Package 'readtext'

May 22, 2017

Version 0.50

Type Package

Title Import and Handling for Plain and Formatted Text Files

Description Functions for importing and handling text files and formatted text files with additional meta-data, such including '.csv', '.tab', '.json', '.xml', '.pdf', '.doc', '.docx', '.xls', '.xlsx', and others.

License GPL-3

Depends R (>= 3.3.1)

Imports utils, stringi, data.table, httr, antiword, pdftools, tibble, jsonlite (>= 0.9.10), streamR, XML, readxl, readODS

Suggests quanteda (>= 0.9.9.24), knitr, rmarkdown, testthat

URL http://github.com/kbenoit/readtext

Encoding UTF-8

BugReports https://github.com/kbenoit/readtext/issues

LazyData TRUE

VignetteBuilder knitr

RoxygenNote 6.0.1

NeedsCompilation no

Author Kenneth Benoit [aut, cre, cph],
Adam Obeng [aut],
Paul Nulty [ctb],
Stefan Müller [ctb]

Maintainer Kenneth Benoit <kbenoit@lse.ac.uk>

Repository CRAN

Date/Publication 2017-05-22 07:20:21 UTC

R topics documented:

`readtext-package` *Import and handling for plain and formatted text files*

Description

A set of functions for importing and handling text files and formatted text files with additional meta-data, such including .csv, .tab, .json, .xml, .xls, .xlsx, and others.

Details

readtext makes it easy to import text files in various formats, including using operating system filemasks to load in groups of files based on glob pattern matches, including files in multiple directories or sub-directories. **readtext** can also read multiple files into R from compressed archive files such as .gz, .zip, .tar.gz, etc. Finally **readtext** reads in the document-level meta-data associated with texts, if those texts are in a format (e.g. .csv, .json) that includes additional, non-textual data.

Package options

`readtext_verbosity` Default verbosity for messages produced when reading files. See `readtext`.

Author(s)

Ken Benoit, Adam Obeng, and Paul Nulty

`as.character.readtext` *return only the texts from a readtext object*

Description

An accessor function to return the texts from a readtext object as a character vector, with names matching the document names.

Usage

```
## S3 method for class 'readtext'
as.character(x, ...)
```

Arguments

x	the readtext object whose texts will be extracted
...	further arguments passed to or from other methods

---

data_char_encodedtexts

encoded texts for testing

Description

data_char_encodedtexts is a 10-element character vector with 10 different encodings

Usage

```
data_char_encodedtexts
```

Format

An object of class character of length 10.

Examples

```
Encoding(data_char_encodedtexts)
data.frame(labelled = names(data_char_encodedtexts),
           detected = encoding(data_char_encodedtexts)$all)
```

---

data_files_encodedtexts

a .zip file of texts containing a variety of differently encoded texts

---

Description

A set of translations of the Universal Declaration of Human Rights, plus one or two other miscellaneous texts, for testing the text input functions that need to translate different input encodings.

Source

The Universal Declaration of Human Rights resources, http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx

Examples

```
## Not run: # unzip the files to a temporary directory
FILEDIR <- tempdir()
unzip(system.file("extdata", "data_files_encodedtexts.zip", package = "readtext"),
      exdir = FILEDIR)

# get encoding from filename
filenames <- list.files(FILEDIR, "\\.txt$")
# strip the extension
filenames <- gsub(".txt$", "", filenames)
parts <- strsplit(filenames, "_")
fileencodings <- sapply(parts, "[", 3)
fileencodings

# find out which conversions are unavailable (through iconv())
cat("Encoding conversions not available for this platform:")
notAvailableIndex <- which(!(fileencodings %in% iconvlist()))
fileencodings[notAvailableIndex]

# try readtext
require(quanteda)
txts <- readtext(paste0(FILEDIR, "/", "*.txt"))
substring(texts(txts)[1], 1, 80) # gibberish
substring(texts(txts)[4], 1, 80) # hex
substring(texts(txts)[40], 1, 80) # hex

# read them in again
txts <- readtext(paste0(FILEDIR,  "/", "*.txt"), encoding = fileencodings)
substring(texts(txts)[1], 1, 80)  # English
substring(texts(txts)[4], 1, 80)  # Arabic, looking good
substring(texts(txts)[40], 1, 80) # Cyrillic, looking good
substring(texts(txts)[7], 1, 80)  # Chinese, looking good
substring(texts(txts)[26], 1, 80) # Hindi, looking good

txts <- readtext(paste0(FILEDIR, "/", "*.txt"), encoding = fileencodings,
                 docvarsfrom = "filenames",
                 docvarnames = c("document", "language", "inputEncoding"))
encodingCorpus <- corpus(txts, source = "Created by encoding-tests.R")
summary(encodingCorpus)

## End(Not run)
```

---

encoding *detect the encoding of texts*

Description

Detect the encoding of texts in a character readtext object and report on the most likely encoding for each document. Useful in detecting the encoding of input texts, so that a source encoding can be (re)specified when inputting a set of texts using `readtext`, prior to constructing a corpus.

Usage

```
encoding(x, verbose = TRUE, ...)
```

Arguments

x	character vector, corpus, or readtext object whose texts' encodings will be detected.
verbose	if FALSE, do not print diagnostic report
...	additional arguments passed to stri_enc_detect

Details

Based on stri_enc_detect, which is in turn based on the ICU libraries. See the ICU User Guide, http://userguide.icu-project.org/conversion/detection.

Examples

```
encoding(data_char_encodedtexts)
# show detected value for each text, versus known encoding
data.frame(labelled = names(data_char_encodedtexts),
           detected = encoding(data_char_encodedtexts)$all)

## Not run: # Russian text, Windows-1251
myreadtext <- readtext("http://www.kenbenoit.net/files/01_er_5.txt")
encoding(myreadtext)

## End(Not run)
```

---

print.readtext *print method for a readtext object*

Description

Print a readtext object in a nicely formatted way.

Usage

```
## S3 method for class 'readtext'
print(x, n = 6L, text_width = 10L, ...)
```

Arguments

x	the readtext object to be printed
n	a single integer, the number of rows of a readtext object to print.
text_width	number of characters to display of the text field
...	not used here

readtext *read a text file(s)*

Description

Read texts and (if any) associated document-level meta-data from one or more source files. The text source files come from the textual component of the files, and the document-level metadata ("docvars") come from either the file contents or filenames.

Usage

```
readtext(file, ignore_missing_files = FALSE, text_field = NULL,
  docvarsfrom = c("metadata", "filenames", "filepaths"), dvsep = "_",
  docvarnames = NULL, encoding = NULL,
  verbosity = getOption("readtext_verbosity"), ...)
```

Arguments

`file` the complete filename(s) to be read. This is designed to automagically handle a number of common scenarios, so the value can be a "glob"-type' wildcard value. Currently available filetypes are:

Single file formats:

`txt` plain text files: So-called structured text files, which describe both texts and metadata: For all structured text filetypes, the column, field, or node which contains the the text must be specified with the `text_field` parameter, and all other fields are treated as docvars.

`json` data in some form of JavaScript Object Notation, consisting of the texts and optionally additional docvars. The supported formats are:

- a single JSON object per file
- line-delimited JSON, with one object per line
- line-delimited JSON, of the format produced from a Twitter stream. This type of file has special handling which simplifies the Twitter format into docvars. The correct format for each JSON file is automatically detected.

`csv,tab,tsv` comma- or tab-separated values

`xml` Basic flat XML documents are supported – those of the kind supported by `xmlToDataFrame`. For xml files, an additional argument `collapse` may be passed through `...` that names the character(s) to use in appending different text elements together.

`pdf` pdf formatted files, converted through `pdftotext`. Requires that xpdf be installed, either through `brew install xpdf` (macOS) or from `http://www.foolabs.com/xpdf/home.html` (Windows).

`doc, docx` Microsoft Word formatted files.

Reading multiple files and file types:

In addition, `file` can also not be a path to a single local file, but also combinations of any of the above types, such as:

a wildcard value any valid pathname with a wildcard ("glob") expression that can be expanded by the operating system. This may consist of multiple file types.

a URL to a remote which is downloaded then loaded

`zip,tar,tar.gz,tar.bz` archive file, which is unzipped. The contained files must be either at the top level or in a single directory. Archives, remote URLs and glob patterns can resolve to any of the other filetypes, so you could have, for example, a remote URL to a zip file which contained Twitter JSON files.

`ignore_missing_files`
: if `FALSE`, then if the file argument doesn't resolve to an existing file, then an error will be thrown. Note that this can happen in a number of ways, including passing a path to a file that does not exist, to an empty archive file, or to a glob pattern that matches no files.

`text_field`
: a variable (column) name or column number indicating where to find the texts that form the documents for the corpus. This must be specified for file types `.csv`, `.json`, and `.xls/.xlsx` files. For XML files, an XPath expression can be specified.

`docvarsfrom`
: used to specify that docvars should be taken from the filenames, when the `readtext` inputs are filenames and the elements of the filenames are document variables, separated by a delimiter (`dvsep`). This allows easy assignment of docvars from filenames such as `1789-Washington.txt`, `1793-Washington`, etc. by `dvsep` or from meta-data embedded in the text file header (`headers`). If `docvarsfrom` is set to `"filepaths"`, consider the full path to the file, not just the filename.

`dvsep`
: separator (a regular expression character string) used in filenames to delimit docvar elements if `docvarsfrom="filenames"` or `docvarsfrom="filepaths"` is used

`docvarnames`
: character vector of variable names for `docvars`, if `docvarsfrom` is specified. If this argument is not used, default docvar names will be used (`docvar1`, `docvar2`, ...).

`encoding`
: vector: either the encoding of all files, or one encoding for each files

`verbosity`
: - 0: output errors only
  - 1: output errors and warnings (default)
  - 2: output a brief summary message
  - 3: output detailed file-related messages

`...`
: additional arguments passed through to low-level file reading function, such as `file`, `fread`, etc. Useful for specifying an input encoding option, which is specified in the same was as it would be give to `iconv`. See the Encoding section of file for details.

Value

a data.frame consisting of a columns `doc_id` and `text` that contain a document identifier and the texts respectively, with any additional columns consisting of document-level variables either found in the file containing the texts, or created through the `readtext` call.

Examples

```
## get the data directory
DATA_DIR <- system.file("extdata/", package = "readtext")

## read in some text data
# all UDHR files
(rt1 <- readtext(paste0(DATA_DIR, "txt/UDHR/*")))

# manifestos with docvars from filenames
(rt2 <- readtext(paste0(DATA_DIR, "txt/EU_manifestos/*.txt"),
                 docvarsfrom = "filenames",
                 docvarnames = c("unit", "context", "year", "language", "party"),
                 encoding = "LATIN1"))

# recurse through subdirectories
(rt3 <- readtext(paste0(DATA_DIR, "txt/movie_reviews/*"),
                 docvarsfrom = "filepaths", docvarnames = "sentiment"))

## read in csv data
(rt4 <- readtext(paste0(DATA_DIR, "csv/inaugCorpus.csv")))

## read in tab-separated data
(rt5 <- readtext(paste0(DATA_DIR, "tsv/dailsample.tsv"), text_field = "speech"))

## read in JSON data
(rt6 <- readtext(paste0(DATA_DIR, "json/inaugural_sample.json"), text_field = "texts"))

## read in pdf data
# UNHDR
(rt7 <- readtext(paste0(DATA_DIR, "pdf/UDHR/*.pdf"),
                 docvarsfrom = "filenames",
                 docvarnames = c("document", "language")))
Encoding(rt7$text)

## read in Word data (.doc)
(rt8 <- readtext(paste0(DATA_DIR, "word/*.doc")))
Encoding(rt8$text)

## read in Word data (.docx)
(rt9 <- readtext(paste0(DATA_DIR, "word/*.docx")))
Encoding(rt9$text)

## use elements of path and filename as docvars
(rt10 <- readtext(paste0(DATA_DIR, "pdf/UDHR/*.pdf"),
                  docvarsfrom = "filepaths", dvsep = "[/_.]"))
```

Index