
# Podstawy systemów UNIX
## Wyrażenia regularne

Autor: Maciej Friedel <maciek@friedel.pl>
Zajęcia prowadzone dla Polskiej Szkoły IT
Wrocław, 2008

# Wyrażenia regularne - definicja

**ang. Regular expressions (regexp)**

Wyrażenia regularne są narzędziem służącym do dopasowywania wzorców. Wykorzystuje się tam gdzie potrzebne jest przeszukanie tekstu i wydobycie z niego informacji pasującej do wzorca.

# Wyrażenia regularne - podział znaków

**Znaki specjalne (metaznaki)**

1. metaznaki rozpoznawane w dowolnym miejscu:

**^, $, *, +, ?, ., (, ), [, ], {, }, **

2. rozpoznawane wewnątrz nawiasów kwadratowych:

**\, ^, -, ]**

**Znaki zwyczajne**

Każdy znak nie będący znakiem specjalnym.
Znak zwyczajny oznaczy tylko i wyłącznie sam siebie.

np.:
"a" – oznacza łańcuch złożony ze znaku a
"a3e" – oznacza łańcuch z kolejno ułożonych znaków a, 3 oraz e

# Wyrażenia regularne - znaki specjalne

**Kropka "."**
**oznacza dowolny znak oprócz znaku nowego wiersza**

np.:
"k.o" – pasuje do słów  kot, kat, kit itd.

**Znak zapytania "?"**
**po symbolu oznacza jedno lub zero wystąpień poprzedzającego wyrażenia**

np.:
"ko?t" – pasuje do słów  kt, kot

# Wyrażenia regularne – znaki specjalne

**Gwiazdka "*"**
**oznacza zero lub więcej wystąpień poprzedzającego wyrażenia**

np.:
"ko*t" - pasuje do wzorców kt, kot, koot, kooot, koooot, ...

**Plus "+"**
**po symbolu oznacza co najmniej jedno wystąpienie poprzedzającego go wyrażenia**

np.:
"ko+t" - pasuje do wzorców kot, koot, kooot, koooot, ...

# Wyrażenia regularne – znaki specjalne

**Pałka "|" to operator OR**

np.:
"a|b|c" - w danym wyrażeniu może wystąpić a, b lub c
"Ala|Ola" - w danym wyrażeniu wystąpić może Ala lub Ola

**Daszek "^" oznacza początek wiersza**
**dolar "$" oznacza koniec wiersza.**

np.:
"^.o.a$" – odpowiada ciągowi dokładnie czterech znaków typu cola, soja, pola itd... występujących samotnie w wierszu
"^Po.*IT$" - Polska Szkoła IT

# Wyrażenia regularne – znaki specjalne

**Podwzorzec może być zamknięty w niepodzielnej grupie za pomocą nawiasów "( )". W ten sposób można użyć gałęzi nie tylko dla całego wzorca, ale też dla jego fragmentów.**

np.:
"Piw(a|ko)" - oznacza Piwa i Piwko

**Zestaw znaków między nawiasami kwadratowymi "[]" oznacza jeden dowolny znak objęty nawiasami kwadratowymi.**

np.:
"[abc]" - oznacza a, b lub c. Można używać także przedziałów: "[a-c]"
"pi[wk]o" - oznacza wzorce piwo i piko

# Wyrażenia regularne – znaki specjalne

**Daszek "^"**
**na początku zestawu oznacza wszystkie znaki oprócz tych z zestawu. Większość znaków specjalnych w tym miejscu traci swoje znaczenie**

np.:
"[^piwo]" wyszukuje łańcuchy w których nie występuje piwo
"pi[^wk]o" wyklucza wyrażenia piwo piko ale nie wyklucza ciągu pilot

**Znaki specjalne poprzedzone odwrotnym ukośnikiem "\" powodują, że poprzedzany znak traci swoje specjalne znaczenie - oznacza sam siebie**

np.:
"\\" - oznacza znak \
".*\.dbf" – pasuje do wszystkich plików z rozszerzeniem .dbf

# Wyrażenia regularne - granice wyrazów

**"/<" - początek słowa**

np.:

"/<pi" - wyrazy zaczynające się od pi (pi, pilot, piwo…)

**"/>" - koniec słowa**

np.:

"ot/>" wyrazy kończące się na ot (mlot, lot, pilot…)

np.:

"\<slowo/>" - słowo jest wyrazem samodzielnym

"\<slow(o|a)/>" to słowo i słowa

# Wyrażenia regularne - liczba powtórzeń

Możliwość precyzyjnego określenia liczby wystąpień danego wyrażenia

- Wyrażenie {N} oznacza dokładnie N wystąpień
- Wyrażenie {N,} co najmniej N wystąpień wyrażenia {0,}=* {1,}=+
- Wyrażenie {,M} co najwyżej M wystąpień wyrażenia
- Wyrażenie {N,M} od N do M wystąpień wyrażenia {0,1}=?

# Przedefiniowane klasy znaków

Rozszerzony zapis przedziałów, wprowadzenie klas znaków np.:

- "[[:digit:]]" oznacza dowolną cyfrę
- "[[:alpha:]]" literę
- "[[:alnum:]]" literę lub cyfrę
- "[[:xdigit:]]" liczby w systemie szesnastkowym
- "[[:lower:]]" małe litery
- "[[:upper:]]" duże litery
- "[[:punct:]]" znaki interpunkcji

# Wyrażenia regularne - przykłady

Kilka prostych przykładów:

- "[[:upper:]]{,4}" oznacza nie więcej niż cztery duże litery
- "\<[[:digit:]]{2,4}>\" oznacza liczby dwu, trzy i czterocyfrowe
- "\<[[:alnum:]]{0,8}\.[[:alnum:]]{3}\>" DOSowe nazwy plików
- "((http(s?):\/\/)|(www\.))([[:alnum:]]+\.[[:alnum:]]+)" - link