网站来源：论坛贴吧的链接和文本

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

http://www.htys123.com

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

http://www.tsz3.com/

<http://bbs.paidai.com>

<http://bbs.tao66.com>

<http://www.maijd.net>（京东）

<http://www.xingqiwang.net/>（鸟人网包含）

<http://www.taobaots.com>

http://www.zuanke8.com

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. 爬取前100个网页的源码(从主页开始采用breadth-first爬取(depth priority设为1, DEPTH\_LIMIT设为3)

记录格式: sitename,pageurl, sourcecode

如鸟人网的一个网页记录：

[www.birdhelper.com，http://www.birdhelper.com/TaskHall/TaskHallIndex](http://www.birdhelper.com，http://www.birdhelper.com/TaskHall/TaskHallIndex)，源码

1. 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）

记录格式: sitename,pageurl,pagecontent

三、爬取网站的所有出站链接

记录格式：sitename, linkurl

如鸟人网www.birdhelper.com有一个出站链接<http://www.kongbaojc168.com/>

那记录为: www.birdhelper.com, <http://www.kongbaojc168.com/>

四、对于每个不同类型的网站，还要对正文内容进行抽取。

抽取的内容不限于淘宝，所有电商都要。对一个网站的所有板块都看一下，先把有权限查看的数据爬下来，有的网站是刷单、刷流量、卖小号、发空包等业务都有，那么这些信息要爬全。

不同网站大概的要求如下，大家可以自己调整，记录格式自定，每条记录添加一个爬取的时间。

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/> 自刷平台，通过操作软件刷单

（13）<http://www.yestao.net/>

（14）<http://www.birdhelper.com/> js写的

（15）<http://s.diyishuadan.com/> 可以看到任务大概信息，但是要接受任务看到具体商品信息需要交保证金

（16）<http://www.qt1351.com>

（17）<http://www.1qishua.com/> 非常严格需要实名认证视频认证绑定买号heihei2017 123456a

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。