统计软件与金融数据

第四章数据导入

罗智超 (ROKIA.ORG)

Contents

通过本章你将学会 . . . 2
第一大坑：Character encoding hell . . . 2
从剪切板读取 . . . 3
从键盘读入 . . . 3
链接方法 file(), url() 等 . . . 3
文件及目录相关函数 . . . 3
读入文本文件 . . . 4
读入固定宽度文件 . . . 4
redLines(),scan() . . . 4
读入数据练习 . . . 4
批量读入多个外部文件 . . . 4
读入 EXCLE 文件 . . . 5
通过 ODBC 访问 . . . 6
访问 ORACLE . . . 6
读入比较大的数据文件 . . . 6
读取其他统计软件数据 . . . 6
访问 ORACLE . . . 6
访问 Sqlite . . . 6
输出数据集 . . . 7
清理内存 . . . 7
网络爬虫 . . . 7
本周“大牛” . . . 7

通过本章你将学会

- 数据导入

–文本文件

–EXCEL 文件

–其他统计软件

–批量导入数据

–数据库

- 数据输出
- 网络爬虫

第一大坑：Character encoding hell

参考博客文章：中文编码问题不再坑

```
# ISO8859-1-->GB2312-->BGK-->GB18030
# ISO8859-1-->UTF-8/UTF-16
# ISO8859-1-->BIG5

# 查看有多少编码类型
codepages <- setNames(iconvlist(), iconvlist())
page(codepages)
# 查看文本文件编码类型
#win:ultraedit
#mac:enca filename
#[checkenc - 自动文本编码识别](http://qinwenfeng.com/cn/checkenc/)
library(devtools)
devtools::install_github("qinwf/checkenc",
                         force = T)
library(checkenc)
checkenc("data/survey2014_student.csv")
sessionInfo()
tau::is.locale() # tests if the components of a vector of character are in the encoding of the current
tau::is.ascii()
tau::is.utf8()  # tests if the components of a vector of character are true UTF-8 strings|

# 文本类型转换
#win:ultraedit
#mac:iconv -f GB2312 -t UTF-8 a.txt > b.txt

# 编码转化
#iconv(x, "ISO_8859-2", "UTF-8")
#fileEncoding="UTF-8"
#Encoding(x)<-"UTF-8"
Encoding(df$var1)<-"UTF-8"

#Windows
#Sys.setlocale(category = "LC_ALL",locale = "English")
#Sys.setlocale(category = "LC_ALL",locale = "Chinese")
```

从剪切板读取

ds<-read.delim("clipboard")

从键盘读入

- scan(), readline(), print(), and cat()

链接方法 file(), url() 等

```
uci <- "http://archive.ics.uci.edu/ml/machine-learning-databases/"
uci <- paste(uci,"echocardiogram/echocardiogram.data",sep="")
ecc <- read.table(uci)

# 查看文本文件前几行数据
# 当一个文本文件比较大的时候，可以通过该方法查看其样本数据
filename<-"data/FFL2_TAQ_IF1102_201102.txt"
conn<-file(filename,encoding = "GB2312")
dl<-readLines(conn,n=10,encoding = "GB2312")
close(con = conn)



df<-read.table(file=filename,sep = ",",
               fileEncoding = "GB2312",
               stringsAsFactors = F,
               nrows = 3,header = F,skipNul = T)


# 练习导入数据 survey2014_student.csv
# 练习导入数据 USIP.txt
# 练习导入数据 rich.txt

df<-read.table(file="data/rich.csv",
                 sep=",",header = T)
df2<-read.csv(file="data/rich.csv",
                 sep=",",header = T)
# 如果数据文件里面出现 # 符号，可能会提示 Error in Scan 错误信息，可以使用 comment.char=TRUE 参数
```

文件及目录相关函数

- file.info() 获取文件信息
- list.dirs()、dir()、list.files()、file.info("."): 返回目录里面的文件信息

```
fl<-dir()
```

- file.exists(): 判断是否存在某文件
- dir.create("newfolder") 创建目录
- dir.create(path="a1/b2/c3",recursive = TRUE) 创建多级目录
- file.rename("tmp", "tmp2") 目录重命名

- unlink("tmp2", recursive = TRUE) 删除目录
- file.create("A.txt") 创建一个空文件
- file.append("A.txt", rep("B.txt", 10)) 合并文件
- readLines("A.txt") 查看文件内容
- getwd() 获取当前工作目录
- setwd() 设定当前工作目录

读入文本文件

- read.table
- read.csv
- read.delim
- read.fwf

```
# 练习 1: 熟悉 read.csv 语法
# 练习 2: 导入"data/ag.csv" 到 R
# 练习 3: 导入 agaricus-lepiota.txt
ag<-read.csv(file="data/ag.csv",header=TRUE)
```

读入固定宽度文件

```
mydata<-read.fwf("data.txt",widths=c(1,4,3))
```

redLines(),scan()

大部分情况下，用 *read.table* 函数可以将文本文件读入 *R*，但有时也有无法使用的时候，如文件中的观察可能是多行的，这时就要使用 *readLines()* 可以用 readLines 交互式的输入数据 *scan() 可以读入更复杂的文件格式

读入数据练习

- 将世界城市列表导入到 R
- 导入数据 "data/cross.txt" 然后将 z=a 的数据输出成 "cross_a.rda"

批量读入多个外部文件

- 方法一：保存成独立文件
- 方法二：合并保存成 list

```
setwd("~/rproject/FinanceData/data/csv")
fileName <- dir()

cls<-c("character","character","character","numeric","numeric","numeric","numeric")
vn<-c("scode","sname","date","lagclose","close","range","turnover","x")
n<-length(fileName)
stock<-list(length=n)
```

```
 stock[[1]]<-read.table(fileName[1],header = T,sep = ",",stringsAsFactors = F,fileEncoding = "GB2312",c


# 方法一

for (i in 1:n){
 stock[[i]]<-read.table(fileName[i],header = T,sep = ",",stringsAsFactors = F,fileEncoding = "GB2312",c

}


# 方法二
for(i in 1:nfile){
assign(paste("s",scode[i], sep=""),read.table(fileName[i],header = T,sep = ",",stringsAsFactors = F,fil
}
# 方法三

fileName <- dir()
cls <- c("character","character","character","numeric","numeric","numeric","numeric")
stockslist<-list(length=n)

stocklist<- lapply(fileName,function(x){
read.table(x,header = T,sep = ",",stringsAsFactors = F,fileEncoding = "GB2312",colClasses = cls ,col.na
} )

lapply(fileName,mean)
allstcok<- do.call(rbind,stocklist)
```

读入 EXCLE 文件

- 远离 EXCEL !!!

```
# Support 64bit system
install.packages("XLConnect")
library("XLConnect")
df = readWorksheetFromFile("data.xls",
                           sheet=1, header=TRUE)
```

```
library(xlsx)#very slow
df<-read.xlsx("excelfile.xlsx",
              sheetIndex=1,header=TRUE,
              colIndex=,rowIndex=)
```

```
install.packages("RODBC")
library(RODBC)
channel <- odbcConnectExcel("myfile.xls")
mydataframe <- sqlFetch(channel, "mysheet")
odbcClose(channel)
```

通过 ODBC 访问

```
# Only support 32-bit system.
library(RODBC)
myconn <-odbcConnect("mydsn", uid="user", pwd="password")
crimedat <- sqlFetch(myconn, Crime)
pundat <- sqlQuery(myconn, "select * from Punishment")
close(myconn)
```

访问 ORACLE

- RJDBC 配置说明

读入比较大的数据文件

- Use data.table library fread()
- 使用 read.table 时明确 colClasses 和 nrows，设置 comment.char=“”

读取其他统计软件数据

- library foreign
- library haven **New**

– 支持 SAS SPSS Stata

- StatTransfer 可以直接将任意统计软件的数据集进行转化

```
library(haven)
aa<-read_sas("data/Fama1973data/fama.sas7bdat")
```

访问 ORACLE

- Using RORACLE package

访问 Sqlite

- Using RSQLite package
- 管理工具 Navicat
- library(“sqldf”)

```
library("RSQLite")
drv <- dbDriver("SQLite")
con <- dbConnect(drv, dbname = "d:/mydb.s3db")
db_u<-dbGetQuery(con, "select * from table1" )
dbDisconnect()
```

输出数据集

```
write.table(dataframe, file = "output.csv",
            sep = ",", col.names = NA,append=TRUE)
save(df,file="data/df.RData")

load("data/df.RData")
```

清理内存

```
ls()
rm(list=ls())
a<-1:10
b<-1:10
rm(a)
#Ctrl+L
```

网络爬虫

- 天气数据爬虫程序
- 爬取Wikipedia article traffic statistics数据

本周“大牛”

- Hadley Wickham 是 RStudio 的首席科学家以及 Rice University 统计系的助理教授。他是著名图形可视化软件包 ggplot2 的开发者，以及其他许多被广泛使用的软件包的作者，代表作品如 dplyr、reshape2 等。
- 统计之都对他的采访