Package 'parseAtekst'

October 10, 2015

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

Version 1.1

Description The Retriever (http://www.retriever-info.com/) online Norwegian media archive ATEKST (C) allows researchers to search through their media archive and download news articles based on that search. Typically one can download up to 100 articles at a time bundled together within a .txt-file. This R-package provides functions for parsing .txt-files downloaded from ATEKST that extracts all articles and imports them into R. The functions 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

This function allows you to parse 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 an directory*

Description

This function allows you to parse all .txt files downloaded from atekst within a directory (including subfolders). If `strict == TRUE` it only tries to parse .txt-files starting with "Utvalgte_dokumenter", otherwise it will try to parse all .txt-files it can find. It is recommended that the directory only contains .txt-files downloaded from Atekst. The function returns a data frame with the headline, paper, date, time, mode (net/print), url, and text for each article. In order to save time when working with large corpuses it is recommended to run the function once and save the resulting data frame as a `RData`-file (using `save()`). 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, strict = TRUE, cores = 1)
```

Arguments

dir	Directory containing atekst .txt files.
recursive	If `TRUE`, the function also parses files within subfolders.
strict	If `TRUE`, the function only parses .txt files with filenames starting with "Utvalgte_dokumenter".
cores	The amount of cores (i.e. parallel processes) that should be utilized.

Examples

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

Index