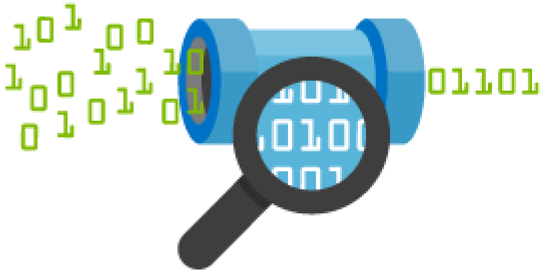
Data wrangling : Cleansing – Regular expressions (3/3)



Data wrangling is the process of importing, cleaning, and transforming raw data into actionable information for analysis. It is a time-consuming process that is estimated to take about 60-80% of analysts' time. In this series, we will go through this process. It will be a brief series with the goal of crafting the reader's skills in data wrangling. This is the fourth part of the series and it aims to cover the cleaning of the data used. In previous parts, we learned how to import, reshape, and transform data. The rest of the series will be dedicated to the data cleansing process. In this post, we will go through the regular expressions, which is a sequence of characters that define a search pattern, mainly
for use in pattern matching with text strings. In particular, we will cover the foundations of regular expression functions.

Before proceeding, it might be helpful to look over the help pages for the grep, sub, gsub, strsplit, .

Moreover, please load the following library.
install.packages("stringr")
library(stringr)

Answers to the exercises are available here. If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1

Find the cars that are Mercedes-Benz (match the pattern

‘Merc’).
Hint: The names of the cars can be retrieved from the command rownames(mtcars)

Exercise 2

Find the cars that are **not** Mercedes-Benz.

Exercise 3

Find the cars that are Mercedes-Benz (match the pattern ‘Merc’), but with a logical output.

Exercise 4

Find the number of Mercedes-Benz in the data set.

Exercise 5

Replace the first ‘a’ of every car with an ‘e’.

Learn more about Text analysis in the online course Text Analytics/Text Mining Using R. In this course you will learn how create, analyse and finally visualize your text based data source. Having all the steps easily outlined will be a great reference source for future work.

Exercise 6

Replace all ‘a’s of every car with ‘

Exercise 7

Separate the brand from the model. (e.g. “Mazda RX4” -> “Mazda” “RX4”).

Exercise 8

Find the cars that are Mercedes-Benz (use the str_detect function).

Exercise 9

Extract the ‘Merc’ string from the cars that contain it.

Exercise 10

Replace the ‘Merc’ string from the cars that contain it with ‘Mercedes-Benz’.