Nicolas Proellochs

Sentiment Analysis

SentimentAnalysis performs a sentiment analysis of textual contents in R. This implementation utilizes various existing dictionaries, such as QDAP, Harvard IV or Loughran-McDonald. Furthermore, it can also create customized dictionaries. The latter uses LASSO regularization as a statistical approach to select relevant terms based on an exogenous response variable.

Overview

The most important functions in SentimentAnalysis are:

- Compute sentiment scores from contents stored in different formats with analyzeSentiment().
- If desired, convert the continuous scores to either binary sentiment classes (negative or positive) or tertiary directions (negative, neutral or positive). This conversion can be done with convertToBinary() or convertToDirection() respectively.
- Compare the calculated sentiment scores with a baseline (i.e. a gold standard). Here, compareToResponse() performs a statistical evaluation, while plotSentimentResponse() enables a visual comparison.
- Generate customized dictionaries with the help of generateDictionary() as part of an advanced analysis. However, this prerequisites a response variable (i.e. the baseline).

To see examples of these functions in use, check out the help pages, the demos and the vignette.

Usage

This section shows the basic functionality of how to perform a sentiment analysis. First, install the package from CRAN. Then load the corresponding package SentimentAnalysis.

```
# install.packages("SentimentAnalysis")

library(SentimentAnalysis)
```

Quick demonstration

This simple example shows how to perform a sentiment analysis of a single string. The result is a two-level factor with levels “positive” and “negative.”

```
# Analyze a single string to obtain a binary response (positive / negative)
sentiment <- analyzeSentiment("Yeah, this was a great soccer game of the German team!")
convertToBinaryResponse(sentiment)$SentimentGI
#> [1] positive
#> Levels: negative positive
```

Small example

The following demonstrates some of the functionality provided by SentimentAnalysis. It also shows its visualization and evaluation capabilities.

```
# Create a vector of strings
documents <- c("Wow, I really like the new lightsabers!",
```

Nicolas Proellochs

```
        "That book was excellent.",
        "R is a fantastic language.",
        "The service in this restaurant was miserable.",
        "This is neither positive or negative.",
        "The waiter forget about my a dessert -- what a poor service!")

# Analyze sentiment
sentiment <- analyzeSentiment(documents)

# Extract dictionary-based sentiment according to the QDAP dictionary
sentiment$SentimentQDAP
#> [1]  0.3333333  0.5000000  0.5000000 -0.3333333  0.0000000 -0.4000000

# View sentiment direction (i.e. positive, neutral and negative)
convertToDirection(sentiment$SentimentQDAP)
#> [1] positive positive positive negative neutral  negative
#> Levels: negative neutral positive

response <- c(+1, +1, +1, -1, 0, -1)

compareToResponse(sentiment, response)
#> Warning in cor(sentiment, response): the standard deviation is zero
#> Warning in cor(x, y): the standard deviation is zero

#> Warning in cor(x, y): the standard deviation is zero
#> Warning in cor(sentiment, response): the standard deviation is zero
#>                             WordCount  SentimentGI  NegativityGI PositivityGI
#> cor                       -0.18569534  0.990011498 -9.974890e-01  0.942954167
#> cor.t.statistic           -0.37796447 14.044046450 -2.816913e+01  5.664705543
#> cor.p.value                0.72465864  0.000149157  9.449687e-06  0.004788521
#> lm.t.value                -0.37796447 14.044046450 -2.816913e+01  5.664705543
#> r.squared                  0.03448276  0.980122766  9.949843e-01  0.889162562
#> RMSE                       3.82970843  0.450102869  1.186654e+00  0.713624032
#> MAE                        3.33333333  0.400000000  1.100000e+00  0.666666667
#> Accuracy                   0.66666667  1.000000000  6.666667e-01  0.666666667
#> Precision                         NaN  1.000000000           NaN          NaN
#> Sensitivity                0.00000000  1.000000000  0.000000e+00  0.000000000
#> Specificity                1.00000000  1.000000000  1.000000e+00  1.000000000
#> F1                                NaN  1.000000000           NaN          NaN
#> BalancedAccuracy           0.50000000  1.000000000  5.000000e-01  0.500000000
#> avg.sentiment.pos.response  3.25000000  0.333333333  8.333333e-02  0.416666667
#> avg.sentiment.neg.response  4.00000000 -0.633333333  6.333333e-01  0.000000000
#>                             SentimentHE NegativityHE PositivityHE SentimentLM
#> cor                           0.4152274 -0.083045480    0.3315938   0.7370455
#> cor.t.statistic               0.9128709 -0.166666667    0.7029595   2.1811142
#> cor.p.value                   0.4129544  0.875718144    0.5208394   0.0946266
#> lm.t.value                    0.9128709 -0.166666667    0.7029595   2.1811142
#> r.squared                     0.1724138  0.006896552    0.1099545   0.5432361
#> RMSE                          0.8416254  0.922958207    0.8525561   0.7234178
#> MAE                           0.7500000  0.888888889    0.8055556   0.6333333
#> Accuracy                      0.6666667  0.666666667    0.6666667   0.8333333
```

```
#> Precision                       NaN         NaN         NaN   1.0000000
#> Sensitivity                   0.0000000  0.000000000    0.0000000   0.5000000
#> Specificity                   1.0000000  1.000000000    1.0000000   1.0000000
#> F1                              NaN         NaN         NaN   0.6666667
#> BalancedAccuracy              0.5000000  0.500000000    0.5000000   0.7500000
#> avg.sentiment.pos.response    0.1250000  0.083333333    0.2083333   0.2500000
#> avg.sentiment.neg.response    0.0000000  0.000000000    0.0000000  -0.1000000
#>                         NegativityLM PositivityLM RatioUncertaintyLM
#> cor                     -0.40804713    0.6305283               NA
#> cor.t.statistic         -0.89389841    1.6247248               NA
#> cor.p.value              0.42189973    0.1795458               NA
#> lm.t.value              -0.89389841    1.6247248               NA
#> r.squared                0.16650246    0.3975659               NA
#> RMSE                     0.96186547    0.7757911        0.9128709
#> MAE                      0.92222222    0.7222222        0.8333333
#> Accuracy                 0.66666667    0.6666667        0.6666667
#> Precision                       NaN         NaN              NaN
#> Sensitivity              0.00000000    0.0000000        0.0000000
#> Specificity              1.00000000    1.0000000        1.0000000
#> F1                              NaN         NaN              NaN
#> BalancedAccuracy         0.50000000    0.5000000        0.5000000
#> avg.sentiment.pos.response   0.08333333    0.3333333        0.0000000
#> avg.sentiment.neg.response   0.10000000    0.0000000        0.0000000
#>                         SentimentQDAP NegativityQDAP PositivityQDAP
#> cor                    0.9865356369   -0.944339551    0.942954167
#> cor.t.statistic         12.0642877257   -5.741148345    5.664705543
#> cor.p.value              0.0002707131    0.004560908    0.004788521
#> lm.t.value              12.0642877257   -5.741148345    5.664705543
#> r.squared                0.9732525629    0.891777188    0.889162562
#> RMSE                     0.5398902495    1.068401367    0.713624032
#> MAE                      0.4888888889    1.011111111    0.666666667
#> Accuracy                 1.0000000000    0.666666667    0.666666667
#> Precision                1.0000000000           NaN            NaN
#> Sensitivity              1.0000000000    0.000000000    0.000000000
#> Specificity              1.0000000000    1.000000000    1.000000000
#> F1                      1.0000000000           NaN            NaN
#> BalancedAccuracy         1.0000000000    0.500000000    0.500000000
#> avg.sentiment.pos.response  0.3333333333    0.083333333    0.416666667
#> avg.sentiment.neg.response -0.3666666667    0.366666667    0.000000000

# Optional visualization: plotSentimentResponse(sentiment$SentimentQDAP, response)
```

Dictionary generation

Research in finance and social sciences nowadays utilizes content analysis to understand human decisions in the face of textual materials. While content analysis has received great traction lately, the available tools are not yet living up to the needs of researchers. This package implements a novel approach named "**dictionary generation" to study tone, sentiment and reception of textual materials.
The approach utilizes LASSO regularization to extract words from documents that statistically feature a positive and negative polarity. This immediately reveals manifold implications for practitioners, finance

research and social sciences: researchers can use R to extract text components that are relevant for readers and test their hypothesis based on these.

Proellochs, Feuerriegel and Neumann (2018): Statistical inferences for polarity identification in natural language, PLOS ONE 13(12):e0209323. DOI: 10.1371/journal.pone.0209323

Proellochs, Feuerriegel and Neumann (2015): Generating Domain-Specific Dictionaries Using Bayesian Learning, Proceedings of the 23rd European Conference on Information Systems (ECIS 2015), Muenster, Germany. DOI: 10.2139/ssrn.2522884

Article Link

https://www.rdocumentation.org/packages/SentimentAnalysis/versions/1.3-5