

爬虫技术初探

——宏观分析小工具系列**一**

分析师 梁中华

执业证书编号: S0740518090002

电话 021-20315056

邮箱 liangzh@r.qlzq.com.cn

研究助理 苏仪

电话

邮箱 suyi@r.qlzq.com.cn

相关报告

- 1 全球领先的高分子材料抗老化助 剂供应商
- 2 全球锂电设备龙头,海外客户加速渗透
- 3 银行角度看 11 月社融数:基建支撑企业中长期贷款

投资要点

- 当前宏观研究中面临的一大问题是数据的限制,可以用于分析的质量较高的数据越来越少,而宏观经济内部结构变化又较大,传统的一些数据还出现了失灵的情况。在这种情况下,将一些新的计算机工具应用到宏观分析中,可以在一定程度上弥补数据不足的缺陷。
- 我们团队的苏仪同学,在数据挖掘、数据库搭建、高性能计算及海量数据分析方面具有专长,接下来会为大家介绍一些技术小工具,欢迎沟通交流,有技术相关需求,也欢迎联系我们。



内容目录

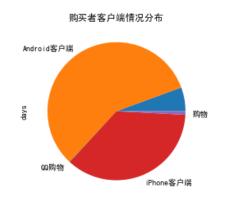
一、爬	虫的合理广泛应用	- 3 -
二、利	用爬虫抓取网页表格数据	- 4 -
三、后:	续使用的相关问题	- 6 -
图.	表目录	
图表 1:	购买者客户端分布情况	- 3 -
图表 2:	购买月份分布图	- 3 -
图表 3:	全国房企破产分布图	- 4 -
图表 4:	run 函数的主要内容	- 4 -
图表 5:	提取关键文本的方法	- 5 -
图表 6.	not csy 函数的主要内容	- 6 -



一、爬虫的合理广泛应用

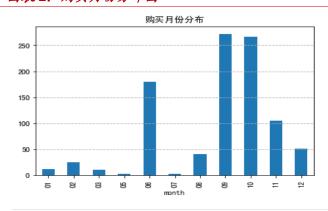
- 爬虫的应用需合法、合理、规范化,大部分抓取的数据只适合用于参考。 近来关于爬虫的新闻有许多,但使用爬虫的主要目的其实是为了更加有 靶向性地快速、高效且持续地在网页上收录及筛选信息,尤其是数据。 在网站上获取的信息,虽然不如行业协会统计的全面,但会对一些细节 问题反映的更加直观。
- 通过在研究中利用爬虫技术获取更新更细的信息,有助于我们更好地做出判断。譬如,在研究茅台销量的时候,从电商平台上抓取一些公布出来的评论信息,也可以对产品做一个侧写。在7月份的时候我们在某电商平台上抓取了500ml53度飞天茅台的用户评论量。在评论区我们可以看到购买者的用户等级、客户端、购买月份及发布的评论。在已有样本中,我们发现购买者使用的客户端中,安卓系统占了57.5%,iPhone占了36%。从购买月份来看,这款茅台在9月份销量最高,其次是10月份,在6月份的时候销量也很突出。

图表 1: 购买者客户端分布情况



来源:京东商城,中泰证券研究所

图表 2: 购买月份分布图



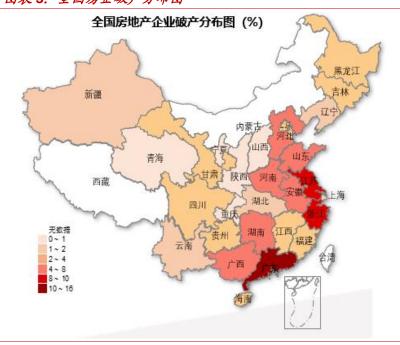
来源:京东商城,中泰证券研究所

■ 但在获取数据时,我们会发现由于网页设置及货品上下架的原因,获得的单一商品的数据在长周期的观察方面会存在断点。并且网站会自动过滤掉一些客户的评论,爬取下来的数据并不完全。这样的数据有一个特性,就是要持续跟踪,积累的数据样本越多,产出的结果越具有借鉴意义。才能看清楚数据背后的意义。我们更倾向于用已有数据计算细分项占比来看趋势,而非精确地定量计算。并且,我们也只是将趋势作为一种参考,并不产生商用价值,还是以官方公布数据为主要数据来源。



- 再比如,考虑到房地产行业在经济中的重要性,我们对人民法院公告网上公布的破产企业文书进行了爬取,筛选出了其中的房地产企业破产文书并进行了统计。根据我们的数据统计,2019年至今我国破产房企已经达到378家,其中广东省破产房企数量最多,高达60家,占比15.87%。浙江、江苏也是房地产企业破产重灾区。
- 但是网站公示的数据量有限,所以我们就需要对该数据做长期持续跟踪 处理。通过月度数据对比,可以直观看到各省份破产企业的增加量,并 做进一步深度研究。





来源:人民法院公告网,中泰证券研究所

■ 此外,我们也可以跟踪一些房产交易网站,统计最新的房产交易、价格 变动状况,能够更加真实、快捷的了解市场走势变化。

二、利用爬虫抓取网页表格数据

- 作为入门,我们先来介绍下如何利用爬虫抓取公开网站表格中的数据。 在爬取过程中会用到两个主要函数,分别是 get_csv 函数和 run 函数。 以统计局公布的 CPI 数据的爬取举例,整个框架中,我们更倾向于首先 读取整个新闻页,筛选出其中的新闻标题和 URL,再进一步筛选出我们 需要的标题并获取 URL,然后访问 URL 并获取具体内容。在解析网页 的时候可以使用 F12 来获取相关信息。
- 首先可以先定义 run 函数,并定义 url 变量。通过 res 变量来获取新闻页的所有内容。然后根据标签内容输入关键词,进行内容检索。在本例中,我们可通过 j.text 来获取文字内容,并通过定义 j.a.attrs 变量来获得我们所需要的信息。在此过程中我们用的解码工具是 BeautifulSoup。

图表 4: run 函数的主要内容



def run():

url = "http://www.stats.gov.cn/tjsj/zxfb/index.html"
res = requests.get(url).content # 获取新闻页内容
bs = BeautifulSoup(res) # 网页代码解析

来源:中泰证券研究所

■ BeautifulSoup 自动将输入文档转换为 Unicode 编码,输出文档转换为 utf-8 编码,因此并不需要考虑编码方式。国家统计局官网上并没有一个指定的编码,因此在抓取数据的过程中 BeautifulSoup 已经够用。当然,当我们遇到有指定编码的网站时,比如猫眼网,BeautifulSoup 就无法自动识别,可以考虑使用 Xpath 和正则表达式。

图表 5: 提取关键文本的方法

```
for j in a:
    if "月份居民消费" in j.text:
    new_url = () + j.a.attrs['href'][]
    file_name = j.a.attrs['href'][]
```

来源:中泰证券研究所

- 另一个重要函数是 get_csv。我们通过这个函数解析获取 CSV 字段格式。
- 在 get_csv 函数中,我们首先要打开 csv 文件,并将爬取到的内容写入 文件。这里我们可以充分使用 Python 的简洁特征,定义格式为打开 csv 文件并调成 "w"写入文件。



图表 6: get_csv 函数的主要内容

```
def get_csv(name, url):
    with open(name + ".csv", "w") as f:
        res = requests.get(url)
        html = res.content
        bs = BeautifulSoup(html)
```

来源:中泰证券研究所

■ 重要的是爬取数据导入 Excel 表格时,要调成一致格式。在导入 Excel 时,可以先把内容 print 一下看格式。一般而言字符串之间会存在空格,可以使用 strip 去掉首位空格,用逗号代替字符串间空格间隔。

三、后续使用的相关问题

- 在爬取数据时,如果追求时效性,可以在服务器上设置爬取时间,这样 在数据公布出来的时间段里,爬虫会反复访问网页以获取相关内容并进 行高效爬取。
- Pytorch 宣布 Python2 即将于 2020 年 1 月之后停止更新维护。Python2 和 Python3 以上版本差异较大,建议安装 Python3 以上版本。在最新推出的 Python3.8 中,新推出的赋值表达式":="操作符使得编写代码更加高效快捷,变量不存在也可以直接在表达式中赋值,不必新定义一个变量然后才能使用。
- 此外,本专题介绍的是从单一网页中提取数据,在整个网站中搜索带有时间序列的同一字段的数据表,并形成 Excel 表格的方法,将在下一个专题中提到。
- 最后,需要声明的是,在明确提示该网页不允许爬取的情况下,尽量不要使用爬虫获取数据并用于商业用途。在利用爬虫获取数据的过程中,一定要做到合法合规。
- 可瓜坦二. 少矶从临利运经汉南西去自证从打接 爬上冲起由当众浬列タ

预览已结束,完整报告链接和二维码如下:

https://www.yunbaogao.cn/report/index/report?reportId=1 8445



