Wyrażenia regularne dla ilościowców

Michał Kaftanowicz

2017-10-31

Wstęp

Co to są wyrażenia regularne?

Oxforddictionaries.com:

"A sequence of symbols and characters expressing a string or pattern to be searched for within a longer piece of text."

IRD:

Reguła decyzyjna wskazująca, jaki fragment tekstu lub wzorzec ma być poszukiwany w tekście.

Dygresja - R i praca z tekstem

```
(x <- "Ala ma kota.")
```

```
## [1] "Ala ma kota."
```

```
grepl(pattern = "Ala", x)
```

```
## [1] TRUE
```

```
grepl(pattern = "Eugeniusz", x)
```

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

Dygresja - R i praca z tekstem

```
gsub(pattern = "Ala", replacement = "Alicja", x)
```

```
## [1] "Alicja ma kota."
```

```
gsub(pattern = "a", replacement = "_", x)
```

```
## [1] "Al_ m_ kot_."
```

```
cat(gsub(pattern = "a", replacement = "_", x))
```

```
## Al_ m_ kot_.
```

Wyrażenia regularne - sposób działania

```
x <- "Ala ma kota. Inaczej: Alicja Kowalska,
urodzona 12 marca 1994 r. (1994-03-12)
w Cendrowicach, gm. Góra Kalwaria,
jest posiadaczką zwierzęcia z rodziny kotowatych."
```

Najprostszy wariant - zwykły fragment tekstu

```
cat(gsub("Ala", "Alicja", x))
```

```
## Alicja ma kota. Inaczej: Alicja Kowalska,
## urodzona 12 marca 1994 r. (1994-03-12)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych.
```

```
cat(gsub("a", "_", x))
```

```
## Al_ m_ kot_. In_czej: Alicj_ Kow_lsk_,
## urodzon_ 12 m_rc_ 1994 r. (1994-03-12)
## w Cendrowic_ch, gm. Gór_ K_lw_ri_,
## jest posi_d_czką zwierzęci_ z rodziny kotow_tych.
```

Kropka zastępuje dowolny symbol

```
cat(gsub("a.", "_", x))
```

```
## Al_m_kot_ In_zej: Alicj_Kow_sk_
## urodzon_12 m_c_1994 r. (1994-03-12)
## w Cendrowic_h, gm. Gór_K_w_i_
## jest posi__zką zwierzęci_z rodziny kotow_ych.
```

Należy z nią uważać

```
cat(gsub(".", "_", x))
```

```
## _____________________________________________________
```

A co, jeśli interesuje nas normalna kropka?

```
cat(gsub("\\.", "_", x))
```

```
## Ala ma kota_ Inaczej: Alicja Kowalska,
## urodzona 12 marca 1994 r_ (1994-03-12)
## w Cendrowicach, gm_ Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych_
```

Zbiory i zakresy

[. . .] wyłapuje dowolny znak ze zbioru wewnątrz nawiasów kwadratowych (można używać zakresów).

[^...] wyłapuje dowolny znak SPOZA zbioru wewnątrz nawiasów kwadratowych (też można używać zakresów).

Przykład użycia zbiorów i zakresów

```
cat(gsub("[A-Z]", "_", x))
```

```
## _la ma kota. _naczej: _licja _owalska,
## urodzona 12 marca 1994 r. (1994-03-12)
## w _endrowicach, gm. _óra _alwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych.
```

```
cat(gsub("[a-z]", "_", x))
```

```
## A__ __ ____. I______: A_____ K_______,
## ________ 12 _____ 1994 _. (1994-03-12)
## _ C___________, __. Gó__ K_______,
## ____ __________ą ______ę___ _ _______ __________.
```

Przykład użycia zbiorów i zakresów

```
cat(gsub("[0-9]", "_", x))
```

```
## Ala ma kota. Inaczej: Alicja Kowalska,
## urodzona __ marca ____ r. (____-__-__)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych.
```

Przykład użycia zbiorów i zakresów

```
cat(gsub("[A-Za-zżźćńółęąśŻŹĆĄŚĘŁÓŃ]", "_", x))
```

```
## ___ __ ____. _______: ______ ________,
## ________ 12 _____ 1994 _. (1994-03-12)
## _ ___________, __. ____ ________,
## ____ __________ _________ _ _______ __________.
```

```
cat(gsub("[A-Za-zżźćńółęąśŻŹĆĄŚĘŁÓŃ]+", "_", x))
```

```
## _ _ _. _: _ _,
## _ 12 _ 1994 _. (1994-03-12)
## _ _, _. _ _,
## _ _ _ _ _ _.
```

Symbole specjalne stawiane za znakiem lub grupą znaków

$*$: 0 lub więcej wystąpień

$+$: co najmniej 1 wystąpienie

?: co najwyżej 1 wystąpienie

$\{$n$\}$: dokładnie n wystąpień

$\{$n,$\}$: co najmniej n wystąpień

$\{$n,m$\}$: między n a m wystąpień

Przykłady

```
cat(gsub("A[a-z]+a", "_", x))
```

```
## _ ma kota. Inaczej: _ Kowalska,
## urodzona 12 marca 1994 r. (1994-03-12)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych.
```

Przykłady

```
cat(gsub("kot[a-z]*", "_", x))
```

```
## Ala ma _. Inaczej: Alicja Kowalska,
## urodzona 12 marca 1994 r. (1994-03-12)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny _.
```

Przykłady

```
cat(gsub("[0-9]{4}-[0-9]{2}-[0-9]{2}",
         "DATA_ISO", x))
```

```
## Ala ma kota. Inaczej: Alicja Kowalska,
## urodzona 12 marca 1994 r. (DATA_ISO)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych.
```

```
cat(gsub("[0-9]{2} [a-zżźćńółęąś]+ [0-9]{4} r\\.",
         "DATA_PL", x))
```

```
## Ala ma kota. Inaczej: Alicja Kowalska,
## urodzona DATA_PL (1994-03-12)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych.
```

Symbole specjalne

^: początek tekstu

```
cat(gsub("A[a-z]+" ,"_", x))
```

```
## _ ma kota. Inaczej: _ Kowalska,
## urodzona 12 marca 1994 r. (1994-03-12)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych.
```

```
cat(gsub("^A[a-z]+" ,"_", x))
```

```
## _ ma kota. Inaczej: Alicja Kowalska,
## urodzona 12 marca 1994 r. (1994-03-12)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny kotowatych.
```

Symbole specjalne

$: koniec tekstu

```
cat(gsub("[a-z]+h.{1}" ,"_", x))
```

```
## Ala ma kota. Inaczej: Alicja Kowalska,
## urodzona 12 marca 1994 r. (1994-03-12)
## w C_ gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny _
```

```
cat(gsub("[a-z]+h.{1}$" ,"_", x))
```

```
## Ala ma kota. Inaczej: Alicja Kowalska,
## urodzona 12 marca 1994 r. (1994-03-12)
## w Cendrowicach, gm. Góra Kalwaria,
## jest posiadaczką zwierzęcia z rodziny _
```

Symbole specjalne

\\b: kraniec (początek lub koniec) wyrazu |: operator logiczny "lub"

```
cat(gsub("[Aa]" ,"_", x))
```

```
## _l_ m_ kot_. In_czej: _licj_ Kow_lsk_,
## urodzon_ 12 m_rc_ 1994 r. (1994-03-12)
## w Cendrowic_ch, gm. Gór_ K_lw_ri_,
## jest posi_d_czką zwierzęci_ z rodziny kotow_tych.
```

```
cat(gsub("\\b[Aa]|[Aa]\\b" ,"_", x))
```

```
## _l_ m_ kot_. Inaczej: _licj_ Kowalsk_,
## urodzon_ 12 marc_ 1994 r. (1994-03-12)
## w Cendrowicach, gm. Gór_ Kalwari_,
## jest posiadaczką zwierzęci_ z rodziny kotowatych.
```

Zastosowania

Czyszczenie danych (również ilościowych)

```
nrs_txt <-
"1000 zł
2  000PLN
3,000
4.000.34
5.000,99
6oOO"
```

Wczytujemy dane

```
nrs <- read.csv(text = nrs_txt,
                stringsAsFactors = FALSE,
                header = FALSE, sep = ';')[[1]]
```

Naiwna próba

```
as.numeric(nrs)
```

```
## Warning: NAs introduced by coercion
```

```
## [1] NA NA NA NA NA NA
```

Na początek pozbądźmy się spacji i liter

```
nrs2 <- gsub("[ A-Za-zżźćńółęąśŻŹĆĄŚĘŁÓŃ]+",
             "", nrs)
cat(nrs2)
```

```
## 1000 2000 3,000 4.000.34 5.000,99 60
```

Falstart! Niektóre zera zostały błędnie rozpoznane przez OCR jako litery “o”. Co teraz?

Spróbujmy ostrożniej:

```
nrs <- gsub(" +|[pPzZ]+[A-Za-zżźćńółęąśŻŹĆĄŚĘŁÓŃ]+$",
            "", nrs)
cat(nrs)
```

```
## 1000 2000 3,000 4.000.34 5.000,99 6oOO
```

Pozbyliśmy się spacji i oznaczeń waluty.

O != 0

```
nrs <- gsub("(o|O){1}",
            "0", nrs)
cat(nrs)
```

```
## 1000 2000 3,000 4.000.34 5.000,99 6000
```

Naprawiliśmy zera błędnie rozpoznane jako “o”.

Kropki dziesiętne, przecinki tysięczne, przecinki dziesiętne, kropki tysięczne?

Teraz będzie trudniej

```
nrs <- gsub("(\\.|\\,)([0-9]{2})$", "decdot\\2", nrs)
cat(nrs)
```

```
## 1000 2000 3,000 4.000decdot34 5.000decdot99 6000
```

Zastosowaliśmy grupowanie za pomocą nawiasów i zwracanie fragmentów wyłapanego tekstu.

Kropki dziesiętne, przecinki tysięczne, przecinki dziesiętne, kropki tysięczne?

```
nrs <- gsub("\\.|\\,", "", nrs)
cat(nrs)
```

```
## 1000 2000 3000 4000decdot34 5000decdot99 6000
```

Pozbywamy się reszty przecinków i kropek.

Kropki dziesiętne, przecinki tysięczne, przecinki dziesiętne, kropki tysięczne?

```
nrs <- gsub("decdot", ".", nrs)
cat(nrs)
```

```
## 1000 2000 3000 4000.34 5000.99 6000
```

Przywracamy kropki dziesiętne na ich właściwe miejsca.

Chwila prawdy

```
as.numeric(nrs)
```

```
## [1] 1000.00 2000.00 3000.00 4000.34 5000.99 6000.00
```

Inne zastosowania

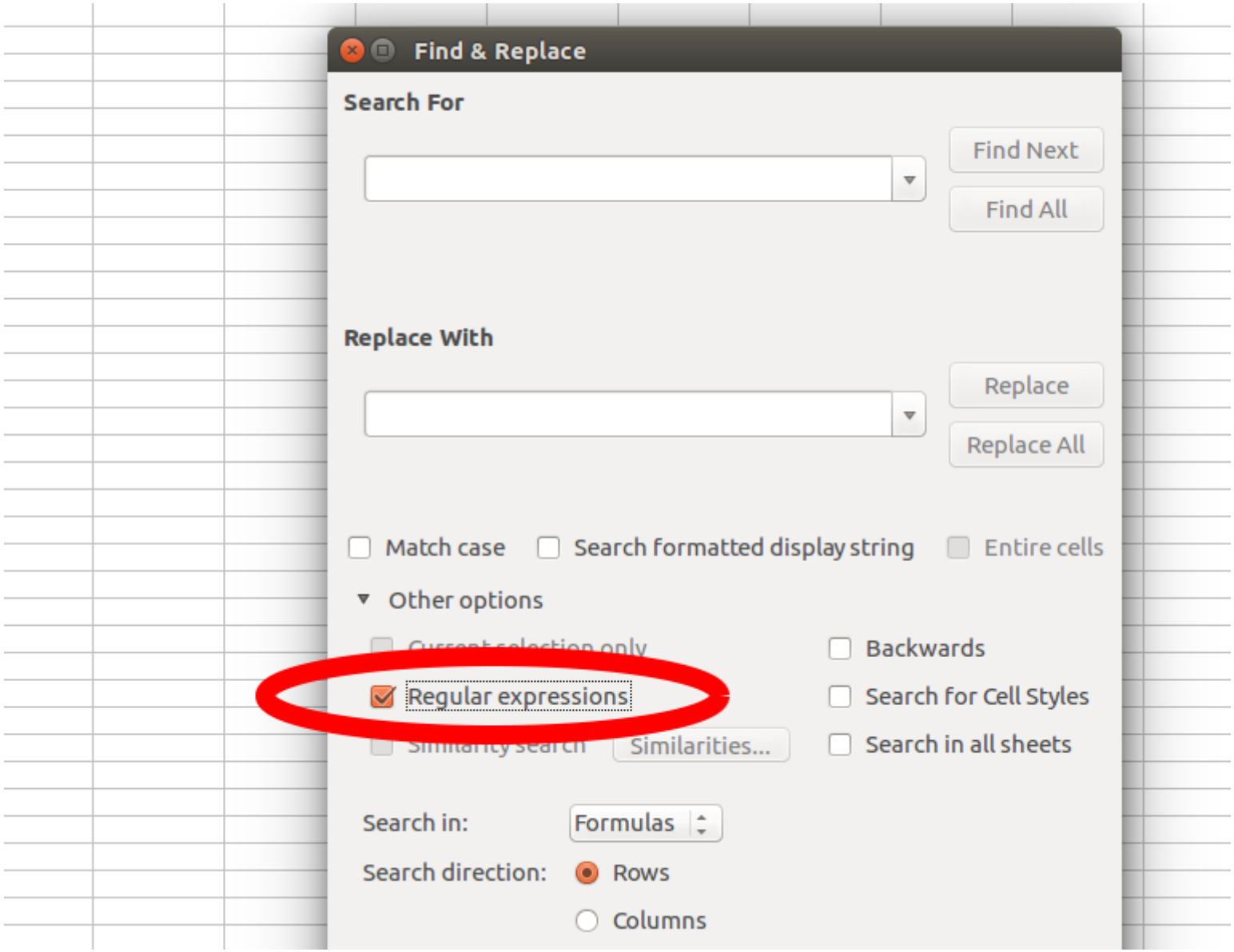

Figure 1: Excel/Calc