Package 'parseAtekst'

November 13, 2015

Title Parse and import articles bundled in .txt-files downloaded from ATEKST

Version 1.2

Description

This R-package provides functions for parsing .txt-files downloaded from ATEKST that imports individual articles (with metadata) into R. They return a data frame with the headline, paper, date and time of publication, mode (net vs print), url and text of each news article.

Depends R (>= 3.0.0), foreach, iterators, doParallel, parallel

License GPL (>= 2)

LazyData true

R topics documented:

`read.atekst` *Parse an atekst .txt file*

Description

Parses an atekst .txt-file. The function returns a data frame with the headline, paper, date, time, mode (net/print), url, and text for each article.

Usage

```
read.atekst(file)
```

Arguments

`file`	Path to a .txt-file downloaded from atekst. Typically called something like `"Utvalgte_dokumenter-100-01.01.2015.txt"`.

Examples

```
corpus <- read.atekst("Utvalgte_dokumenter-100-01.01.2015.txt")
save(corpus, file = "atekst-corpus.RData")
```

`read.atekst.dir` *Parse all atekst .txt files in a directory*

Description

Parse all .txt files downloaded from atekst within a directory (including subfolders). It can use a pattern (regex) to identify files. The function returns a data frame with the headline, paper, date, time, mode (net/print), url, and text for each article. In order to speed it up it is possible to run it in parallel by setting `parallel` to `TRUE` and setting `cores`. When working with large corpuses it is recommended to run the function once and save the resulting data frame as a `.RData`-file. That way it can be loaded (using `load()`) into R in a fraction of the time it takes to parse the whole corpus.

Usage

```
read.atekst.dir(dir, recursive = TRUE,
  regex = "^Utvalgte_dokumenter.*.txt$", parallel = FALSE, cores = 1)
```

Arguments

`dir`	Directory containing atekst .txt files.
`recursive`	If `TRUE`, the function also parses files within subfolders.
`regex`	Regular expression (pattern) to use for selecting files to parse.
`parallel`	If `TRUE` it will try to do it in parallel (using the packages `foreach`, `iterators`, `doParallel` and `parallel`).
`cores`	The amount of cores to use (if `parallel` is `TRUE`).

Examples

```
corpus <- read.atekst.dir("some/directory")
save(corpus, file = "atekst-corpus.RData")
```

Index