数据来源（论坛贴吧）

<http://www.tb58.net/>

<http://www.htys123.com/forum.php>

<http://www.zhenshua.com/>

http://www.tsz3.com/

<http://bbs.paidai.com>

http://bbs.tao66.com

<http://www.maijd.net>

<http://www.xingqiwang.net/>（鸟人网包含）

统一要求：

爬虫编写：采用scrapy1.3.3，python 2.7， 存储到mysql数据库

爬虫策略：如果网站有反爬虫，注意一下U-A，访问频次，IP地址、cookie等等；一般电商黑灰产网站都需要登录才能获取更多信息，所以爬虫要模拟登录。提前用邮箱小号注册一下。如果需要绑定淘宝小号的话，也可以用邮箱注册小号，手机验证就行。

爬取内容：

对于每个网站，

1)爬取网页的html源码(从主页开始采用breadth-first爬取，depth设为3)

2)爬取网站内全部网页的正文内容（即所有对用户可见的文本）

3）爬取网站的所有出站链接

4)对于每个不同类型的网站，还要对正文内容进行抽取。（推荐xpath）

不同网站抽取的要求如下：

1. 小号交易

（1）<http://www.money668.com/>

抽取 账号类型、数量、价格

2.空包空单服务

（2）[http://www.88dan.com/](http://www.88dan.com/?uid=393)

（3）<http://www.ttkongbao.com/>

（4）http://www.666zkb.com/

抽取发货地址、快递类型、单号（没有单号看看有没有数量）、价格

3.流量平台

（5）<http://txip8.com>

（6）http://www.ip114.cc/

（7）<http://www.liuliang51.com/>

抽取不同类型的流量服务（淘宝PC端访客流量等）、具体描述（收藏、加购、浏览等）、任务完成时间、价格等，下图是一个例子：



（8）<http://www.liuliangbao.cn/>（软件）

对于这个网站只要做1）2）3）

4.刷单平台

（9）<http://www.taoshuashua.net/>

（10）http://www.dagonglang.com/

（11）<http://www.milioo.com>

（12）http://www.taoshiye.net/

（13）<http://www.yestao.net/>

（14）<http://www.birdhelper.com/>

（15）http://s.diyishuadan.com/

（16）<http://www.qt1351.com>

（17）<http://www.1qishua.com/>

<http://www.nuomart.com.cn/> 刷单网站，有任务的

<http://www.5v186.com/> 淘宝京东阿里

<http://www.jddan.cn/products/> 京东

<http://www.jingdongshuadan168.com/>没有看到具体任务，可能要联系QQ

<http://www.jdyhb.com/> 没有看到具体任务，可能要联系QQ

抽取刷单任务数、任务佣金、商品名称链接、商品价格、刷单步骤、卖家店铺或账号（京东淘宝等）、对刷手的要求，刷手数量等,如下图：



大部分网站都需要登录才能爬到任务信息等，不一定每个网站都有这些信息，可以自己增加或减少。账号自己能用小号注册的话就用小号，不能的话再找我解决一下。

其他：

购物兼刷单网站

<http://www.51lingla.com/>

空包兼刷流量

<http://www.kongbao100.com/>

<http://www.kongbaojc168.com/>

这两个网站极其相似，几乎一模一样

刷单刷流量