Web Data Collection with R

File Manipulation

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT

Why ?

Functions

Why ?

Functions

Patterns

function	description
`list.files()`	list files
`list.dirs()`	list directories
`file.info()`	information on file
`file.size()`	check size of file
`file.exists()`	check existence of file
`file.copy()`	copy file
`file.remove()`	delete file
`file.rename()`	rename file
`file.mtime()`	get modiefied time of file
`readLines()`	read file into char. vector
`writeLines()`	write character vector into file
`write()`	write to file, e.g. append to file
`dir.create()`	creates a directory

use case 1 - write file if not existent

```
fname <- "testfile.txt"

if( !file.exists(fname)  ) {
  writeLines( as.character(Sys.time()), fname )
}

file.mtime(fname)
```

```
## [1] "2016-02-27 18:08:23 CET"
```

```
readLines(fname)
```

```
## [1] "2016-02-27 18:08:23"
```

use case 2 - append to file

```
for( i in 1:3 ){
  atime <- substr(as.character(Sys.time()),12,20)
  mtime <- substr(as.character(file.mtime(fname)),12,20)
  write(
    paste0(
      " append : ", atime, "   mtime : ", mtime,
      collapse = ""
    ),
    fname,
    append = TRUE
  )
  Sys.sleep(1)
}
```

use case 2 - append to file

```
file.mtime(fname)
```

```
## [1] "2016-02-27 18:08:25 CET"
```

```
readLines(fname)
```

```
## [1] "2016-02-27 18:08:23"
## [2] " append : 18:08:23   mtime : 18:08:23"
## [3] " append : 18:08:24   mtime : 18:08:23"
## [4] " append : 18:08:25   mtime : 18:08:24"
```

use case 3 - get file name from URL

```
url      <- "http://www.r-datacollection.com/materials/http
destfile <- basename(url)
download.file(url = url, destfile = destfile)
```

```
destfile
```

```
## [1] "index.html"
```

```
file.exists(destfile)
```

```
## [1] TRUE
```

```
readLines(destfile, warn=FALSE)
```

```
## [1] "<html>"
## [2] " <head>"
## [3] "  <title>List of Files</title>"
## [4] " </head>"
## [5] " <body>"
```

use case 4 - handling files names in a loop

```
for ( i in 1:10) {
  content <- paste(sample(letters,10),collapse = "")
  fname   <-
    paste(
      "f_",
      Sys.time(), " ",
      stringr::str_pad(i, 3, "left", 0),
      ".test",
      collapse = "", sep=""
    )
  writeLines(content, fname)
}
```

use case 4 - handling files names in a loop

```
list.files(pattern = "\\d.test$")
```

```
##  [1] "f_2016-02-27 18:08:27 001.test"
##  [2] "f_2016-02-27 18:08:27 002.test"
##  [3] "f_2016-02-27 18:08:27 003.test"
##  [4] "f_2016-02-27 18:08:27 004.test"
##  [5] "f_2016-02-27 18:08:27 005.test"
##  [6] "f_2016-02-27 18:08:27 006.test"
##  [7] "f_2016-02-27 18:08:27 007.test"
##  [8] "f_2016-02-27 18:08:27 008.test"
##  [9] "f_2016-02-27 18:08:27 009.test"
## [10] "f_2016-02-27 18:08:27 010.test"
```

use case 5 - zipping

```
filelist <- list.files(pattern = ".txt$|html$|test$")
zip(zipfile = "archive.zip", files = filelist)
file.remove(filelist[file.exists(filelist)])
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [8] TRUE TRUE TRUE TRUE TRUE
```

```
list.files()
```

```
## [1] "archive.zip"
## [2] "file_manipulation.pdf"
## [3] "file_manipulation.Rmd"
## [4] "file_manipulation.tex"
```