Package 'tableParser'

August 1, 2025

Title Parse Tabled Content

Date 2025-07-30

Version 1.0.1

Maintainer Ingmar Böschen <ingmar.boeschen@uni-hamburg.de>

Description Functions to parse content from HTML-encoded tables to a human readable text format by simulating the experience of a screen reader for visually impaired users. 'tableParser' contains several functions to work with HTML-encoded tables, as well as native character matrixes. The functions *table2matrix()*, table2text() and table2stats() can be applied on documents in HTML, HML, XML, CERMXML, as well as DOCX and PDF file format. The table extraction from DOCX files is performed with the function docx2matrix(), tables in PDF documents are extracted with the 'tabulapdf' package. Table captions and footers cannot be extracted from tables in DOCX and PDF documents, which reduces the decoding capabilities. The textual representation of characters in matrix content can be unified with unifyMatrix() before parsing. The function table2stats() extracts tabled statistical results. The function then unifies the parsed text and processes it with JATSdecoder::standardStats() to extract all statistical standard results and, if possible, check the reported and coded p-values for consistency. Due to the variability and complexity of table structures, parsing accuracy may vary. To minimize parsing errors, it is recommended to work with simple, accessible, barrier-free table structures.

Depends R (>= 4.1)

Imports utils,
tabulapdf,
xml2,
JATSdecoder

License GPL-3

URL https://github.com/ingmarboeschen/tableParser

BugReports https://github.com/ingmarboeschen/tableParser/issues

Language en-US

Encoding UTF-8

RoxygenNote 7.3.2

R topics documented:

`docx2matrix` *docx2matrix*

Description

Extracts tables from DOCX documents and returns list of character matrices.

Usage

```
docx2matrix(x, replicate = TRUE)
```

Arguments

`x`	File path to a DOCX input file with tables.
`replicate`	Logical. If TRUE, replicates content when splitting connected cells.

Value

List with extracted tables as character matrices.

get.caption *get.caption*

Description

Extracts the content of <caption>-tag/s.

Usage

```
get.caption(x, rm.html = TRUE, sentences = FALSE, letter.convert = TRUE)
```

Arguments

x	A vector with HTML coded tables.
rm.html	logical. If TRUE, all HTML tags are removed, <sub> converts to '_', <sup> to '^'.
sentences	logical. If TRUE, a sentence vector is returned.
letter.convert	logical. If TRUE, hexadecimal letters are converted to Unicode und unified with JATSdecoder::letter.convert.

Value

A character vector with the extracted caption text and NULL for no caption text

get.footer *get.footer*

Description

Extracts the content of <table-wrap-foot>-tag/s.

Usage

```
get.footer(x, rm.html = TRUE, sentences = FALSE, letter.convert = TRUE)
```

Arguments

x	A vector with HTML coded tables.
rm.html	logical. If TRUE, all HTML tags are removed, <sub> converts to '_', <sup> to '^'.
sentences	logical. If TRUE, a sentence vector is returned.
letter.convert	logical. If TRUE, hexadecimal letters are converted to Unicode und unified with JATSdecoder::letter.convert.

Value

A character vector with the extracted footer text and NULL for no footer text

get.HTML.tables *get.HTML.tables*

Description

Extracts HTML tables as vector of HTML coded tables from plain HTML code, HTML, HML, XML or CERMXML files.

Usage

```
get.HTML.tables(x)
```

Arguments

x	HTML, HML, XML or CERMXML file or character object with HTML-encoded content.

Value

Character vector with one HTML-encoded table per cell.

legendCodings *legendCodings*

Description

Extracts the coding of p-values, brackets, abbreviations, superscripts and the reported sample size/s with 'N=number' from tables caption and footer notes/text.

Usage

```
legendCodings(x)
```

Arguments

x	An HTML coded table or plain textual input of table caption and/or footer text.

Value

A list with detected p-value and superscript codings, abbreviations and reported sample size/s.

Examples

```
x<-"+ p>.05, ^**p<.01, SSq, Sum of Squares, ^a t-test, n=120.
POS: perceived organizational support, JP; job performance.
Numbers in parenthesis are standard errors.
Bold values indicate significance at p<.05."
legendCodings(x)
```

matrix2text *matrix2text*

Description

Converts character matrices to text.

Usage

```
matrix2text(
  x,
  legend = NULL,
  unifyMatrix = TRUE,
  correctComma = FALSE,
  expandAbbreviations = TRUE,
  superscript2bracket = TRUE,
  standardPcoding = FALSE,
  addDF = TRUE,
  rotate = FALSE,
  split = FALSE
)
```

Arguments

x	A character matrix or list of character matrices.
legend	A list with table legend codes extracted from table caption and/or footer with tableParser::legendCodings().
unifyMatrix	Logical. If TRUE, matrix cells are unified for better post processing.
correctComma	Logical. If TRUE and unifyMatrix=TRUE, decimal sign commas are converted to dots.
expandAbbreviations	Logical. If TRUE, detected abbreviations are expanded to label detected in table caption/footer with tableParser::legendCodings().
superscript2bracket	Logical. If TRUE, detected superscript codings are inserted inside parentheses.
standardPcoding	Logical. If TRUE, and no other detection of p-value coding is detected, standard coding of p-values is assumed to be: * p<.05, ** p<.01 and *** p<.001.
addDF	Logical. If TRUE, detected sample size N in caption/footer is inserted as degrees of freedom (N-2) to r- and t-values that are reported without degrees of freedom.
rotate	Logical. If TRUE, matrix/matrices is rotated before parsing.
split	Logical. If TRUE, matrix/matrices are split for multi-model tables.

Value

Character vector with a parsed and human readable form of the input table. The result vector can be further processed with standardStats() to extract and structure the statistical standard test results only.

Examples

```
# some random data
x<-rnorm(100)
y<-x+rnorm(100)
# a model result table...
mod<-round(summary(lm(y~x))$coefficients,3)
rnames<-c("",rownames(mod))
cnames<-colnames(mod)
mod<-rbind(cnames,mod)
mod<-cbind(rnames,mod)
# ...as character result matrix
x<-unname(mod);x
## parse matrix to text vector
# -as is
matrix2text(x,unifyMatrix=FALSE)
# -with unified content
matrix2text(x,unifyMatrix=TRUE)
## processing of a matrix with two header lines
x<-rbind(c("","A","A","B","B"),x);x
matrix2text(x,unifyMatrix=FALSE)
## processing of a matrix with two header lines and naming columns
x<-cbind(c("","","C","D"),x);x
matrix2text(x,unifyMatrix=FALSE)
```

---

parseMatrixContent *parseMatrixContent*

Description

Parses character matrix content into a text vector. This is the basic function of tableParser, which is implemented in matrix2text(), table2text() and table2stats(). Row and column names are parsed to cell content with operators, that depend on the cell content. Numeric cells are parsed with "=", text cells with ":". Detected codings for (e.g. p-values, abbreviations) from tables legend text can be used to extend the tabled content to a fully written out form.

Usage

```
parseMatrixContent(
  x,
  legend = NULL,
  standardPcoding = TRUE,
  expandAbbreviations = TRUE,
```

```
  superscript2bracket = FALSE,
  addDF = TRUE
)
```

Arguments

x	A character matrix or list with a character matrix as first and only element.
legend	The tables caption/footer notes as character vector.
standardPcoding	Logical. If TRUE, and no other detection of p-value coding is detected, standard coding of p-values is assumed to be: * p<.05, ** p<.01 and *** p<.001.
expandAbbreviations	Logical. If TRUE, detected abbreviations are expanded to label detected in table caption/footer with tableParser::legendCodings().
superscript2bracket	Logical. If TRUE, detected superscript codings are inserted inside parentheses.
addDF	Logical. If TRUE, detected sample size N in caption/footer is inserted as degrees of freedom (N-2) to r- and t-values that are reported without degrees of freedom.

Value

A text vector with the parsed matrix content.

`table2matrix` *table2matrix*

Description

Extracts tables from HTML, HML, XML, CERMXML, DOCX, PDF files or plain HTML code to a list of character matrices.

Usage

```
table2matrix(
  x,
  unifyMatrix = FALSE,
  letter.convert = TRUE,
  greek2text = FALSE,
  replicate = FALSE,
  repNums = FALSE,
  rm.html = FALSE,
  rm.empty.row.col = FALSE,
  collapseHeader = TRUE,
  header2colnames = FALSE
)
```

Arguments

x	File path to a DOCX, PDF or HTML-encoded file, or text with HTML code.
unifyMatrix	Logical. If TRUE, matrix cells are unified for better post processing (see: unifyMatrixContent()).
letter.convert	Logical. If TRUE hex codes will be unified and converted to Unicode with JATSdecoder::letter.convert().
greek2text	Logical. If TRUE and 'letter.convert=TRUE', converts and unifies various Greek letters to a text based form (e.g. 'alpha', 'beta').
replicate	Logical. If TRUE the content of cells with row/col span > 1 are replicated in all connected cells, if FALSE, the value will only be placed to the first of the connected cell.
repNums	Logical. If TRUE cells with numbers, that have row/col span > 1 are replicated in every connected cell.
rm.html	Logical. If TRUE all HTML tags are removed, except <sub> and <sup> , </break> is converted to space.
rm.empty.row.col	Logical. If TRUE empty rows/columns are removed from output.
collapseHeader	Logical. If TRUE header cells are collapsed for each column if header has 2 or more lines.
header2colnames	Logical. If TRUE and 'collapseHeader=TRUE' first table row is used for column names and removed from table.

Value

List with detected tables as character matrices.

Examples

```
x<-readLines("https://en.wikipedia.org/wiki/R_(programming_language)",warn=FALSE)
table2matrix(x,rm.html=TRUE)
```

---

table2stats *table2stats*

Description

Extracts tabulated statistical results from documents in XML, HTML, HML, DOCX or PDF format.

Usage

```
table2stats(
  x,
  standardPcoding = FALSE,
  correctComma = FALSE,
  expandAbbreviations = TRUE,
  superscript2bracket = TRUE,
  stats.mode = "all",
  checkP = FALSE,
  alpha = 0.05,
  criticalDif = 0.02,
  alternative = "undirected",
  estimateZ = FALSE,
  T2t = FALSE,
  addTableName = TRUE,
  rm.na.col = TRUE
)
```

Arguments

x	Input. Either a file path to an XML, HTML, HML, DOCX or PDF file or matrix object or vector of plain HTML coded tables.
standardPcoding	Logical. If TRUE, and no other detection of coding is detected, then standard coding of p-values is assumed to be * p<.05, ** p<.01 and ***p<.001.
correctComma	Logical. If TRUE, decimal sign commas are converted to dots.
expandAbbreviations	Logical. If TRUE, detected abbreviations are expanded to label from table caption/footer.
superscript2bracket	Logical. If TRUE, detected superscript codings are inserted inside parentheses.
stats.mode	Select a subset of test results by p-value checkability for output. One of: c("all", "checkable", "computable", "uncomputable").
checkP	Logical. If TRUE, detected p-values and recalculated p-values will be checked for consistency.
alpha	Numeric. Defines the alpha level to be used for error assignment.
criticalDif	Numeric. Sets the absolute maximum difference in reported and recalculated p-values for error detection.
alternative	Character. Select test sidedness for recomputation of p-values from t-, r- and beta-values. One of c("undirected", "directed"). If "directed" is specified, p-values for directed null-hypothesis are added to the table but still require a manual inspection on consistency of the direction.
estimateZ	Logical. If TRUE, detected beta-/d-values are divided by the reported standard error "SE" to estimate Z-values ("Zest") for observed beta/d and computation of p-values. Note: This is only valid, if Gauss-Marcov assumptions are met and a sufficiently large sample size is used. If a Z- or t-value is detected in a report of

	a beta-/d-coefficient with SE, no estimation will be performed, although set to TRUE.
T2t	Logical. If TRUE, capital letter T is treated as t-statistic.
addTableName	Logical. If TRUE, table number is added in front of the eaxtracted results.
rm.na.col	Logical. If TRUE, removes all columns with only NA.

Value

A data.frame object with the extracted statistical standard results, recalculated p-values and a rudimentary, optional consistency check for reported p-values (if 'checkP=TRUE').

`table2text` *table2text*

Description

Parses tabled content from HTML coded content or HTML, DOCX or PDF file to human readable text vector. Before parsing, header lines are collapsed and connected cells are broken up.

Usage

```
table2text(
  x,
  unifyMatrix = TRUE,
  unifyStats = FALSE,
  expandAbbreviations = TRUE,
  superscript2bracket = TRUE,
  standardPcoding = FALSE,
  addDF = TRUE,
  correctComma = FALSE,
  addDescription = TRUE,
  addTableName = TRUE
)
```

Arguments

x	A vector with HTML tables or a single file path to an HTML, XML, CERMXML, HML, PDF or DOCX file.
unifyMatrix	Logical. If TRUE, matrix cells are unified for better post processing.
unifyStats	Logical. If TRUE, output is unified for better post processing (e.g.: "p-value"->"p").
expandAbbreviations	Logical. If TRUE, detected abbreviations are expanded to label from table caption/footer.
superscript2bracket	Logical. If TRUE, detected superscript codings are inserted inside parentheses.

`standardPcoding`
: Logical. If TRUE, and no other detection of coding is detected, standard coding of p-values is assumed to be * p<.05, ** p<.01 and ***p<.001.

`addDF`
: Logical. If TRUE, detected sample size N in caption/footer is inserted as degrees of freedom (N-2) to r- and t-values that are reported without degrees of freedom.

`correctComma`
: Logical. If TRUE and unifyMatrix=TRUE, decimal sign commas are converted to dots.

`addDescription`
: Logical. If TRUE, table caption and footer are added before the extracted table content for better readability.

`addTableName`
: Logical. If TRUE, table number is added before the parsed text lines.

Value

List with parsed tabled content as elements. The text vector in each list element can be further processed with JATSdecoder::standardStats() to extract and structure the statistical standard test results.

`tableClass` *tableClass*

Description

Classifies matrix content to either 'tabled results', 'correlation', 'matrix', 'text', 'vector', 'model with model statistics', 'multi model with model statistics'.

Usage

```
tableClass(x, legend = NULL)
```

Arguments

`x`
: A character matrix

`legend`
: A text vector with tables caption and/or footer.

Value

A character object of length=1 with the tables class.

unifyMatrixContent *unifyMatrixContent*

Description

Unifies textual and numerical content of character matrices. Unifies hyphens, spaces, hexadecimal and greek letters and performs space and comma corrections. Big marks in numbers are removed. HTML tags <sup> and <sub> are converted to '^' and '_' respectively. All other HTML tags are removed.

Usage

```
unifyMatrixContent(
  x,
  letter.convert = TRUE,
  greek2text = TRUE,
  text2num = TRUE,
  correctComma = FALSE
)
```

Arguments

x	a character matrix.
letter.convert	Logical. If TRUE hexadecimal coded letters will be unified and converted to Unicode with JATSdecoder::letter.convert().
greek2text	Logical. If TRUE and 'letter.convert=TRUE', converts and unifies various Greek letters to a text based form (e.g. 'alpha', 'beta').
text2num	Logical. If TRUE, textual representations of numbers (words, exponents, fractions) are converted to digit numbers.
correctComma	Logical. If TRUE, commas used as numeric separator are converted to dots.

unifyStats *unifyStats*

Description

Unifies many textual representations of statistical results in text vectors created with table2text().
This uniformisation is needed for a more precise extraction of standard results with JATSdecoder::standardStats().

Usage

```
unifyStats(x)
```

Arguments

x	A text string with parsed tabled results.

Value

A unified text string.

Index