Package 'tm.plugin.lexisnexis'

June 5, 2018

Type Package

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

Version 1.4.0

Date 2018-06-05

Imports utils, NLP, tm (>= 0.6), xml2, ISOcodes

Description Provides a 'tm' Source to create corpora from articles exported from the 'LexisNexis' content provider as HTML files. It is able to read both text content and meta-data information (including source, date, title, author and pages). Note that the file format is highly unstable: there is no warranty that this package will work for your corpus, and you may have to adjust the code to adapt it to your particular format.

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 2018-06-05 17:09:23 UTC

R topics documented:

`tm.plugin.lexisnexis-package` *A plug-in for the tm text mining framework to import articles from LexisNexis*

Description

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

Details

Typical usage is to create a corpus from HTML files exported from LexisNexis (here called `myLexisNexisArticles.html`). Setting `language=NA` allows the language to be set automatically from the information provided by Factiva:

```
# Import corpus
source <- LexisNexisSource("myLexisNexisArticles.html")
corpus <- Corpus(source, readerControl = list(language = NA))

# See how many articles were imported
corpus

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

Currently, only HTML files saved in English and French are supported. Please send the maintainer examples of LexisNexis files in your language if you want it to be supported.

See `link{LexisNexisSource}` for more details and real examples.

Author(s)

Milan Bouchet-Valat <nalimilan@club.fr>

References

http://www.lexisnexis.com/

`LexisNexisSource` *LexisNexis Source*

Description

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

Usage

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

Arguments

`x`	Either a character identifying the file or a connection.
`encoding`	A character giving the encoding of `x`. It will be ignored unless the HTML input does not include this information, which should normally not happen with files exported from LexisNexis.

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.
- `intro`: The short introduction accompanying the article.
- `section`: The part of the newspaper containing the article.
- `subject`: One or several keywords defining the subject.
- `coverage`: One or several keywords identifying the covered regions.
- `company`: One or several keywords identifying the covered companies.
- `stocksymbol`: One or several keywords identifying the stock exchange symbols of the covered companies.
- `industry`: One or several keywords identifying the covered industries.
- `type`: The type of source from which the document originates.
- `wordcount`: The number of words in the article.
- `publisher`: The publisher of the newspaper.
- `rights`: The copyright information associated with the article.
- `language`: This information is set automatically if `readerControl = list(language = NA)` is passed (see the example below). Else, the language specified manually is set for all articles. If omitted, the default, "en", is used.

Please note that dates are not guaranteed to be parsed correctly if the machine from which the HTML file was exported uses a locale different from that of the machine where it is read.

Currently, only HTML files saved in English and French are supported. Please send the maintainer examples of LexisNexis files in your language if you want it to be supported.

Value

An object of class `LexisNexisSource` which extends the class `Source` representing set of articles from LexisNexis.

Author(s)

Milan Bouchet-Valat

See Also

`readLexisNexisHTML` for the function actually parsing individual articles.

`getSources` to list available sources.

Examples

```
library(tm)
file <- system.file("texts", "lexisnexis_test_en.html",
                    package = "tm.plugin.lexisnexis")
corpus <- Corpus(LexisNexisSource(file))

# See the contents of the documents
inspect(corpus)

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

---

readLexisNexisHTML *Read in a LexisNexis article in the HTML format*

Description

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

Usage

```
readLexisNexisHTML(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.

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