河北师大软件学院 *@Software College*

前端开发与HTML5
程序设计基础

王岩

2.11正则表达式

实际应用：抓取数据


豆瓣douban
首页
浏览发现
移动应用 new
游戏
线上活动
搜索你感兴趣的内容和人...
文科生转码农的几点建议
Casa Nova 2016-05-03 07:22:15
刚才看到友邻的广播，想到自己这一年多的经历，忍不住想说两句。
先介绍一下自己的背景，从本科到PhD，都是文科专业——先是英语，然后语言学。两年前误打误撞回国找了工作，跑到一家商业咨询公司做文本分析。读书的时候，只学过shell和R，工作之后，才开始写Python。
其实在豆瓣上，在生活中，和我有类似经历的朋友不少，大家或多或少地都迷茫过，所以我斗胆在这里写下一些建议。
首先，我们要明确优势。虽然出身文科，可是我们有自己的优势。前司去年招了一个读理论语言学出身的硕士做自然语言开发，来的时候一行代码都不会写，可是她对句法的见解很独到。四个月时间，边学Python，边把自己的想法实现，最后自己用Python写成了一个汉语句法分析工具。这种长时间积攒的功底，恐怕不是其他项目能够比拟的。所以，尽管从文科转码农，我们的优势还是很明显的。
其次，我们要摆正心态。很简单，我们的peer都是理工科背景出身的，写过的代码比我们多得多，而且难免有时心直口快。所以我们被他们打击是很正常的，时不时地被批代码烂都是家常便饭了。当然，大多数的时候，他们只是说我们代码烂，没有其他意思，所以要摆正心态，不要太在意，努力把代码写好就是了。
再次，我们要找到合适的环境。我选的第一家公司虽然是商业咨询公司，但里面的数据处理团队是刚组建的，而且有几个牛人，所以相当于在一个创业的环境里工作。这样的好处在于，第一，有人可以带你，第二，有足够多的活可以干。反面的例子也有，我们最近面试了一个非常优秀的姑娘，硕士在英国最好的学校读了应用语言学，回国在一家跨国科技巨头做搜索引擎监测。问她为何想跳槽，她说她最大的顾虑就是，在大公司里分工太细，她能学到的太少，生怕今后跳槽找不到其他工作。所以我的建议是，找一家有牛人的小公司，这样我们会成长得很快
Casa Nova (Edinburgh, United Kingdom)
拍更好的照片 写更好的代码 过更难的路线 做更好的人 谢绝转载 谢绝转载...
Casa Nova的最新日记 · · · · · · (全部)
初期创业公司招聘指南 (6人喜欢)
再征一次 (87人喜欢)
2016年2月读书记 (2人喜欢)
豆瓣征友文写作指南 (6人喜欢)
2016年1月读书记 (3人喜欢)
Casa Nova的日记标签 · · · · · · (全部)
學術 21
工具 13
workflow 12
tips 10
寫作 9
...

实际应用：数据校验


1
创建账户
2
填写账户信息
3
企业实名认证
注册成功
个人账户
企业账户
账户名
wwwww
邮箱格式不正确，请重新输入
验证码
>>
请按住滑块，拖动到最右边
我同意支付宝服务协议
下一步
港澳台及海外企业注册>

引言

- 字符串的查找、替换、校验等操作，都需要一种更方便、快捷的方法表示出需要的字符规则（字符模式）
- 正则表达式：表示一些特定的字符规则（字符模式）

正则表达式简介

- 正则表达式
- 按一定规则书写，用来描述模式的特殊字符串
- 主要用来依据模式查找、替换、校验等
- 广泛应用在多种编程语言中，如JavaScript，Java，PHP等

PHP中的正则表达式

- Perl兼容的正则表达式
- POSIX风格的正则表达式(主要用在Unix系统中)

使用正则表达式

使用正则表达式

- 创建正则表达式
  - 例：QQ号码的正则表达式：/^[1-9][0-9]{4,10}$/
- 使用正则表达式进行字符串操作
  - 匹配：preg_match()
  - 替换：preg_replace()
  - 拆分：preg_split()
  - 过滤：preg_grep()
- 处理操作结果

模式匹配

- preg_match()
  - preg_match($pattern, $string [,$matches]);
  - 该函数返回匹配的次数：0次或1次
    - Demo2-11-1：判断一个字符串是不是QQ号码的格式（5-11位数字组成，且首位不能是0）
      - QQ号码正则表达式：/^[1-9][0-9]{4,10}$/
- preg_match_all()
  - preg_match_all($pattern, $string , $matches);

替换

- preg_replace()
  - preg_replace($pattern, $replacement, $string[, $limit]);
- Demo2-11-2：移除页面上“/* … */”多行注释内容
  - 多行注释的正则表达式　/(\/*)+(.|\n|\r)+(*\/)/

拆分

- preg_split()
  - preg_split($pattern, $subject[, $limit]);
  - 返回被分割后的数组
  - Demo2-11-3：在数字表达式中取操作数
    - 正则表达式为 /[+-*/]/

过滤数组

- preg_grep()
  - preg_grep ($pattern, $input);
  - 返回一个数组，其中包括了 input 数组中与给定的 pattern 模式相匹配的单元
  - Demo2-11-4：得到数组中扩展名为”.txt”的文件名
    - 文件扩展名正则表达式 /\.txt$/

正则表达式语法

正则表达式基本语法

- 正则表达式是按一定规则书写，用来描述模式的特殊字符串
- 语法：
  - 1）写在/ /之间
  - 2）使用引号引起来
- 如：判断字符串$str是否为有效的QQ号码(5 – 11位)
  - $patern=“/^[1-9][0-9]{4,10}$/”

普通字符

- 原字符
  - 如‘/abc/’、’/123/’……
- 元字符
  - (、[、{、\、|、^、$、？、*、.、+
  - 如何匹配‘？’
    - \?
- 转义字符：\0、\t、\f、\n等

字符类

- 如何匹配子串：’a1b’、’a2b’、’a3b’ ‘/(a1b)|(a2b)|(a3b)/’
- 字符类：[ ]，选择其中的任何一个 ‘/a[123]b)/’
- 字符范围类：’/[1-9] /’、’/[a-zA-Z]/’
- 反义字符类：’/[^1-9]/’

练习

- 一年中的月份(01-09或10-12)
- 每一月的某一天(01-09或10-29或30-31)
- 除换行符和回车符外的其它任意字符
- 任何ASCII码单字符(字母数字下划线)
- 任何ASCII码数字

预定义字符类

Perl风格字符类	说明
.（点号）	等价于[^\n\r]
\w	任何ASCII单字字符，等价于[a-zA-Z0-9_]
\W	任何非ASCII单字字符，等价于[^a-zA-Z0-9_]
\d	数字，等价于[0-9]
\D	除了数字之外的任何字符，等价于[^0-9]
\s	空白符，等价于[\t\n\x0B\f\r]
\S	非空白符，等价于[^\s]

重复类数量词

- 如何匹配
  - Google
  - Gooogle
  - Goooogle
  - Gooooogle
  - …

重复类数量词

重复类数量词	说明
{n}	匹配前一项恰好n次
{n,m}	匹配前一项n<=x<=m次
{n,}	匹配前一项至少n次
?	匹配前一项0次或1次
*	匹配前一项0次或多次
+	匹配前一项1次或多次

定界符

定位符	说明
^	匹配一行的开头
$	匹配一行的结尾

其他定界符

Perl风格锚	说明
\b	单词边界
\B	非单词边界
\A	字符串的开始
\Z	字符串的末尾或在结尾的\n前
\z	字符串的末尾

后缀选项

- 把单个字符选项放在正则表达式模式的后面来修改匹配的解释或行为。
  - 例如：'/cat/i'

修饰符	说明
i	不区分大小写的匹配
x	从模式中删除空白符和注释

子模式

- 作用：
  - 把子模式表达式当作一个整体
  - 表示子模式，方便后续引用
- 例：“PHPer love PHP”匹配PHPer或者PHP
  - 正则表达式：‘/PHP(er)?/’

反向引用

- 思考：如何解决引号匹配问题？？？
  - 正则表达式：/[‘”].*[‘”]/　　有什么问题?
- 子模式匹配的结果可以被后续引用，使用“\1”,“\2”分别引用第1个、第2个子模式所匹配的字符串
  - 如：/([‘\”]).*\1/
- Demo2-11-5

理解正则表达式

写正则表达式

- 验证电子邮箱地址
  - 分析目标的规则（可能出现的情况）
  - 必须完全匹配（不能是部分匹配）
    - 必须有@
    - @符号之前可能是字母数字下滑线、加号减号、点（.）
    - @符号之后必须至少有一个点（.），此外可以有字母数字或下划线或减号
- Demo2-11-6：逐步写出验证电子邮箱的正则表达式

读正则表达式

- 读一个正则表达式：
  - URL：/^http(s)?:\/\/([\w-]+\.)+[\w]+(\/[\w-.\/?%&=]*)?$/
    - 分析目标规则（可能出现的情况）
    - 从前往后逐步分析正则表达式

实例

- 抓取豆瓣网站的一篇文章的标题、作者及内容
- Demo2-11-7

作业

- 抓取果壳网站上，“小组”应用中的帖子：标题、作者、发表时间及帖子内容，并在本页面中显示。

谢谢！