

DownloadHelper 用户手册

Ethan Deng

Version 2.0

September 15, 2014

1 功能介绍

本 R 文件（DownloadHelper.R 简称 DH）主要功能是批量下载相近文件名或者有文件名列表的文件。原则上 DH 支持所有的文件格式，经过测试的文件类型有：PDF 文件，Excel 文件，txt 纯文本文件，LaTeX 文件，R 文件，Matlab 文件，MP3 文件。DH 定义了四个函数，用户只需要使用最外层的函数即可批量下载相似地址的文件。

2 函数说明

DH 定义了如下四个函数：

- `pkgtest`：检测是否已经安装某个宏包，参数为宏包名（参数：字符常量 `x`）
- `transform`：数值索引转换为统一位数格式（参数 1: `index`；参数 2: `ndigits`）
- `downloader`：下载文件函数 1.0 版本，参数及示例见后文。
- `tdownloader`：下载文件函数 2.0 版本，参数及示例见后文。

2.1 `pkgtest` 函数说明

其中 `pkgtest` 用于检测函数是否已经安装，如果没有，则进行安装。由于文件下载需要借助 `downloader` 宏包，所以在使用 `downloader` 之前需要检测宏包是否安装，当然用户不需要自行检察。`downloader` 函数自动调用 `pkgtest` 函数。示例如下

```
1 pkgtest("xtable")
2 # output as follows
3 [1] "The required packages have been installed"
```

2.2 `transform` 函数说明

有时候我们会碰到文件的下载地址可能如下形式 `~/chap002.pdf`，此地址文件索引包含的是 002 而不是 2，为了考虑这种情况，我们定义了 `transform` 函数，用以把普通数字转换成统一位数（不够位数补零）的数字格式，也即 2 转化为 002。这个函数有两个参数，分别为 `index` 和 `ndigits`，第一个参数表示需要转换的数字，可以是向量。第二个参数是转换后数字的位数（也即转换后数字的长度，比如 002 位数为 3）。示例如下

```

1 index <- 1:24
2 transform(index,ndigits=3)
3 # output as follows
4 [1] "001" "002" "003" "004" "005" "006" "007" "008" "009" "010" "011" "012"
5 [13] "013" "014" "015" "016" "017" "018" "019" "020" "021" "022" "023" "024"

```

2.3 downloader 函数说明

- **urlpre**: 字符参数, 下载文件网址不含文件名部分, 必选参数;
- **index**: 文件名索引, 不含拓展名, 必选参数, 可以为字符向量或者数值向量;
- **transform**: 逻辑参数, 补零转换, 默认 **transform = FALSE**;
- **ndigits**: 数值参数, 总位数, 比如需要得到 012, 则 **ndigits=3**, 默认 **ndigits = 2**;
- **filepre**: 字符参数, 文件名前缀, 比如 chap, 默认为空;
- **filetype**: 字符参数, 文件类型, 文件拓展名, 默认 **filetype="pdf"**。

我们用一个例子来解释上面的参数, 比如, 现在要下载 <http://www.math.upenn.edu/~guffin/teaching/fall10/lectures/lecture-01.pdf>, 我们对这几个参数一一解释。

urlpre 是指下载文件地址的前部分, 大部分情况下, 都是最后一个斜杠的之前的内容 (含斜杠)。在我们的这个例子中, **urlpre** = "http://www.math.upenn.edu/~guffin/teaching/fall10/lectures/"。

index 是指我们需要下载文件名的索引, 比如这里文件名分别为 01.pdf-32.pdf, 则 **index = 1:32**。当然, 可能有时候并不是顺序, 则可以使用向量表示, 可以为字符向量或者数值向量。字符向量比如 **index = c("hello", "rintro", "fname-esim")**, 也就是文件名的一个列表 (不含拓展名)。

transform 是个逻辑值参数, 可选项有 **TRUE** 和 **FALSE**, 前者表示进行数字格式转换, 比如此例中, 由于下载地址中是 01.pdf, 需要将 1 转换成 01 这类形式, 所以这个例子中 **transform = TRUE**。

ndigits 表示需要转换的位数, 当然, 只在 **transform = TRUE** 时候才会生效。在我们这个例子中, 由于转换之后为 01, 所以位数为 2, **ndigits = 2**。

filepre 表示文件名前缀, 字符参数, 在我们这个例子中, 文件名前缀参数为 **filepre = "lecture-"**。

filetype 字符参数, 表示文件类型, 准确来说应该是文件名后缀, 在本例中, **filetype = "pdf"**, 注意必须要用双引号。

所以综合起来完整的例子如下

```

1 source("downloadhelper.R")
2 urlpre <- "http://www.math.upenn.edu/~guffin/teaching/fall10/lectures/"
3 index <- 2:4
4 downloader(urlpre,index,ndigits=2,transform=TRUE,filepre="lecture-",filetype="pdf")

```

2.4 tdownloader 函数说明

tdownloader 函数是 **downloader** 的升级版, 本质上 **tdownloader** 使用 **downloader** 函数。使用 **tdownloader** 下载文件需要使用到两个参数, 分别是 **url** 和 **index**。

- **url**: 文件下载示例网址, 具有代表性的网址。

- index: 文件名索引, 同 downloader 函数参数。

tdownloader 目前仅支持三类地址下载, 见使用示例。

3 使用示例

3.1 downloader 下载示例

PDF 文件下载

相关网页: <http://www.math.upenn.edu/~guffin/teaching/fall10/>

```
1 source("downloadhelper.R")
2 urlpre <- "http://www.math.upenn.edu/~guffin/teaching/fall10/lectures/"
3 index <- 2:4
4 downloader(urlpre,index,ndigits=2,transform=TRUE,filepre="lecture-",filetype="pdf")
```

R 文件下载

相关网页: <http://www.rni.helsinki.fi/~pek/r-koulutus/index.html>

```
1 source("downloadhelper.R")
2 urlpre <- "http://www.rni.helsinki.fi/~pek/r-koulutus/"
3 index <- c("hello","rintro","fname-esim")
4 downloader(urlpre,index,filetype="R")
```

txt 文件下载

相关网页: <http://www1.umn.edu/statsoft/doc/statnotes/>

```
1 source("downloadhelper.R")
2 urlpre <- "http://www1.umn.edu/statsoft/doc/statnotes/"
3 index <- 1:8
4 downloader(urlpre,index,transform=TRUE,ndigits=2,filepre="stat",filetype="txt")
```

LaTeX 文件下载

相关网页: <https://gitorious.org/pkuthss/pkuthss/>

```
1 source("downloadhelper.R")
2 urlpre <- "https://gitorious.org/pkuthss/pkuthss/raw/1bb0c28e36e7ddb7d6db000
3     9a1d154e1e3ef4c6c:doc/chap/"
4 index <- 1:3
5 downloader(urlpre,index,filetype="tex",filepre="chap")
```

Excel 表格下载

相关网页: <http://www.pbc.gov.cn/publish/html/>

```
1 source("downloadhelper.R")
2 urlpre <- "http://www.pbc.gov.cn/publish/html/"
3 index <- 5:8
4 downloader(urlpre,index,transform=TRUE,filetype="xls",filepre="2012s")
```

Matlab m 文件下载

相关网页: <http://www.rni.helsinki.fi/~pek/software.html>

```

1 source("downloadhelper.R")
2 urlpre <- "http://www.rni.helsinki.fi/~pek/software/smoothing/"
3 index <- c("llrev","llrcvev","llrssev","llrrev","llrrcvev","llrrssev","knnev","knncvev","knnssev","demo")
4 downloader(urlpre,index,filetype="m")

```

MP3 文件下载

相关网页: <http://download.dogwood.com.cn/online/grechjx/index.html>

```

1 source("downloadhelper.R")
2 urlpre <- "http://download.dogwood.com.cn/online/grechjx/"
3 urlindex <- 4:6
4 downloader(urlpre,index,ndigits=2,transform=TRUE,filepre="WordList",filetype="mp3")

```

3.2 tdownloader 下载示例

索引类型 1: 常规字符文件列表

```

1 source("downloadhelper.R")
2 url <- "http://www.rni.helsinki.fi/~pek/software/smoothing/llrev.m"
3 index <- c("llrev","llrcvev","llrssev","llrrev","llrrcvev","llrrssev","knnev","knncvev","knnssev","demo")
4 tdownloader(url,index)

```

索引类型 2: 含字符 + 常规数字文件

```

1 url <- "https://gitorious.org/pkuthss/pkuthss/source/1bb0c28e36e7ddb7d6db0009a1d154e1e3ef4c6c"
2       ":doc/chap/chap1.tex"
3 index <- 1:3
4 tdownloader(url,index)

```

索引类型 3: 含字符 + 统一格式数字文件

```

1 source("downloadhelper.R")
2 url <- "http://www1.umn.edu/statsoft/doc/statnotes/stat01.txt"
3 index <- 1:8
4 tdownloader(url,index)

```

4 更新日志

- 2014-09-06: 创立文件, 内含 `pkgtest\transform\downloader` 函数
- 2014-09-07: 新增 `tdownloader` 函数, 减少 4 个参数 (版颖协助)。
- 2014-09-15: 增加中文 PDF 说明。

5 源码

```
1 # Copy Right by Ethan Deng (http://ddswhu.com)
2 # Email: ddswhu@gmail.com
3 # Last Modification: 2014-9-15
4 # detect whether the downloader pkg has been installed
5 # if not, then install the pkg
6 # otherwise, return the information that the pkg is not installed
7 pkgtest <- function(x="downloader") {
8   if (x %in% rownames(installed.packages()) == FALSE){
9     install.packages(x)
10   } else {print("The required packages have been installed")}
11 }
12
13 # transform the number index to special character
14 # index which has the same number of digits
15 # Example: 1:13 -> 01-09,10,11,12,13
16 transform <- function(index,ndigits=2){
17   nullnum <- 10^{ndigits}
18   cindex <- as.character(index + nullnum)
19   dindex <- substr(cindex,2,ndigits+2)
20   return(dindex)
21 }
22
23 downloader <- function(urlpre, index, transform=FALSE, ndigits = 2, filepre = "", filetype = "pdf") {
24   # test whether the pkg is installed
25   # load the pkg
26   pkgtest("downloader")
27   library(downloader)
28   # If the transform is set to be TRUE and class of index
29   # is not character, then the dindex should be transformed
30   # using transform function defined
31   if (transform == TRUE & !class(index) == "character"){
32     dindex <- transform(index,ndigits)
33   } else {
34     dindex <- index
35   }
36   # sometimes, we have filepre such as "chap" or "chapter"
37   # filetype can be "pdf","csv","xlsx","mp3"
38   fileindex <- paste(filepre,dindex,".",filetype,sep="")
39
40   # to create the full path of the files
41   # urlpre indicates the url prefix
42   urlindex <- paste(urlpre,fileindex,sep="")
43 }
```

```

44 # binary files and text file list
45 textfilelist <- c("csv","txt","R","tex","r","m")
46
47 # mode is set to be "w" short for write(default)
48 # if the file is text files, then it should be set
49 # to be wb, short for write binary files
50 if (filetype %in% textfilelist == FALSE){
51     mode <- "wb"
52 } else {
53     mode <- "w"
54 }
55
56 # download process
57 for (i in 1:length(index)) {
58     download(url = urlindex[i], destfile = fileindex[i], mode = mode)
59 }
60 }
61
62 tdownloader <- function(url,index){
63     medfile <- strsplit(url,"\\",fixed=FALSE)
64     finalfile <- medfile[[1]][length(medfile[[1]])]
65     urlpre <- substr(url,1,nchar(url)-nchar(finalfile)) # option 1
66     filesplit <- strsplit(finalfile,"\\.",fixed=F)[[1]]
67     filetype <- filesplit[2] # option 6
68     fchar <- strsplit(filesplit[1],"[0-9]+",fixed=F)
69     filepre <- ifelse(is.numeric(index), fchar[[1]], "")
70     # option 5
71     fnum <- strsplit(filesplit[1],"[a-zA-Z]*",fixed=F)
72     fnumb <- substr(filesplit[1],nchar(fchar[[1]])+1,nchar(fchar[[1]])+length(fnum[[1]])-1)
73     transform <- !nchar(fnumb)==nchar(as.numeric(fnumb)) # option 3
74     ndigits <- nchar(fnumb) # option 4
75     downloader(urlpre,index,transform,ndigits,filepre,filetype)
76 }

```