

# Web Data Collection with R

## File Manipulation

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

**Why ?**

**Functions**

**Why ?**

# Functions

## Patterns

function	description
<code>list.files()</code>	list files
<code>list.dirs()</code>	list directories
<code>file.info()</code>	information on file
<code>file.size()</code>	check size of file
<code>file.exists()</code>	check existence of file
<code>file.copy()</code>	copy file
<code>file.remove()</code>	delete file
<code>file.rename()</code>	rename file
<code>file.mtime()</code>	get modified time of file
<code>readLines()</code>	read file into char. vector
<code>writeLines()</code>	write character vector into file
<code>write()</code>	write to file, e.g. append to file
<code>dir.create()</code>	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] "2016-03-01 00:22:13 CET"
```

```
readLines(fname)
```

```
## [1] "2016-03-01 00:22:13"
```

## 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] "2016-03-01 00:22:15 CET"
```

```
readLines(fname)
```

```
## [1] "2016-03-01 00:22:13"
```

```
## [2] " append : 00:22:13      mtime : 00:22:13"
```

```
## [3] " append : 00:22:14      mtime : 00:22:13"
```

```
## [4] " append : 00:22:15      mtime : 00:22:14"
```



## 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_2016-03-01 00:22:17 001.test"  
## [2] "f_2016-03-01 00:22:17 002.test"  
## [3] "f_2016-03-01 00:22:17 003.test"  
## [4] "f_2016-03-01 00:22:17 004.test"  
## [5] "f_2016-03-01 00:22:17 005.test"  
## [6] "f_2016-03-01 00:22:17 006.test"  
## [7] "f_2016-03-01 00:22:17 007.test"  
## [8] "f_2016-03-01 00:22:17 008.test"  
## [9] "f_2016-03-01 00:22:17 009.test"  
## [10] "f_2016-03-01 00:22:17 010.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 TRUE
```

```
list.files()
```

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