

正则表达式初识之前世今生

Tom

正则表达式的历史

1、神经生理学

2、数学科学（正则集的代数）

3、计算搜索算法（Unix上的编辑器ed）

正则表达式的概念

使用**单个字符串**来描述、匹配一系列符合某个**句法规则**的字符串，是一门表达式语言。

```
\(0\d{2}\)[-]?d{8}|0\d{2}[-]?d{8}
```

```
\b(\w+)\b\s+\1\b
```

```
\b\w*q[u]\w*\b
```

```
[\u4e00-\u9fa5]
```

```
http://([\w-]+\.)+[\w-]+(\/[\w-.\/?%&=]*)?
```

```
^(\d{1,2}|1\d\d|2[0-4]\d|25[0-5])\.(\d{1,2}|1\d\d|2[0-4]\d|25[0-5])\.(\d{1,2}|1\d\d|2[0-4]\d|25[0-5])\.(\d{1,2}|1\d\d|2[0-4]\d|25[0-5])
```

正则表达式引擎

1、DFA: awk (大多数版本)、egrep (大多数版本)、flex、lex、MySQL、Procmail

文本主导；快，但特性较少；最长的左子正则式优先匹配

2、传统NFA: Java .Net PCRE library、Perl、PHP、Python、Ruby

表达式主导：相对慢，支持lazy和backreference等特性；最左子正则式优先匹配

3、POSIX NFA: mawk、Mortice Learn System's utilities、GUN Emacs

举例

匹配一个通用的URL地址

```
((http|https|ws|wss):\\|\\/)(([a-zA-Z0-9\\._-]+\\.[a-zA-Z]{2,6})|([0-9]{1,3}\\.[0-9]{1,3}\\.[0-9]{1,3}))(:[0-9]{1,5})*(\\/[a-zA-Z0-9\\&%_\\.\\-]*)?
```


学习正则的几个原则

- 1、多看、多练、多思考
- 2、多结合实际场景
- 3、学好之后，威力无穷，一劳永逸