Package ‘lexicon’

March 21, 2019

Title Lexicons for Text Analysis

Version 1.2.1

Maintainer Tyler Rinker <tyler.rinker@gmail.com>

Description A collection of lexical hash tables, dictionaries, and word lists.

Depends R (>= 3.2.2)

Imports data.table, syuzhet (>= 1.0.1)

License GPL-3

LazyData TRUE

Encoding UTF-8

RoxygenNote 6.1.1

BugReports https://github.com/trinker/lexicon/issues?state=open

URL https://github.com/trinker/lexicon

Collate 'available_data.R' 'cliches.R' 'common_names.R'
'constraining_loughran_mcdonald.R' 'freq_first_names.R'
'freq_last_names.R' 'function_words.R' 'grady_augmented.R'
'hash_emoticons.R' 'hash_grady_pos.R' 'hash_internet_slang.R'
'hash_lemmas.R' 'hash_nrc_emotion.R' 'hash_sentiment_emojis.R'
'hash_sentiment_huliu.R' 'utils.R' 'hash_sentiment_jockers.R'
'hash_sentiment_jockers_rinker.R'
'hash_sentiment_loughran_mcdonald.R' 'hash_sentiment_nrc.R'
'hash_sentiment_senticnet.R' 'hash_sentiment_sentiword.R'
'hash_sentiment_slangsd.R' 'hash_sentiment_socal_google.R'
'hash_valence_shifters.R' 'key_contractions.R'
'key_corporate_social_responsibility.R' 'key_grade.R'
'key_ratings.R' 'key_regressive_imagery.R' 'lexicon-package.R'
'modal_loughran_mcdonald.R' 'nrc_emotions.R'
'pos_action_verb.R' 'pos_df_irregular_nouns.R'
'pos_df_pronouns.R' 'pos_interjections.R' 'pos_preposition.R'
'profanity_alvarez.R' 'profanity_arr_bad.R'
'profanity_banned.R' 'profanity_racist.R'
'profanity_zac_anger.R' 'sw_dolch.R' 'sw_fry_100.R'
'sw_fry_1000.R' 'sw_fry_200.R' 'sw_fry_25.R' 'sw_jockers.R'

'sw_loughran_mcdonald.R' 'sw_lucene.R' 'sw_mallet.R'
'sw_python.R'

NeedsCompilation no

Author Tyler Rinker [aut, cre, cph],
University of Notre Dame [dtc, cph],
Department of Knowledge Technologies [dtc, cph],
Unicode, Inc. [dtc, cph],
John Higgins [dtc, cph],
Grady Ward [dtc],
Heiko Possel [dtc],
Michal Boleslav Mechura [dtc, cph],
Bing Liu [dtc],
Minqing Hu [dtc],
Saif M. Mohammad [dtc],
Peter Turney [dtc],
Erik Cambria [dtc],
Soujanya Poria [dtc],
Rajiv Bajpai [dtc],
Bjoern Schuller [dtc],
SentiWordNet [dtc, cph],
Liang Wu [dtc, cph],
Fred Morstatter [dtc, cph],
Huan Liu [dtc, cph],
Grammar Revolution [dtc, cph],
Vidar Holen [dtc, cph],
Alejandro U. Alvarez [dtc, cph],
Stackoverflow User user2592414 [dtc, cph],
BannedWordList.com [dtc, cph],
Apache Software Foundation [dtc, cph],
Andrew Kachites McCallum [dtc, cph],
Alireza Savand [dtc, cph],
Zact Anger [dtc, cph],
Titus Wormer [dtc, cph],
Colin Martindale [dtc, cph],
John Wiseman [dtc, cph],
Nadra Pencle [dtc, cph],
Irina Malaescu [dtc, cph]

Repository CRAN

Date/Publication 2019-03-21 10:40:03 UTC

R topics documented:

`available_data` *Get Available* **lexicon** *Data*

Description

See available **lexicon** data a data.frame.

Usage

```
available_data(regex = NULL, ...)
```

Arguments

`regex`	A regex to search for within the data columns.
`...`	Other arguments passed to `grep`.

Value

Returns a data.frame

Examples

```
available_data()
available_data('hash_')
available_data('hash_sentiment')
available_data('python')
available_data('prof')
available_data('English')
available_data('Stopword')
```

`cliches` *Common Cliches*

Description

A dataset containing a character vector of cliches.

Usage

```
data(cliches)
```

Format

A character vector with 697 elements

License

References

https://github.com/dunckr/retext-cliches

`common_names` *First Names (U.S.)*

Description

A dataset containing 1990 U.S. census data on first names.

Usage

```
data(common_names)
```

Format

A character vector with 5493 elements

References

http://www.census.gov

`constraining_loughran_mcdonald` *Loughran-McDonald Constraining Words*

Description

A dataset containing a character vector of Loughran & McDonald's (2016) constraining words list.

Usage

```
data(constraining_loughran_mcdonald)
```

Format

A character vector with 184 elements

License

The original authors note the data is available for non-commercial, research use: "The data compilations provided on this website are for use by individual researchers.". For more details see: https://sraf.nd.edu/textual-analysis/resources/#Master

Copyright

Copyright holder University of Notre Dame

References

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54(4), 1187-1230. doi: 10.2139/ssrn.2504147

https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary

---

emojis_sentiment *Emoji Sentiment Data*

Description

A slightly modified version of Novak, Smailovic, Sluban, & Mozetic's (2015) emoji sentiment data. The authors used Twitter data and 83 coders to rate each of the the emoji uses as negative, neutral, or positive to form a probability distribution (p_-, p_0, p_+) (http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0144296&type=printable).. The sentiment score is calculated via the authors' formula: $\frac{\sum(-1*p_-,0*p_0,p_+)}{\sum(p_-,p_0,p_+)}$.

Usage

```
data(emojis_sentiment)
```

Format

A data frame with 734 rows and 10 variables

Details

- byte. Byte code representation of emojis
- name. Description of the emoji
- id. An id for the emoji
- sentiment. Sentiment score of the emoji
- polarity. The direction of the sentiment
- category. A category for the emoji
- frequency. How often the emoji occurred in Novak et. al.'s (2015) data
- negative. How often Novak et al. (2015) observed the emoji being used negatively
- neutral. How often Novak et al. (2015) observed the emoji being used neutrally
- positive. How often Novak et al. (2015) observed the emoji being used positively

Copyright

2015 - Department of Knowledge Technologies

References

Novak, P. K., Smailovic, J., Sluban, B., and Mozetic, I. (2015) Sentiment of emojis. PLoS ONE 10(12). doi:10.1371/journal.pone.0144296

http://kt.ijs.si/data/Emoji_sentiment_ranking/index.html

https://creativecommons.org/licenses/by-sa/4.0/

`freq_first_names` *Frequent U.S. First Names*

Description

A dataset containing frequent first names based on the 1990 U.S. census.

Usage

```
data(freq_first_names)
```

Format

A data frame with 5494 rows and 3 variables

Details

- Name. A first name
- prop. The proportion within the sex
- sex. The sex corresponding to the name

References

https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html

---

`freq_last_names` *Frequent U.S. Last Names*

Description

A dataset containing frequent last names based on the 1990 U.S. census.

Usage

```
data(freq_last_names)
```

Format

A data frame with 14,840 rows and 2 variables

Details

- Surname. A last name
- prop. The proportion

References

https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html

---

`function_words` *Function Words*

Description

A vector of function words from John and Muriel Higgins's list used for the text game ECLIPSE. The list is augmented with additional contractions from `key_contractions`.

Usage

```
data(function_words)
```

Format

A character vector with 350 elements

Copyright

John Higgins

References

`http://myweb.tiscali.co.uk/wordscape/museum/funcword.html`

`grady_augmented` *Augmented List of Grady Ward's English Words and Mark Kantrowitz's Names List*

Description

A dataset containing a vector of Grady Ward's English words augmented with Mark Kantrowitz's names list, other proper nouns, and contractions.

Usage

```
data(grady_augmented)
```

Format

A character vector with 122,806 elements

Details

A dataset containing a vector of Grady Ward's English words augmented with proper nouns (U.S. States, Countries, Mark Kantrowitz's Names List, and months) and contractions. That dataset is augmented for spell checking purposes.

References

Moby Thesaurus List by Grady Ward

`hash_emojis` *Emoji Description Lookup Table*

Description

A dataset containing ASCII byte code representation of emojis and their accompanying description (from unicode.org).

Usage

```
data(hash_emojis)
```

Format

A data frame with 734 rows and 2 variables

Details

- x. Byte code representation of emojis
- y. Emoji description

References

http://www.unicode.org/emoji/charts/full-emoji-list.html

---

hash_emojis_identifier *Emoji Identifier Lookup Table*

Description

A dataset containing ASCII byte code representation of emojis and their accompanying identifier (for use in the **textclean** or **sentimentr** packages).

Usage

```
data(hash_emojis_identifier)
```

Format

A data frame with 734 rows and 2 variables

Details

- x. Byte code representation of emojis
- y. Emoji description

COPYRIGHT AND PERMISSION NOTICE

Copyright (c) 1991-2018 Unicode, Inc. All rights reserved. Distributed under the Terms of Use in http://www.unicode.org/copyright.html.

Permission is hereby granted, free of charge, to any person obtaining a copy of the Unicode data files and any associated documentation (the "Data Files") or Unicode software and any associated documentation (the "Software") to deal in the Data Files or Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, and/or sell copies of the Data Files or Software, and to permit persons to whom the Data Files or Software are furnished to do so, provided that either (a) this copyright and permission notice appear with all copies of the Data Files or Software, or (b) this copyright and permission notice appear in associated Documentation.

THE DATA FILES AND SOFTWARE ARE PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE BE LIABLE FOR ANY CLAIM, OR ANY SPECIAL INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THE DATA FILES OR SOFTWARE.

Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in these Data Files or Software without prior written authorization of the copyright holder.

References

http://www.unicode.org/emoji/charts/full-emoji-list.html

---

hash_emoticons *Emoticons*

Description

A **data.table** key containing common emoticons (adapted from Wikipedia's Page semi-protected 'List of emoticons').

Usage

```
data(hash_emoticons)
```

Format

A data.table with 144 rows and 2 variables

Details

- x. The graphic representation of the emoticon
- y. The meaning of the emoticon

License

https://creativecommons.org/licenses/by-sa/3.0/legalcode

References

https://en.wikipedia.org/wiki/List_of_emoticons

Examples

```
## Not run:
library(data.table)
hash_emoticons[c(':-(', '0:)')]

## End(Not run)
```

`hash_grady_pos` *Grady Ward's Moby Parts of Speech*

Description

A dataset containing a hash lookup of Grady Ward's parts of speech from the Moby project. The words with non-ASCII characters removed.

`grady_pos_feature` - A function for augmenting `hash_grady_pos` with 3 additional columns: (1) `n_pos` - the number of parts of speech a word has, (2) `space` - logical; indicating if a word contains a space, & (3) `primary` - logical; indicating if this is the most likely part of speech given the word.

Usage

```
data(hash_grady_pos)

grady_pos_feature(data)
```

Arguments

`data`	This should be `lexicon::hash_grady_pos`.

Format

A data frame with 246,691 rows and 3 variables

Details

- word. The word.
- pos. The part of speech; one of :Adjective, Adverb, Conjunction, Definite Article, Interjection, Noun, Noun Phrase, Plural, Preposition, Pronoun, Verb (intransitive), Verb (transitive), or Verb (usu participle). Note that the first part of speech for a word is its primary use; all other uses are secondary.

Source

Originally downloaded from: http://icon.shef.ac.uk/Moby

Examples

```
## Not run:
library(data.table)

hash_grady_pos <- grady_pos_feature(hash_grady_pos)
hash_grady_pos['dog']
hash_grady_pos[primary == TRUE, ]
hash_grady_pos[primary == TRUE & space == FALSE, ]

## End(Not run)
```

hash_internet_slang *List of Internet Slang and Corresponding Meanings*

Description

A dataset containing Internet slang terms and corresponding meaning. The data set is an augmented version of http://www.smart-words.org/abbreviations/text.html.

Usage

```
data(hash_internet_slang)
```

Format

A data frame with 175 rows and 2 variables

Details

- x. The slang term.
- y. The meaning.

References

Possel, H. (n.d.). English language smart words. Retrieved from http://www.smart-words.org

http://www.smart-words.org/abbreviations/text.html hash_lemmas *Lemmatization List*

Description

A dataset based on Mechura's (2016) English lemmatization list. This data set can be useful for join style lemma replacement of inflected token forms to their root lemmas. While this is not a true morphological analysis this style of lemma replacement is fast and typically still robust.

Usage

```
data(hash_lemmas)
```

Format

A data frame with 41,531 rows and 2 variables

Details

- token. An inflected token with affixes
- lemma. A base form

## ODC Open Database License (ODbL)

### Preamble

The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Database while maintaining this same freedom for others. Many databases are covered by copyright, and therefore this document licenses these rights. Some jurisdictions, mainly in the European Union, have specific rights that cover databases, and so the ODbL addresses these rights, too. Finally, the ODbL is also an agreement in contract for users of this Database to act in certain ways in return for accessing this Database.

Databases can contain a wide variety of types of content (images, audiovisual material, and sounds all in the same database, for example), and so the ODbL only governs the rights over the Database, and not the contents of the Database individually. Licensors should use the ODbL together with another license for the contents, if the contents have a single set of rights that uniformly covers all of the contents. If the contents have multiple sets of different rights, Licensors should describe what rights govern what contents together in the individual record or in some other way that clarifies what rights apply.

Sometimes the contents of a database, or the database itself, can be covered by other rights not addressed here (such as private contracts, trade mark over the name, or privacy rights / data protection rights over information in the contents), and so you are advised that you may have to consult other documents or clear other rights before doing activities not covered by this License.

——

The Licensor (as defined below)

and

You (as defined below)

agree as follows:

### 1.0 Definitions of Capitalised Words

"Collective Database" - Means this Database in unmodified form as part of a collection of independent databases in themselves that together are assembled into a collective whole. A work that constitutes a Collective Database will not be considered a Derivative Database.

"Convey" - As a verb, means Using the Database, a Derivative Database, or the Database as part of a Collective Database in any way that enables a Person to make or receive copies of the Database or a Derivative Database. Conveying does not include interaction with a user through a computer network, or creating and Using a Produced Work, where no transfer of a copy of the Database or a Derivative Database occurs. "Contents" - The contents of this Database, which includes the information, independent works, or other material collected into the Database. For example, the contents of the Database could be factual data or works such as images, audiovisual material, text, or sounds.

"Database" - A collection of material (the Contents) arranged in a systematic or methodical way and individually accessible by electronic or other means offered under the terms of this License.

"Database Directive" - Means Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended or succeeded.

"Database Right" - Means rights resulting from the Chapter III ("sui generis") rights in the Database Directive (as amended and as transposed by member states), which includes the Extraction and Re-utilisation of the whole or a Substantial part of the Contents, as well as any similar rights available in the relevant jurisdiction under Section 10.4.

"Derivative Database" - Means a database based upon the Database, and includes any translation, adaptation, arrangement, modification, or any other alteration of the Database or of a Substantial part of the Contents. This includes, but is not limited to, Extracting or Re-utilising the whole or a Substantial part of the Contents in a new Database.

"Extraction" - Means the permanent or temporary transfer of all or a Substantial part of the Contents to another medium by any means or in any form.

"License" - Means this license agreement and is both a license of rights such as copyright and Database Rights and an agreement in contract.

"Licensor" - Means the Person that offers the Database under the terms of this License.

"Person" - Means a natural or legal person or a body of persons corporate or incorporate.

"Produced Work" - a work (such as an image, audiovisual material, text, or sounds) resulting from using the whole or a Substantial part of the Contents (via a search or other query) from this Database, a Derivative Database, or this Database as part of a Collective Database.

"Publicly" - means to Persons other than You or under Your control by either more than 50 activities (such as contracting with an independent consultant).

"Re-utilisation" - means any form of making available to the public all or a Substantial part of the Contents by the distribution of copies, by renting, by online or other forms of transmission.

"Substantial" - Means substantial in terms of quantity or quality or a combination of both. The repeated and systematic Extraction or Re-utilisation of insubstantial parts of the Contents may amount to the Extraction or Re-utilisation of a Substantial part of the Contents.

"Use" - As a verb, means doing any act that is restricted by copyright or Database Rights whether in the original medium or any other; and includes without limitation distributing, copying, publicly performing, publicly displaying, and preparing derivative works of the Database, as well as modifying the Database as may be technically necessary to use it in a different mode or format.

"You" - Means a Person exercising rights under this License who has not previously violated the terms of this License with respect to the Database, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

Words in the singular include the plural and vice versa.

### 2.0 What this License covers

2.1. Legal effect of this document. This License is:

a. A license of applicable copyright and neighbouring rights;

b. A license of the Database Right; and

c. An agreement in contract between You and the Licensor.

2.2 Legal rights covered. This License covers the legal rights in the Database, including:

a. Copyright. Any copyright or neighbouring rights in the Database. The copyright licensed includes any individual elements of the Database, but does not cover the copyright over the Contents independent of this Database. See Section 2.4 for details. Copyright law varies between jurisdictions, but is likely to cover: the Database model or schema, which is the structure, arrangement, and organisation of the Database, and can also include the Database tables and table indexes; the data entry and output sheets; and the Field names of Contents stored in the Database;

b. Database Rights. Database Rights only extend to the Extraction and Re-utilisation of the whole or a Substantial part of the Contents. Database Rights can apply even when there is no copyright over the Database. Database Rights can also apply when the Contents are removed from the Database and are selected and arranged in a way that would not infringe any applicable copyright; and

c. Contract. This is an agreement between You and the Licensor for access to the Database. In return you agree to certain conditions of use on this access as outlined in this License.

2.3 Rights not covered.

a. This License does not apply to computer programs used in the making or operation of the Database;

b. This License does not cover any patents over the Contents or the Database; and

c. This License does not cover any trademarks associated with the Database.

2.4 Relationship to Contents in the Database. The individual items of the Contents contained in this Database may be covered by other rights, including copyright, patent, data protection, privacy, or personality rights, and this License does not cover any rights (other than Database Rights or in contract) in individual Contents contained in the Database. For example, if used on a Database of images (the Contents), this License would not apply to copyright over individual images, which could have their own separate licenses, or one single license covering all of the rights over the images.

### 3.0 Rights granted

3.1 Subject to the terms and conditions of this License, the Licensor grants to You a worldwide, royalty-free, non-exclusive, terminable (but only under Section 9) license to Use the Database for the duration of any applicable copyright and Database Rights. These rights explicitly include commercial use, and do not exclude any field of endeavour. To the extent possible in the relevant jurisdiction, these rights may be exercised in all media and formats whether now known or created in the future.

The rights granted cover, for example:

a. Extraction and Re-utilisation of the whole or a Substantial part of the Contents;

b. Creation of Derivative Databases;

c. Creation of Collective Databases;

d. Creation of temporary or permanent reproductions by any means and in any form, in whole or in part, including of any Derivative Databases or as a part of Collective Databases; and

e. Distribution, communication, display, lending, making available, or performance to the public by any means and in any form, in whole or in part, including of any Derivative Database or as a part of Collective Databases.

3.2 Compulsory license schemes. For the avoidance of doubt:

a. Non-waivable compulsory license schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

b. Waivable compulsory license schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

c. Voluntary license schemes. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

3.3 The right to release the Database under different terms, or to stop distributing or making available the Database, is reserved. Note that this Database may be multiple-licensed, and so You may have the choice of using alternative licenses for this Database. Subject to Section 10.4, all other rights not expressly granted by Licensor are reserved.

### 4.0 Conditions of Use

4.1 The rights granted in Section 3 above are expressly made subject to Your complying with the following conditions of use. These are important conditions of this License, and if You fail to follow them, You will be in material breach of its terms.

4.2 Notices. If You Publicly Convey this Database, any Derivative Database, or the Database as part of a Collective Database, then You must:

a. Do so only under the terms of this License or another license permitted under Section 4.4;

b. Include a copy of this License (or, as applicable, a license permitted under Section 4.4) or its Uniform Resource Identifier (URI) with the Database or Derivative Database, including both in the Database or Derivative Database and in any relevant documentation; and

c. Keep intact any copyright or Database Right notices and notices that refer to this License.

d. If it is not possible to put the required notices in a particular file due to its structure, then You must include the notices in a location (such as a relevant directory) where users would be likely to look for it.

4.3 Notice for using output (Contents). Creating and Using a Produced Work does not require the notice in Section 4.2. However, if you Publicly Use a Produced Work, You must include a notice associated with the Produced Work reasonably calculated to make any Person that uses, views, accesses, interacts with, or is otherwise exposed to the Produced Work aware that Content was obtained from the Database, Derivative Database, or the Database as part of a Collective Database, and that it is available under this License.

a. Example notice. The following text will satisfy notice under Section 4.3:

Contains information from DATABASE NAME, which is made available here under the Open Database License (ODbL).

DATABASE NAME should be replaced with the name of the Database and a hyperlink to the URI of the Database. "Open Database License" should contain a hyperlink to the URI of the text of this License. If hyperlinks are not possible, You should include the plain text of the required URI's with the above notice.

4.4 Share alike.

a. Any Derivative Database that You Publicly Use must be only under the terms of:

i. This License;

ii. A later version of this License similar in spirit to this License; or

iii. A compatible license.

If You license the Derivative Database under one of the licenses mentioned in (iii), You must comply with the terms of that license.

b. For the avoidance of doubt, Extraction or Re-utilisation of the whole or a Substantial part of the Contents into a new database is a Derivative Database and must comply with Section 4.4.

c. Derivative Databases and Produced Works. A Derivative Database is Publicly Used and so must comply with Section 4.4. if a Produced Work created from the Derivative Database is Publicly Used.

d. Share Alike and additional Contents. For the avoidance of doubt, You must not add Contents to Derivative Databases under Section 4.4 a that are incompatible with the rights granted under this License.

e. Compatible licenses. Licensors may authorise a proxy to determine compatible licenses under Section 4.4 a iii. If they do so, the authorised proxy's public statement of acceptance of a compatible license grants You permission to use the compatible license.

4.5 Limits of Share Alike. The requirements of Section 4.4 do not apply in the following:

a. For the avoidance of doubt, You are not required to license Collective Databases under this License if You incorporate this Database or a Derivative Database in the collection, but this License still applies to this Database or a Derivative Database as a part of the Collective Database;

b. Using this Database, a Derivative Database, or this Database as part of a Collective Database to create a Produced Work does not create a Derivative Database for purposes of Section 4.4; and

c. Use of a Derivative Database internally within an organisation is not to the public and therefore does not fall under the requirements of Section 4.4.

4.6 Access to Derivative Databases. If You Publicly Use a Derivative Database or a Produced Work from a Derivative Database, You must also offer to recipients of the Derivative Database or Produced Work a copy in a machine readable form of:

a. The entire Derivative Database; or

b. A file containing all of the alterations made to the Database or the method of making the alterations to the Database (such as an algorithm), including any additional Contents, that make up all the differences between the Database and the Derivative Database.

The Derivative Database (under a.) or alteration file (under b.) must be available at no more than a reasonable production cost for physical distributions and free of charge if distributed over the internet.

4.7 Technological measures and additional terms

a. This License does not allow You to impose (except subject to Section 4.7 b.) any terms or any technological measures on the Database, a Derivative Database, or the whole or a Substantial part of the Contents that alter or restrict the terms of this License, or any rights granted under it, or have the effect or intent of restricting the ability of any person to exercise those rights.

b. Parallel distribution. You may impose terms or technological measures on the Database, a Derivative Database, or the whole or a Substantial part of the Contents (a "Restricted Database") in contravention of Section 4.74 a. only if You also make a copy of the Database or a Derivative Database available to the recipient of the Restricted Database:

i. That is available without additional fee;

ii. That is available in a medium that does not alter or restrict the terms of this License, or any rights granted under it, or have the effect or intent of restricting the ability of any person to exercise those rights (an "Unrestricted Database"); and

iii. The Unrestricted Database is at least as accessible to the recipient as a practical matter as the Restricted Database.

c. For the avoidance of doubt, You may place this Database or a Derivative Database in an authenticated environment, behind a password, or within a similar access control scheme provided that You do not alter or restrict the terms of this License or any rights granted under it or have the effect or intent of restricting the ability of any person to exercise those rights.

4.8 Licensing of others. You may not sublicense the Database. Each time You communicate the Database, the whole or Substantial part of the Contents, or any Derivative Database to anyone else in any way, the Licensor offers to the recipient a license to the Database on the same terms and conditions as this License. You are not responsible for enforcing compliance by third parties with this License, but You may enforce any rights that You have over a Derivative Database. You are solely responsible for any modifications of a Derivative Database made by You or another Person at Your direction. You may not impose any further restrictions on the exercise of the rights granted or affirmed under this License.

### 5.0 Moral rights

5.1 Moral rights. This section covers moral rights, including any rights to be identified as the author of the Database or to object to treatment that would otherwise prejudice the author's honour and reputation, or any other derogatory treatment:

a. For jurisdictions allowing waiver of moral rights, Licensor waives all moral rights that Licensor may have in the Database to the fullest extent possible by the law of the relevant jurisdiction under Section 10.4;

b. If waiver of moral rights under Section 5.1 a in the relevant jurisdiction is not possible, Licensor agrees not to assert any moral rights over the Database and waives all claims in moral rights to the fullest extent possible by the law of the relevant jurisdiction under Section 10.4; and

c. For jurisdictions not allowing waiver or an agreement not to assert moral rights under Section 5.1 a and b, the author may retain their moral rights over certain aspects of the Database.

Please note that some jurisdictions do not allow for the waiver of moral rights, and so moral rights may still subsist over the Database in some jurisdictions.

### 6.0 Fair dealing, Database exceptions, and other rights not affected

6.1 This License does not affect any rights that You or anyone else may independently have under any applicable law to make any use of this Database, including without limitation:

a. Exceptions to the Database Right including: Extraction of Contents from non-electronic Databases for private purposes, Extraction for purposes of illustration for teaching or scientific research, and Extraction or Re-utilisation for public security or an administrative or judicial procedure.

b. Fair dealing, fair use, or any other legally recognised limitation or exception to infringement of copyright or other applicable laws.

6.2 This License does not affect any rights of lawful users to Extract and Re-utilise insubstantial parts of the Contents, evaluated quantitatively or qualitatively, for any purposes whatsoever, including creating a Derivative Database (subject to other rights over the Contents, see Section 2.4). The repeated and systematic Extraction or Re-utilisation of insubstantial parts of the Contents may however amount to the Extraction or Re-utilisation of a Substantial part of the Contents.

### 7.0 Warranties and Disclaimer

7.1 The Database is licensed by the Licensor "as is" and without any warranty of any kind, either express, implied, or arising by statute, custom, course of dealing, or trade usage. Licensor specifically disclaims any and all implied warranties or conditions of title, non-infringement, accuracy or completeness, the presence or absence of errors, fitness for a particular purpose, merchantability, or otherwise. Some jurisdictions do not allow the exclusion of implied warranties, so this exclusion may not apply to You.

### 8.0 Limitation of liability

8.1 Subject to any liability that may not be excluded or limited by law, the Licensor is not liable for, and expressly excludes, all liability for loss or damage however and whenever caused to anyone by any use under this License, whether by You or by anyone else, and whether caused by any fault on the part of the Licensor or not. This exclusion of liability includes, but is not limited to, any special, incidental, consequential, punitive, or exemplary damages such as loss of revenue, data, anticipated profits, and lost business. This exclusion applies even if the Licensor has been advised of the possibility of such damages.

8.2 If liability may not be excluded by law, it is limited to actual and direct financial loss to the extent it is caused by proved negligence on the part of the Licensor.

### 9.0 Termination of Your rights under this License

9.1 Any breach by You of the terms and conditions of this License automatically terminates this License with immediate effect and without notice to You. For the avoidance of doubt, Persons who have received the Database, the whole or a Substantial part of the Contents, Derivative Databases, or the Database as part of a Collective Database from You under this License will not have their licenses terminated provided their use is in full compliance with this License or a license granted under Section 4.8 of this License. Sections 1, 2, 7, 8, 9 and 10 will survive any termination of this License.

9.2 If You are not in breach of the terms of this License, the Licensor will not terminate Your rights under it.

9.3 Unless terminated under Section 9.1, this License is granted to You for the duration of applicable rights in the Database.

9.4 Reinstatement of rights. If you cease any breach of the terms and conditions of this License, then your full rights under this License will be reinstated:

a. Provisionally and subject to permanent termination until the 60th day after cessation of breach;

b. Permanently on the 60th day after cessation of breach unless otherwise reasonably notified by the Licensor; or

c. Permanently if reasonably notified by the Licensor of the violation, this is the first time You have received notice of violation of this License from the Licensor, and You cure the violation prior to 30 days after your receipt of the notice.

Persons subject to permanent termination of rights are not eligible to be a recipient and receive a license under Section 4.8.

9.5 Notwithstanding the above, Licensor reserves the right to release the Database under different license terms or to stop distributing or making available the Database. Releasing the Database under different license terms or stopping the distribution of the Database will not withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

### 10.0 General

10.1 If any provision of this License is held to be invalid or unenforceable, that must not affect the validity or enforceability of the remainder of the terms and conditions of this License and each remaining provision of this License shall be valid and enforced to the fullest extent permitted by law.

10.2 This License is the entire agreement between the parties with respect to the rights granted here over the Database. It replaces any earlier understandings, agreements or representations with respect to the Database.

10.3 If You are in breach of the terms of this License, You will not be entitled to rely on the terms of this License or to complain of any breach by the Licensor.

10.4 Choice of law. This License takes effect in and will be governed by the laws of the relevant jurisdiction in which the License terms are sought to be enforced. If the standard suite of rights granted under applicable copyright law and Database Rights in the relevant jurisdiction includes additional rights not granted under this License, these additional rights are granted in this License in order to meet the terms of this License.

References

Mechura, M. B. (2016). *Lemmatization list: English (en)* [Data file]. Retrieved from http://www.lexiconista.com

---

hash_nrc_emotions *NRC Emotion Table*

Description

A **data.table** dataset containing a filtered version of Mohammad & Turney', P. D.'s (2010) emotion word list as lookup table.

Usage

```
data(hash_nrc_emotions)
```

Format

A data frame with 8265 rows and 2 variables

Details

- token. A search token indicating emotion.
- emotion. An accompanying emotion assocatiated with the token.

References

http://www.purl.com/net/lexicons

Mohammad, S. M. & Turney, P. D. (2010) Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon, In Proceeding of Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 26-34.

`hash_sentiment_emojis` *Emoji Sentiment Polarity Lookup Table*

Description

A dataset containing an emoji identifier key and sentiment value. This data comes from Novak, Smailovic, Sluban, & Mozetic's (2015) emoji sentiment data. The authors used Twitter data and 83 coders to rate each of the the emoji uses as negative, neutral, or positive to form a probability distribution (p_-, p_0, p_+) (`http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0144296&type=printable`).. The sentiment score is calculated via the authors' formula: $\frac{\sum(-1*p_-, 0*p_0, p_+)}{\sum(p_-, p_0, p_+)}$. This polarity lookup table differs from the other ones included in the **lexicon** package in the the first column are not words but identifiers. These identifiers are found in the `emojis_sentiment` data set. The typical use case is to utilize the **textclean** or **sentimentr** packages' `replace_emoji` to swap out emojis for a more computer friendly identifier.

Usage

```
data(hash_sentiment_emojis)
```

Format

A data frame with 734 rows and 2 variables

Details

- x. Words
- y. Sentiment

Copyright

2015 - Department of Knowledge Technologies

References

Novak, P. K., Smailovic, J., Sluban, B., and Mozetic, I. (2015) Sentiment of emojis. PLoS ONE 10(12). doi:10.1371/journal.pone.0144296

http://kt.ijs.si/data/Emoji_sentiment_ranking/index.html

https://creativecommons.org/licenses/by-sa/4.0/

`hash_sentiment_huliu` *Hu Liu Polarity Lookup Table*

Description

A **data.table** dataset containing an augmented version of Hu & Liu's (2004) positive/negative word list as sentiment lookup values.

Usage

```
data(hash_sentiment_huliu)
```

Format

A data frame with 6874 rows and 2 variables

Details

- x. Words
- y. Sentiment values (+1, 0, -1.05, -1, -2), -2 indicate phrasing that is always negative (e.g., 'too much fun' and 'too much evil' both denote negative though the following word is positive and negative respectively).

References

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004). Seattle, Washington.

Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. National Conference on Artificial Intelligence.

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

`hash_sentiment_jockers` *Jockers Polarity Lookup Table*

Description

A **data.table** dataset containing a modified version of Jocker's (2017) sentiment lookup table used in **syuzhet**.

Usage

```
hash_sentiment_jockers
```

Format

An object of class `data.table` (inherits from `data.frame`) with 10738 rows and 2 columns.

Details

- x. Words
- y. Sentiment values ranging between -1 and 1.

References

Jockers, M. L. (2017). Syuzhet: Extract sentiment and plot arcs from Text. Retrieved from https://github.com/mjockers/syuzhet

`hash_sentiment_jockers_rinker` *Combined Jockers & Rinker Polarity Lookup Table*

Description

A **data.table** dataset containing a combined and augmented version of Jockers (2017) & Rinker's augmented Hu & Liu (2004) positive/negative word list as sentiment lookup values.

Usage

```
data(hash_sentiment_jockers_rinker)
```

Format

A data frame with 11,710 rows and 2 variables

Details

- x. Words
- y. Sentiment

References

Jockers, M. L. (2017). Syuzhet: Extract sentiment and plot arcs from Text. Retrieved from https://github.com/mjockers/syuzhet

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004). Seattle, Washington.

---

hash_sentiment_loughran_mcdonald

Loughran-McDonald Polarity Table

Description

A **data.table** dataset containing an filtered version of Loughran & McDonald's (2016) positive/negative financial word list as sentiment lookup values.

Usage

```
data(hash_sentiment_loughran_mcdonald)
```

Format

A data frame with 2,702 rows and 2 variables

Details

- x. Words
- y. Sentiment values

License

The original authors note the data is available for non-commercial, research use: "The data compilations provided on this website are for use by individual researchers.". For more details see: https://sraf.nd.edu/textual-analysis/resources/#Master

Copyright

Copyright holder University of Notre Dame

References

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54(4), 1187-1230. doi: 10.2139/ssrn.2504147

`https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary`

`hash_sentiment_nrc` *NRC Sentiment Polarity Table*

Description

A **data.table** dataset containing a filtered version of Mohammad & Turney', P. D.'s (2010) positive/negative word list as sentiment lookup values.

Usage

```
data(hash_sentiment_nrc)
```

Format

A data frame with 5468 rows and 2 variables

Details

- x. Words
- y. Sentiment values (+1, -1)

License

The original authors note the data is available for non-commercial use: "If interested in commercial use of any of these lexicons, send email to Saif M. Mohammad (Senior Research Officer at NRC and creator of these lexicons): saif.mohammad@nrc-cnrc.gc.ca and Pierre Charron (Client Relationship Leader at NRC): Pierre.Charron@nrc-cnrc.gc.ca. A nominal one-time licensing fee may apply."

References

http://www.purl.com/net/lexicons

Mohammad, S. M. & Turney, P. D. (2010) Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon, In Proceeding of Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 26-34.

Examples

```
## Not run:
library(data.table)
hash_sentiment_nrc[c('happy', 'angry')]

## End(Not run)
```

`hash_sentiment_senticnet`

Augmented SenticNet Polarity Table

Description

A **data.table** dataset containing an augmented version of Cambria, Poria, Bajpai,& Schuller's (2016) positive/negative word list as sentiment lookup values.

Usage

```
data(hash_sentiment_senticnet)
```

Format

A data frame with 23,626 rows and 2 variables

Details

- x. Words
- y. Sentiment values

Original Publication Credit Statement: Thank you for using SenticNet 4!

Please acknowledge the authors by citing the following publication in any research work or presentation containing results obtained in whole or in part through the use of SenticNet 4:

Cambria, E., Poria, S., Bajpai, R. and Schuller, B. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: COLING, pp. 2666-2677, Osaka (2016))

References

Cambria, E., Poria, S., Bajpai, R. and Schuller, B. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: COLING, pp. 2666-2677, Osaka (2016) http://sentic.net/downloads

`hash_sentiment_sentiword`

Augmented Sentiword Polarity Table

Description

A **data.table** dataset containing an augmented version of Baccianella, Esuli and Sebastiani's (2010) positive/negative word list as sentiment lookup values. This list has be restructured to long format. A polarity value was assigned by taking the difference between the original data set's negative and positive attribution (`PosScore - NegScore`). All rows with a zero polarity were removed from the data set as well as any duplicated in the valence shifter's data set.

Usage

```
data(hash_sentiment_sentiword)
```

Format

A data frame with 20,093 rows and 2 variables

Details

- x. Words
- y. Sentiment values

License

https://creativecommons.org/licenses/by-sa/3.0/legalcode

References

Baccianella S., Esuli, A. and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. International Conference on Language Resources and Evaluation.

https://sentiwordnet.isti.cnr.it

`hash_sentiment_slangsd`

SlangSD Sentiment Polarity Table

Description

A **data.table** dataset containing a filtered version of Wu, Morstatter, & Liu's (2016) positive/negative slang word list as sentiment lookup values. All words containing other than `"[a-z ']"` have been removed as well as any neutral words.

Usage

```
data(hash_sentiment_slangsd)
```

Format

A data frame with 48,277 rows and 2 variables

Details

- x. Words
- y. Sentiment values (+1, -1)

Original Licensing: The dictionary is free to use. If you use it for an academic publication, we ask that you cite it using the citation below. If it is used in anything other than an academic publication, we ask that you provide a credit and link to SlangSD.com.

articleDBLP:journals/corr/Wu-etal16, author = Liang Wu and Fred Morstatter and Huan Liu, title = SlangSD: Building and Using a Sentiment Dictionary of Slang Words for Short-Text Sentiment Classification, journal = CoRR, volume = abs/1608.05129, year = 2016, url = http://arxiv.org/abs/1608.05129, timestamp = Wed, 17 Aug 2016 23:32:57 GMT

References

Wu, L., Morstatter, F., and Liu, H. (2016). SlangSD: Building and using a sentiment dictionary of slang words for short-text sentiment classification. CoRR. abs/1168.1058. 1-15.

http://slangsd.com

---

hash_sentiment_socal_google

SO-CAL Google Polarity Table

Description

A **data.table** dataset containing a version of Taboada, Brooke, Tofiloski, Voll, & Stede's (2011) positive/negative word list as sentiment lookup values.

Usage

```
data(hash_sentiment_socal_google)
```

Format

A data frame with 3272 rows and 2 variables

Details

- x. Words
- y. Sentiment values

License

The original license states: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. https://creativecommons.org/licenses/by-nc-sa/4.0/

References

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2). 267-307.

https://github.com/sfu-discourse-lab/SO-CAL

`hash_valence_shifters` *Valence Shifters*

Description

A **data.table** dataset containing a vector of valence shifter words that can alter a polarized word's meaning and a numeric key for negators (1), amplifiers [intensifier] (2), de-amplifiers [downtoners] (3), and adversative conjunctions (4).

Usage

```
data(hash_valence_shifters)
```

Format

A data frame with 140 rows and 2 variables

Details

Valence shifters are words that alter or intensify the meaning of the polarized words and include negators and amplifiers. Negators are, generally, adverbs that negate sentence meaning; for example the word like in the sentence, "I do like pie.", is given the opposite meaning in the sentence, "I do not like pie.", now containing the negator not. Amplifiers (intensifiers) are, generally, adverbs or adjectives that intensify sentence meaning. Using our previous example, the sentiment of the negator altered sentence, "I seriously do not like pie.", is heightened with addition of the amplifier seriously. Whereas de-amplifiers (downtoners) decrease the intensity of a polarized word as in the sentence "I barely like pie"; the word "barely" deamplifies the word like. Adversative conjunction trump the previous clause (e.g., “He’s a nice guy but not too smart.”).

- x. Valence shifter
- y. Number key value corresponding to:

Valence Shifter	**Value**
Negator	1
Amplifier (intensifier)	2
De-amplifier (downtoner)	3
Adversative Conjunction	4

`key_contractions` *Contraction Conversions*

Description

A dataset containing common contractions and their expanded form.

Usage

```
data(key_contractions)
```

Format

A data frame with 70 rows and 2 variables

Details

- contraction. The contraction word
- expanded. The expanded form of the contraction

`key_corporate_social_responsibility` *Nadra Pencle and Irina Mălăescu's Corporate Social Responsibility Dictionary*

Description

A dataset containing Pencle & Mălăescu's Corporate Social Responsibility (CSR) Dictionary. The Corporate Social Responsibility Dictionary is a text analysis coding taxonomy that was used to predict initial public offerings for new companies. This particular list was taken from http://www.catscanner.net/dictionaries.php.

Usage

```
data(key_corporate_social_responsibility)
```

Format

A data frame with 1,421 rows and 3 variables

Details

- dimension. One of: "human_rights", "employee", "social_and_community", or "environment"
- regex. An associated search regex
- token. An associated word/token

References

Pencle, N. and Mălăescu, I. (2016) What's in the words? Development and validation of a multidimensional dictionary for CSR and application using prospectuses. Journal of Emerging Technologies in Accounting, 13(2), 109-127.
http://www.catscanner.net/dictionaries.php

`key_grade` *Grades Data Set*

Description

A dataset containing common grades.

Usage

```
data(key_grade)
```

Format

A data frame with 15 rows and 2 variables

Details

- x. The graphic representation of the grade
- y. The meaning of the grade

`key_rating` *Ratings Data Set*

Description

A dataset containing common ratings.

Usage

```
data(key_rating)
```

Format

A data frame with 35 rows and 2 variables

Details

- x. The graphic representation of the rating
- y. The meaning of the rating

key_regressive_imagery *Colin Martindale's English Regressive Imagery Dictionary*

Description

A dataset containing Colin Martindale's (1975, 1990) English Regressive Imagery Dictionary (RID). The Regressive Imagery Dictionary (RID) is a text analysis coding taxonomy that can be used to measure the degree to which a text is *primordial* vs. *conceptual*. This acts as a proxy for assessing the illuctioner's mental thinking in producing the text. This dictionary is essentially a bucketed grouping of regexes' The main level of bucketing is *thinking* and is either *primordial* vs. *conceptual*. Under the primordial group is the *primary* process group while the conceptual thinking includes *secondary* and *emotional* process groups. These can be further broken into categories and subcategories (subcategories for primary process only). Comparing the percentages of the buckets provides insight into the writer's thinking. This particular list was taken from https://github.com/jefftriplett/rid.py.

Usage

```
data(key_regressive_imagery)
```

Format

A data frame with 3,151 rows and 5 variables

Details

- thinking. Either primordial or conceptual
- process. One of three: primary (5 categories & 29 subcategories), secondary (7 categories), or emotional (7 categories)
- category. A level of bucketing lower than process
- subcategory. A level of bucketing lower than category (only applies to rimary process)
- regex. An associated search regex

License

The data set was extracted from https://github.com/jefftriplett/rid.py. Below is the license from Wiseman's project.

References

Martindale, C. (1975). Romantic progression: The psychology of literary history. Washington, D.C.: Hemisphere.
Martindale, C. (1976). Primitive mentality and the relationship between art and society. Scientific Aesthetics, 1, 5218.
Martindale, C. (1977). Syntactic and semantic correlates of verbal tics in Gilles de la Tourette's syndrome: A quantitative case study. Brain and Language, 4, 231-247.
Martindale, C. (1990). The clockwork muse: The predictability of artistic change. New York: Basic Books.
https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/regressive-imagery-dictionary/

`key_sentiment_jockers` *Jockers Sentiment Key*

Description

A dataset containing an imported version of Jocker's (2017) sentiment lookup table used in **syuzhet**.

Usage

```
key_sentiment_jockers
```

Format

An object of class `data.frame` with 10748 rows and 2 columns.

Details

- word. Words
- value. Sentiment values ranging between -1 and 1.

References

Jockers, M. L. (2017). Syuzhet: Extract sentiment and plot arcs from Text. Retrieved from https://github.com/mjockers/syuzhet

lexicon *Lexicons for Text Analysis*

Description

A collection of lexical hash tables, dictionaries, and word lists.

modal_loughran_mcdonald *Loughran-McDonald Modal List*

Description

A dataset containing a character vector of Loughran & McDonald's (2016) modal list. Wikipedia states: "A modal verb is a type of verb that is used to indicate modality - that is: likelihood, ability, permission and obligation."

Usage

```
data(modal_loughran_mcdonald)
```

Format

A data frame with 61 rows and 2 variables

Details

- modal.
- strength.

License

The original authors note the data is available for non-commercial, research use: "The data compilations provided on this website are for use by individual researchers.". For more details see: https://sraf.nd.edu/textual-analysis/resources/#Master

Copyright

Copyright holder University of Notre Dame

References

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54(4), 1187-1230. doi: 10.2139/ssrn.2504147

https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary

`nrc_emotions` *NRC Emotions*

Description

A **data.table** dataset containing Mohammad & Turney', P. D.'s (2010) emotions word list as a binary table.

Usage

```
data(nrc_emotions)
```

Format

A data frame with 14182 rows and 9 variables

Details

- term. A term
- anger. Counts of anger anger
- anticipation. Counts of anticipation
- disgust. Counts of disgust
- fear. Counts of fear
- joy. Counts of joy
- sadness. Counts of sadness
- surprise. Counts of surprise
- trust. Counts of trust

License

The original authors note the data is available for non-commercial use: "If interested in commercial use of any of these lexicons, send email to Saif M. Mohammad (Senior Research Officer at NRC and creator of these lexicons): saif.mohammad@nrc-cnrc.gc.ca and Pierre Charron (Client Relationship Leader at NRC): Pierre.Charron@nrc-cnrc.gc.ca. A nominal one-time licensing fee may apply."

References

http://www.purl.com/net/lexicons

Mohammad, S. M. & Turney, P. D. (2010) Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon, In Proceeding of Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 26-34.

pos_action_verb *Action Word List*

Description

A dataset containing a vector of action words. This is a subset of the Moby project: Moby Part-of-Speech.

Usage

```
data(pos_action_verb)
```

Format

A character vector with 1569 elements

Details

From Grady Ward's Moby project: "This second edition is a particularly thorough revision of the original Moby Part-of-Speech. Beyond the fifteen thousand new entries, many thousand more entries have been scrutinized for correctness and modernity. This is unquestionably the largest P-O-S list in the world. Note that the many included phrases means that parsing algorithms can now tokenize in units larger than a single word, increasing both speed and accuracy." Originally downloaded from: http://icon.shef.ac.uk/Moby

pos_df_irregular_nouns *Irregular Nouns Word Dataframe*

Description

A dataset containing a `data.frame` of irregular noun singular and plural forms from Arizona Department of Education (https://cms.azed.gov) and augmented with selected common nouns from Wikipedia's "English Plurals" (https://en.wikipedia.org/wiki/English_plurals).

Usage

```
data(pos_df_irregular_nouns)
```

Format

A data frame with 124 rows and 2 variables https://cms.azed.gov/home/GetDocumentFile?id=54de1d89aadebe14a8707103

https://en.wikipedia.org/wiki/English_plurals

Details

- singular. The singular form of the noun
- plural. The plural form of the noun

License

The Wikipedia data is Creative Commons. See https://creativecommons.org/licenses/by-sa/3.0/ for License information.

`pos_df_pronouns` *Pronouns*

Description

A dataset containing pronouns categorized by type, singular, point_of_view, and use. Note that 'you', and 'yours' appear twice because 'you' can be singular or plural.

Usage

```
data(pos_df_pronouns)
```

Format

A data frame with 34 rows and 5 variables

Details

- pronoun. The pronoun.
- type. The pronoun type; either "personal", "reflexive", or "possessive".
- singular. logical. If `TRUE` the pronoun is singular, otherwise it's plural.
- point_of_view. The point of view; either "first", "second", or "third".

References

http://www.english-grammar-revolution.com/list-of-pronouns.html

`pos_interjections` *Interjections*

Description

Vidar Holen's dataset containing a character vector of common interjections compiled from: http://www.vidarholen.net/conte

Usage

```
data(pos_interjections)
```

Format

A character vector with 139 elements

References

http://www.vidarholen.net/contents/interjections/

`pos_preposition` *Preposition Words*

Description

A dataset containing a vector of common prepositions.

Usage

```
data(pos_preposition)
```

Format

A character vector with 162 elements

`profanity_alvarez` *Alejandro U. Alvarez's List of Profane Words*

Description

A dataset containing a character vector of profane words from Alejandro U. Alvarez.

Usage

```
data(profanity_alvarez)
```

Format

A character vector with 438 elements

TermsOfUse

https://archive.org/about/terms.php

References

`https://web.archive.org/web/20130704010355/http://urbanoalvarez.es:80/blog/2008/04/04/bad-words-list/`

`profanity_arr_bad` *Stackoverflow user2592414's List of Profane Words*

Description

A dataset containing a character vector of profane words from Stackoverflow user2592414.

Usage

```
data(profanity_arr_bad)
```

Format

A character vector with 343 elements

License

The Stackoverflow data is Creative Commons. See https://creativecommons.org/licenses/by-sa/3.0/ for License information.

References

`https://stackoverflow.com/a/17706025/1000343`

`profanity_banned` *bannedwordlist.com's List of Profane Words*

Description

A dataset containing a character vector of profane words from bannedwordlist.com.

Usage

```
data(profanity_banned)
```

Format

A character vector with 77 elements

Disclaimer

From the original author: "These lists are free to download. You may use them for any purpose you wish and may copy, modify and distribute them freely. The swear words lists are provided "as-is" without any warranty or guarantee whatsoever. Don't blame me when the users of your forum, blog or community find more creative ways of offending people."

References

http://www.bannedwordlist.com

`profanity_racist` *Titus Wormer's List of Racist Words*

Description

A dataset containing a character vector of racist words from Titus Wormer.

Usage

```
data(profanity_racist)
```

Format

A character vector with 470 elements

License

References

https://github.com/words/profanities

`profanity_zac_anger` *Zac Anger's List of Profane Words*

Description

A dataset containing a character vector of profane words from Zac Anger.

Usage

```
data(profanity_zac_anger)
```

Format

A character vector with 3,076 elements

License

The original authors note the data allows the following: "Everyone is permitted to copy and distribute verbatim or modified copies of this license document, and changing it is allowed as long as the name is changed." https://github.com/zacanger/profane-words/blob/master/LICENSE.md

References

https://github.com/zacanger/profane-words

sw_dolch *Leveled Dolch List of 220 Common Words*

Description

Edward William Dolch's list of 220 Most Commonly Used Words by reading level.

Usage

```
data(sw_dolch)
```

Format

A character vector with 220 elements

Details

Dolch's Word List made up 50-75% of all printed text in 1936.

- Word. The word
- Level. The reading level of the word

References

Dolch, E. W. (1936). A basic sight vocabulary. Elementary School Journal, 36, 456-460.

sw_fry_100 *Fry's 100 Most Commonly Used English Words*

Description

A stopword list containing a character vector of stopwords.

Usage

```
data(sw_fry_100)
```

Format

A character vector with 100 elements

Details

Fry's Word List: The first 25 make up about one-third of all printed material in English. The first 100 make up about one-half of all printed material in English. The first 300 make up about 65% of all printed material in English.

References

Fry, E. B. (1997). Fry 1000 instant words. Lincolnwood, IL: Contemporary Books.

`sw_fry_1000` *Fry's 1000 Most Commonly Used English Words*

Description

A stopword list containing a character vector of stopwords.

Usage

```
data(sw_fry_1000)
```

Format

A character vector with 1000 elements

Details

Fry's 1000 Word List makes up 90% of all printed text.

References

Fry, E. B. (1997). Fry 1000 instant words. Lincolnwood, IL: Contemporary Books.

`sw_fry_200` *Fry's 200 Most Commonly Used English Words*

Description

A stopword list containing a character vector of stopwords.

Usage

```
data(sw_fry_200)
```

Format

A character vector with 200 elements

Details

Fry's Word List: The first 25 make up about one-third of all printed material in English. The first 100 make up about one-half of all printed material in English. The first 300 make up about 65% of all printed material in English.

References

Fry, E. B. (1997). Fry 1000 instant words. Lincolnwood, IL: Contemporary Books.

sw_fry_25 *Fry's 25 Most Commonly Used English Words*

Description

A stopword list containing a character vector of stopwords.

Usage

```
data(sw_fry_25)
```

Format

A character vector with 25 elements

Details

Fry's Word List: The first 25 make up about one-third of all printed material in English. The first 100 make up about one-half of all printed material in English. The first 300 make up about 65% of all printed material in English.

References

Fry, E. B. (1997). Fry 1000 instant words. Lincolnwood, IL: Contemporary Books.

sw_jockers *Matthew Jocker's Expanded Topic Modeling Stopword List*

Description

A dataset containing a character vector of Jocker's stopwords he used for topic modeling. He later resorted to eliminating everything but nouns: http://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/.

Usage

```
data(sw_jockers)
```

Format

A character vector with 5,902 elements

References

http://www.matthewjockers.net/materials/uwm-2013

`sw_loughran_mcdonald_long` *Loughran-McDonald Long Stopword List*

Description

A dataset containing a character vector of Loughran & McDonald's (2016) long stopword list.

Usage

```
data(sw_loughran_mcdonald_long)
```

Format

A character vector with 570 elements

License

The original authors note the data is available for non-commercial, research use: "The data compilations provided on this website are for use by individual researchers.". For more details see: https://sraf.nd.edu/textual-analysis/resources/#Master

Copyright

Copyright holder University of Notre Dame

References

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54(4), 1187-1230. doi: 10.2139/ssrn.2504147

`https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary`

`sw_loughran_mcdonald_short` *Loughran-McDonald Short Stopword List*

Description

A dataset containing a character vector of Loughran & McDonald's (2016) short stopword list.

Usage

```
data(sw_loughran_mcdonald_short)
```

Format

A character vector with 121 elements

License

The original authors note the data is available for non-commercial, research use: "The data compilations provided on this website are for use by individual researchers.". For more details see: https://sraf.nd.edu/textual-analysis/resources/#Master

References

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54(4), 1187-1230. doi: 10.2139/ssrn.2504147

`https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary`

`sw_lucene` *Lucene Stopword List*

Description

A dataset containing a character vector of Lucene's stopwords used in `StopAnalyzer.ENGLISH_STOP_WORDS_SE`.

Usage

```
data(sw_lucene)
```

Format

A character vector with 33 elements

Details

Lucene's License:

Licensed to the Apache Software Foundation (ASF) under one or more contributor license agreements. See the NOTICE file distributed with this work for additional information regarding copyright ownership. The ASF licenses this file to You under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

References

http://lucene.apache.org/core/4_0_0/analyzers-common/org/apache/lucene/analysis/core/StopFilter.html

---

`sw_mallet` *MALLET Stopword List*

Description

A stopword list containing a character vector of stopwords.

Usage

```
data(sw_mallet)
```

Format

A character vector with 523 elements

Details

From MAchine Learning for LanguagE Toolkit

Common Public License Version 1.0 (CPL) (NOTE: This license has been superseded by the Eclipse Public License)

(text)

"Program" means the Contributions distributed in accordance with this Agreement.

"Recipient" means anyone who receives the Program under this Agreement, including all Contributors.

2. GRANT OF RIGHTS

a) Subject to the terms of this Agreement, each Contributor hereby grants Recipient a non-exclusive, worldwide, royalty-free copyright license to reproduce, prepare derivative works of, publicly display, publicly perform, distribute and sublicense the Contribution of such Contributor, if any, and such derivative works, in source code and object code form.

b) Subject to the terms of this Agreement, each Contributor hereby grants Recipient a non-exclusive, worldwide, royalty-free patent license under Licensed Patents to make, use, sell, offer to sell, import and otherwise transfer the Contribution of such Contributor, if any, in source code and object code form. This patent license shall apply to the combination of the Contribution and the Program if, at the time the Contribution is added by the Contributor, such addition of the Contribution causes such combination to be covered by the Licensed Patents. The patent license shall not apply to any other combinations which include the Contribution. No hardware per se is licensed hereunder.

c) Recipient understands that although each Contributor grants the licenses to its Contributions set forth herein, no assurances are provided by any Contributor that the Program does not infringe the patent or other intellectual property rights of any other entity. Each Contributor disclaims any liability to Recipient for claims brought by any other entity based on infringement of intellectual property rights or otherwise. As a condition to exercising the rights and licenses granted hereunder, each Recipient hereby assumes sole responsibility to secure any other intellectual property rights needed, if any. For example, if a third party patent license is required to allow Recipient to distribute the Program, it is Recipient's responsibility to acquire that license before distributing the Program.

d) Each Contributor represents that to its knowledge it has sufficient copyright rights in its Contribution, if any, to grant the copyright license set forth in this Agreement.

3. REQUIREMENTS

A Contributor may choose to distribute the Program in object code form under its own license agreement, provided that:

a) it complies with the terms and conditions of this Agreement; and

b) its license agreement:

i) effectively disclaims on behalf of all Contributors all warranties and conditions, express and implied, including warranties or conditions of title and non-infringement, and implied warranties or conditions of merchantability and fitness for a particular purpose;

ii) effectively excludes on behalf of all Contributors all liability for damages, including direct, indirect, special, incidental and consequential damages, such as lost profits;

iii) states that any provisions which differ from this Agreement are offered by that Contributor alone and not by any other party; and

iv) states that source code for the Program is available from such Contributor, and informs licensees how to obtain it in a reasonable manner on or through a medium customarily used for software exchange.

When the Program is made available in source code form:

a) it must be made available under this Agreement; and

b) a copy of this Agreement must be included with each copy of the Program.

Contributors may not remove or alter any copyright notices contained within the Program.

Each Contributor must identify itself as the originator of its Contribution, if any, in a manner that reasonably allows subsequent Recipients to identify the originator of the Contribution.

4. COMMERCIAL DISTRIBUTION

Commercial distributors of software may accept certain responsibilities with respect to end users, business partners and the like. While this license is intended to facilitate the commercial use of the Program, the Contributor who includes the Program in a commercial product offering should do so in a manner which does not create potential liability for other Contributors. Therefore, if a Contributor includes the Program in a commercial product offering, such Contributor ("Commercial Contributor") hereby agrees to defend and indemnify every other Contributor ("Indemnified Contributor") against any losses, damages and costs (collectively "Losses") arising from claims, lawsuits and other legal actions brought by a third party against the Indemnified Contributor to the extent caused by the acts or omissions of such Commercial Contributor in connection with its distribution of the Program in a commercial product offering. The obligations in this section do not apply to any claims or Losses relating to any actual or alleged intellectual property infringement. In order to qualify, an Indemnified Contributor must: a) promptly notify the Commercial Contributor in writing of such claim, and b) allow the Commercial Contributor to control, and cooperate with the Commercial Contributor in, the defense and any related settlement negotiations. The Indemnified Contributor may participate in any such claim at its own expense.

For example, a Contributor might include the Program in a commercial product offering, Product X. That Contributor is then a Commercial Contributor. If that Commercial Contributor then makes performance claims, or offers warranties related to Product X, those performance claims and warranties are such Commercial Contributor's responsibility alone. Under this section, the Commercial Contributor would have to defend claims against the other Contributors related to those performance claims and warranties, and if a court requires any other Contributor to pay any damages as a result, the Commercial Contributor must pay those damages.

5. NO WARRANTY

EXCEPT AS EXPRESSLY SET FORTH IN THIS AGREEMENT, THE PROGRAM IS PROVIDED ON AN "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, EITHER EXPRESS OR IMPLIED INCLUDING, WITHOUT LIMITATION, ANY WARRANTIES OR CONDITIONS OF TITLE, NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Each Recipient is solely responsible for determining the appropriateness of using and distributing the Program and assumes all risks associated with its exercise of rights under this Agreement, including but not limited to the risks and costs of program errors, compliance with applicable laws, damage to or loss of data, programs or equipment, and unavailability or interruption of operations.

6. DISCLAIMER OF LIABILITY

EXCEPT AS EXPRESSLY SET FORTH IN THIS AGREEMENT, NEITHER RECIPIENT NOR ANY CONTRIBUTORS SHALL HAVE ANY LIABILITY FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING WITHOUT LIMITATION LOST PROFITS), HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OR DISTRIBUTION OF THE PROGRAM OR THE EXERCISE OF ANY RIGHTS GRANTED HEREUNDER, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. GENERAL

If any provision of this Agreement is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Agreement, and without further action by the parties hereto, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

If Recipient institutes patent litigation against a Contributor with respect to a patent applicable to software (including a cross-claim or counterclaim in a lawsuit), then any patent licenses granted by that Contributor to such Recipient under this Agreement shall terminate as of the date such litigation is filed. In addition, if Recipient institutes patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Program itself (excluding combinations of the Program with other software or hardware) infringes such Recipient's patent(s), then such Recipient's rights granted under Section 2(b) shall terminate as of the date such litigation is filed.

All Recipient's rights under this Agreement shall terminate if it fails to comply with any of the material terms or conditions of this Agreement and does not cure such failure in a reasonable period of time after becoming aware of such noncompliance. If all Recipient's rights under this Agreement terminate, Recipient agrees to cease use and distribution of the Program as soon as reasonably practicable. However, Recipient's obligations under this Agreement and any licenses granted by Recipient relating to the Program shall continue and survive.

Everyone is permitted to copy and distribute copies of this Agreement, but in order to avoid inconsistency the Agreement is copyrighted and may only be modified in the following manner. The Agreement Steward reserves the right to publish new versions (including revisions) of this Agreement from time to time. No one other than the Agreement Steward has the right to modify this Agreement. IBM is the initial Agreement Steward. IBM may assign the responsibility to serve as the Agreement Steward to a suitable separate entity. Each new version of the Agreement will be given a distinguishing version number. The Program (including Contributions) may always be distributed subject to the version of the Agreement under which it was received. In addition, after a new version of the Agreement is published, Contributor may elect to distribute the Program (including its Contributions) under the new version. Except as expressly stated in Sections 2(a) and 2(b) above, Recipient receives no rights or licenses to the intellectual property of any Contributor under this Agreement, whether expressly, by implication, estoppel or otherwise. All rights in the Program not expressly granted under this Agreement are reserved.

This Agreement is governed by the laws of the State of New York and the intellectual property laws of the United States of America. No party to this Agreement will bring a legal action under this Agreement more than one year after the cause of action arose. Each party waives its rights to a jury trial in any resulting litigation.

References

http://mallet.cs.umass.edu

`sw_python` *Python Stopword List*

Description

A dataset containing a character vector of Python's stopwords.

Usage

```
data(sw_python)
```

Format

A character vector with 174 elements

Details

Copyright (c) 2014, Alireza Savand, Contributors All rights reserved.

Original Author License: Copyright (c) 2014, Alireza Savand, Contributors All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

* Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

* Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

* Neither the name of the organization nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

References

https://pypi.python.org/pypi/stop-words

Index