爱数据 lovedata.cn

Python网络爬虫实战

讲师：简杨 Jan Yang

1 Python环境搭建及IDE

2 爬虫基础知识

3 初学爬取网站

4 爬取大规模数据

5 反爬与反反爬

6 其他常见问题

Python环境搭建及IDE

01

part one

1. Anaconda及第三方库安装
2. Pycharm及环境配置
3. Mongodb安装和配置

爬虫基础知识

02

part two

1. 网址地址和构成
2. 网页源代码和结构
3. 爬虫基本原理
4. HTTP状态码

网址地址和构成

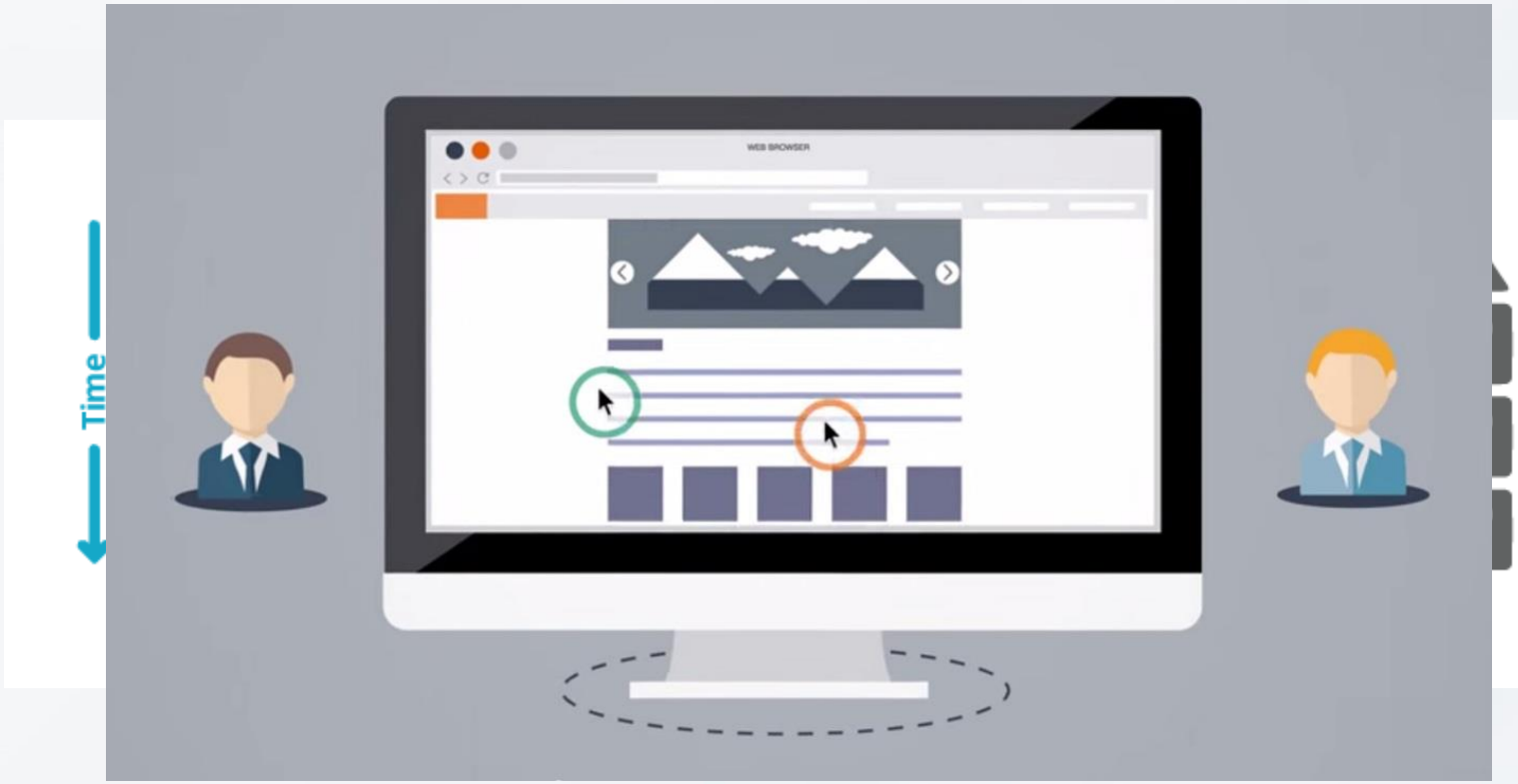
协议 域名 路径 文件 参数

http://tieba.baidu.com/page/3797994694.html?see_lz=1

网页源代码和结构

<html>	整个网页
<head>	头部
<body>	主体
	某块区域框架
<h1>	不同字号标题
	列表
<img>	图片
<a>	放置链接
<p>	文字内容

爬虫基本原理



模拟人通过浏览器批量对网站url点击获取网站的数据

HTTP状态码

200：服务器成功返回网页

404：请求的网页不存在

403：服务器拒绝请求

开始爬取网站

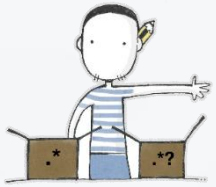
03

part three

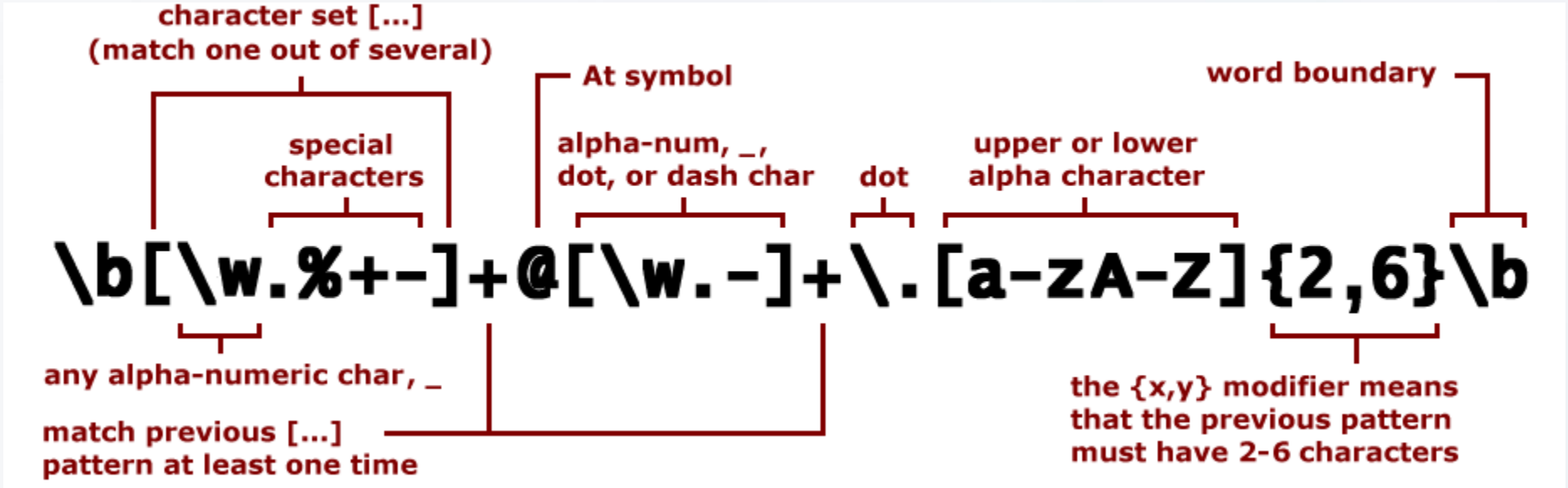
1. 使用正则表达式：下载百度贴吧图片
2. 第三方库：爬取贴吧多页码图片
3. 传递URL参数：买粮网搜索列表
4. 爬取JSON响应内容：第六感别墅度假别墅列表
5. 处理POST请求：获取有道翻译翻译结果

Python自带库

- Python2.X : urllib & urllib2
- Python3.X : urllib


.*
.*?

正则表达式


character set [...]
(match one out of several)
special
characters
At symbol
alpha-num, _,
dot, or dash char
dot
upper or lower
alpha character
word boundary
\b[\w.%+-]+@[\w.-]+\.[a-zA-Z]{2,6}\b
any alpha-numeric char, _
match previous [...]
pattern at least one time
the {x,y} modifier means
that the previous pattern
must have 2-6 characters

常用正则表达式

. 除**换行符“\n”外**的任意字符，但**DOTALL模式**下能匹配换行符

\ **转义字符**，使后一个字符改变原来的意思

\d 数字：**[0-9]**

{m} 匹配前一个字符**m**次

.+? 匹配除换行符外的任意字符一次或者无限次，**非贪婪匹配**

(...) 将被括起来的表达式作为**分组**，多次使用编号+1

第三方库

HTTP请求

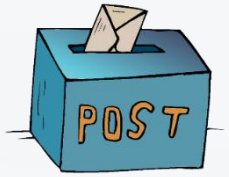
解析HTML

传递URL参数

http://www.mailiangwang.com/biz/list?keyword=玉米&pageid=2

爬取JSON响应内容

```
{ "users":[
            {
                "firstName":"Ray",
                "lastName":"Villalobos",
                "joined": {
                    "month":"January",
                    "day":12,
                    "year":2012
                }
            },
            {
                "firstName":"John",
                "lastName":"Jones",
                "joined": {
                    "month":"April",
                    "day":28,
                    "year":2010
                }
            }
    ]}
```



处理POST请求

有道 youdao 翻译

- 利用有道翻译获取翻译结果

爬取大规模数据

04

part three

1. 使用Mongodb数据库
2. 编写多进程爬虫
3. 实战：爬取第六感别墅度假

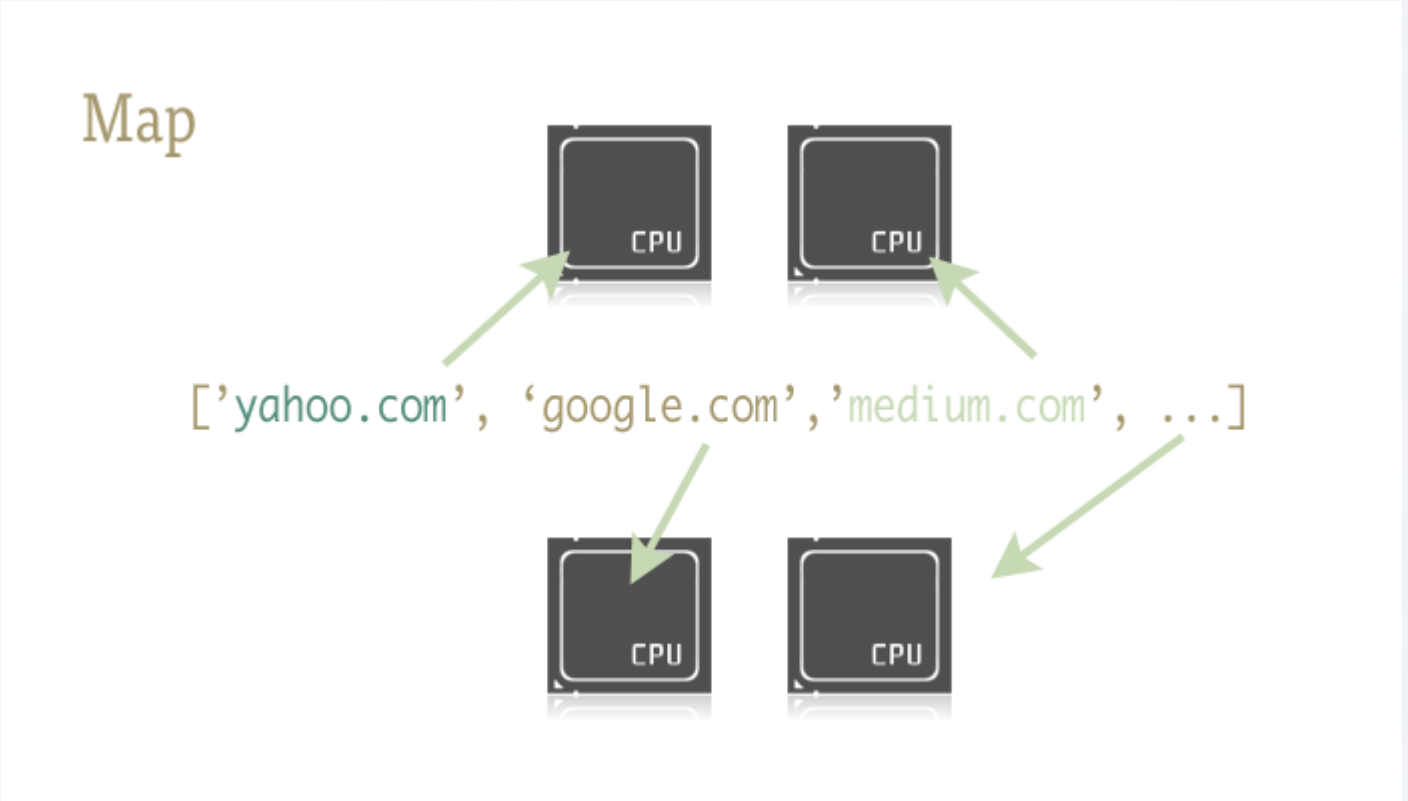
存储数据

python™

mongoDB

Python multiprocessing

编写多进程爬虫



大规模爬取工作流

列表页

详情页

评论页

爬虫实战练习

SENSE LUXURY
第六感别墅度假

- 爬取第六感别墅度假

反爬与反反爬

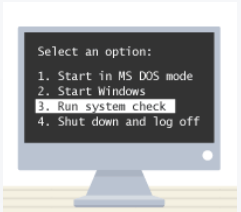
05

part five

1. 降低请求网站频率

2. 定制请求头

3. 使用代理IP访问目标网站



爬虫实战练习

- 访问CSDN博客
- 使用代理IP访问并检验IP

其他常见问题

06

part six

1 编码问题

2 错误与异常

3 设计断点续爬

4 数据导出及读取

Q&A