

# Package ‘tm.plugin.alceste’

July 2, 2014

**Type** Package

**Title** Import texts from files in the Alceste format using the tm text mining framework

**Version** 1.1

**Date** 2014-05-31

**Imports** NLP, tm (>= 0.6)

**Suggests** stringi

**Description** This package provides a tm Source to create corpora from a corpus prepared in the format used by the Alceste application (i.e. a single text file with inline meta-data). It is able to import both text contents and meta-data (starred) variables.

**License** GPL (>= 2)

**URL** <https://r-forge.r-project.org/projects/r-temis/>

**BugReports** [https://r-forge.r-project.org/tracker/?group\\_id=1437](https://r-forge.r-project.org/tracker/?group_id=1437)

**Author** Milan Bouchet-Valat [aut, cre]

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

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2014-06-11 23:47:21

## R topics documented:

tm.plugin.europresse-package . . . . .	2
AlcesteSource . . . . .	3
readAlceste . . . . .	4

<b>Index</b>	<b>5</b>
--------------	----------

tm.plugin.europresse-package

*A plug-in for the tm text mining framework to import corpora from Alceste files*

---

## Description

This package provides a tm Source to create corpora from files formatted in the format used by the Alceste application.

## Details

Typical usage is to create a corpus from an Alceste file prepared manually (here called myAlcesteCorpus.txt). Frequently, it is necessary to specify the encoding of the texts via link{AlcesteSource}'s encoding argument.

```
# Import corpus
source <- europresseSource("myAlcesteCorpus.txt")
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{AlcesteSource} for more details and real examples.

## Author(s)

Milan Bouchet-Valat <nalimilan@club.fr>

## References

<http://www.image-zafar.com/en/alceste-software>

---

AlcesteSource

*Alceste Source*

---

## Description

Construct a source for an input containing a set of texts saved in the Alceste format in a single text file.

## Usage

```
AlcesteSource(x, encoding = "auto")
```

## Arguments

x	Either a character identifying the file or a connection.
encoding	A character string: if non-empty declares the encoding used when reading the file, so the character data can be re-encoded. See the ‘Encoding’ section of the help for <a href="#">file</a> . The default, “auto”, uses <code>stri_enc_detect</code> to try to guess the encoding; this may fail, in which case the native encoding is used.

## Details

Several texts are saved in a single Alceste-formatted file, separated by lines starting with “\*\*\*” or digits, followed by starred variables (see links below). These variables are set as document meta-data that can be accessed via the [meta](#) function.

Currently, “theme” lines starting with “-\*” are ignored.

## Value

An object of class `AlcesteSource` which extends the class `Source` representing set of articles from Alceste.

## Author(s)

Milan Bouchet-Valat

## See Also

[http://www.image-zafar.com/sites/default/files/telechargements/formatage\\_alceste.pdf](http://www.image-zafar.com/sites/default/files/telechargements/formatage_alceste.pdf) (in French) about the Alceste format

[readAlceste](#) for the function actually parsing individual articles.

[getSources](#) to list available sources.

## Examples

```
library(tm)
file <- system.file("texts", "alceste_test.txt",
                    package = "tm.plugin.alceste")
corpus <- Corpus(AlcesteSource(file))

# See the contents of the documents
inspect(corpus)

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

---

readAlceste	<i>Read in a text in the Alceste format</i>
-------------	---

---

## Description

Read in a text in the Alceste format using starred variables.

## Usage

```
readAlceste(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

AlcesteSource, [3](#)

eoi.AlcesteSource (AlcesteSource), [3](#)

file, [3](#)

getElem.AlcesteSource (AlcesteSource), [3](#)

getReaders, [4](#)

getSources, [3](#)

meta, [3](#)

readAlceste, [3](#), [4](#)

stri\_enc\_detect, [3](#)

tm.plugin.europresse

(tm.plugin.europresse-package),

[2](#)

tm.plugin.europresse-package, [2](#)