Package 'tableParser'

June 4, 2025

Title Parse Tabled Content

Date 2025-05-07

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 appled 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 table2matrix(), tables in PDF documents are extracted with the 'tabulapdf' package. The textual representation of characters in matrix content can be unified with unifyMatrix() before parsing. The function table2stats() extracts tabled statistical results. The function further unifies the parsed text, which is then processed with JATSdecoder::standardStats(), in order to extract all statistical standard results and check the reported p-values for consistency. Due to the variability in table structures and complexity, parsing accuracy may vary. For best results, it is recommended to work with simple, accessible, and barrier-free table structures to minimize parsing errors.

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 return list of character matrices.

Usage

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

Arguments

`x`	File path of a docx input file.
`replicate`	Logical. If TRUE, replicates content when splitting connected cells.

Value

List with extracted matrices.

flattenList *flattenList flatten multi level list to simple list*

Description

flattenList flatten multi level list to simple list

Usage

```
flattenList(x)
```

Arguments

x	a list with listed elements

Value

single level list

get.caption *get.caption Extracts the content of a <caption>-tag.*

Description

get.caption Extracts the content of a <caption>-tag.

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 Extracts the content of <table-wrap-footer>-tag.*

Description

get.footer Extracts the content of <table-wrap-footer>-tag.

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.tables *get.tables*

Description

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

Usage

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

Arguments

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

Value

Character vector with one plain HTML-encoded table per cell.

legendCodings *legendCodings*

Description

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

Usage

```
legendCodings(x)
```

Arguments

x	An HTML coded table or plain textual input.

Value

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

matrix2text *matrix2text*

Description

Convert character matrices to text.

Usage

```
matrix2text(
  x,
  legend = NULL,
  unifyMatrix = TRUE,
  expandAbbreviations = TRUE,
  standardPcoding = FALSE,
  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.
expandAbbreviations	Logical. If TRUE, detected abbreviations are expanded to label detected in table caption/footer with tableParser::legendCodings().
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.
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 straight forward 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)
x<-unname(mod)
# ...as character result matrix
# parse matrix to text
matrix2text(x,unifyMatrix=FALSE)
```

---

parseMatrixContent *parseMatrixContent*

Description

Function to parse content from a character matrix into a text vector. This is the basic funtion of tableParser, that is implementent in matrix2text() and table2text()

Usage

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

Value

A text vector.

---

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.rows = 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.rows	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 HTML tables as matrices.

Examples

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

---

table2stats *table2stats*

Description

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

Usage

```
table2stats(
  x,
  standardPcoding = FALSE,
  expandAbbreviations = 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 filepath 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.
expandAbbreviations	Logical. If TRUE, detected abbreviations are expanded to label from table caption/footer.
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 and 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 text.

Usage

```
table2text(
  x,
  unifyMatrix = TRUE,
  unifyStats = FALSE,
  expandAbbreviations = TRUE,
  standardPcoding = FALSE,
  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 unifiedfor 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.
`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.
`addTableName`	Logical. If TRUE, table number is added before the parsed text lines.

Value

A List with parsed table content per HTML table. The result vector in each list element can be further processed with standardStats() to extract and structure the statistical standard test results only.

`tableClass` *tableClass Classifies matrix content to either 'tabled content', 'correlation', or 'text'*

Description

tableClass Classifies matrix content to either 'tabled content', 'correlation', or 'text'

Usage

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

Arguments

`x`	A character matrix
`legend`	A text string from tables caption and/or footer

Value

A character object with the tables class.

`unifyMatrixContent` *unifyMatrixContent Unifies content of character matrices. E.g.: comas in big numbers and HTML tags are removed. Performs space corrections and unifies hyphens and spaces.*

Description

unifyMatrixContent Unifies content of character matrices. E.g.: comas in big numbers and HTML tags are removed. Performs space corrections and unifies hyphens and spaces.

Usage

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

Arguments

x	a character matrix.
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').
text2num	Logical. If TRUE, textual representations of numbers (words, exponents, fractions) are converted to digit numbers.

---

unifyStats *unifyStats Unifies textual representations of statistical results.*

Description

unifyStats Unifies textual representations of statistical results.

Usage

```
unifyStats(x)
```

Arguments

x	A text string as vector.

Value

A unified text string.

Index