# 第四阶段 Python数据分析

# Python 爬虫4

## 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requests爬取数据

### 【重点】

* 利用requests爬取数据

### 【难点】

* 利用requests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. 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TXT和EXCEL文件

### 【知识点2】使用requests爬取数据

1. requests库介绍
   1. 概念及使用原因

requests是Python的一个HTTP客户端库，跟urllib类似，python的标准库urllib提供了大部分需要的HTTP功能，但是一个简单的功能就需要一大堆代码。

1. requests使用

*import requests*

*#发出请求*

*#requests.get(url, params=None, \*\*kwargs)*

*response = requests.get(‘http://www.baidu.com’)*

#url：请求网址

#params：字典型请求参数

#kwargs：其他参数，如headers={'User-Agent': 'Mozilla/4.0','content-type': 'text/html; charset=utf-8'}

#返回响应结果（response是返回的响应对象)

#response.url获取请求地址

#response.text或r.content 获取响应内容。本方法requests会自动将内容转码。大多数unicode字体都会无缝转码。

#response.encoding来获取或设置网页编码

print(response.text)

#设置传递参数

#headers设置request-header

#proxies设置代理服务器

*r=requests.get("http://www.baidu.com/", params= language,headers={'User-Agent': 'Mozilla/4.0','content-type': 'text/html; charset=utf-8'})*

*proxies1 = {"https": "27.40.149.116:61234"}*

*r = requests.get("http://www.baidu.com",headers={'User-agent': 'Mozilla/4.0','content-type': 'text/html; charset=utf-8'},proxies=proxies1)*

## 自主学习作业讲解

无

## 课程总结

#### 使用正则表达式处理爬取的网页数据

#### 使用requests爬取网页数据

## 下次自主学习任务布置