Wprowadzenie do wyrażeń regularnych

Beata Pańczyk

Plan wykładu

- Zastosowania wyrażeń regularnych
- Składnia RegExp
- Znaki specjalne
- Powtarzalność
- Podwyrażenia
- Klasy znaków

Źródła:

- http://regexlib.com/
- http://www.rexv.org/
- http://osteele.com/tools/rework/
- http://maciek.lasyk.info/regexp_checker.html

Wprowadzenie do wyrażeń regularnych

- **wyrażenia regularne** (ang. regular expressions, regexp lub regexes)- sposób opisywania tekstu poprzez dopasowywanie wzorców
- **zastosowanie** - bardziej złożone dopasowywanie ciągów (np. sprawdzanie poprawności dat, zastępowanie jednych fragmentów tekstu innymi, pobieranie fragmentów z większych bloków tekstu)
- 2 style składni wyrażeń regularnych: POSIX i Perl
- w połączeniu z konkretnym narzędziem, w którym zostały zaimplementowane, umożliwiają różnorodne sposoby przetwarzania tekstu.

Wyrażenia regularne - składnia

- **Znaki zwyczajne**
  a, b, c, z, A, B, C, Z, 0, 1, 2, 9, ,, !, _, ...
- **Znaki specjalne (metaznaki)**
  - metaznaki rozpoznawane w dowolnym miejscu wzorca poza nawiasami kwadratowymi []
    ^, $, ., *, ?, +, [,], {, }, (,), \
  - metaznaki rozpoznawane w **klasach znaków** (część wzorca ujęta w nawiasy kwadratowe [], pasująca do dokładnie jednego znaku, bez względu na liczbę tworzących ją znaków)
    ^, -, \,]

Elementy wyrażeń regularnych: dopasowywanie tekstu

- Wszystkie znaki oprócz znaków specjalnych, określają same siebie, np:
  k - określa łańcuch złożony ze znaku k
- Kolejne znaki oznaczają, że znaki te muszą wystąpić w łańcuchu dokładnie w takiej samej kolejności, np:
  kot pozwala znaleźć łańcuch kot w dowolnym miejscu wiersza
  RegExp oznacza RegExp

Elementy wyrażeń regularnych: znaki specjalne

- **Kropka .** – oznacza dowolny znak, z wyjątkiem znaku nowego wiersza, np:
  - **r.k** – pasuje do rok, rak, ryk itp.
  - **.a.** – do mak, rak, lat itp.
  - **.o.a** – do lola, cola, wola, kolacja itp.
- **rozgałęzianie** [|] – oznacza „lub”, „OR” i pozwala na łączenie wielu wyrażeń w jedno, do którego pasuje dowolne z wyrażeń składowych np:
  - **(gif)|(jpg)** - pasuje do gif lub jpg
  - **(U|u)(l|lica) (3-go|Trzeciego) Maja** – pasuje do ul 3-go Maja, Ul 3-go Maja, ulica 3-go Maja, Ulica 3-go Maja, ul Trzeciego Maja, Ul Trzeciego Maja, ulica Trzeciego Maja, Ulica Trzeciego Maja

Elementy wyrażeń regularnych: znaki specjalne - grupowanie

- Podwzorzec może być zamknięty w niepodzielnej grupie za pomocą nawiasów ( ). W ten sposób można użyć gałęzi nie tylko dla całego wzorca, ale również dla jego fragmentów np:
  - **Fizy(cy|k)** – pasuje do Fizycy i Fizyk
- Zestaw znaków między nawiasami kwadratowymi oznacza dowolny znak objęty nawiasami kwadratowymi, np:
  - **[1234], [1-4]** – oznacza 1 lub 2 lub 3 lub 4
  - **pi[wk]o** – pasuje do piwo i piko
  - **[a-z]** - dopasowywane znaki ograniczamy do zbioru małych liter
  - **[aeiouy]** - wyliczanie elementów zbioru (samogłoski)
  - **[a-zA-Z]** - wszystkie małe i duże litery

Kotwiczenie

- **kotwiczenie:**
  ^ - stosowany na początku wyrażenia regularnego w celu wskazania, że musi się ono pojawić na początku szukanego ciągu
  $ - stosowany na końcu wyrażenia regularnego, które musi się pojawić na końcu szukanego ciągu np.
  - **kot** – dopasuje łańcuch kot w dowolnym miejscu wiersza
  - **^kot** – pasuje wtedy, gdy mamy początek wiersza, po którym od razu występuje litera k, po niej od razu litera o, a po niej od razu litera t
  - **^kot$** – pasuje jeśli wiersz zawiera początek, po którym od razu znajdują się znaki kot, a po nich od razu koniec wiersza
  - **^paweł, gif$, ^[a-z]$** - (pasuje do każdego pojedynczego znaku a-z, jako osobnego ciągu

Zakotwiczenia

- **^ (daszek)** oznacza ***nie***, kiedy jest umieszczony w []
  - [^a-z] - każdy znak, który nie pochodzi z zakresu a-z
- Większość znaków specjalnych w tym miejscu traci swoje znaczenie, np:
  - [^piwo] – pasuje do wszystkich łańcuchów w których nie występuje słowo piwo
  - pi[^wk]o – pasuje np. do pinokio, ale wyklucza słowa: piwo oraz piko
- UWAGA - ze względu na to, że zarówno [-] jak i [^] mają specjalne znaczenie, to aby:
  - dopasować daszek [^] - nie należy umieszczać go na początku;
  - dopasować minus [-] – należy umieścić go jako ostatni znak w zakresie;
  - zamiast [^%$#@!] - należy zastosować [%$#@!^]
  - zamiast [a-c], chcąc dopasować ‘a’, ‘c’ lub ‘–’ należy zastosować [ac-].

Klasy znaków

- **[...]** - pojedynczy znak podany lub zawierający się w określonym zakresie
- **[^...]** - pojedynczy znak, który nie został podany lub nie zawiera się w określonym zakresie
- **[[:klasa:]]** - klasa znaków POSIX

Predefiniowane klasy znaków POSIX

- [[:alnum:]] - znaki alfanumeryczne
- [[:alpha:]] - znaki alfabetu
- [[:lower:]] - małe litery
- [[:upper:]] - duże litery
- [[:digit:]] - liczby dziesiętne
- [[:xdigit:]] - liczby szesnastkowe
- [[:punct:]] - znaki przestankowe
- [[:blank:]] - tabulatory i spacje
- [[:space:]] - pusta przestrzeń
- [[:cntrl:]] - znaki kontrolne
- [[:print:]] - wszystkie możliwe do wyświetlenia znaki
- [[:graph:]] - wszystkie możliwe do wyświetlenia znaki poza spacjami

Powtarzalność i podwyrażenia

- **powtarzalność** można określić stosując znaki specjalne:
  ***** - wzór może powtórzyć się zero bądź więcej razy
  + - wzór może powtórzyć się jeden bądź więcej razy
  ? - wzór może wystąpić jeden bądź zero razy
  np.
  [[:alnum:]]+ - co najmniej jeden znak alfanumeryczny
  ko?t pasuje do kt, kot
  ko*t pasuje do kt, kot, koot, kooooot, ...
  ko+t pasuje do kot, koot, kooooot, ...
- wyrażenie można rozdzielić na **podwyrażenia** stosując nawiasy jak w zwykłych wyrażeniach arytmetycznych np.
  (bardzo)*dużo pasuje do 'dużo', 'bardzo dużo', 'bardzo bardzo dużo'

Powtarzalność

- **podwyrażenia policzalne** - ilość powtórzeń danego ciągu można określić stosując nawiasy klamrowe
- Wyrażenie {X} oznacza dokładnie X wystąpień
- Wyrażenie {X,} co najmniej X wystąpień, czyli przykładowo {0,} = *, {1,} = +
- Wyrażenie {,X} co najwyżej X wystąpień
- Wyrażenie {X,Y} oznacza Y dopasowań (jeśli to możliwe), ale do powodzenia wystarczy mu już X, np: {0,1} = ?
- np.:
  {3} - dokładnie 3 powtórzenia
  {2,4} - od dwu do czterech powtórzeń
  {2,} - co najmniej dwa powtórzenia
  np. **(bardzo){2,3}** - pasuje do 'bardzo bardzo', 'bardzo bardzo bardzo'

Znaki specjalne - zestawienie

. dopasowanie do każdego znaku oprócz nowej linii
(początek podciągu
) koniec podciągu
{ początek minimalnego/maksymalnego kwantyfikatora
} koniec minimalnego/maksymalnego kwantyfikatora

W nawiasach kwadratowych wyrażeń POSIX stosuje się:
**** poprzedza znak specjalny (np. \\b, \\\n)
^ NOT jeśli użyte przed wyrażeniem
- określenie zakresu znaków

Znaki predefiniowane

- \r znak powrotu karetki
- \n znak nowej linii
- \t tabulator horyzontalny
- \v tabulacja pionowa
- \0 znak NUL
- \s odstęp (skrót dla [\f\n\r\t\v\u00A0\u2028\u2029])
- \S znak inny niż odstęp (skrót dla [^\f\n\r\t\v\u00A0\u2028\u2029])
- \w znak wyrazu (skrót dla [a-zA-Z0-9_])
- \W znak inny niż znak wyrazu (skrót dla [^a-zA-Z0-9_])
- \d liczba (skrót dla [0-9])
- \D znak inny niż liczba (skrót dla [^0-9])
- \c**X** znak ctrl+X. Np: \cm oznacza control-M
- \x**hh** znak o kodzie *hh* (w systemie hexadecymalnym)
- \u**hhhh** znak Unicode o kodzie *hhhh* (w systemie hexadecymalnym)
- **dopasowywanie specjalnych znaków literowych**:
  **** - należy umieścić przed znakiem specjalnym
  np. \\, \}, \$,\.

Zastosowania

- Przetwarzanie tekstu:
  - walidacja formularzy
  - poprawianie pomyłek
  - masowa zmiana wyrazów w tekście
  - wyciąganie pewnych wyrazów pasujących do wzorca z tekstu
  - konwertowanie adresów www, generowanych dynamicznie na statyczne
  - i wiele innych operacji związanych z tekstem...

Zastosowanie wyrażeń regularnych

- sprawdzenie poprawności adresu pocztowego postaci: **user@serwer.domena** za pomocą wyrażenia regularnego:
  ^[a-zA-Z0-9_]+@[a-zA-Z0-9\-]+\.[a-zA-Z0-9\-\.]+$
- znaczenie podwyrażeń:

^[a-zA-Z0-9_]+	początek ciągu to przynajmniej jedna litera, cyfra lub _ (albo kombinacja tych znaków
@	znak @
[a-zA-Z0-9\-]+	znaki alfanumeryczne i łączniki
\.	znak .
[a-zA-Z0-9\-\.]+$	litery, cyfry i łączniki oraz ewentualnie więcej kropek, i tak do końca ciągu

Przykładowe wyrażenia regularne

- www_reg = "^(((f|ht)tpp?s?):\/\/|www\.)[\w\-_]+(\.[\w\-_]+)+([\w\-\.,@?^=%&:\/~\+#]*[\w\-\@?^=%&\/~\+#])?";
- email_reg = "^([\w_]+)-*\.?)+@[\w](([\w]+)-?_?\.?)+([a-z]{2,4})$";
- imie_reg = "^[a-zA-ZąćęłńóśżźĄĆĘŁŃÓŚŻŹ]{2,20}$";
- nazwisko_reg = "^[a-zA-ZąćęłńóśżźĄĆĘŁŃÓŚŻŹ]{2,40}$";
- login_reg = "^[a-zA-Z0-9_-]{3,15}$";
- miasto_reg = "^[a-zA-ZąćęłńóśżźĄĆĘŁŃÓŚŻŹ\s]{2,50}$";
- tel_reg = "^([1-9]{1,1}[0-9]{1,1}(-)?[1-9]{1,1}[0-9]{6,6})|([1-9]{1,1}[0-9]{8,8})$";