

Package ‘tm.plugin.mail’

May 14, 2012

Title Text Mining E-Mail Plug-In

Version 0.0-5

Date 2012-05-14

Author Ingo Feinerer

Maintainer Ingo Feinerer <feinerer@logic.at>

Enhances tm (>= 0.5)

Imports tm (>= 0.5)

Description A plug-in for the tm text mining framework providing mail handling functionality.

License GPL (>= 2)

Repository CRAN

Date/Publication 2012-05-14 16:22:39

R topics documented:

convert_mbox_eml	2
MailDocument	2
readMail	3
removeCitation	4
removeMultipart	5
removeSignature	5
threads	6

Index	7
--------------	----------

convert_mbox_eml	<i>Convert E-Mails From mbox Format To eml Format</i>
------------------	---

Description

Convert e-mails from mbox (i.e., several mails in a single box) format to eml (i.e., every mail in a single file) format.

Usage

```
convert_mbox_eml(mbox, dir)
```

Arguments

mbox	A character or connection describing the mbox location.
dir	A character describing the output directory.

Value

No explicit return value. As a side product the directory dir contains the e-mails in eml format.

Author(s)

Ingo Feinerer

MailDocument	<i>E-Mail Document</i>
--------------	------------------------

Description

Construct an object representing an electronic mail document.

Usage

```
MailDocument(x, author = character(0), datetimestamp = as.POSIXlt(Sys.time()), tz = "GMT"), description
```

Arguments

x	Object of class list containing the content.
author	Object of class character containing the author names.
datetimestamp	Object of class POSIXlt containing the date and time when the document was written.
description	Object of class character containing additional text information.
header	Object of class character containing the mail header.

heading	Object of class <code>character</code> containing the title or a short heading.
id	Object of class <code>character</code> containing an identifier.
origin	Object of class <code>character</code> containing information on the source and origin of the text.
language	Object of class <code>character</code> containing the language of the text (preferably in ISO 639-2 format).
localmetadata	Object of class <code>list</code> containing local meta data in form of tag-value pairs.

Author(s)

Ingo Feinerer

See Also

[PlainTextDocument](#)

readMail	<i>Read In an E-Mail Document</i>
----------	-----------------------------------

Description

Return a function which reads in an electronic mail document.

Usage

```
readMail(DateFormat = "%d %B %Y %H:%M:%S", ...)
```

Arguments

<code>DateFormat</code>	The format of the Date header in the mail document.
<code>...</code>	Arguments for the generator function.

Details

Formally this function is a function generator, i.e., it returns a function (which in turn reads in a mail document) with a well-defined signature, but can access passed over arguments (e.g., to specify the format of the Date header in the e-mail via `DateFormat`) via lexical scoping.

Value

A function with the signature `elem, language, id`:

<code>elem</code>	A list with the two named elements <code>content</code> and <code>uri</code> . The first element must hold the document to be read in, the second element must hold a call to extract this document. The call is evaluated upon a request for load on demand.
<code>language</code>	A character vector giving the text's language.

`id` A character vector representing a unique identification string for the returned text document.

The function returns a MailDocument representing content.

Author(s)

Ingo Feinerer

See Also

[strptime](#) for date format specifications.

Examples

```
require("tm")
newsgroup <- system.file("mails", package = "tm.plugin.mail")
news <- Corpus(DirSource(newsgroup), readerControl = list(reader = readMail))
inspect(news)
```

removeCitation	<i>Remove E-Mail Citations</i>
----------------	--------------------------------

Description

Remove citations, i.e., lines beginning with >, from an e-mail message.

Usage

```
## S3 method for class 'MailDocument'
removeCitation(x)
```

Arguments

`x` A mail document.

See Also

[removeMultipart](#) to remove non-text parts from multipart e-mail messages, and [removeSignature](#) to remove signature lines from e-mail messages.

Examples

```
require("tm")
newsgroup <- system.file("mails", package = "tm.plugin.mail")
news <- Corpus(DirSource(newsgroup), readerControl = list(reader = readMail))
news[[6]]
removeCitation(news[[6]])
```

removeMultipart	<i>Remove Non-Text Parts From E-Mails</i>
-----------------	---

Description

Remove non-text parts from multipart e-mail messages.

Usage

```
## S3 method for class 'MailDocument'  
removeMultipart(x)
```

Arguments

x A mail document.

Author(s)

Ingo Feinerer

See Also

[removeCitation](#) to remove e-mail citations, and [removeSignature](#) to remove signature lines from e-mail messages.

removeSignature	<i>Remove E-Mail Signatures</i>
-----------------	---------------------------------

Description

Remove signature lines from an e-mail message.

Usage

```
## S3 method for class 'MailDocument'  
removeSignature(x, marks = character(0))
```

Arguments

x A mail document.
marks Signature identifications marks (in form of regular expression patterns). Note that the official signature start mark -- (dash dash blank) is always considered.

Author(s)

Ingo Feinerer

See Also

[removeCitation](#) to remove e-mail citations, and [removeMultipart](#) to remove non-text parts from multipart e-mail messages.

Examples

```
require("tm")
newsgroup <- system.file("mails", package = "tm.plugin.mail")
news <- Corpus(DirSource(newsgroup), readerControl = list(reader = readMail))
news[[5]]
removeSignature(news[[5]], marks = "^+[+]*[+]$")
```

threads

E-Mail Threads

Description

Extract threads (i.e., chains of messages on a single subject) from e-mail documents.

Usage

```
threads(x)
```

Arguments

x A corpus consisting of e-mails (MailDocuments).

Details

This function uses a one-pass algorithm for extracting the thread information. I.e., reply mails appearing before their corresponding base mails are not detected, and are tagged with thread id NA and depth 2.

Value

A list with the two named components ThreadID and ThreadDepth, listing a thread and the level of replies for each mail in the corpus x.

Author(s)

Ingo Feinerer

Examples

```
require("tm")
newsgroup <- system.file("mails", package = "tm.plugin.mail")
news <- Corpus(DirSource(newsgroup), readerControl = list(reader = readMail))
sapply(news, ID)
lapply(news, function(x) grep("In-Reply-To", attr(x, "Header"), value = TRUE))
threads(news)
```

Index

`as.PlainTextDocument.MailDocument`
`(MailDocument)`, [2](#)

`convert_mbox_eml`, [2](#)

`MailDocument`, [2](#)

`PlainTextDocument`, [3](#)

`readMail`, [3](#)

`removeCitation`, [4](#), [5](#), [6](#)

`removeMultipart`, [4](#), [5](#), [6](#)

`removeSignature`, [4](#), [5](#), [5](#)

`strptime`, [4](#)

`threads`, [6](#)