Przetwarzanie napisów w R

Bartłomiej Tartanus

Wyrażenia regularne

Regular Expression

Wygodne i zręczne narzędzie do znajdowania, zamieniania oraz ekstrakcji **tekstów**

Regular Expression

+ przenośne - działają w wielu językach
+ zwięzłe - krótkie wyrażenie potrafi dużo zdziałać

- mogą być mało intuicyjne
- ciężko zdebugować

Regex - wzorzec

[A-Z]+-\d{1,3}

AB-30

XYZ-5

F-16

Znaki specjalne

. \ | ( ) [ ] { } ^ $ * + ? –

Znaki specjalne

. (kropka) oznacza dowolny znak

^ (daszek) oznacza początek tekstu

$ (dolar) oznacza koniec tekstu

[ ] (nawias kwadratowy) zbiór znaków

^ (operator dopełnienia) negacja

Grupy znaków

[abc] - dopasuje znak a,b lub c

[a-x] dopasuje znak z zakresu od a do x

[0-5] dopasuje cyfrę od 0 do 5

[^2-4] dopasuje każdy ZNAK oprócz 2,3 i 4

[a-cA-C] dopasuje a,b,c,A,B lub C

Grupy znaków - skróty

\d [0-9]

\w[a-zA-Z0-9_]

\s białe znaki + [\t\r\n]

\D [^0-9]

\W [^a-zA-Z0-9_]

\S [^białeznaki \t\r\n]

Kwantyfikatory

Każdy kwantyfikator działa tylko na poprzedzające go wyrażenie!

Kwantyfikatory są zachłanne albo leniwe.

Kwantyfikatory

Kwantyfikator	Znaczenie	Zachłanny
?	powtórz 0 lub 1 raz	Tak
*	powtórz 0 lub wiele razy	Tak
+	powtórz 1 lub wiele razy	Tak
{n}	powtórz n razy	N/D
{n, m}	powtórz od n do m razy	Tak
{n, }	powtórz przynajmniej n razy	Tak
*?	powtórz 0 lub więcej razy	Nie
+?	powtórz 1 lub więcej razy	Nie
{n, m}?	powtórz od n do m razy	Nie
{n, }?	powtórz przynajmniej n razy	Nie

Podwzorce

^Ala ma (.*?)$

^Ala ma (kota|psa|chomika)$

^Ala ma ((kot(eczk)?|p(s|iesk)|chomi(cz)?k)a)$

Co jest nie tak?

Data w formacie `dd.mm.yyyy` - 22.09.2016

[0-9]{2}.[0-9]{2}.[0-9]{4}

[1-31].[1-12].[0001-9999]