06 Textmanipulation mit stringR

Martin Hanewald

2019-02-19

Packages

```
library(tidyverse)
library(knitr)
```

Aufgabe: Zähle alle Vorkommnisse von *fruit* in *sentences*

```
data(fruit)
head(fruit)
#> [1] "apple"       "apricot"     "avocado"     "banana"      "bell pepper"
#> [6] "bilberry"
data(sentences)
head(sentences)
#> [1] "The birch canoe slid on the smooth planks."
#> [2] "Glue the sheet to the dark blue background."
#> [3] "It's easy to tell the depth of a well."
#> [4] "These days a chicken leg is a rare dish."
#> [5] "Rice is often served in round bowls."
#> [6] "The juice of lemons makes fine punch."
```

Ziel

fruit	count
star	7
fig	5
pear	5
apple	3
bell	3
grape	2
nut	2
rock	2
pepper	1
orange	1
lemon	1
peach	1
plum	1
purple	1

Lösung

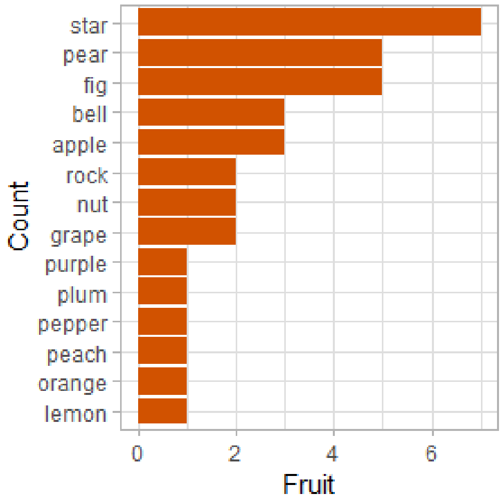
```
words <- fruit %>% str_split(" ") %>% unlist()

ans <- list()

for(w in words){
    ans[[w]] <- str_detect(tolower(sentences), tolower(w)) %>% sum()
}

result <- ans %>% as_tibble() %>%
    gather(fruit, ct) %>%
    filter(ct > 0) %>%
    arrange(desc(ct))

result %>%
    ggplot(aes(fruit %>% fct_reorder(ct), ct)) + geom_col(fill='#d15200') + coord_flip() +
    theme_light() + labs(x = 'Count', y = 'Fruit')
```



Alternative mit *apply*

```
### without for loop

fruit %>%
    str_split(" ") %>%
    unlist() %>%
    unique() %>%
    sapply(function(x) str_detect(tolower(sentences), x) %>% sum(), simplify = F) %>%
    as_tibble() %>%
    gather(fruit, count) %>%
    filter(count > 0) %>%
    arrange(desc(count))
#> # A tibble: 14 x 2
#>     fruit  count
#>     <chr>  <int>
#>   1 star       7
#>   2 fig        5
#>   3 pear       5
#>   4 apple      3
#>   5 bell       3
#>   6 grape      2
#>   7 nut        2
#>   8 rock        2
#>   9 pepper      1
#>  10 orange      1
#>  11 lemon       1
#>  12 peach       1
#>  13 plum        1
#>  14 purple      1
```