# 第四阶段 Python数据分析

# Python 爬虫3

## 网校相关视频

Python爬虫开发

第一章网络爬虫基础

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* 利用正则表达式爬取数据

### 【重点】

* 利用正则表达式爬取数据

### 【难点】

* 利用正则表达式爬取数据

## 复习巩固作业讲解

## 本次课程任务讲解

### 【知识点1】使用正则表达式爬取数据

1. 正则表达式的概念

正则表达式是一个特殊的字符序列，它能帮助你方便的检查一个字符串是否与某种模式匹配。

1. Python中的Re模块

提供了对正则表达式的支持

返回pattern对象：

***re.compile(string[,flag])***

匹配函数：

***re.match(pattern, string[, flags]) re.search(pattern, string[, flags])***

***re.split(pattern, string[, maxsplit])***

***re.findall(pattern, string[, flags])***

***re.finditer(pattern, string[, flags])***

***re.sub(pattern, repl, string[, count])***

***re.subn(pattern, repl, string[, count])***

1. re.compile

***compile(pattern[,flags] )***

根据包含正则表达式的字符串创建模式对象

flag参数：匹配模式，取值可以使用按位或运算符’|’表示同时生效，比如re.I | re.M。可选值有：

（1）re.I(全拼：IGNORECASE): 忽略大小写

（2）re.M(全拼：MULTILINE): 多行模式，改变'^'和'$'的行为

（3）re.S(全拼：DOTALL): 点任意匹配模式，改变'.'的行为

（4）re.L(全拼：LOCALE): 使预定字符类 \w \W \b \B \s \S 取决于当前区域设定

（5）re.U(全拼：UNICODE): 使预定字符类 \w \W \b \B \s \S \d \D 取决于unicode定义的字符属性

（6）re.X(全拼：VERBOSE): 详细模式。这个模式下正则表达式可以是多行，忽略空白字符，并可以加入注释。

1. 匹配
   1. ***re.match()：***从字符串的起始位置匹配一个模式，如果不是起始位置匹配成功的话返回None
2. 正则表达式语法：

正则表达式通常通过特殊的语法来表示。方法如下：

* 1. ^表示字符串的开始，$表示字符串的末尾，经常省略
  2. 字母和数字表示他们自身。一个正则表达式模式中的字母和数字匹配同样的字符串。
  3. 其它模式特殊符号描述如下：







1. 使用正则表达式：
   1. **re*.search(pattern, string[, flags])***

re.search函数会在字符串内查找模式匹配,只要找到第一个匹配然后返回，如果字符串没有匹配，则返回None。

re.match与re.search的区别

re.match只匹配字符串的开始，如果字符串开始不符合正则表达式，则匹配失败，函数返回None；而re.search匹配整个字符串，直到找到一个匹配。

match和search一旦匹配成功，就是一个match object对象，而match object对象有以下方法：

group() 返回被 RE 匹配的字符串，是字符串中的部分内容

start() 返回匹配开始的位置

end() 返回匹配结束的位置

span() 返回一个元组包含匹配 (开始,结束) 的位置

group() 返回re整体匹配的字符串，可以一次输入多个组号，对应组号匹配的字符串。

groups()返回所有括号匹配的字符，以元组的形式。

* 1. ***re.findall(pattern, string[, flags])***

遍历匹配，可以获取字符串中所有匹配的字符串，返回一个列表（由匹配到的所有子字符串构成）。(在网页爬取中最常用)

* 1. ***re.finditer(pattern, string[, flags])***

搜索string，返回一个顺序访问每一个匹配结果（Match对象）的迭代器。找到 RE 匹配的所有子串，并把它们作为一个迭代器返回。

* 1. ***re.split(pattern, string[, maxsplit])***

maxsplit用于指定最大分割次数，不指定将全部分割。

按照能够匹配的子串将string分割后返回列表。可以使用re.split来分割字符串，如：re.split(r'\s+', text)；将字符串按空格分割成一个单词列表。

* 1. ***re.sub(pattern, repl, string, count)***

使用re替换string中每一个匹配的子串后返回替换后的字符串。

* 1. ***re.subn(pattern, repl, string, count=0, flags=0)***

返回替换的结果和替换次数

1. 使用正则表达式获取数据的范例：
   1. 常用正则表达式
   2. 判断密码是否有效（8位，包括大写字母、小写字母、数字）
   3. 使用正则表达式获取网页数据
2. 综合案例
   1. 利用正则表达式获取天猫官网商品分类的链接，以xls文件保存
   2. 爬取证券之星网站上沪深A股股票信息，并写入到EXCEL文件中
   3. 爬取快代理上的代理服务器，并写入EXCEL文件
   4. 爬取51job上的数据分析师职位信息，写入EXCEL文件

## 自主学习作业讲解

无

## 课程总结

#### 使用代理服务器

#### Python中的异常处理

#### 使用正则表达式处理爬取的网页数据

## 下次自主学习任务布置

编写代码实现下列需求

爬取51job上的职位信息，并写入到文本和excel文件中



