

R语法基础



多了去的YangXuLei (/u/9d619ad71963) [+ 关注](#)

2016.12.22 18:45* 字数 3454 阅读 160 评论 2 喜欢 2

(/u/9d619ad71963)

一、基础

R是一种语法非常简单的表达式语言(expression language),大小写敏感。可以在R环境下使用的命名字符集依赖于R所运行的系统和国家(系统的locale 设置)、允许数字、字母、“.”和“_”

1.命名

命名必须以“.”或者字母开头，以“.”开头时第二个字符不允许是数字。

2.基本命令

基本命令要么是表达式 (expressions) ， 要么就是赋值 (assignments) 。

表达式：命令将被解析，并将结果显示在屏幕上，同时清空该命令所占内存。

赋值：命令将被解析，并把值传给变量，但结果不会自动显示在屏幕上。

命令可以被“;”隔开或者另起一行。基本命令可以通过大括弧{}放在一起构成一个复合表达式。

注释：一行中以井号“#”开头 换行：如果一条命令在一行结束的时候在语法上还不完整，换行提示符，默认是+

3. 基本的对象

R创建和控制的实体被称为对象。它们可以是变量，数组，字符串，函数，或者其他通过这些实体定义的一般性的结构。

矩阵 (matrix) 或者更为一般的数组 (array) 是多维的广义向量。实际上，它们就是向量，而且可以同时被两个或者更多个索引引用，并且以特有的方式显示出来。

因子 (factor) 为处理分类数据提供了一种有效方法。

列表 (list) 是一种泛化 (generalform) 的向量。它没有要求所有元素是同一类型，许多时候它本身就是向量和列表类型。列表为统计计算的结果返回提供了一种便利的方法。

数据框 (data frame) 是和矩阵类似的一种结构。在数据框中，列可以是不同的对象。可以把数据框看作是一个行表示观测个体并且 (可能) 同时拥有数值变量和分类变量的‘数据矩阵’。许多实验数据都可以很好的用数据框描述，处理方式是分类变量而响应值是数值变量。

函数 (function) 是可以保存在项目工作空间的R对象。该对象为R提供了一个简单而又便利的功能扩充方法。当编写你自己的函数在R会话过程中，对象是通过名字创建和保存的。objects(),ls()可以显示当前会话的对象名字。rm()可以删除对象。

对象持久化

R会话中创建的所有对象可以永久地保存在一个文件中以便于以后的R会话调用。在每一次R会话结束的时候，你可以保存当前所有可用的对象。如果你想这样做，这些对象将会写入当前目录下一个叫.RData的文件中，并且所有在这次会话中用过的命令行都会被保存在.Rhistory的文件中。当R再次在同一目录下启动，这些对象将从这个文件中重新导入工作空间。同时，相关的历史命令文件也会被导入。

二、向量和赋值

向量是由一串有序数值构成的序列

```
x <- c(10.4, 5.6, 3.1, 6.4, 21.7);x
[1] 10.4 5.6 3.1 6.4 21.7
```

函数c()完成的赋值语句，这里的函数c()可以有任意多个参数，而它返回的值则是一个把这些参数首尾相连形成的向量。

赋值也可以用函数assign()实现。

```
assign("x", c(10.4, 5.6, 3.1, 6.4, 21.7));x
[1] 10.4 5.6 3.1 6.4 21.7
```

赋值符<-,>可以看作是该命令一个语义上的缩写。

```
c(10.4, 5.6, 3.1, 6.4, 21.7) -> x;x
[1] 10.4 5.6 3.1 6.4 21.7
```

1.向量运算

在算术表达式中使用向量将会对该向量的每一个元素都进行同样的算术运算。出现在同一个表达式中的向量最好是长度一致。如果他们的长度不一样，该表达式的值将是一个和其中最向量等长的向量。表达式中短的向量会被循环使用以达到最长向量的长度。对于一个常数就是简单的重复。

```
x <- 1:5
y <- 1:3
v <- 2*x + y + 1;v
[1] 4 7 10 10 13
```

2.逻辑向量

逻辑向量元素可以被赋予的值,有TRUE , FALSE 和NA 逻辑向量可以由条件式 (conditions) 产生 temp <- x > 13

```
temp <- 1 < 3;temp
```

```
[1] TRUE
```

3.字符向量

字符向量就是字符串，可以用双引号和单引号作分割符。

paste():可以把单独的字符连成字符串，可以有任意多的参数。参数中的任何数字都将被显式地强制转换成字符串，而且以同样的方式在终端显示。默认的分隔符是单个的空格符。

修改分隔符换成空字符串""

```
labs <- paste(c("X","Y"), 1:10, sep="");labs
[1] "X1" "Y2" "X3" "Y4" "X5" "Y6" "X7" "Y8" "X9" "Y10"
```

4.索引向量：通过索引值可以选择和修改一个数据集的子集

一个向量的子集元素可以通过向量名后面的方括号中加入索引向量得到。如果一个表达式的结果是向量，则我们可以直接在表达式的末尾方括号中加入索引向量以得到结果向量的子向量。

逻辑向量:索引向量必须和被挑选元素的向量长度一致。向量中对应索引向量元素为TRUE的元素将会被选中，而那些对应FALSE的元素则被忽略。

```
x <- c(TRUE, TRUE, FALSE, TRUE, FALSE)
y <- x[x];y
[1] TRUE TRUE TRUE
```

正整数向量:索引向量必须是1,2,...,length(x)的子向量。索引向量中索引对应的元素将会被选中，并且在结果向量中的次序和索引向量中的次序一致。这种索引向量可以是任意长度的，结果向量的长度和索引向量完全一致。

```
x <- 1:10
x[1:5]
[1] 1 2 3 4 5
```

负整数向量:这种索引向量指定被排除的元素而不是包括进来。

```
x <- 1:10
x[-(1:5)]
[1] 6 7 8 9 10
```

字符串向量:这可能仅仅用于一个对象可以用names属性来识别它的元素。名字向量的子向量可以像上面第二条提到的正整数标签一样使用。

```
fruit <- c(5, 10, 1, 20)
names(fruit) <- c("orange", "banana", "apple", "peach")
lunch <- fruit[c("apple", "orange")];lunch
apple orange
1 5
```

三、运算符

算术运算符：+ - * /

逻辑运算符：< , <= , > , >= , ==, !=, &, |, !

数学函数：log , exp , sin , cos , tan , sqrt , max , min , range , length , sum , prod , var

注：var(x): 等价于 $\text{sum}((x - \text{mean}(x))^2) / (\text{length}(x) - 1)$

四、控制语句

1.条件语句:if

```
if (expr1) expr2 else expr3
```

2.循环控制：for,repeat,while

```
for (name in expr1) expr2
```

其中name 是循环变量，expr1是一个向量表达式，而expr2常常是根据虚拟变量name而设计的成组表达式。在name访问expr1所有可以取到的值时，expr2都会运行。

例如使用for和if：

```
for (i in 1:5) {
```

```
  if (i %% 2 == 0) {
```



```
print(paste(i, "is even"))
```

```
} else {
```

```
print(paste(i, "is odd"))
```

```
}
```

- }

```
[1] "1 is odd"
```

```
[1] "2 is even"
```

```
[1] "3 is odd"
```

```
[1] "4 is even"
```

```
[1] "5 is odd"
```

警告：相比其他程序语言，R代码里面很少使用for().执行效率很低

repeat expr

while (condition) expr

while例子：

```
x <- 1:5
sum <- 0
i <- 1
while (i <= length(x)) {
```

```
sum <- sum + x[i]
```

```
i <- i + 1
```

- }

```
sum
[1] 15
repeat例子：
```

```
x <- 1:5
sum <- 0
i <- 1
repeat {
```

```
if (i > length(x)) {
```

```
break
```

```
}
```

```
sum <- sum + x[i]
```

```
i <- i + 1
```

```
• }
```

```
sum
```

```
[1] 15
```

关键字break：可以用于结束任何循环，甚至是非常规的。它是结束repeat循环的唯一办法。关键字next：可以用来结束一次特定的循环，然后直接跳入"下一次"循环，类似大家熟悉的continue。

五、生成正则序列

1:30语句等价于向量c(1,2,...,29,30)，30:1语句可用来产生一个逆向的数列。seq：数列生成中最为常用的工具

```
seq(1,5,1)
```

```
[1] 1 2 3 4 5
```

rep：把一个数的完整拷贝多次，保持数列顺序

```
x <- 1:3
```

```
rep(x, times=2)
```

```
[1] 1 2 3 1 2 3
```

六、缺损值

在某些情况下，向量的元素可能有残缺。当一个元素或者值在统计的时候"不可得到"（not available）或者"值丢失"（missing value），相关位置可能会被保留并且赋予一个特定的值NA。任何含有NA数据的运算结果都将是NA。函数is.na(x)返回一个和x同等长度的向量。它的某个元素值为TRUE 当且仅当x中对应元素是NA。

```
z <- c(1:3,NA); ind <- is.na; ind
```

```
[1] FALSE FALSE FALSE TRUE
```

第二种"缺损"值，也称为非数值NaN(Not a Number)

0/0 或 Inf

七、对象

1.内在属性：模式和长度

R操作的实体是对象。实数或复数向量，逻辑向量和字符串向量之类的对象属于"原子"型的对象，因为它们的元素都是一样的类型或模式。R的对象类型包括数值型，复数型，逻辑型,字符型和原生型。向量必须保证它的所有元素是一样的模式。因此任何给定的向量必须明确属于逻辑性，数值型，复数型，字符型或者原生型。

列表是任何模式的对象的有序序列。列表被认为是一种"递归"结构而不是原子结构,因为它们的元素可以以它们各自的方式单独列出。函数和表达式也是递归结构。

所有对象都有模式(mode)和长度(length)两个内在属性

```
x <- 1:3
mode(x)
[1] "numeric"
length(x)
[1] 3
2.外部属性
```

attributes(x):给出对象当前定义的非内在属性的列表。 attr(object, name): 可以用来设置对象的外部属性。

```
z<-c(1:3,NA)
attr(z, "name") <- "abc"
attributes(z)
$name
[1] "abc"
3.对象的类属性
```

R里面的所有对象都属于类(class), 可以通过函数class(x)查看。 对于简单的向量, 类是对应的模式(mode): "numeric", "logical", "character" 或者"list", 其他的类型, 像"matrix", "array", "factor" 和"data.frame" 就可能是其他值。

八、因子(factor)

假定我们有一份来自澳大利亚所有州和行政区的信息样本以及他们各自所在地的州名。

```
state <- c("tas", "sa", "qld", "nsw", "nsw", "nt")
在字符向量中, "有序"意味着以字母排序的。
```

创建因子factor :

```
statef <- factor(state)
statef
[1] tas sa qld nsw nsw nt
Levels: nsw nt qld sa tas
levels():可以用来得到因子的水平 ( levels ) 。
```

```
levels(statef)
[1] "nsw" "nt" "qld" "sa" "tas"
函数tapply()和不规则数组
```

为计算样本中每个州的平均收入, 我们可以用函数tapply() :

```
incomes <- c(60, 49, 40, 61, 64, 60)
incmeans <- tapply(incomes, statef, mean)
incmeans
nsw nt qld sa tas
62.5 60.0 40.0 49.0 60.0
```

函数tapply() 可以用来处理一个由多个分类因子决定的向量下标组合。

九、小结

R语言中使用的命名字符集依赖于R所运行的系统和国家(系统的locale 设置)、允许数字、字母、“.”和“_”，命名必须以“.”或者字母开头，以“.”开头时第二个字符不允许是数字，<-，->，R语言的基本对象有矩阵、因子、列表、数据框和函数，assign()都可以赋值，c()可以有任意多个参数，而它返回的值则是一个把这些参数首尾相连形成的向量，paste()可以把单独的字符连成字符串，可以有任意多的参数，运算符有算术运算符、逻辑运算符、数学函数，控制语句有if...else,for,while,repeat,seq()是数列生成中最为常用的工具，rep()把一个数的完整拷贝多次，并保持数列顺序，在某些情况下，向量的元素可能有残缺。当一个元素或者值在统计的时候“不可得到”(not available) 或者“值丢失”(missing value)，相关位置可能会被保留并且赋予一个特定的值NA，对象有内在属性和外在属性，因子中常用levels()和tapply()操作。

数据科学 (nb/3752668) 举报文章 © 著作权归作者所有



多了去的YangXuLei (/u/9d619ad71963)

写了 93655 字，被 61 人关注，获得了 139 个喜欢

(/u/9d619ad71963)

+ 关注

知识障碍患者,知识&代码搬运工,CS在读,搞了软件研究起了数学,搭上了机器学习车却没入门,日语N6英...


赞赏支持

喜欢 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-like-button) | 2




更多分享

(http://cwb.assets.jianshu.io/notes/images/7863774



下载简书 App ▶

随时随地发现和创作内容



(/apps/download?utm_source=nbc)



登录 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-comment-form)

评论

智慧如你，不想发表一点想法 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-nocomments-text) 咩~

R语言学习笔记-文本挖掘之字符处理 (1) (/p/5bd0846b0d7d?utm_campa...

在挖掘分析的过程当中对字符串的处理是极为重要的，且出现也较为频繁，R语言作为当前最为流行的开源数据分析和可视化平台，虽然文本的处理并不是它的强项，但是R语言还是包含大量的字符串操作工具，本章...



Crazy_Rabbit (/u/5e2d5533a717?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

R语言基础 (/p/5ca6a04ec0fe?utm_campaign=maleskine&utm_content=...

本节概述R语言的基本知识。用稍微简单的方式讲述，而将一些细节问题轻轻带过是有必要的。虽然包含一节关于编程的简单介绍，但是这部分内容的重点是介绍交互界面使用中的一些有用的东西，而不是具体编程。..



出版圈郭志敏 (/u/54b59ee78c42?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/690f8fbd6390?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

《R的极客理想》系列图书作者张丹：用R语言把数据玩出花样 (/p/690f8fbd...

感谢关注天善智能，走好数据之路↑↑ 欢迎关注天善智能，我们是专注于商业智能BI，大数据，数据分析领域的垂直社区，学习，问答、求职一站式搞定！ 本文作者：天善智能社区专家张丹 前言 作为数据分析师，...



天善智能 (/u/474e739f391e?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

R_数据处理_dplyr&reshape2&tidyr (/p/259067f0ff54?utm_campaign=m...

dplyr：轻松地处理data.frame, data.table以及多种数据库为基础的数据，实现选择、变换、分组等 plyr：轻松地在vector, list, data.frame之间做分组变换，实现拆分、变换、合并的操作 reshape2：横向、纵向做数...



Jerry_Cai (/u/1998962ab991?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

R_字符串处理_stringr (/p/3205695d623c?utm_campaign=maleskine&ut...

R_字符串处理_stringr stringr：对一个单元格内的 [元素] 进行操作，而非把一个单元格看做一个整体来对待。 文本文件的读写 这里的文本文件指的是非表格格式的文件，如纯文本文件，html文件 获取文件目录信息...



Jerry_Cai (/u/1998962ab991?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/62748bbd1fd8?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

要不是为了见你，才懒得参加同学会 (/p/62748bbd1fd8?utm_campaign=m...

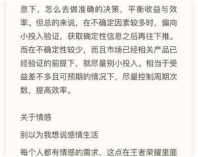
如果不是莎莎的怂愿，毫无营养的同学聚会我是绝对绝对不想去参加的。相隔多年未曾谋面，一见面却互相寒暄，话题永远绕不开：“现在做什么，收入多少，准备结婚了吗……” 打探彼此的近况似乎不是真心关心，逐...




瓶子是个深井冰 (/u/a94d33b29ecb?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/072c62e62511?)




utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
是日随记 (/p/072c62e62511?utm_campaign=maleskine&utm_content=n...

 橙柠呀呀呀 (/u/627b7c10a8ea?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

我的童年伙伴J (/p/0e8e5c47b502?utm_campaign=maleskine&utm_cont...

J是我的童年伙伴，比我大一岁，小时候我和他一样喜欢上山摘果，喜欢下河捕鱼，不同的是我是女孩子，他是男孩子，我成绩好考上了大学，而他，初中毕业就外出谋生了。去年年底我决定要结婚了，按照习俗，...

 余有木木夕 (/u/761ecabb2a0c?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/9be06711953e?)



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
诗|近与远 (/p/9be06711953e?utm_campaign=maleskine&utm_content=n..


远方还在期待探险者足迹的临近，近处的杨柳却已衰颓。树桩还在沐浴斜阳的余辉，丰富的年轮镌刻了太多孤独的记忆。老人坐在树桩上眺望大海，在水天一处是梦开始的地方。那蓬勃的心跳已经持续了八十多...

 金忆潇阳 (/u/3b4a4a12c157?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

欺负没有理由 (/p/5833ecf5a86c?utm_campaign=maleskine&utm_conten...

小学同桌一直是男生，很羡慕两个女生坐一起。恰巧同桌是个闷，成绩平平人也呆。欺负起来特别带感。总爱在眼保健操时扔他东西，趁他低头捡东西时将窗户推到他上方，看他抬头时被撞。若是坐到四组，便把他..

 醉茶觞 (/u/bf5572f24226?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)