

在R语言中如何导入各种格式的数据

丑高武

- 引言
- 搜索资料
- 导入文件的一些基本函数
 - 选择路径
 - 基本导入函数
 - 设置路径
 - load
 - scan
 - readLines
 - read.table family
 - read.fwf
- 从clipboard导入文件
- 导入基本类型的文件，txt和csv
- 如何导入excel文件

引言

在学习工作中，我们会遇到各种格式的数据，例如excel文件，csv文件，txt文件，数据库文件等等。不同的文件在R中有不同的导入方式，甚至同一个文件在R语言中也有不同的导入方式。如何在R语言中正确的导入这些文件呢？

本教程所涉及的文件已经上传到对应的data文件夹中。

搜索资料

进入网站

<http://www.aol.com/> (<http://www.aol.com/>)

在搜索栏输入: how to import xxx file in r. where xxx is replaced by the kind of file you need to import.

导入文件的一些基本函数

选择路径

导入文件的第一步，是知道文件放在哪里。有两种方法可以正确的找到你的文件：

- 一.获取R的当前路径，然后把文件放到当前路径下；
- 二.找到你目前文件放置的路径。

首先介绍第一种方法，在R语言中`getwd()`函数可以帮你获取当前路径

```
getwd()
```

```
## [1] "/Users/chougaowu/Documents/R code/Asset_pricing_hw"
```

记住此路径，并且把文件置于该路径之下即可。

第二种方法

windows用户右键所需导入的文件，在属性栏中找到文件的地址，例如：

C:\Program Files (x86)\R-3.2.2

把其中的\改为/，即：

C:/Program Files (x86)/R-3.2.2

修改符号的时候一定要注意符号需要在英文输入法下，不然会出错。

mac用户找到正确的英文路径的方法稍显麻烦，在下面的链接贴出了如何显示mac英文路径的方法。

<http://carlme.blog.163.com/blog/static/18371632720125773115533/>
(<http://carlme.blog.163.com/blog/static/18371632720125773115533/>)

mac下还有一种方法，打开你的终端(terminal)，将所要打开的文件拖拽进去，即可获取该文件的路径。
例如：

/Users/chougaowu/Desktop/wiserclub导入导出所需数据/2015年常用利率日序列mac.csv

基本导入函数

找到了文件之后，再来介绍一下基本的导入函数。

设置路径

```
# 设置文件路径。  
# 设置为你把文件下载的位置。  
path <- "/Users/chougaowu/Desktop/wiserclub导入导出所需数据"  
  
# 找到需要导入的文件名，例如：2015年常用利率日序列mac.csv  
# 两种方式创建完整的路径  
full_path <- file.path(path, "2015年常用利率日序列mac.csv")  
full_path <- paste(path, "2015年常用利率日序列mac.csv", sep = "/")
```

load

Reload datasets written with the function save()

载入由R保存的文件，例如rda数据。

```
load(file = "/path/file_name")
```

```
full_path <- file.path(path, "2015年常用利率日序列mac.rda")  
load(file = full_path)
```

scan

Read data into a vector or list from the console or file

将各种格式的数据导入成一个向量。

scan is a pretty low-level function and in almost every case you're better off using read.table() or read.csv()

```
scan(file = "/path/file_name", # 选择文件路径
      what = "character", # 指定读入数据的类型
      sep = ",", # 指定分隔符号
      skip = 1 # 跳过文件的前多少行
    )
```

readLines

Read some or all text lines from a connection.

将文件一行行读入。

```
readLines(con = "/path/file_name")
```

```
full_path <- file.path(path, "帝国总裁的丑妻mac.txt")
zongcai <- readLines(con = full_path)
str(zongcai)
```

```
## chr [1:25276] "声明：本书为爱奇艺电子书(www.i7wu.cn)的用户上传至其在本站的存储空间，本站只提供TXT全集电子书存储服务以及免费下载服务，以下作品内" | __truncated__ ...
```

```
zongcai[100:115]
```

```
## [1] "  依然在想她叫什么？ "
## [2] "  家在那里？ "
## [3] "  她为什么会掉落海中？ "
## [4] "  脸上为什么会有这么丑的黑色胎记？ "
## [5] "  有没有人焦急的在找她？ "
## [6] "  双手捂住脑袋，蓝七七强迫自己一而再再而三的深想，直到她扯着头发，头皮发痛，脑袋还
  是一阵空荡荡。"
## [7] "  她有丝气馁，心情灰暗到了极点，落魄踱着步子，漫无目的缓缓向前走。"
## [8] "  蓝七七整个人就像没了灵魂的躯壳一般，闲散游荡，因为，她不知道该去哪里，而那里又是
  属于她该去的地方....."
## [9] "  当她走到游轮拐角的尽头，蓦然地，冒出一个莽撞不安的男人，与她相撞在了一起。"
## [10] "  蓝七七差点跌倒，却被那沾着酒气的男人拽住了手腕。"
## [11] "  “跟我走！”男人有着一张极其英俊的脸，脸形棱角分明，鼻尖英挺，薄薄地唇很性感，棕色
  的眼睛剔透如琉璃，只是隐藏着一丝常人不易觉察的冰冷与低沉。"
## [12] "  奇怪的是，那男人额头挥汗如雨，眼眸猩红，仿佛在抑制着什么，他修长的手不断扯着领口
  的白衬衫，好像很难受。"
## [13] "  “我不认识你.....放开我，我不去！”蓝七七使命甩开江霄云紧抓住她手腕的手，想摆脱他的钳
  制。"
## [14] "  女人的直接告诉她，这个男人很危险！"
## [15] "  "
## [16] "5 危险的男人2"
```

read.table famliy

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

最常用的导入文件方式，根据分隔符号，将文件导入为dataframe，即一张“表”。

```
read.table(file = , # 符合格式的各种文件
           header = TRUE, # 是否把文件的第一行输入为“表头”，即列名
           sep = " ", # 文件以什么分隔符号作为分隔
           row.names = TRUE, # 是否把文件的第一列输入为行名
           na.strings = " ", # 用什么符号替代NA值
           colClasses = c("character","integer"), # 指定不同列的类型，提高导入速度
           stringsAsFactors = FALSE, # 是否把文件全部按照factor的格式导入，一般都需要
           改为FALSE
           encoding = "UTF-8" # 编码格式，一般不需要修改。
           )
```

```
read.csv() # 导入逗号分隔文件
```

```
read.delim() # 导入以tab分隔的文件
```

read.fwf

Read a table of fixed width formatted data into a data.frame.

按照固定宽度导入文件。

```
read.fwf(file = ,
         widths = )
```

从clipboard导入文件

用户一个最原始的需求是，我复制了一堆“东西”，如何粘贴到R语言中。

一个最简单的方法是直接赋值给一个变量。

首先在网站或者其他文件中找到需要粘贴的内容粘贴好。

```
xuxiang <- "起底徐翔：曾被指股市第一恶庄
```

11月1日晚，一则关于泽熙徐翔被逮捕的消息迅猛传开。直到当晚23:50，新华网终于刊登了一则消息《泽熙公司总经理徐翔涉嫌获取股市内幕操纵股价》。

该消息称，记者从公安部获悉，泽熙投资管理有限公司法定代表人、总经理徐翔等人通过非法手段获取股市内幕信息，从事内幕交易、操纵股票交易价格，其行为涉嫌违法犯罪，近日被公安机关依法采取刑事强制措施。目前，相关案侦工作正在严格依法进行中。

徐翔，上海泽熙投资管理有限公司总经理。1976年出生，17岁带3万入市，对股市很有天分，19岁时曾被上海黑帮胁迫，替黑帮操盘，相传两大黑帮为争他做操盘手而打了一次。几年后，90年底末成为浙江龙头老大，被称为浙江敢死队“总舵主”，而叶大户当时只是“堂主”。目前他的个人资产已经达到40亿元。

从3万炒到40亿的投资传奇

1993年，还是个高中生的徐翔带着父母给的几万元本钱进入股票市场。“那个时候，宁波老百姓家里几万元很正常，而且家里就我一个小孩。”徐翔说。

18岁，徐翔放弃高考，专心投资，并得到家人同意。

徐翔对股市热点的把握非常敏感，个人资产迅速累积，成为营业部高手。

“我学习股票，看书，听券商培训，也看国外投资方法。三人行必有我师，对宏观经济，也懂一点。”徐翔表示。到20世纪90年代后期，徐翔和几个以短线擅长的朋友被市场冠名“宁波涨停板敢死队”，投资风格彪悍，被“敢死队”选中的股票，大进大出，高起高落。这批人以银河证券宁波解放南路营业部为核心，成员20人左右，核心是徐翔、徐晓、张建斌等。徐翔在成名后更被媒体封为“涨停板敢死队总舵主”。

三大短线高手中，徐翔是其中的头号人物，也是最年轻的一位；2号人物姓吴，35岁左右，这两位不仅经历相似，连炒股水平也是旗鼓相当；3号人物徐海鸥的，1975年出生，上大学时就开始炒股，1997年在北京商学院毕业后没找工作，直接回宁波专职炒股。在做散户时，徐海鸥与徐翔结识，便从散户室搬到4楼，为小他1岁的徐翔做助手，并逐渐成长为“敢死队”的核心成员。

名气渐响之后，徐翔的一举一动开始为外界所注意，特别其在营业部的仓位、持股，很容易被泄露，有人专门分析其持股以跟风投资。徐翔对此颇为苦恼，称很希望阳光私募能争取到机构席位，摆脱跟风者。

2005年，徐翔从宁波迁到上海，经历了A股的一波大牛市，2009年成立泽熙投资，经过三年的发展，目前总管理规模在70亿元左右，规模仅次于目前阳光私募中管理资金规模排名第一的上海重阳投资。“我们70亿的投资中，有至少40至50只股票。”徐翔说。

```
str(xuxiang)
```

```
## chr "起底徐翔：曾被指股市第一恶庄\n    11月1日晚，一则关于泽熙徐翔被逮捕的消息迅猛传  
开。直到当晚23:50，新华网终于刊登了一则消息《泽”| __truncated__
```

如果利用我们刚刚提到的函数，也可以实现赋值。

```
# under windows
xuxiang <- readLines("clipboard")
xuxiang <- read.table("clipboard")

# under mac OS
xuxiang <- readLines(pipe("pbpaste"))
xuxiang <- read.table(file = pipe("pbpaste"), header = F)
```

导入基本类型的文件，txt和csv

```
# 两个文件需要导入
# 2015年常用利率日序列mac.txt 和 2015年常用利率日序列mac.csv
# 获取全路径
full_path_txt <- file.path(path,"2015年常用利率日序列windows.txt") # windows
full_path_csv <- file.path(path,"2015年常用利率日序列windows.csv") # windows
full_path_txt <- file.path(path,"2015年常用利率日序列mac.txt") # mac OS
full_path_csv <- file.path(path,"2015年常用利率日序列mac.csv") # mac OS

# 导入数据
interest_data <- read.csv(full_path_csv,header = T,blank.lines.skip = T,strings
AsFactors = F)
interest_data <- read.table(full_path_txt,"\t",header = T,stringsAsFactors = F)
```

如何导入excel文件

导入excel的方式多种多样，详见：

<http://www.milanor.net/blog/?p=779> (<http://www.milanor.net/blog/?p=779>)

有时候选择太多也不是一件好事，下面仅摘要一种方法。

```
# 安装并且载入XLConnect包
install.packages("XLConnect")
library(XLConnect)

full_path_xlsx <- file.path(path,"2015年常用利率日序列both.xlsx")
full_path_xls <- file.path(path,"2015年常用利率日序列both.xls")

workbook <- loadWorkbook(full_path_xlsx)
interest_df <- readWorksheet(workbook, sheet = "Sheet1", header = TRUE)
head(interest_df)

workbook <- loadWorkbook(full_path_xls)
interest_df <- readWorksheet(workbook, sheet = "Sheet1", header = TRUE)
head(interest_df)
```