淘宝数据抓取需求规格说明书

# 流程概述

1. 登陆淘宝首页www.taobao.com，选择“淘宝服务 -> 更多>”；
2. 在淘宝更多服务中选择“淘宝排行榜”；



1. 进入淘宝排行榜后，选择商品的种类，目前我们需要抓取的种类有：

* 数码家电 -> 手机/配件 -> 热门手机；
* 数码家电 -> 电脑/配件 -> 笔记本；
* 母婴 -> 宝宝食品 -> 婴儿奶粉；
* 母婴 -> 宝宝用品 -> 纸尿裤；



1. 选择其中的一个商品种类如热门手机，点击“销售热门”，选择top10的热门手机；



1. 选择其中的一款热门手机，如“Sony Ericsson/索尼爱立信 LT26i“，点击进入后，需要抓取 page1-4的卖家的信息，针对每一个卖家需要抓取的信息有：

* 卖家名称（下图绿色标出）
* 抓取卖家是否有全球购、金牌卖家等等的标志（下图绿色标出）
* 卖家报价；
* 运费；
* 信用卡支付；
* 卖家地址；
* 最近成交笔数；
* 评价条数；
* 消费者保障；
* 假一赔三；
* 七天退换；
* Page rank；(在搜索结果中的页数(1,2,3或者4)以及在其所在页的排名位置)

注意：不抓取tmall的卖家；

1. 选择其中的一个卖家如“正港 Sony Ericsson/索尼爱立信 LT26i/LT26 质量保证+惊喜4重奏”，点击进入后，需要抓取的信息主要有以下方面：
2. 卖家提供的宝贝基本信息，主要是抓取以下方面的信息：

* 卖家报价最低价；
* 卖家报价最高价；
* 物流运费；
* 30天售出；
* 评价；
* 宝贝类型；
* 浏览次数；
* 支付类型；
* 服务；



1. 点击页面右侧的钻石标记，进入该卖家的信用等级详细页面，此页面基本所有的信息我们都需要抓取下来；



1. 页面下侧的评价详情，我们需要获取第一条评价和最后一条评价的时间；

时间的格式是 **（年.月.日）**



1. 页面下侧的成交记录，我们选择其中的一条成交记录如“tb5954577”，点击进入后，需要抓取**买家**昵称下方的地理位置信息，如果没有位置信息，做标0处理；

在点击进入买家信息页面前，需要记录：

* + 买家自己的信用级别（鼠标移动到名字后面就能显示）
  + 买家的评价得分
  + 拍下价格
  + 数量
  + 付款时间
  + 款式和型号

在点击进入买家信息页面后，需要记录：

* + 买家地址（省份，市，（如果还有市区））
  + 买家性别（在头像下有‘他的好友“或者是”她的好友“。如果没有这类信息，就标”0”）





# 尚未明确的地方

* 进入淘宝排行榜后，选择商品的种类，目前我们需要抓取的种类有：
* 数码家电 -> 手机/配件 -> 热门手机；
* 数码家电 -> 电脑/配件 -> 笔记本；
* 改为： [淘宝排行榜](http://top.taobao.com/index.php)->[化妆品](http://top.taobao.com/level2.php?cat=TR_HZP)->[美容护肤](http://top.taobao.com/level3.php?cat=TR_MRHF)->洁面
* 在“流程概述“步骤4中，当我们选择婴儿奶粉的销售热门，选择某个具体的奶粉点击进入后，和之前的热门手机显示页面不一样，这个页面没有列出商家，而是直接显示了某商家销售该奶粉的信息页面；改为化妆品就不会出现这样的问题。但在化妆品类目中有几样要注意的：
  1. 在商家列表中，请帮我们抓取他们是否有“全球购”或者“香港商家”等的标志。



* 1. 在宝贝页面上抓取化妆品的规格和容量。



* 在“流程概述“步骤5中，针对某个具体的商品，我们抓取page1-4中不含tmall的商家，意味着每个商品对应的商家数量可能不同？另外page rank的意思是指这个商家显示在第几页吗？显示该商家在搜索结果中的页数(1,2,3或者4)以及在其所在页的排名位置
* 在“流程概述“步骤6 c)中，关于用户评价的时间格式是采用淘宝页面上显示的格式2012.03.08？可以
* 我们从淘宝上抓取下来的数据保存在什么地方？是已经确定好的，还是由我们进行设计？ 最好存成excel的形式。

# 开发计划

根据技术，时间等因素，综合考虑，项目的开发时间大约为两个月。

|  |  |  |
| --- | --- | --- |
| 时间段 | 任务 | 备注 |
| 3.19 至3.25 | 确定项目需求 | 详细描述项目的需求 |
| 3.26 至4.16 | 设计 | 概要设计和详细设计 |
| 4.17 至4.30 | 编码 | 项目编码 |
| 5.1 至5.17 | 调试 | Debug |
| 5.18至5.19 | 交付 | 交付使用 |