filepath of data
dataset_ag_news(return_path = TRUE)

# Access both training and testing dataset
train <- dataset_ag_news(split = "train")
test <- dataset_ag_news(split = "test")

## End(Not run)
```

`dataset_dbpedia` *DBpedia Ontology Dataset*

Description

DBpedia ontology dataset classification dataset. It contains 560,000 training samples and 70,000 testing samples for each of 14 nonoverlapping classes from DBpedia.

Usage

```
dataset_dbpedia(
  dir = NULL,
  split = c("train", "test"),
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = FALSE
)
```

Arguments

`dir` Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.

`split` Character. Return training ("train") data or testing ("test") data. Defaults to "train".

`delete` Logical, set `TRUE` to delete dataset.

`return_path` Logical, set `TRUE` to return the path of the dataset.

`clean` Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.

`manual_download`
Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

Details

The classes are

- Company
- EducationalInstitution
- Artist
- Athlete
- OfficeHolder
- MeanOfTransportation
- Building

- NaturalPlace
- Village
- Animal
- Plant
- Album
- Film
- WrittenWork

Value

A tibble with 560,000 or 70,000 rows for "train" and "test" respectively and 3 variables:

class Character, denoting the class class

title Character, title of article

description Character, description of article

Source

https://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf

https://www.dbpedia.org/

https://github.com/srhrshr/torchDatasets/raw/master/dbpedia_csv.tar.gz

See Also

Other topic: dataset_ag_news(), dataset_trec()

Examples

```
## Not run:
dataset_dbpedia()

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

# Deleting dataset
dataset_dbpedia(delete = TRUE)

# Returning filepath of data
dataset_dbpedia(return_path = TRUE)

# Access both training and testing dataset
train <- dataset_dbpedia(split = "train")
test <- dataset_dbpedia(split = "test")

## End(Not run)
```

dataset_imdb *IMDB Large Movie Review Dataset*

Description

The core dataset contains 50,000 reviews split evenly into 25k train and 25k test sets. The overall distribution of labels is balanced (25k pos and 25k neg).

Usage

```
dataset_imdb(
  dir = NULL,
  split = c("train", "test"),
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = FALSE
)
```

Arguments

`dir`	Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.
`split`	Character. Return training ("train") data or testing ("test") data. Defaults to "train".
`delete`	Logical, set `TRUE` to delete dataset.
`return_path`	Logical, set `TRUE` to return the path of the dataset.
`clean`	Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.
`manual_download`	Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

Details

In the entire collection, no more than 30 reviews are allowed for any given movie because reviews for the same movie tend to have correlated ratings. Further, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing movie-unique terms and their associated with observed labels. In the labeled train/test sets, a negative review has a score <= 4 out of 10, and a positive review has a score >= 7 out of 10. Thus reviews with more neutral ratings are not included in the train/test sets. In the unsupervised set, reviews of any rating are included and there are an even number of reviews > 5 and <= 5.

When using this dataset, please cite the ACL 2011 paper

InProceedings{maas-EtAl:2011:ACL-HLT2011,
author = {Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng,

Andrew Y. and Potts, Christopher},
title = {Learning Word Vectors for Sentiment Analysis},
booktitle = {Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies},
month = {June},
year = {2011},
address = {Portland, Oregon, USA},
publisher = {Association for Computational Linguistics},
pages = {142–150},
url = {http://www.aclweb.org/anthology/P11-1015} }

Value

A tibble with 25,000 rows and 2 variables:

Sentiment Character, denoting the sentiment

text Character, text of the review

Source

http://ai.stanford.edu/~amaas/data/sentiment/

Examples

```
## Not run:
dataset_imdb()

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

# Deleting dataset
dataset_imdb(delete = TRUE)

# Returning filepath of data
dataset_imdb(return_path = TRUE)

# Access both training and testing dataset
train <- dataset_imdb(split = "train")
test <- dataset_imdb(split = "test")

## End(Not run)
```

---

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,
  clean = FALSE,
  manual_download = 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.

`clean` Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.

`manual_download`
Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

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

```
## Not run:
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)

## End(Not run)
```

dataset_trec *TREC dataset*

Description

The TREC dataset is dataset for question classification consisting of open-domain, fact-based questions divided into broad semantic categories. It has both a six-class (TREC-6) and a fifty-class (TREC-50) version. Both have 5,452 training examples and 500 test examples, but TREC-50 has finer-grained labels. Models are evaluated based on accuracy.

Usage

```
dataset_trec(
  dir = NULL,
  split = c("train", "test"),
  version = c("6", "50"),
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = FALSE
)
```

Arguments

`dir` Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.

`split` Character. Return training ("train") data or testing ("test") data. Defaults to "train".

`version` Character. Version 6("6") or version 50("50"). Defaults to "6".

`delete` Logical, set `TRUE` to delete dataset.

`return_path` Logical, set `TRUE` to return the path of the dataset.

`clean` Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.

`manual_download`
Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

Details

The classes in TREC-6 are

- ABBR - Abbreviation
- DESC - Description and abstract concepts
- ENTY - Entities
- HUM - Human beings
- LOC - Locations
- NYM - Numeric values

the classes in TREC-50 can be found here `https://cogcomp.seas.upenn.edu/Data/QA/QC/definition.html`.

Value

A tibble with 5,452 or 500 rows for "train" and "test" respectively and 2 variables:

class Character, denoting the class

text Character, question text

Source

`https://cogcomp.seas.upenn.edu/Data/QA/QC/`

`https://trec.nist.gov/data/qa.html`

See Also

Other topic: `dataset_ag_news()`, `dataset_dbpedia()`

Examples

```
## Not run:
dataset_trec()

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

# Deleting dataset
```

```
dataset_trec(delete = TRUE)

# Returning filepath of data
dataset_trec(return_path = TRUE)

# Access both training and testing dataset
train_6 <- dataset_trec(split = "train")
test_6 <- dataset_trec(split = "test")

train_50 <- dataset_trec(split = "train", version = "50")
test_50 <- dataset_trec(split = "test", version = "50")

## End(Not run)
```

embedding_glove *Global Vectors for Word Representation*

Description

The GloVe pre-trained word vectors provide word embeddings created using varying numbers of tokens.

Usage

```
embedding_glove6b(
  dir = NULL,
  dimensions = c(50, 100, 200, 300),
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = FALSE
)

embedding_glove27b(
  dir = NULL,
  dimensions = c(25, 50, 100, 200),
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = FALSE
)

embedding_glove42b(
  dir = NULL,
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = FALSE
)

embedding_glove840b(
  dir = NULL,
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = FALSE
)
```

Arguments

`dir` Character, path to directory where data will be stored. If `NULL`, user_cache_dir will be used to determine path.

`dimensions` A number indicating the number of vectors to include. One of 50, 100, 200, or 300 for glove6b, or one of 25, 50, 100, or 200 for glove27b.

`delete` Logical, set `TRUE` to delete dataset.

`return_path` Logical, set `TRUE` to return the path of the dataset.

`clean` Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.

`manual_download` Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

Details

Citation info:

InProceedings{pennington2014glove,
author = {Jeffrey Pennington and Richard Socher and Christopher D.
Manning},
title = {GloVe: Global Vectors for Word Representation},
booktitle = {Empirical Methods in Natural Language Processing (EMNLP)},
year = 2014
pages = {1532-1543}
url = {http://www.aclweb.org/anthology/D14-1162}
}

Value

A tibble with 400k, 1.9m, 2.2m, or 1.2m rows (one row for each unique token in the vocabulary) and the following variables:

token An individual token (usually a word)

d1, d2, etc The embeddings for that token.

Source

`https://nlp.stanford.edu/projects/glove/`

References

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

Examples

```
## Not run:
embedding_glove6b(dimensions = 50)

# Custom directory
embedding_glove42b(dir = "data/")

# Deleting dataset
embedding_glove6b(delete = TRUE, dimensions = 300)

# Returning filepath of data
embedding_glove840b(return_path = TRUE)

## End(Not run)
```

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,
  clean = FALSE,
  manual_download = 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.
clean	Logical, set TRUE to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.
manual_download	Logical, set TRUE if you have manually downloaded the file and placed it in the folder designated by running this function with return_path = TRUE.

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

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

See Also

Other lexicon: `lexicon_bing()`, `lexicon_loughran()`, `lexicon_nrc_eil()`, `lexicon_nrc_vad()`, `lexicon_nrc()`

Examples

```
## Not run:
lexicon_afinn()

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

# Deleting dataset
lexicon_afinn(delete = TRUE)

# Returning filepath of data
lexicon_afinn(return_path = TRUE)

## End(Not run)
```

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,
  clean = FALSE,
  manual_download = 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.
`clean`	Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.
`manual_download`	Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

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`

See Also

Other lexicon: `lexicon_afinn()`, `lexicon_loughran()`, `lexicon_nrc_eil()`, `lexicon_nrc_vad()`, `lexicon_nrc()`

Examples

```
## Not run:
lexicon_bing()

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

# Deleting dataset
lexicon_bing(delete = TRUE)

# Returning filepath of data
lexicon_bing(return_path = TRUE)

## End(Not run)
```

`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,
  clean = FALSE,
  manual_download = 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.
`clean`	Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.
`manual_download`	Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

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/loughranmcdonald-master-dictionary/

See Also

Other lexicon: lexicon_afinn(), lexicon_bing(), lexicon_nrc_eil(), lexicon_nrc_vad(), lexicon_nrc()

Examples

```
## Not run:
lexicon_loughran()

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

# Deleting dataset
lexicon_loughran(delete = TRUE)

# Returning filepath of data
lexicon_loughran(return_path = TRUE)

## End(Not run)
```

---

lexicon_nrc *NRC word-emotion association lexicon*

Description

General purpose English sentiment/emotion lexicon. This lexicon labels words with six possible sentiments or emotions: "negative", "positive", "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", or "trust". The annotations were manually done through Amazon's Mechanical Turk.

Usage

```
lexicon_nrc(
  dir = NULL,
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = 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.
`clean`	Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.
`manual_download`	Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

Details

License required for commercial use. Please contact Saif M. Mohammad (saif.mohammad@nrc-cnrc.gc.ca).

Citation info:

This dataset was published in Saif Mohammad and Peter Turney. (2013), “Crowdsourcing a Word-Emotion Association Lexicon.” Computational Intelligence, 29(3): 436-465.

article{mohammad13,
author = {Mohammad, Saif M. and Turney, Peter D.},
title = {CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON},
journal = {Computational Intelligence},
volume = {29},
number = {3},
pages = {436-465},
doi = {10.1111/j.1467-8640.2012.00460.x},
url = {https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00460.x},
eprint = {https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00460.x},
year = {2013}
}

Value

A tibble with 13,901 rows and 2 variables:

word An English word

sentiment Indicator for sentiment or emotion: "negative", "positive", "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", or "trust"

Source

`http://saifmohammad.com/WebPages/lexicons.html`

See Also

Other lexicon: `lexicon_afinn()`, `lexicon_bing()`, `lexicon_loughran()`, `lexicon_nrc_eil()`, `lexicon_nrc_vad()`

Examples

```
## Not run:
lexicon_nrc()

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

# Deleting dataset
lexicon_nrc(delete = TRUE)

# Returning filepath of data
lexicon_nrc(return_path = TRUE)

## End(Not run)
```

lexicon_nrc_eil *NRC Emotion Intensity Lexicon (aka Affect Intensity Lexicon) v0.5*

Description

General purpose English sentiment/emotion lexicon. The NRC Affect Intensity Lexicon is a list of English words and their associations with four basic emotions (anger, fear, sadness, joy).

Usage

```
lexicon_nrc_eil(
  dir = NULL,
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = 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.
clean	Logical, set TRUE to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.
manual_download	Logical, set TRUE if you have manually downloaded the file and placed it in the folder designated by running this function with return_path = TRUE.

Details

For a given word and emotion X, the scores range from 0 to 1. A score of 1 means that the word conveys the highest amount of emotion X. A score of 0 means that the word conveys the lowest amount of emotion X.

License required for commercial use. Please contact Saif M. Mohammad (saif.mohammad@nrc-cnrc.gc.ca).

Citation info:

Details of the lexicon are in this paper. Word Affect Intensities. Saif M. Mohammad. In Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018), May 2018, Miyazaki, Japan.

inproceedings{LREC18-AIL,
author = {Mohammad, Saif M.},
title = {Word Affect Intensities},
booktitle = {Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)},
year = {2018},
address={Miyazaki, Japan}
}

Value

A tibble with 5.814 rows and 3 variables:

term An English word

score Value between 0 and 1

AffectDimension Indicator for sentiment or emotion: ("anger", "fear", "sadness", "joy")

Source

https://saifmohammad.com/WebPages/AffectIntensity.htm

See Also

Other lexicon: lexicon_afinn(), lexicon_bing(), lexicon_loughran(), lexicon_nrc_vad(), lexicon_nrc()

Examples

```
## Not run:
lexicon_nrc_eil()

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

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

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

## End(Not run)
```

lexicon_nrc_vad *The NRC Valence, Arousal, and Dominance Lexicon*

Description

The NRC Valence, Arousal, and Dominance (VAD) Lexicon includes a list of more than 20,000 English words and their valence, arousal, and dominance scores. For a given word and a dimension (V/A/D), the scores range from 0 (lowest V/A/D) to 1 (highest V/A/D). The lexicon with its fine-grained real- valued scores was created by manual annotation using best–worst scaling. The lexicon is markedly larger than any of the existing VAD lexicons. We also show that the ratings obtained are substantially more reliable than those in existing lexicons.

Usage

```
lexicon_nrc_vad(
  dir = NULL,
  delete = FALSE,
  return_path = FALSE,
  clean = FALSE,
  manual_download = 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.

`clean` Logical, set `TRUE` to remove intermediate files. This can greatly reduce the size. Defaults to FALSE.

`manual_download` Logical, set `TRUE` if you have manually downloaded the file and placed it in the folder designated by running this function with `return_path = TRUE`.

Details

License required for commercial use. Please contact Saif M. Mohammad (saif.mohammad@nrc-cnrc.gc.ca).

Citation info:

Details of the NRC VAD Lexicon are available in this paper:

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Saif M. Mohammad. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, July 2018.

inproceedings{vad-acl2018,
title={Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words},
author={Mohammad, Saif M.},
booktitle={Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)},
year={2018},
address={Melbourne, Australia}
}

Value

A tibble with 20.007 rows and 4 variables:

word An English word

Valence valence score of the word

Arousal arousal score of the word

Dominance dominance score of the word

Source

https://saifmohammad.com/WebPages/nrc-vad.html

See Also

Other lexicon: lexicon_afinn(), lexicon_bing(), lexicon_loughran(), lexicon_nrc_eil(), lexicon_nrc()

Examples

```
## Not run:
lexicon_nrc_vad()

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

# Deleting dataset
lexicon_nrc_vad(delete = TRUE)

# Returning filepath of data
lexicon_nrc_vad(return_path = TRUE)

## End(Not run)
```

Index