Hands-On Data Science with R
Strings in R

Graham.Williams@togaware.com

4th June 2016

Visit http://HandsOnDataScience.com/ for more Chapters.

In this module we introduce tools available in R for handling and processing strings.

```
# Load the required packages from local library into R session.

library(rattle)   # Weather dataset.
library(stringi)  # The string concat operator %s+%.
library(stringr)  # String manipulation.
library(magrittr) # Pipelines for data processing: %>% %T>% %<>%.
```

As we work through this chapter, new R commands will be introduced. Be sure to review the command's documentation and understand what the command does. You can ask for help using the `?` command as in:

```
?read.csv
```

We can obtain documentation on a particular package using the *help=* option of `library()`:

```
library(help=rattle)
```

This chapter is intended to be hands on. To learn effectively, you are encouraged to have R running (e.g., RStudio) and to run all the commands as they appear here. Check that you get the same output, and you understand the output. Try some variations. Explore.

1 Test

Load a dataset as strings, one line is a string, returning a vector of strings, using the function base::readLines().

```
dsname  <- "weather" # Dataset name.
ftype   <- "csv"     # Source dataset file type.
csvname <- dsname %s+% "." %s+% ftype
ds <-
  csvname %>%
  system.file(ftype, ., package="rattle") %>%
  readLines()
```

A sample of the data.

```
head(ds)
```

```
## [1] "\"Date\",\"Location\",\"MinTemp\",\"MaxTemp\",\"Rainfall\",\"Evaporat...
## [2] "2007-11-01,\"Canberra\",8,24.3,0,3.4,6.3,\"NW\",30,\"SW\",\"NW\",6,20...
## [3] "2007-11-02,\"Canberra\",14,26.9,3.6,4.4,9.7,\"ENE\",39,\"E\",\"W\",4,...
## [4] "2007-11-03,\"Canberra\",13.7,23.4,3.6,5.8,3.3,\"NW\",85,\"N\",\"NNE\"...
## [5] "2007-11-04,\"Canberra\",13.3,15.5,39.8,7.2,9.1,\"NW\",54,\"WNW\",\"W\...
## [6] "2007-11-05,\"Canberra\",7.6,16.1,2.8,5.6,10.6,\"SSE\",50,\"SSE\",\"ES...
....
```

Find those strings that contain a specific pattern using base::grep().

```
grep('ENE', ds)
```

```
##  [1]   3  10  23  26  28  36  37  42  43  49  50  54  68  69  71  76  86
## [18]  91  97 101 103 106 108 109 110 118 129 132 133 135 138 145 160 171
## [35] 176 215 222 278 303 304 310 323 341 348 351 357 365
```

```
grep('ENE', ds, value=TRUE)
```

```
##  [1] "2007-11-02,\"Canberra\",14,26.9,3.6,4.4,9.7,\"ENE\",39,\"E\",\"W\",4...
##  [2] "2007-11-09,\"Canberra\",8.8,19.5,0,4,4.1,\"S\",48,\"E\",\"ENE\",19,1...
##  [3] "2007-11-22,\"Canberra\",16.4,19.4,0.4,9.2,0,\"E\",26,\"ENE\",\"E\",6...
##  [4] "2007-11-25,\"Canberra\",15.4,28.4,0,4.4,8.1,\"ENE\",33,\"SSE\",\"NE\...
##  [5] "2007-11-27,\"Canberra\",13.3,22.2,0.2,6.6,2.3,\"ENE\",39,\"E\",\"E\"...
##  [6] "2007-12-05,\"Canberra\",14.5,21.8,0,8.4,9.8,\"ENE\",43,\"ESE\",\"E\"...
....
```


Generated 2016-06-04 16:39:40+08:00

2 Concatenate Strings

The concatenate operation `stringi::%s+%`.

```
"abc" %s+% "def"
## [1] "abcdef"
"abc" %s+% "def" %s+% "ghi"
## [1] "abcdefghi"
c("abc", "def", "ghi", "jkl") %s+% c("mno")
## [1] "abcmno" "defmno" "ghimno" "jklmno"
c("abc", "def", "ghi", "jkl") %s+% c("mno", "pqr")
## [1] "abcmno" "defpqr" "ghimno" "jklpqr"
c("abc", "def", "ghi", "jkl") %s+% c("mno", "pqr", "stu", "vwx")
## [1] "abcmno" "defpqr" "ghistu" "jklvwx"
```

3 Generate Random Text

Useful in testing using `stringi::stri_rand_lipsum()`.

```
stri_rand_lipsum(2)
```

```
## [1] "Lorem ipsum dolor sit amet, accumsan amet quis arcu phasellus facilis...
## [2] "Laoreet scelerisque hendrerit metus integer purus nec. Purus parturie...
```

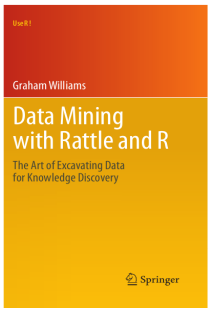
4 Command Summary

This chapter has introduced, demonstrated and described the following R packages, functions, commands, operators, and datasets:

5 Exercises

6 Further Reading

The Rattle Book, published by Springer, provides a comprehensive introduction to data mining and analytics using Rattle and R. It is available from Amazon. Other documentation on a broader selection of R topics of relevance to the data scientist is freely available from http://datamining.togaware.com, including the Datamining Desktop Survival Guide.



This chapter is one of many chapters available from http://HandsOnDataScience.com. In particular follow the links on the website with a * which indicates the generally more developed chapters.

We list below further resources that augment the material we have presented in this chapter.

-
-
- Handling and Processing Strings in R, a freely available ebook by Gaston Sanchez from 2013.
- http://www.rexamine.com/2013/04/properly-internationalized-regular-expressions-in-r/

7 References

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Williams GJ (2009). "Rattle: A Data Mining GUI for R." *The R Journal*, **1**(2), 45–55. URL http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf.

Williams GJ (2011). *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Use R! Springer, New York.

Williams GJ (2015). *rattle: Graphical User Interface for Data Mining in R*. R package version 3.4.2, URL http://rattle.togaware.com/.

This document, sourced from StringsO.Rnw bitbucket revision 152, was processed by KnitR version 1.13 of 2016-05-09 and took 0.8 seconds to process. It was generated by gjw on theano running Ubuntu 14.04.4 LTS with Intel(R) Core(TM) i7-3517U CPU @ 1.90GHz having 4 cores and 3.9GB of RAM. It completed the processing 2016-06-04 16:39:40.

Draft Only

Generated 2016-06-04 16:39:40+08:00