R Navigational Tricks

HDS, Mike Kozlowski

August 2, 2018

R Navigational Tricks

Finding files, setting directories and the like are some of the tasks that can often eat up a lot of time, and you wind up repeating them over and over.

I have found some tricks or methods in R that I find useful, in terms of locating files or directories and working with them. These are sort of shortcuts that save time.

The number one tool for me is the file.choose() function, that opens a file browser window, to let you select files. I used the variable name "stemp" over and over to mean string temporary

Here I use it to navigate to a file I might want to work with, or to a directory I might want to work with

Try selecting the Red Wine file to see this function work

```
stemp=file.choose()
stemp
```

```
## [1] "C:\\Users\\Mike\\Documents\\DAT511\\1-25 class\\winequality-red.csv"
```

We could use stemp in a file read

mydata=read.table(stemp, header=TRUE, sep=",")

or we could use file.choose() directly in the file read

mydata=read.table(file.choose())

We can do a bit of cut and paste of the text in stemp to extract just the directory name, this is handy, for example to change the current working directory

setwd("/Users/hsheets/Dropbox/Data_Analytics/Data_Cleaning/Lectures_Fall_2018/Example_data/Horse_Colic"")

we can get a list of the files in a directory from inside R, which turns out to be a handy thing to be able to do

Change the directory name below to one on your computer and then run the cell

You can use the "Properties" pop up menu on Windows to see the path of a file and cut and paste it into the cell below

```
t=dir("C:\\Users\\Mike\\Documents\\DAT511\\1-25 class")
t
```

```
##  [1] "1-25-class-example.html"      "1-25 class example.pdf"
##  [3] "1-25 class example.Rmd"       "2020-2021_Assessment_Roll.csv"
##  [5] "Completed Files"              "Data_Frame_example.Rmd"
##  [7] "Data_Frames_2_3.Rmd"          "package_mgmt.Rmd"
##  [9] "R_file_Navigation_tricks.Rmd" "Tows.csv"
## [11] "Vector_Calculations_2_2.html" "Vector_Calculations_2_2.Rmd"
## [13] "winequality-red.csv"
```

We could loop through the files, and load each one sequentially using t

the variable t has a list of all the files in the directory

We could momentarily change the working directory, do some operation and then return to our initial working directory

Again, pick another directory name, not the one I used

```
startDir=getwd()
setwd("C:\\Users\\Mike\\Documents\\DAT511\\1-25 class")
dir()
```

```
##  [1] "1-25-class-example.html"      "1-25 class example.pdf"
##  [3] "1-25 class example.Rmd"       "2020-2021_Assessment_Roll.csv"
##  [5] "Completed Files"              "Data_Frame_example.Rmd"
##  [7] "Data_Frames_2_3.Rmd"          "package_mgmt.Rmd"
##  [9] "R_file_Navigation_tricks.Rmd" "Tows.csv"
## [11] "Vector_Calculations_2_2.html" "Vector_Calculations_2_2.Rmd"
## [13] "winequality-red.csv"
```

```
setwd(startDir)
getwd()
```

```
## [1] "C:/Users/Mike/Documents/DAT511/1-25 class"
```

One of the tricks we will use is to get the library path, where R looks for packages, and then modify it, so we can store packages on a remote drive, such as the H: drive at Canisius

curlib=.libPaths() myHdrive="H:/sheets" newpath=c(curlib,mnHdrive) .libPaths(newpath)

This adds "H:/sheets" to the library search path used to find packages

R versions

R.Version() give you a lot of information about the version of R you are running and the environment. This can be important, as some functions do change with the operating system in use.

For example, on a Windows computer, the function windows() opens a large graphics window for plots, which is helpful. On a Mac, the command is quartz(), and it does require the quartz() drivers on the Mac

Sys.info() gives you a lot of information about the system you are working on

A quick peak at files

Sometimes it helps greatly to be able to take a quick look at a data file, to see what the format looks like. Excel does okay at this, but Excel can change the cell format on you, so what appears in Excel isn't always what is stored in a text file. Many software tools will use the formatting codes that may be embedded in a file to alter how the file is displayed, so what you see in the tool isn't what's actually contained in the file.

This can make it hard to find errors.

The readr package has the function read_lines() which reads in a text file, line by line and returns it as list, each element of the list is one line in the file. We can thus look at the file line by line to see what we are dealing with. This is very handy when you run into file loading or converting errors with more sophisticated file I/O tools

try running this on the redwine data file

```
require(readr)
```

```
## Loading required package: readr
```

```
temp=read_lines("C:\\Users\\Mike\\Documents\\DAT511\\1-25 class\\winequality-red.csv",n_max=
5)
temp[1]
```

```
## [1] "\"fixed acidity\";\"volatile acidity\";\"citric acid\";\"residual sugar\";\"chloride
s\";\"free sulfur dioxide\";\"total sulfur dioxide\";\"density\";\"pH\";\"sulphates\";\"alcoh
ol\";\"quality\""
```

```
temp[2]
```

```
## [1] "7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5"
```

```
temp[3]
```

```
## [1] "7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5"
```

```
temp[4]
```

```
## [1] "7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5"
```

If we clicked on an html file, like a web page, we would see all the html "tags" used to format the html document

Later in the semester, we will look at tools to extract information from text files, and on how to process text.

Using computers to process text is called Natural Language Processing, NLP. ChatGPT and other text transformers are really changing fast and making the news right now.

If we really need to get down into the weeds, we can look at the binary values in a file

```
temp=readBin("C:\\Users\\Mike\\Documents\\DAT511\\1-25 class\\winequality-red.csv",what="ra
w", n=400)
temp
```

```
##    [1] 22 66 69 78 65 64 20 61 63 69 64 69 74 79 22 3b 22 76 6f 6c 61 74 69 6c 65
##   [26] 20 61 63 69 64 69 74 79 22 3b 22 63 69 74 72 69 63 20 61 63 69 64 22 3b 22
##   [51] 72 65 73 69 64 75 61 6c 20 73 75 67 61 72 22 3b 22 63 68 6c 6f 72 69 64 65
##   [76] 73 22 3b 22 66 72 65 65 20 73 75 6c 66 75 72 20 64 69 6f 78 69 64 65 22 3b
##  [101] 22 74 6f 74 61 6c 20 73 75 6c 66 75 72 20 64 69 6f 78 69 64 65 22 3b 22 64
##  [126] 65 6e 73 69 74 79 22 3b 22 70 48 22 3b 22 73 75 6c 70 68 61 74 65 73 22 3b
##  [151] 22 61 6c 63 6f 68 6f 6c 22 3b 22 71 75 61 6c 69 74 79 22 0a 37 2e 34 3b 30
##  [176] 2e 37 3b 30 3b 31 2e 39 3b 30 2e 30 37 36 3b 31 31 3b 33 34 3b 30 2e 39 39
##  [201] 37 38 3b 33 2e 35 31 3b 30 2e 35 36 3b 39 2e 34 3b 35 0a 37 2e 38 3b 30 2e
##  [226] 38 38 3b 30 3b 32 2e 36 3b 30 2e 30 39 38 3b 32 35 3b 36 37 3b 30 2e 39 39
##  [251] 36 38 3b 33 2e 32 3b 30 2e 36 38 3b 39 2e 38 3b 35 0a 37 2e 38 3b 30 2e 37
##  [276] 36 3b 30 2e 30 34 3b 32 2e 33 3b 30 2e 30 39 32 3b 31 35 3b 35 34 3b 30 2e
##  [301] 39 39 37 3b 33 2e 32 36 3b 30 2e 36 35 3b 39 2e 38 3b 35 0a 31 31 2e 32 3b
##  [326] 30 2e 32 38 3b 30 2e 35 36 3b 31 2e 39 3b 30 2e 30 37 35 3b 31 37 3b 36 30
##  [351] 3b 30 2e 39 39 38 3b 33 2e 31 36 3b 30 2e 35 38 3b 39 2e 38 3b 36 0a 37 2e
##  [376] 34 3b 30 2e 37 3b 30 3b 31 2e 39 3b 30 2e 30 37 36 3b 31 31 3b 33 34 3b 30
```