# 随堂笔记-正则表达式

# 主要内容

* 正则表达式
* 扩展：爬虫入门

# 课程内容

## 需求案例

开发了一个项目，项目中需要用户对象[昵称、姓名、账号、密码、邮箱、手机..]

我们需要用户可以通过**昵称/邮箱/账号/手机** + 密码都能登录系统。

需求：**验证用户输入的是手机号码[158|156|188]开头**

|  |
| --- |
| 代码开发： |
| 提示1：用户可能输入长度不合法的手机号码 1568288082270  提示2：用户可能输入非数字的手机号码 1568280827o  提示3：用户的操作，无规则的！ |
| 代码设计：用户输入的只能是字符串  目标：判断用户输入的字符串，是否以156|158|188开头的11位数字  是：合法，可用  否：不合法，非法账号  核心：[长度]->[开头]->[数字] |

## 正则表达式-what

什么是正则表达式

正则表达式，主要是通过定义正确的规则从目标字符串中(非常高效的)匹配符合规则的数据；正则表达式是一个独立的基础，python语言支持正则操作。

正则表达式的重点

高效率的处理(查询、匹配、替换)**字符串**

## 正则表达式-why

正则表达式的由来

显示生活中，什么样的一种存在是操作最为频繁的？

为什么要用正则表达式

就是为了更加高效的处理程序中出现最频繁的字符串数据的

## 正则表达式-how

正则表达式-语法

|  |
| --- |
| python中re模块，主要用来进行正则表达式的操作 1. 定义正则表达式 【普通字符串】 2. 从目标字符串中提取数据 【正则规则】  正则规则：语法 1) 普通字符：也是正则表达式的一种字符，匹配当前字符本身  2) ^:表示匹配字符串的开头位置  3) $:表示匹配字符串的结束的位置  4) .:表示匹配任意一个字符[换行、制表符..除外..]  5) \*:表示匹配任意一个字符出现了0次~n次  6) +:表示匹配任意一个字符出现1次~n次  7) ?:表示匹配一个字符出现了0次~1次  8) {m}:表示匹配一个字符出现了m次  {m,n}:表示匹配一个字符至少出现m次最多出现n次  {m,}:表示匹配任意一个字符至少出现m次  {,n}:表示匹配任意一个字符最多出现n次   9) [0-9]:表示匹配任意一个0~9之间的数字  [a-z]:表示匹配任意一个小写字母  [A-Z]:表示匹配任意一个大写字母    [0-3]:表示匹配任意一个0~3的数字  [a-zA-Z]:表示匹配任意一个字母  [abc]:表示匹配a或者b或者c  [^a-z]:尖角符号出现在方括号中，表示取反：匹配非小写字母的字符    [0-120]:表示匹配0~1 或者2 或者0--不能操作数字范围，如果要判断数字范围请使用比较运算符  10) ()表示分组，将一组数据单独区分出来  (abc)：表示匹配abc字符，并且单独存放为一组  为了更加方便的操作正则表达式，出现了一些特殊符号，用于简化正则表达式的操作 11) \d: 和[0-9]相同的意义,匹配任意一个数字  \D: 和[^0-9]相同的意义，匹配任意一个非数字   12) \s: 匹配任意一个空白字符  \S: 匹配任意一个非空白字符  13) \b: 匹配任意一个单词边界  14) \w: 匹配任意一个数字、字母或者下划线，和[a-zA-Z0-9\_]相同的意义   15) \：转义字符  . -> 匹配-> \. 匹配一个普通的字符. |

## 正则表达式-where

## 爬虫

python2.x[使用较多]

urllib2[爬虫基础模块]

requests[爬虫基础模块-第三方模块-生产]

python3.x[主流]

urllib[爬虫基础模块]

requests[爬虫基础模块-第三方模块-生产]

scrapy爬虫框架 – scrapyd 部署爬虫程序 – gerapy 图形化界面操作爬虫

scrapy-redis爬虫分布式框架

celery + rabbitMQ分布式爬虫框架

---

urllib:操作网络资源的一个内建标准模块

urllib.request 模块，包含在urllib模块中的request模块，用于操作网络请求