采集软件定制需求20141225

目标网站：m.taobao.com

语言:Python 2.7.X

数据:CVS/exsl

运行环境：ubuntu sever X64 14.04

依赖包：如有依赖包，请注明版本

**需求分解：**

全局功能：

1. 可自定义搜索关键词
2. 可同时运行多个采集任务
3. 每个采集任务可配置有效时间段，重复执行频率，或在有效时间内，采集完成后多长时间内循环采集
4. 每次采集结果需比对已有数据，如已保存数据中一项发生改变，则重新开始采集，数据以最新为准
5. 不需要GUI界面，命令行即可，需打印程序运行结果信息，包含每条采集耗时，如报错，请注明原因。

**需采集内容：**

入口(只包含“淘宝店铺”类型结果)：http://shop.m.taobao.com/shop/shop\_search.htm?spm=0.0.0.0&sort=default&fx=0&lp=0&loc=&pds=taobao%23h%23shopsearch&type=all&searchfrom=&my=0&\_input\_charset=utf-8&base64=1&jf=0&q=ye3M5cjp&paytype=&isb=0&olu=

采集所有分页店铺，得到店铺shop\_ID.

根据shop\_ID,转到每个店铺首页，例：

http://shop.m.taobao.com/shop/shop\_index.htm?spm=0.0.0.0&shop\_id=35213196&sid=1cde69550b00fdd142c75879022fc803

在这个页面的源代码中可获取到userId

然后根据得到的userId访问店铺资料页面，例：

http://h5.m.taobao.com/weapp/view\_page.htm?page=shop/about&userId=54199406&weapp\_encode=1

需保存信息项如下：

|  |  |
| --- | --- |
| 项目 | 说明 |
| 搜索关键词 | 例：身体乳 |
| 入库时间 | 例：2015-03-03 15:55:02 |
| 更新时间 | 例：2015-03-03 18:55:02 |
| shop\_ID | 例：35213196 |
| userId | 例：54199406 |
| 店铺地址 | 例：http://shop35213196.taobao.com |
| 店铺名称 | 例：巴巴爸爸 Barpa 护肤 |
| 粉丝数 | 例：33.19万粉丝 |
| 全部宝贝数 | 例：722 |
| 上新数 | 例：44 |
| 好评率 | 例：99.75% |
| 所在地 | 例：广东广州（此处需将省市两级分开） |
| 开店时间 | 例：2007-12-23 11:17:39 |
| 掌柜名(旺旺名) | 例：kinn\_ning |
| 服务电话 | 例： 400-8048808 |

注：

以上项目，如有空值出现，请留空。

感恩，请报价！

另请考虑效率问题。