Вероятности и статистика с R

Асен Чорбаджиев

October 19, 2017

1 Текст

Основният тип (class of object) на текстови данни в R е character. За изписване на стринг от character се използват единични или двойни кавички:
'a character string using single quotes'
"a character string using double quotes" Позволено е използването на единични кавички в стринг означен с двойни кавички или стринг от двойни кавички в стринг с единични:
"The 'R' project for statistical computing"
'The "R" project for statistical computing'
Но изпозването на подстринг с двойни кавички в стринг с двойни или такъв с единични в стренг с единични е грешно:
"This "is" totally unacceptable"
'This 'is' absolutely wrong'
Функцията за създаване на вектор от стрингове е character(). За основните операции за работа със стрингове се използват следните 'C'-style функции:

- paste(..., sep = " ", collapse = NULL) - Функцията взима един или няколко обекта, превръща ги в "character" и ги конкатенира за получаването на един или няколко стринга:
  PI = paste("The value of pi is", pi)
  IR = paste("Just", "Do", "It", sep = ".")
- *print(my_string)* - Принтиране на текст.
- cat() - конкатенира обекти и ги записва или на екран или на файл.
- format() - Форматиране на R object, като ги интерпретира за стрингове:
  format(pi, digits = 2)
- sprintf() - C-style format
- nchar() - Брой символи в стринг.

- tolower(), toupper() - конвертира от главни към малки букви и обратно.
- substring() ,substr() - връща субстринг:
  substr("abcdef", 2, 4)
  substring("abcdef", 2, 4)
- sort() сортиране
- intersect(), setdiff() - съвпадения или разлики:
  set1 = c("some", "random", "few", "words")
  set2 = c("some", "many", "none", "few")
  intersect(set1, set2)
  setdiff(set1, set2)

2 Времеви данни

Много често наличните данни са свързани с времеви измервания. За работа с такъв тип данни в R се използва date/time формат за чиято основа е взет UNIX POSIX. Обикновенно създаването на подобен тип структури става посредством превръщането на текстови стринг до date/time. Като е важно да се запомни, че в R подредбата се подразбира, че започва от най-ранната дата последователно до най-късната. Това обръщане става с оператора as.POSIXct():
start.date <- as.POSIXct("2004-01-01 00:00", tz = "GMT")
end.date <- as.POSIXct("2004-12-31 23:00", tz = "GMT")
Вече създаден, обектът подлежи да сравнителни операции. Например, селектирането на подмножество на елементи от df по време съгласно наличните записи с колона за време datetime:
df[df.datetime>=start.date & df.datetime <= end.date,]
Когато се налага преформатиране на date/time, основната функция е C-type format() и strptime(), като основният разделител за шаблона е "%". Например,
format(datetime, "%Y − %m − %d %H − %M − %S")
Други полезни функционалности за работа с времеви данни:

- difftime(t1,t2, units='mins') = t1-t2 in minutes
- as.POSIXlt() - calendar time.
- Ако обектът lt е от тип POSIXlt то следният дъстъп е позволен:
  lt$year
  lt$hour=lt$hour+1
- date()

3 Цикли и проверки

Логическите проверки в R не се различават от останалите езици. За целта се използват операторите if() и else(), с единственото условия зя проверка за наличие на NULL логически състояние

Масовото използване на цикли не е удачно в R поради неефективност. Но когато това се налага синтаксисът е следния:
for(i in 1:n)
{ ... *Body* ... }
while(true)
{ ... *Body* ... }

4 Функции

Основната причина за създаването на собствени функции в R е дефинирането на функционалност за многократно използване и/или структуриране на кода. Синтаксисът е сравнително прост:
func_name<−function(input_params,...)
{ .....
BODY
.....
return (ret)}

Важно е да се запомни, че локално дефинираната променлива предефинира глобалната такава само в тялото на функцията:
x=5
foo=function(a)
{
x= 4
return (a+x)
}
foo(2)
x

5 Упражнения:

- От таблицата mtcars да се извади списък от всичките марки автомобили с 6 цилиндъра и 4 скорост. Да се провери дали измежду тях са Merc 280 и Merc 280C ?

- **Зад. 2.** : Да се свалят данните от Daily share prices (time series).:
  1.1 Да се извади графика на цените и цените с обемите във времева серия от данни.
  1.2 Да се извадят котировките само във времевия интервал 12:00 - 14:00

- **Зад. 3.** Да се напише фукция за решение на задача 4 от Week 1.