Package 'doc2concrete'

June 19, 2020

Type Package

Title Measuring Concreteness in Natural Language

Version 0.4.6

Author Mike Yeomans

Maintainer Mike Yeomans <mk.yeomans@gmail.com>

Description Models for detecting concreteness in natural language. This package is built in support of Yeomans (2020) <doi:10.17605/OSF.IO/DYZN6>, which reviews linguistic models of concreteness in several domains. Here, we provide an implementation of the best-performing domain-general model (from Brysbaert et al., (2014) <doi:10.3758/s13428-013-0403-5>) as well as two pre-trained models for the feedback and plan-making domains.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Depends R (>= 2.10)

Imports tm, quanteda, ggplot2, parallel, glmnet, stringr, dplyr, english, textstem, SnowballC, textclean

RoxygenNote 7.1.0

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2020-06-19 11:00:02 UTC

R topics documented:

`adviceModel` *Pre-trained Concreteness Detection Model for Advice*

Description

This model was pre-trained on 3289 examples of feedback on different tasks (e.g. writing a cover letter, boggle, workplace annual reviews). All of those documents were annotated by research assistants for concreteness, and this model simulates those annotations on new documents.

Usage

```
adviceModel
```

Format

A pre-trained glmnet model

Source

Yeomans (2020). A Concrete Application of Open Science for Natural Language Processing.

`adviceNgrams` *Pre-trained advice concreteness features*

Description

For internal use only. This dataset demonstrates the ngram features that are used for the pre-trained adviceModel.

Usage

```
adviceNgrams
```

Format

A (truncated) matrix of ngram feature counts for alignment to the pre-trained advice glmnet model.

Source

Yeomans (2020). A Concrete Application of Open Science for Natural Language Processing.

bootstrap_list *Concreteness mTurk Word List*

Description

Word list from Paetzold & Specia (2016). A list of 85,942 words where concreteness was imputed using word embeddings.

Usage

```
bootstrap_list
```

Format

A data frame with 85,942 rows and 2 variables.

Word character text of a word with an entry in this dictionary

Conc.M predicted concreteness score for that word (from 100-700)

Source

#' Paetzold, G., & Specia, L. (2016, June). Inferring psycholinguistic properties of words. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 435-440).

doc2concrete *Concreteness Scores*

Description

Detects linguistic markers of concreteness in natural language. This function is the workhorse of the doc2concrete package, taking a vector of text documents and returning an equal-length vector of concreteness scores.

Usage

```
doc2concrete(
  texts,
  domain = c("open", "advice", "plans"),
  wordlist = NULL,
  stop.words = TRUE,
  number.words = TRUE,
  shrink = FALSE,
  fill = FALSE,
  num.mc.cores = 1
)
```

Arguments

texts	character A vector of texts, each of which will be tallied for concreteness.
domain	character Indicates the domain from which the text data was collected (see details).
wordlist	Dictionary to be used. Default is the Brysbaert et al. (2014) list.
stop.words	logical Should stop words be kept? Default is TRUE
number.words	logical Should numbers be converted to words? Default is TRUE
shrink	logical Should open-domain concreteness models regularize low-count words? Default is FALSE.
fill	logical Should empty cells be assigned the mean rating? Default is TRUE.
num.mc.cores	numeric number of cores for parallel processing - see parallel::detectCores(). Default is 1.

Details

In principle, concreteness could be measured from any english text. However, the definition and interpretation of concreteness may vary based on the domain. Here, we provide a domain-specific pre-trained classifier for concreteness in advice & feedback data, which we have empirically confirmed to be robust across a variety of contexts within that domain (Yeomans, 2020).

There are many domains where such pre-training is not yet possible. Accordingly, we provide support for two off-the-shelf concreteness "dictionaries" - i.e. document-level aggregations of word-level scores. We found that that have modest (but consistent) accuracy across domains and contexts. However, we still encourage researchers to train a model of concreteness in their own domain, if possible.

Value

A vector of concreteness scores, with one value for every item in ‘text‘.

References

Yeomans, M. (2020). Concreteness, Concretely. Working Paper.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46(3), 904-911.

Paetzold, G., & Specia, L. (2016, June). Inferring psycholinguistic properties of words. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 435-440).

Examples

```
data("feedback_dat")

doc2concrete(feedback_dat$feedback, domain="open")
```

```
cor(doc2concrete(feedback_dat$feedback, domain="open"),feedback_dat$concrete)
```

`feedback_dat` *Personal Feedback Dataset*

Description

A dataset containing responses from people on Mechanical Turk, writing feedback to a recent collaborator, that were then scored by other Turkers for feedback specificity.

Usage

```
feedback_dat
```

Format

A data frame with 171 rows and 2 variables:

feedback character text of feedback from writers

concrete numeric average specificity score from readers

Source

Blunden, H., Green, P., & Gino, F. (2018).

"The Impersonal Touch: Improving Feedback-Giving with Interpersonal Distance."

Academy of Management Proceedings, 2018.

`mturk_list` *Concreteness mTurk Word List*

Description

Word list from Brysbaert, Warriner & Kuperman (2014). A list of 39,954 words that have been hand-annotated by crowdsourced workers for concreteness.

Usage

```
mturk_list
```

Format

A data frame with 39,954 rows and 2 variables.

Word character text of a word with an entry in this dictionary

Conc.M average concreteness score for that word (from 1-5)

Source

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46(3), 904-911.

`planModel` *Pre-trained Concreteness Detection Model for Plan-Making*

Description

This model was pre-trained on 5,172 examples of pre-course plans from online courses at HarvardX. Each plan was annotated by research assistants for concreteness, and this model simulates those annotations on new plans.

Usage

```
planModel
```

Format

A pre-trained glmnet model

Source

Yeomans (2020). A Concrete Application of Open Science for Natural Language Processing.

`planNgrams` *Pre-trained plan concreteness features*

Description

For internal use only. This dataset demonstrates the ngram features that are used for the pre-trained planModel.

Usage

```
planNgrams
```

Format

A (truncated) matrix of ngram feature counts for alignment to the pre-trained planning glmnet model.

Source

Yeomans (2020). A Concrete Application of Open Science for Natural Language Processing.

Index