Package 'tm.plugin.europresse'

March 29, 2025

Type Package

Title Import Articles from 'Europresse' Using the 'tm' Text Mining
Framework

Version 1.4.1

Date 2025-03-28

Imports utils, NLP, tm (>= 0.6), XML

Description Provides a 'tm' Source to create corpora from
articles exported from the 'Europresse' content provider as
HTML files. It is able to read both text content and meta-data
information (including source, date, title, author and pages).

License GPL (>= 2)

URL https://github.com/nalimilan/R.TeMiS

BugReports https://github.com/nalimilan/R.TeMiS/issues

NeedsCompilation no

Author Milan Bouchet-Valat [aut, cre]

Maintainer Milan Bouchet-Valat <nalimilan@club.fr>

Repository CRAN

Date/Publication 2025-03-28 23:00:02 UTC

Contents

tm.plugin.europresse-package

A plug-in for the tm text mining framework to import articles from Europresse

Description

This package provides a tm Source to create corpora from articles exported from the Europresse content provider as HTML files.

Details

Typical usage is to create a corpus from HTML files exported from Europresse (here called myEuropresseArticles.html). Frequently, it is necessary to specify the encoding of the texts via link{EuropresseSource}'s encoding argument.

```
# Import corpus
source <- EuropresseSource("myEuropresseArticles.html")
corpus <- Corpus(source)

# See how many articles were imported
corpus

# See the contents of the first article and its meta-data
inspect(corpus[1])
meta(corpus[[1]])
```

See link{EuropresseSource} for more details and real examples.

Author(s)

Milan Bouchet-Valat <nalimilan@club.fr>

References

https://www.europresse.com/

EuropresseSource	*Europresse Source*

Description

Construct a source for an input containing a set of articles exported from Europresse in the HTML format.

Usage

```
EuropresseSource(x, encoding = "UTF-8")
```

Arguments

x	Either a character identifying the file or a connection.
encoding	A character giving the encoding of x. Files exported from Europresse often specify an incorrect encoding, in which case you will need to find out the correct one.

Details

This function imports the body of the articles, but also sets several meta-data variables on individual documents:

- datetimestamp: The publication date.
- heading: The title of the article.
- origin: The newspaper the article comes from.
- section: If available, the part of the newspaper containing the article.
- pages: If available, the pages where the article appeared.

Please note that it commonly happens that the encoding specified in Europresse HTML files does not correspond to the one actually used in the text: in that case, you will need to find out the correct encoding and specify it manually.

Value

An object of class EuropresseSource which extends the class Source representing set of articles from Europresse.

Author(s)

Milan Bouchet-Valat

See Also

readEuropresseHTML2 for the function actually parsing individual articles.

getSources to list available sources.

Examples

```
library(tm)
file <- system.file("texts", "europresse_test2.html",
                    package = "tm.plugin.europresse")
corpus <- Corpus(EuropresseSource(file))

# See the contents of the documents
inspect(corpus)

# See meta-data associated with first article
meta(corpus[[1]])
```

---

readEuropresseHTML   *Read in a Europresse article in the HTML format*

Description

Read in an article exported from Europresse in the HTML format.

Usage

```
readEuropresseHTML1(elem, language, id)
readEuropresseHTML2(elem, language, id)
```

Arguments

elem	A list with the named element content which must hold the document to be read in.
language	A character vector giving the text's language. If set to NA, the language will automatically be set to the value reported in the document (which is usually correct).
id	A character vector representing a unique identification string for the returned text document.

Details

readEuropresseHTML1 reads documents in the old format, while readEuropresseHTML2 reads documents in the new one. EuropresseSource automatically chooses the correct reader based on the structure of the file.

Value

A PlainTextDocument with the contents of the article and the available meta-data set.

Author(s)

Milan Bouchet-Valat

See Also

[getReaders](#) to list available reader functions.

Index