Package 'docxtractr'

August 29, 2015

Title Extract Data Tables from Microsoft Word Documents

Version 0.1.0.9000

Maintainer Bob Rudis <bob@rudis.net>

Description Microsoft Word docx files provide an XML structure that is fairly straightforward to navigate, especially when it applies to Word tables. The docxtractr package provides tools to determine table count/structure and extract/clean tables from Microsoft Word docx documents.

Depends R (>= 3.0.0)

License MIT + file LICENSE

LazyData true

Suggests testthat

Imports tools, xml2, dplyr, utils

NeedsCompilation no

Author Bob Rudis [aut, cre]

Repository CRAN

Date/Publication 2015-08-29 13:15:01

R topics documented:

assign_colnames *Make a specific row the column names for the specified data.frame*

Description

Many tables in Word documents are in twisted formats where there may be labels or other oddities mixed in that make it difficult to work with the underlying data. This function makes it easy to identify a particular row in a scraped `data.frame` as the one containing column names and have it become the column names, removing it and (optionally) all of the rows before it (since that's usually what needs to be done).

Usage

```
assign_colnames(dat, row, remove = TRUE, remove_previous = remove)
```

Arguments

`dat`	can be any `data.frame` but is intended for use with ones retuned by this package
`row`	numeric value indicating the row number that is to become the column names
`remove`	remove row specified by `row` after making it the column names? (Default: `TRUE`)
`remove_previous`	remove any rows preceeding `row`? (Default: `TRUE` but will be assigned whatever is given for `remove`).

Value

`data.frame`

See Also

`docx_extract_all`, `docx_extract_tbl`

Examples

```
# a "real" Word doc
real_world <- read_docx(system.file("examples/realworld.docx", package="docxtractr"))
docx_tbl_count(real_world)

# get all the tables
tbls <- docx_extract_all(real_world)

# make table 1 better
assign_colnames(tbls[[1]], 2)

# make table 5 better
assign_colnames(tbls[[5]], 2)
```

docxtractr *docxtractr is an R pacakge for extracting tables out of Word documents (docx)*

Description

Microsoft Word docx files provide an XML structure that is fairly straightforward to navigate, especially when it applies to Word tables. The docxtractr package provides tools to determine table count + table structure and extract tables from Microsoft Word docx documents.

Author(s)

Bob Rudis (@hrbrmstr)

docx_describe_tbls *Returns a description of all the tables in the Word document*

Description

This function will attempt to discern the structure of each of the tables in `docx` and print this information

Usage

```
docx_describe_tbls(docx)
```

Arguments

`docx`	`docx` object read with `read_docx`

Examples

```
complx <- read_docx(system.file("examples/complex.docx", package="docxtractr"))
docx_tbl_count(complx)
docx_describe_tbls(complx)
```

docx_extract_all *Extract all tables from a Word document*

Description

This function makes no assumptions about an

Usage

```
docx_extract_all(docx, guess_header = TRUE, trim = TRUE)
```

Arguments

docx	docx object read with read_docx
guess_header	should the function make a guess as to the existense of a header in a table? (Default: TRUE)
trim	trim leading/trailing whitespace (if any) in cells? (default: TRUE)

Value

list of data.frames or an empty list if no tables exist in docx

See Also

assign_colnames, docx_extract_tbl

Examples

```
# a "real" Word doc

real_world <- read_docx(system.file("examples/realworld.docx", package="docxtractr"))
docx_tbl_count(real_world)

# get all the tables
tbls <- docx_extract_all(real_world)
```

docx_extract_tbl *Extract a table from a Word document*

Description

Given a document read with read_docx and a table to extract (optionally indicating whether there was a header or not and if cell whitepace trimming is desired) extract the contents of the table to a data.frame.

Usage

```
docx_extract_tbl(docx, tbl_number = 1, header = TRUE, trim = TRUE)
```

Arguments

docx	docx object read with read_docx
tbl_number	which table to extract (defaults to 1)
header	assume first row of table is a header row? (default; TRUE)
trim	trim leading/trailing whitespace (if any) in cells? (default: TRUE)

Value

data.frame

See Also

docx_extract_all, docx_extract_tbl, assign_colnames

Examples

```
doc3 <- read_docx(system.file("examples/data3.docx", package="docxtractr"))
docx_extract_tbl(doc3, 3)
```

---

docx_tbl_count *Get number of tables in a Word document*

Description

Get number of tables in a Word document

Usage

```
docx_tbl_count(docx)
```

Arguments

docx	docx object read with read_docx

Value

numeric

Examples

```
complx <- read_docx(system.file("examples/complex.docx", package="docxtractr"))
docx_tbl_count(complx)
```

print.docx *Display information about the document*

Description

Display information about the document

Usage

```
## S3 method for class 'docx'
print(x, ...)
```

Arguments

x	docx object
...	ignored

read_docx *Read in a Word document for table extraction*

Description

Local file path or URL pointing to a .docx file.

Usage

```
read_docx(path)
```

Arguments

path	path to the Word document

Examples

```
doc <- read_docx(system.file("examples/data.docx", package="docxtractr"))
class(doc)
## Not run:
# from a URL
budget <- read_docx(
"http://www.anaheim.net/docs_agend/questys_pub/MG41925/AS41964/AS41967/AI44538/DO44539/1.DOCX")

## End(Not run)
```

Index