

Zend Framework! Programming

Shopping application

Regular expressions

Regular expressions (Biểu thức chính qui) bắt nguồn từ ngôn ngữ Perl và hiện nay nó hầu như có trong tất cả các ngôn ngữ lập trình. Là một phần rất quan trọng quá trình xử lý chuỗi và hỗ trợ cho các lập trình viên giảm bớt những dòng mã trong quá trình xử lý chuỗi bằng những biểu thức ngắn gọn nhưng đem lại kết quả như sự mong đợi. Nhưng để sử dụng tốt Regular expressions thì không phải dễ dàng nếu chúng ta không đi đúng hướng và hiểu hết các kí hiệu của nó. Regular Expression thường sử dụng trong những trường hợp sau:

- Kiểm tra giá trị các phần tử của Form
- Xử lý yêu cầu phức tạp trong chuỗi như bóc tách, thay đổi nội dung, loại bỏ các ký tự không cần thiết.

Giáo trình: Zend Framework! Programming

Chuyên đề: Shopping application

Bản quyền: ZendVN group
Biên soạn: Phạm Vũ Khánh
Email: vukhanh2212@gmail.com
Điện thoại: 0908.893326
Website: www.zend.vn
Cập nhật: 07-2010

A. Một số hàm hỗ trợ Regular expressions

1. Tham chiếu

Trong lập trình có 2 loại biến chúng ta thường sử dụng truyền giá trị vào hàm đó là biến tham trị và biến tham chiếu.

- Biến tham trị: Khi truyền giá trị vào hàm, mọi sự thay đổi của biến tham trị trong thân hàm không ảnh hưởng đến biến ban đầu.
- Biến tham chiếu: Khi truyền giá trị vào hàm, thì giá trị của biến tham trị sẽ thay đổi bởi nội dung xử lý trong thân hàm được gọi.

Ví dụ:

```
<?php
    $pheapCong = pheapCong(10,5,$c);
    echo "<br>" . $pheapCong;
    echo "<br>" . $c;

    function pheapCong($a = 0, $b = 0, &$c = 0){
        $tong = $a + $b;
        $c = $tong * 10;
        return $tong;
    }

?>
```

2. preg_match

Hàm preg_match trả về giá trị 1 nếu tìm được pattern trong chuỗi và trả về 0 nếu không tìm được. Hàm này trả về một mảng có một phần tử trong chuỗi

```
preg_match ($pattern, $subject, &$matches)
```

\$pattern : chứa mẫu cần tìm kiếm, nó được xem như 1 chuỗi.

\$subject : chứa chuỗi nguồn.

&\$matches: Mảng tham biến này chứa phần tử được tìm thấy

Ví dụ:

```
<?php
    $subject = "abcdefasdasdasd";
    $pattern = '#a#';
    echo preg_match($pattern, $subject, $matches) . '<br>';
    print_r($matches);

?>
```

3. preg_match_all

Hàm preg_match trả về giá trị 1 nếu tìm được pattern trong chuỗi và trả về 0 nếu không tìm được. Hàm này trả về một mảng có nhiều phần tử trong chuỗi

```
preg_match_all($pattern, $subject, &$matches);
```

Ví dụ:

```
<?php
$subject = "abcdefasdasdasd";
$pattern = '#a#';
echo preg_match_all($pattern, $subject, $matches) . '<br>';
echo '<pre>';
print_r($matches);
echo '</pre>';
?>
```

Output:

```
4
Array
(
    [0] => Array
        (
            [0] => a
            [1] => a
            [2] => a
            [3] => a
        )
)
```

Hai hàm này thường dùng để kiểm tra chuỗi nhập vào có đúng với yêu cầu hay không hoặc sử dụng để bóc tách các chuỗi có cấu trúc phức tạp.

B. Các ký hiệu Regular expressions

1. Tìm chuỗi

<i>\$subject</i>	Hello, world!
<i>\$pattern</i>	Hello
<i>\$matches</i>	Hello, world!

2. Ký tự ^ và \$

Ký tự mũ (^): Tìm giá trị ở đầu chuỗi

Ký tự dola (\$): Tìm giá trị ở cuối chuỗi

Ví dụ: Tìm chữ 'who' ở đầu chuỗi:

<i>\$subject</i>	who is who
<i>\$pattern</i>	^who
<i>\$matches</i>	who is who

Ví dụ: Tìm chữ 'who' ở cuối chuỗi:

<i>\$subject</i>	who is who
<i>\$pattern</i>	who\$
<i>\$matches</i>	who is who

3. Ký tự \

*Khi tìm kiếm chuỗi có một số giá trị đặc biệt trong cú pháp của REX chúng ta sẽ phải sử dụng dấu *

Ví dụ 1: Tìm ký hiệu '\$' trong chuỗi:

\$subject	\$12\$ \- \$25\$
\$pattern	\$
\$matches	\$12\$ \- \$25\$ (Not Found)

Ví dụ 2: Tìm ký hiệu '\$' trong chuỗi:

\$subject	\$12\$ \- \$25\$
\$pattern	\\$
\$matches	\$12\$ \- \$25\$

Ví dụ 3: Tìm ký hiệu '\$' ở đầu chuỗi:

\$subject	\$12\$ \- \$25\$
\$pattern	^\\$
\$matches	\$12\$ \- \$25\$

Ví dụ 4: Tìm ký hiệu '\$' ở cuối chuỗi:

\$subject	\$12\$ \- \$25\$
\$pattern	\\$
\$matches	\$12\$ \- \$25\$

Ví dụ 5: Tìm ký hiệu '\' trong chuỗi:

\$subject	\$12\$ \- \$25\$
\$pattern	\\
\$matches	\$12\$ \- \$25\$

4. Ký tự .

Ký tự là dấu chấm là đại diện cho một ký tự bất kỳ

Ví dụ 1:

\$subject	Regular expressions are powerful!!!
\$pattern	.
\$matches	Regular expressions are powerful!!!

Ví dụ 2:

\$subject	Regular expressions are powerful!!!
\$pattern
\$matches	Regular expressions are powerful!!!

Ví dụ 3:

\$subject	O.K.
\$pattern	.
\$matches	O.K.

Ví dụ 4: Tìm ký tự chấm (.) trong chuỗi

\$subject	O.K.
\$pattern	\.
\$matches	O K

5. Ký tự []

Ví dụ 1: Tìm tập hợp các ký tự oyu trong chuỗi

\$subject	How do you do?
\$pattern	[oyu]
\$matches	How do you do?

Ví dụ 2:

\$subject	How do you do?
\$pattern	[dH].
\$matches	How do you do?

Ví dụ 3:

\$subject	How do you do?
\$pattern	[owy][yow]
\$matches	How do you do?

6. Ký tự gạch ngang (-)

Ví dụ 1: Lấy các ký tự từ C-K

\$subject	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789
\$pattern	[C-K]
\$matches	ABCDEFGHIJKLMN

Ví dụ 2:

\$subject	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789
\$pattern	[CDEFGHIJK]
\$matches	ABCDEFGHIJK

Ví dụ 3:

\$subject	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789
\$pattern	[a-d]
\$matches	abcd

Ví dụ 4:

\$subject	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789
\$pattern	[2-6]
\$matches	23456

Ví dụ 5:

\$subject	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789
-----------	--

\$pattern	[C-Ka-d2-6]
\$matches	ABCDEF FGHIJK LMNOPQRSTUVWXYZ abc defghijklmnopqrstuvwxyz01 23456 789

7. Phủ định (^)

Nếu một ký tự ^ đứng trước một ký tự hay một tập hợp có nghĩa là phủ định của ký hiệu hay tập hợp đó

Ví dụ 1:

\$subject	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789
\$pattern	[^CDghi45]
\$matches	AB C DEFGHIJKLMNOPQRSTUVWXYZ abc defghijklmnopqrstuvwxyz012345 6789

Ví dụ 2:

\$subject	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789
\$pattern	[^W-Z]
\$matches	ABCDEF FGHIJKLMNOPQRSTU VWXYZabcdefghijklmnopqrstuvwxyz0123456789

8. Sự lựa chọn (A|B|C)

Ví dụ 1:

\$subject	Monday Tuesday Friday
\$pattern	(on ues rida)
\$matches	Mon day Tues day Frida y

Ví dụ 2:

\$subject	Monday Tuesday Friday
\$pattern	(Mon Tues Fri)day
\$matches	Monday Tuesday Friday

Ví dụ 3:

\$subject	Monday Tuesday Friday
\$pattern	..(id esd nd)ay
\$matches	Monday Tuesday Friday

9. Ký tự * + ?

*: 0 hoặc nhiều lần xuất hiện
+: 1 hoặc nhiều lần xuất hiện
?: 0 hoặc 1 lần xuất hiện

Ví dụ 1:

\$subject	aabc abc bc
\$pattern	a*b
\$matches	aabc abc bc

Ví dụ 2:

\$subject	aabc abc bc
\$pattern	a+b
\$matches	aabc abc bc

Ví dụ 3:

\$subject	aabc abc bc
\$pattern	a?b
\$matches	aabc abc bc

10. Một số ví dụ về kí hiệu *

Ví dụ 1:

\$subject	-@- *** -- "*" -- *** -@-
\$pattern	.*
\$matches	-@- *** -- "*" -- *** -@-

Ví dụ 2:

\$subject	-@- *** -- "*" -- *** -@-
\$pattern	-A*-
\$matches	-@- *** -- "*" -- *** -@-

Ví dụ 3:

\$subject	-@- *** -- "*" -- *** -@-
\$pattern	[-@]*
\$matches	-@- *** -- "*" -- *** -@-

11. Một số ví dụ về kí hiệu +

Ví dụ 1:

\$subject	-@ @ @- * * * - "*" - * * * -@ @ @-
\$pattern	*+
\$matches	-@ @ @- * * * - "*" - * * * -@ @ @-

Ví dụ 2:

\$subject	-@ @ @- * * * - "*" - * * * -@ @ @-
\$pattern	-@+-
\$matches	-@ @ @- * * * - "*" - * * * -@ @ @-

Ví dụ 3:

\$subject	-@ @ @- * * * - "*" - * * * -@ @ @-
\$pattern	[^]+
\$matches	-@ @ @- * * * - "*" - * * * -@ @ @-

12. Một số ví dụ về kí hiệu ?

Ví dụ 1:

\$subject	--XX-@-XX-@@-XX-@@@-XX-@@@@-XX-@@-@@-
\$pattern	-X?XX?X
\$matches	-XX-@-XX-@@-XX-@@@-XX-@@@@-XX-@@-@@-

Ví dụ 2:

\$subject	--XX-@-XX-@@-XX-@@@-XX-@@@@-XX-@@-@@-
\$pattern	-@?@?@?-
\$matches	-XX-@-XX-@@-XX-@@@-XX-@@@@-XX-@@-@@-

Ví dụ 3:

\$subject	--XX-@-XX-@@-XX-@@@-XX-@@@@-XX-@@-@@-
\$pattern	[^@]@?@
\$matches	--XX-@-XX-@@-XX-@@@-XX-@@@@-XX-@@-@@-

13. Số lần xuất hiện của ký tự {n,m}

Ví dụ 1:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	.{5}
\$matches	One ring to bring them all and in the darkness bind them

Ví dụ 2:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	[els]{1,3}
\$matches	One ring to bring them all and in the darkness bind them

Ví dụ 3:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	[a-z]{3,}
\$matches	One ring to bring them all and in the darkness bind them

14. Ví dụ kết hợp * + ? {n,m}

Ví dụ 1:

\$subject	AA ABA ABBA ABBBA
\$pattern	AB*A
\$matches	AA ABA ABBA ABBBA

Ví dụ 2:

\$subject	AA ABA ABBA ABBBA
\$pattern	AB{0,}A
\$matches	AA ABA ABBA ABBBA

Ví dụ 3:

\$subject	AA ABA ABBA ABBBA
\$pattern	AB+A
\$matches	AA ABA ABBA ABBBA

Ví dụ 4:

\$subject	AA ABA ABBA ABBBA
\$pattern	AB{1,}A
\$matches	AA ABA ABBA ABBBA

Ví dụ 5:

\$subject	AA ABA ABBA ABBBA
\$pattern	AB?A
\$matches	AA ABA ABBA ABBBA

15. Ví dụ kết hợp . * + ?

- a. *: 0 hoặc nhiều lần xuất hiện
- b. +: 1 hoặc nhiều lần xuất hiện
- c. ?: 0 hoặc 1 lần xuất hiện

Ví dụ 1:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	r.*
\$matches	One ring to bring them all and in the darkness bind them

Ví dụ 2:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	r.*?
\$matches	One ring to bring them all and in the dark ness bind them

Ví dụ 3:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	r.+
\$matches	One ring to bring them all and in the darkness bind them

Ví dụ 4:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	r.+?
\$matches	One ring to bring them all and in the dark ness bind them

Ví dụ 5:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	r.?
\$matches	One ring to bring them all and in the dark ness bind them

Ví dụ 6:

\$subject	One ring to bring them all and in the darkness bind them
\$pattern	r.??
\$matches	One ring to bring them all and in the darkness bind them

16. Tập hợp các ký tự \w tương đương [A-z0-9_]

- a. *: 0 hoặc nhiều lần xuất hiện
- b. +: 1 hoặc nhiều lần xuất hiện
- c. ?: 0 hoặc 1 lần xuất hiện

Ví dụ 1:

\$subject	A1 B2 c3 d_4 e:5 ffGG77--__--
\$pattern	\w
\$matches	A1 B2 c3 d_4 e:5 ffGG77--__--

Ví dụ 2:

\$subject	A1 B2 c3 d_4 e:5 ffGG77--__--
\$pattern	\w*
\$matches	A1 B2 c3 d_4 e:5 ffGG77--__--

Ví dụ 3:

\$subject	A1 B2 c3 d_4 e:5 ffGG77--__--
\$pattern	[a-z]\w*
\$matches	A1 B2 c3 d_4 e:5 ffGG77--__--

Ví dụ 4:

\$subject	A1 B2 c3 d_4 e:5 ffGG77--__--
\$pattern	\w{5}
\$matches	A1 B2 c3 d_4 e:5 ffGG77--__--

Ví dụ 5:

\$subject	A1 B2 c3 d_4 e:5 ffGG77--__--
\$pattern	[A-z0-9_]
\$matches	A1 B2 c3 d_4 e:5 ffGG77--__--

17. Tập hợp các ký tự \W tương đương [^A-z0-9_] (phủ định của \w)

- *: 0 hoặc nhiều lần xuất hiện
- +: 1 hoặc nhiều lần xuất hiện
- ?: 0 hoặc 1 lần xuất hiện

Ví dụ 1:

\$subject	AS_34:AS11.23 @#\$ %12^*
-----------	--------------------------

\$pattern	\W
\$matches	AS_34:AS11.23 @\$ %12^*

Ví dụ 2:

\$subject	AS_34:AS11.23 @\$ %12^*
\$pattern	\w
\$matches	AS_34:AS11.23 @\$ %12^*

Ví dụ 3:

\$subject	AS_34:AS11.23 @\$ %12^*
\$pattern	[^A-z0-9_]
\$matches	AS_34:AS11.23 @\$ %12^*

18. \s tập hợp những ký tự khoảng trắng - \S tập hợp những ký tự không phải khoảng trắng

- *: 0 hoặc nhiều lần xuất hiện
- +: 1 hoặc nhiều lần xuất hiện
- ?: 0 hoặc 1 lần xuất hiện

Ví dụ 1:

\$subject	When young was mountain under moon;
\$pattern	\s
\$matches	When young was mountain under moon;

Ví dụ 2:

\$subject	When young was mountain under moon;
\$pattern	\S
\$matches	When young was mountain under moon;

19. \d tập hợp những ký tự từ 0 đến 9 [0-9] - \D tập hợp không thuộc từ 0-9 [^0-9]

- *: 0 hoặc nhiều lần xuất hiện
- +: 1 hoặc nhiều lần xuất hiện
- ?: 0 hoặc 1 lần xuất hiện

Ví dụ 1:

\$subject	Page 123; published: 1234 id=12#24@112
\$pattern	\d
\$matches	Page 123; published: 1234 id=12#24@112

Ví dụ 2:

\$subject	Page 123; published: 1234 id=12#24@112
\$pattern	\D

\$matches	Page 123; published: 1234 id=12#24@112
-----------	--

Ví dụ 3:

\$subject	Page 123; published: 1234 id=12#24@112
\$pattern	[0-9]
\$matches	Page 123; published: 1234 id=12#24@112

20. \A tìm từ đầu chuỗi nguồn - \A tương đương với dấu ^; \Z tìm từ cuối chuỗi nguồn -\Z với dấu \$

Ví dụ 1:

\$subject	Ere iron was found or tree was hewn.
\$pattern	\A...
\$matches	Ere iron was found or tree was hewn.

Ví dụ 2:

\$subject	Ere iron was found or tree was hewn.
\$pattern	...\Z
\$matches	Ere iron was found or tree was hewn.

21. (?=<patter>)

Ví dụ 1:

\$subject	AAAX---aaax---111
\$pattern	\w+(?=X)
\$matches	AAAX---aaax---111

Ví dụ 2:

\$subject	AAAX---aaax---111
\$pattern	\w+
\$matches	AAAX---aaax---111

22. (?!<patter>)

Ví dụ 1:

\$subject	AAAX---AAA
\$pattern	AAA(?!X)
\$matches	AAAX---AAA

Ví dụ 2:

\$subject	AAAX---AAA
\$pattern	AAA
\$matches	AAAX---AAA

23. (?<=<patter>)

Ví dụ 1:

<i>\$subject</i>	<p class="Title">This is a title</p>
<i>\$pattern</i>	(?<=<p class=\"Title\">).*(<\/p>)
<i>\$matches</i>	This is a title

24. \b vị trí ở biên của một từ

Ví dụ 1:

<i>\$subject</i>	Ere iron was found or tree was hewn, When young was mountain under moon;
<i>\$pattern</i>	\b.
<i>\$matches</i>	Ere iron was found or tree was hewn, When young was mountain under moon;

Ví dụ 2:

<i>\$subject</i>	Ere iron was found or tree was hewn, When young was mountain under moon;
<i>\$pattern</i>	. \b
<i>\$matches</i>	Ere iron was found or tree was hewn, When young was mountain under moon;

Ví dụ 3:

<i>\$subject</i>	The captain wore his cap and cape proudly as he sat listening to the recap of how his crew saved the men from a capsized vessel.
<i>\$pattern</i>	\bcap
<i>\$matches</i>	The captain wore his cap and cape proudly as he sat listening to the recap of how his crew saved the men from a capsized vessel.

Ví dụ 4:

<i>\$subject</i>	The captain wore his cap and cape proudly as he sat listening to the recap of how his crew saved the men from a capsized vessel.
<i>\$pattern</i>	\bcap
<i>\$matches</i>	The captain wore his cap and cape proudly as he sat listening to the recap of how his crew saved the men from a capsized vessel.

25. \B vị trí không ở biên của một từ

Ví dụ 1:

<i>\$subject</i>	Ere iron was found or tree was hewn, When young was mountain under moon;
<i>\$pattern</i>	\B.
<i>\$matches</i>	Ere iron was found or tree was hewn, When young was mountain under moon;

Ví dụ 2:

\$subject	Ere iron was found or tree was hewn, When young was mountain under moon;
\$pattern	.\B
\$matches	Ere iron was found or tree was hewn, When young was mountain under moon;

Special Sequences (Chuỗi đặt biệt)

\w	- Những ký tự bình thường (a-z 0-9 _)
\W	- Những ký tự khác những ký tự bình thường.
\s	- Ký tự khoảng trắng (space, tab CRLF)
\S	- Những ký tự khác các ký tự khoảng trắng
\d	- Số (0-9)
\D	- Những ký tự không phải là số
.	- Những ký tự bất kỳ ngoại trừ ký tự xuống dòng

Meta Characters (Ký tự meta)

^	- Bắt đầu
\$	- Kết thúc
[- Bắt đầu của một tập hợp
]	- Kết thúc của một tập hợp
	- Dấu phân chia (a b) matches a or b
(- Bắt đầu mẫu phụ (subpattern)
)	- Kết thúc mẫu phụ (subpattern)
\	- Escape character

Quantifiers (Số lượng)

n*	- 0 đến n lần xuất hiện
n+	- 1 đến n lần xuất hiện
n?	- 0 đến 1 lần xuất hiện
{n}	- số lần xuất hiện
{n,}	- Xuất hiện ít nhất n lần (n: 0~n)
{n,m}	- Xuất hiện trong khoảng từ n đến m lần

Point based assertions (Điểm xuất hiện)

\b	- Vị trí biên của một từ
\B	- Không phải vị trí biên của một từ
\A	- Vị trí đầu của một chủ đề
\Z	- Vị trí kết thúc hoặc xuống dòng của một chủ đề
\z	- Vị trí cuối của một chủ đề

C. Một số ứng dụng trong thực tế của Regular expressions

1. Kiểm tra giá trị đầu vào.

Để kiểm tra một giá trị nào đó có đúng với những theo yêu cầu hay không thì Regular Expression được xem là một sự lựa chọn tốt nhất. Nhưng trước khi viết một biểu thức Regular Expression chúng ta cần xác định điều kiện kiểm tra

Ví dụ 1: Kiểm tra địa chỉ email với những điều kiện cho trước

- Địa chỉ email phải bắt đầu là một ký tự
- Độ dài tối thiểu của email là 3 ký tự và độ dài tối đa là 32 ký tự
- Địa chỉ email là tập hợp các ký tự a-z, từ 0-9 và có thể có các ký tự như dấu chấm (.), dấu gạch dưới (_)
- Tên miền của email có thể là tên miền cấp 1 (vd: zend.vn) hoặc cấp 2 (vd: zend.com.vn)

Ví dụ: Kiểm tra địa chỉ email

```
<?php
    $subject = "zendvn@yahoo.com";
    $pattern = '#^[a-z][a-z0-9_\.]{2,}@[a-z0-9]{3,}(\.[a-z]{2,4}){1,2}$#';
    if(preg_match_all($pattern, $subject, $matches)==1){
        echo 'Đây là địa chỉ email';
    }
?>
```