Package ‘tm.plugin.mail’

September 10, 2009

Title Text Mining E-Mail Plug-In

Version 0.0-1

Date 2009-09-10

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 2009-09-10 19:52:47

R topics documented:

`convert_mbox_eml` *Convert E-Mails From mbox Format To eml Format*

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` *E-Mail Document*

Description

Construct an object representing an electronic mail document.

Usage

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

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

Author(s)

Ingo Feinerer

See Also

`PlainTextDocument`

`readMail` *Read In an E-Mail Document*

Description

Return a function which reads in an electronic mail document.

Usage

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

Arguments

`DateFormat`	The format of the Date header in the mail document.
`...`	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`:

`elem`	A `list` with the two named elements `content` and `uri`. 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.
`language`	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` *Remove E-Mail Citations*

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 *Remove Non-Text Parts From E-Mails*

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 *Remove E-Mail Signatures*

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