正则表达式学习笔记

阿左 [1]　　Nobody [2]

December 9, 2012

[1]感谢档
[2]感谢郭嘉

Contents

List of Figures

List of Tables

Abstract

Regex study note

摘要

Regex study note

Part I

基本概念

Chapter 1

元字符（Metacharacters）

1.1 基本元字符

1.1.1 任意字符

点号“.”匹配任意一个字符。

1.1.2 行开始与结束

脱字符与美元符分别代表行的开始与结束位置。注意这两个元字符只表示两个特殊的位置，位置上是没有字符的。

```
^This is a line.$
```

匹配空白行：

```
^$
```

匹配所有的行（因为所有的行都有一个开头）：

```
^
```

1.1.3 单词分界符（Word Boundaries）

“\<”与“\>”匹配单词的开始与结束。注意匹配的是位置，而不是字符。

1.2 字符范围（Character Classes）

“[...]”可以定义一个位置上可以出现的字符的范围。“<H1>”、“<H2>”、“<H3>”可以用：“<H[123]>”来表示。

“[^...]”表示排除指定字符。没有列出来的任何字符都可以。

表达式“q[^u]”匹配不了单词“Iraq”，因为表达式的意义不是“q”后面没有u，而是“q”后面要“有”一个字符，这个字符不能是“u”，其他的都行。

可以用连字符来表示连续的字符：“[0-9a-zA-Z_!.?]”；只有在也只有连字符是特殊字符。后面的下划线、问号、点号等都是普通字符。

如果连字符在开头，那也表示普通字符，不表示连续字符。

```
1 echo '-123456789' | egrep '[a-b]'   # not match
2 echo '-123456789' | egrep '[-ab]'   # match
```

1.3 选择结构（Alternation）

括号构成子表达式，“|”表示逻辑“或”。

```
1 Jeffrey|Jeffery
2 Jeff(re|er)y
```

1.4 重复控制

1.4.1 选项元素（Optional Items）

“?”表示前一个元素是可选的。“July?”可以匹配“Jul”或“July”。

1.4.2 其他量词：重复出现（Other Quantifier: Repetition）

“+”表示前一元素出现一次以上；“*”表示前一元素可不出现或出现多次。

1.4.3 区间量次（Interval Quantifier）

“{min,max}”规定重复出现的次数：

```
1 echo '1234567890' | egrep '[0-9]{8,15}'
```

1.5 括号与反向引用（Parentheses and Backreferences）

在很多版本的正则表达式中，括号中的子表达式能“记住”匹配的内容。“verb|/num|”可以代表第几个子表达式匹配的内容。如，要查找重复的单词：

```
1 echo 'that that' | egrep '\<([A-Za-z]+) +\1\>'
```

“([a-z])([0-9])\1\2”这个表达式中，“\1”表示第一个表达式“[0-9]”匹配的内容；“\2”表示第二个表达式“[0-9]”匹配的内容。

1.6 转义元字符

反斜线“\”实现元字符的转义。大多数正则工具会把字符范围“[...]”中的“\”作为普通字符。

Part II

语言与工具

Chapter 2

egrep

在邮件中查找发信人与主题的例子：

```
egrep '^(From|Subject):' ./*
```

忽略大小写：

```
egrep -i '^(From|Subject):' ./*
```