Package: readOffice (via r-universe)

August 21, 2024

Type Package

Title Read Text Out of Modern Office Files

Version 0.3.0

Author Mark Ewing

Maintainer Mark Ewing <b.mark@ewingsonline.com>

URL https://github.com/bmewing/readOffice

BugReports https://github.com/bmewing/readOffice/issues

Description Reads in text from 'unstructured' modern Microsoft Office files (XML based files) such as Word (.docx) and PowerPoint (.pptx). This does not read in structured data (from Excel or Access) as there are many other great packages to that do so already.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports xml2 (>= 1.0.0), rvest (>= 0.3.2), purrr (>= 0.2.2), magrittr (>= 1.5)

RoxygenNote 6.0.1

Repository https://bmewing.r-universe.dev

RemoteUrl https://github.com/bmewing/readoffice

RemoteRef HEAD

RemoteSha 77ff96bb826dc7a2fedd638ad09d1204ffd35459

Contents

`read_docx` *Read data from a Modern Word File*

Description

Read data from a Modern Word File

Usage

```
read_docx(docx, tables = T, drawings = T, diagrams = T)
```

Arguments

`docx`	The .docx file to read
`tables`	Should tables be processed from the document?
`drawings`	Should drawings be processed from the document?
`diagrams`	Should diagrams be processed from the document?

Details

Only accepts one file at a time and only .docx files. Modifying file extensions will not work.

List is comprised of named elements, one per 'section' (sections are recognized after a page break). If tables exist in the document and are processed, then the named list elements will be lists containing the text of paragraphs, drawings (if present and processed) and matrices holding the table structure. Otherwise, the list elements will contain vectors of the text processed.

Diagrams are typically what Microsoft calls 'SmartArt'

Value

Named list with document contents

Examples

```
read_docx(docx = system.file('extdata','example.docx',package='readOffice'))
read_docx(docx = system.file('extdata','example.docx',package='readOffice'),diagrams=FALSE)
```

read_pptx *Read data from a Modern PowerPoint File*

Description

Read data from a Modern PowerPoint File

Usage

```
read_pptx(pptx, tables = T, drawings = T, diagrams = T)
```

Arguments

pptx	The .pptx file to read
tables	Should tables be processed from the document?
drawings	Should drawings be processed from the document?
diagrams	Should diagrams be processed from the document?

Details

Only accepts one file at a time and only .pptx files. Modifying file extensions will not work.

The returned list contains named lists of the elements on the slide, each element of which is either a data.frame or a matrix containing the text and minor details about the structure on the page.

Data frames will contain the text in addition to the following columns: "Bulleted" indicates if the text is part of a bulleted or numbered list on the slide. "Hierarchy" indicates the tabbed depth of the element in a bulleted or numbered list (NA if not bulleted).

Alternatively, returns a matrix for tables on the slide.

Value

List containing slide elements.

Examples

```
read_pptx(system.file('extdata','example.pptx',package='readOffice'))
read_pptx(system.file('extdata','example.pptx',package='readOffice'),diagrams=FALSE)
```

Index