

sb_rmd_ex_07

Sander Bossier

2024-01-16

1.

Consider the character vector `x`:

```
x <- c("Swan swam over the pond, Swim swan swim!",  
      "Swam swam back again - Well swum swan!")
```

Split the strings in `x` into words, removing any *commas* or *hyphens*. Call the result `y`.

```
y <- strsplit(  
  x,  
  split = ",? -? ?"  
)  
y
```

```
## [[1]]  
## [1] "Swan" "swam" "over" "the" "pond" "Swim" "swan" "swim!"  
##  
## [[2]]  
## [1] "Swam" "swam" "back" "again" "Well" "swum" "swan!"
```

2.

The `.rda` file `hafu` in the RStudio Project `Text` contains a data set `hafu` on manga characters. In the `Father` and `Mother` columns, some values have question marks after the country name, indicating that the author was uncertain about the nationality of the parent:

```
load("hafu.rda")  
head(hafu[, -ncol(hafu)], n = 10)
```

##	Year	Series	Character	Gender	Father
## 1	1963	Yuki no Taiyou	Sanae	F	Japanese
## 2	1964	Cyborg 009	Joe Shimamura	M	American
## 3	1967	Lupin III	Lupin III	M	French?
## 4	1967	Nekome Kozou	Cat-Eyed Boy	M	Japanese
## 5	1972	Gatchaman	Jun the Swan	F	<NA>
## 6	1974	Great Mazinger	Jun Hono	F	American
## 7	1975	Ichigo Monogatari	Ichigo	F	Scandinavian?
## 8	1976	Kochira Katsushika-ku...	Reiko Katherine Akimoto	F	Japanese?
## 9	1977	Haikara-san ga Tooru	Shinobu Iijyuin	M	Japanese
## 10	1979	Mobile Suit Gundam	Amuro Ray	M	Japanese
##	Mother	Eyes	Hair		
## 1	American	<NA>	<NA>		
## 2	Japanese	brown	brown		
## 3	Japanese?	black	black		

```
## 4    Fantasy brown  <NA>
## 5      <NA> green green
## 6    Japanese black black
## 7    Japanese? brown brown
## 8      French?  blue blond
## 9      German  blue blond
## 10   Canadian  blue brown
```

Create two vectors `FatherUncertain` and `MotherUncertain` denoting whether or not there was a question mark in the `Father` or `Mother` column, respectively.

```
FatherUncertainSB <- endsWith(
  x = as.character(hafu$Father),
  suffix = "?"
)
MotherUncertainSB <- endsWith(
  x = as.character(hafu$Mother),
  suffix = "?"
)

FatherUncertain <- with(hafu, grepl("\\?", Father))
MotherUncertain <- with(hafu, grepl("\\?", Mother))

identical(FatherUncertainSB, FatherUncertain)
```

```
## [1] FALSE
```

```
identical(MotherUncertainSB, MotherUncertain)
```

```
## [1] FALSE
```

3.

Create a new data frame `hafu2` by removing the question marks from the `Father` and `Mother` columns in `hafu` (i.e. replace them by an empty string).

```
hafu2 <- within(hafu, {
  Father <- gsub(pattern = "\\?", replacement = "", x = Father)
  Mother <- gsub(pattern = "\\?", replacement = "", x = Mother)
})
```

Challenges

4.

The `.rda` file **Dataline** contains a vector `Dataline` that contains a single value. This value is the result of a webscraper. It contains 1500 fields, separated by a semicolon: