Does the string contain multiple substrings?

Authored by

stats writer

June 23, 2024

This inquiry refers to whether a given string contains more than one instance of a specific sequence of characters, known as a substring. By determining if there are multiple substrings present, the overall structure and composition of the string can be better understood and analyzed. This can be useful in various applications, such as data processing and text manipulation.

R: Check if String Contains Multiple Substrings

You can use the following methods in R to check if a string contains multiple substrings:

Method 1: Check if String Contains One of Several Substrings

df$contains_any <- apply(sapply(find_strings, grepl, df$team), 1, any)

This particular syntax checks if each string in the team column contains *any* of the strings specified in the vector of strings called find_strings.

Method 2: Check if String Contains Several Substrings

df$contains_any <- apply(sapply(find_strings, grepl, df$team), 1, all)

This particular syntax checks if each string in the team column contains all of the strings specified in the

vector of strings called find_strings.

The following examples show how to use each method in practice with the following data frame in R:

```
#create data frame
df = data.frame(team=c('Good East Team', 'Good West Team', 'Great East Team',
'Great West Team', 'Bad East Team', 'Bad West Team'),
points=c(93, 99, 105, 110, 85, 88))

#view data frame
df

team points
1 Good East Team 93
2 Good West Team 99
3 Great East Team 105
4 Great West Team 110
5 Bad East Team 85
6 Bad West Team 88
```

Example 1: Check if String Contains One of Several Substrings

We can use the following syntax to check if each string in the team column contains either the substring

"Good" or "East":

#define substrings to look for
find_strings <- c('Good', 'East')

#check if each string in team column contains either substring
df$good_or_east <- apply(sapply(find_strings , grepl, df$team), 1, any)

#view updated data frame
df

team points good_or_east
1 Good East Team 93 TRUE
2 Good West Team 99 TRUE
3 Great East Team 105 TRUE
4 Great West Team 110 FALSE
5 Bad East Team 85 TRUE
6 Bad West Team 88 FALSE

The new good_or_east column returns the following values:

TRUE if team contains "Good" or "East"FALSE if team

contains neither "Good" nor "East"

Example 2: Check if String Contains Several Substrings

We can use the following syntax to check if each string in the team column contains the substring "Good" and "East":

```
#define substrings to look for
find_strings <- c('Good', 'East')

#check if each string in team column contains either substring
df$good_and_east <- apply(sapply(find_strings , grepl, df$team), 1, all)

#view updated data frame
df

team points good_and_east
1 Good East Team 93 TRUE
2 Good West Team 99 FALSE
3 Great East Team 105 FALSE
4 Great West Team 110 FALSE
5 Bad East Team 85 FALSE
6 Bad West Team 88 FALSE
```

TRUE if team contains "Good" and "East"FALSE if team doesn't contain "Good" and "East"

Notice that only one TRUE value is returned since there is only one team name that contains the substring "Good" and the substring "East."