*爬虫：spider 网络蜘蛛*

*爬虫：模仿浏览器，根据自己制定的规则，下载需要的资源*

*爬虫分为聚焦爬虫和搜索爬虫*

*聚焦爬虫：只针对某个网站的爬取*

*搜索爬虫：针对全网络进行爬取（搜索引擎）*

*模仿浏览器的模块：request,urllib2,urllib3,httpclient*

*自己制定规则：re,bs4(Beautifulsoup),xpath*

*潜规则：类名的首字母大写*

*object:基类（基础的类）*

*爬虫爬取糗事百科文字*

*# 爬虫三部曲 1.分析网址  
# class 糗事\_spider(object):  
# def 发送请求(self,page):  
# url = 'https://www.qiushibaike.com/text/page/{}/'.format(page)  
 #伪装成浏览器  
# head = {  
# 'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/42.0.2311.135 Safari/537.36 Edge/12.10240'}  
# #发送请求  
# res = requests.get(url,headers=head) #get 请求  
# #读取响应 1.text 以字符串的方式读取 2.content 以字节的方式读取  
# html=(res.content.decode('utf-8')) #decode('编码方式')方式解码 查看网站什么编码方式 就用什么方式解码  
# # 返回结果并赋值  
# return html  
# #爬虫三部曲 2.制定规则,获取资源  
# def guolv(self,abc):  
# #将正则表达式编译  
# shuju = []  
# patt = re.compile(r'<div class="content">(.\*?)</span>',re.S)  
# #将编译后的结果到字符串中去查找  
# items = patt.findall(abc)  
# for i in items:  
# i =i.replace('<span>','').replace('<br/>','').strip()  
# shuju.append(i)  
# return shuju  
# # print(len(shuju))  
# # print(i)  
# #爬虫三部曲 3.保存  
# def sava(self,qwe):  
# with open('a.txt','a',encoding='utf-8') as f:  
# for i in qwe:  
# f.write(i+'\n')  
# qiu = 糗事\_spider()  
# for i in range(1,6):  
# html = qiu.发送请求(page=i)  
# shuju = qiu.guolv(abc=html)  
# qiu.sava(shuju)*

*反爬*

*1.通过headers Ures-Agent*

*refere(防盗链)*

*2.验证码 登录 cookie*

*识别*

*3.拉黑IP地址：检测访问的频率 解决办法（设置代理IP）*

*4.数据混淆*

*5．行为分析*