Package 'tidylo'

May 25, 2020

Type Package

Title Weighted Tidy Log Odds Ratio

Version 0.1.0

Description How can we measure how the usage or frequency of some feature, such as words, differs across some group or set, such as documents? One option is to use the log odds ratio, but the log odds ratio alone does not account for sampling variability; we haven't counted every feature the same number of times so how do we know which differences are meaningful? Enter the weighted log odds, which 'tidylo' provides an implementation for, using tidy data principles. In particular, here we use the method outlined in Monroe, Colaresi, and Quinn (2008) <doi:10.1093/pan/mpn018> to weight the log odds ratio by a prior. By default, the prior is estimated from the data itself, an empirical Bayes approach, but an uninformative prior is also available.

License MIT + file LICENSE

URL http://github.com/juliasilge/tidylo

BugReports http://github.com/juliasilge/tidylo/issues

Imports dplyr, rlang

Suggests covr, ggplot2, janeaustenr, knitr, rmarkdown, stringr, testthat (>= 2.1.0), tidytext

VignetteBuilder knitr

Encoding UTF-8

LazyData TRUE

RoxygenNote 7.1.0

NeedsCompilation no

Author Tyler Schnoebelen [aut], Julia Silge [aut, cre, cph] (<https://orcid.org/0000-0002-3671-836X>), Alex Hayes [aut] (<https://orcid.org/0000-0002-4985-5160>)

Maintainer Julia Silge <julia.silge@gmail.com>

Repository CRAN

Date/Publication 2020-05-25 19:10:03 UTC

R topics documented:

`bind_log_odds` *Bind the weighted log odds to a tidy dataset*

Description

Calculate and bind posterior log odds ratios, assuming a multinomial model with a Dirichlet prior. The Dirichlet prior parameters are set using an empirical Bayes approach by default, but an uninformative prior is also available. Assumes that data is in a tidy format, and adds the weighted log odds ratio as a column. Supports non-standard evaluation through the tidyeval framework.

Usage

```
bind_log_odds(tbl, set, feature, n, uninformative = FALSE, unweighted = FALSE)
```

Arguments

`tbl`	A tidy dataset with one row per `feature` and `set`.
`set`	Column of sets between which to compare features, such as documents for text data.
`feature`	Column of features for identifying differences, such as words or bigrams with text data.
`n`	Column containing feature-set counts.
`uninformative`	Whether or not to use an uninformative Dirichlet prior. Defaults to `FALSE`.
`unweighted`	Whether or not to return the unweighted log odds, in addition to the weighted log odds. Defaults to `FALSE`.

Details

The arguments `set`, `feature`, and `n` are passed by expression and support rlang::quasiquotation; you can unquote strings and symbols. Grouping is preserved but ignored.

The default empirical Bayes prior inflates feature counts in each group by total feature counts across all groups. This is like using a moment based estimator for the parameters of the Dirichlet prior. Note that empirical Bayes estimates perform well on average, but can have some surprising properties. If you are uncomfortable with empirical Bayes estimates, we suggest using the uninformative prior.

The weighted log odds computed by this function are also z-scores for the log odds; this quantity is useful for comparing frequencies across sets but its relationship to an odds ratio is not straightforward after the weighting.

The dataset must have exactly one row per set-feature combination for this calculation to succeed. Read Monroe et al (2008) for more on the weighted log odds ratio.

Value

The original tidy dataset with up to two additional columns.

- `weighted_log_odds`: The weighted posterior log odds ratio, where the odds ratio is for the feature distribution within that set versus all other sets. The weighting comes from variance-stabilization of the posterior.
- `log_odds` (optional, only returned if requested): The posterior log odds without variance stabilization.

References

1. Monroe, B. L., Colaresi, M. P. & Quinn, K. M. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Polit. anal. 16, 372-403 (2008). https://doi.org/10.1093/pan/mpn018
2. Minka, T. P. Estimating a Dirichlet distribution. (2012). https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf

Examples

```
library(dplyr)

gear_counts <- mtcars %>%
  count(vs, gear)

gear_counts

# find the number of gears most characteristic of each engine shape `vs`

regularized <- gear_counts %>%
  bind_log_odds(vs, gear, n)

regularized

unregularized <- gear_counts %>%
  bind_log_odds(vs, gear, n, uninformative = TRUE, unweighted = TRUE)

# these log odds will be farther from zero
# than the regularized estimates
unregularized
```

Index