**分布式爬虫**

**——设计思路**

**组长：赵淑晨**

**组员：钟凯华**

**赵晨阳**

**指导老师：任志考**

# 语言和工具

开发语言：Python2.7

开发工具：pycharm

存储：mongodb

缓存：redis

# 分布式系统结构设计

整个分布式系统采用“主人——工人”式设计思路，主服务器做为“主人”，主要提供“工作场所”，即消息队列，“主人”还要时刻与“工人”保持联系，在系统中主要通过socket通信来完成。“工人”分为三类：“抓取工人”（服务器1），“去重工人”（服务器2），“解析工人”（服务器3）。“抓取工人”负责从过滤url队列中取出url，然后进行抓取，将抓取后的内容放到html队列中，“去重工人”负责从url队列中取出url，利用去重算法进行过滤，重复的url丢弃，不重复的url放入过滤器和过滤url队列中，“解析工人”负责从html队列中取出html，进行分类，然后进行结构化解析，将有用的信息和url进行提取，信息存入数据库中，url放入url队列中。



# 各模块详细设计

## “主人”

“主人”时刻保持对“工人”的监听，有“工人”来“报道”，就为他建立一个通信连接，并提供“工作场所”给他。开始“工作”后，“主人”每隔一段时间便向“工人”发送消息，询问“工人”的工作状态，并对“工人”返回的消息进行接收，接收的信息进行处理，各类“工人”的工作状态进行记录，显示每类“工人”的工作数量。

## “工人”

“工人”要与“主人”进行联系，联系成功后告诉“主人”自己的身份，听从“主人”的安排，收到“主人”的命令后进行相应的操作，比如“开始工作”、“结束工作”等。

### “抓取工人”

“抓取工人”从特殊url队列、一级url队列、二级url队列中取出url，抓取json格式数据或html网页数据分别放入json或html队列中。特殊url队列中的是拼接的json请求，一级url队列中的是商品列表url或商品url，二级url队列中的是其他url，“抓取工人”先从特殊url队列中取url，进行抓取后放入json队列，再从一级url队列中取url，进行抓取后放入html队列，最后从二级url队列中取url，抓取后也放入html队列。



### “去重工人”

“去重工人”从url过滤队列中取出url，利用去重算法过滤后，对url的类别进行判断，商品列表url和商品url放入一级url队列，其他url放入二级url队列。

去重算法采用的是一个布隆过滤器，原理是先生成一个m位的bit数组，然后将数组的每一个bit位都初始化为0，url经过hash算法映射后生成k个值，然后将每个值对应位数组的位置为1，每次对url进行hash后，判断生成的k个值对应的位数组的位是否为1，如果都为1说明该url重复，不都为1说明该url不重复，bit数组不能继续储存后生成新的每位都为0的bit数组，用list对所有的位数组进行储存，判断时需要对list中所有的位数组进行遍历。



### “解析工人”

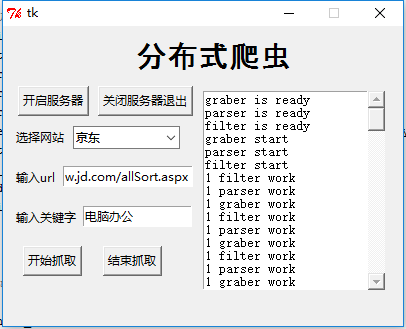
“解析工人”从json队列、html队列中取出json或html，进行解析，json解析后的数据直接存入数据库中，html解析后的数据存入数据库中，url放入url过滤队列中，拼接的json请求放入特殊url队列中。“解析工人”可以进行模块化提取，分为京东网页解析、苏宁网页解析、淘宝网页解析三类，网页解析原则为提取商品列表页的内容、商品全部分类页的商品分类url和其他页面的url。京东商品列表页提取商品名称和商品id，商品名称存储到数据库，商品id进行商品店铺、商品价格和商品评论的json请求拼接后将拼接的url放入特殊url队列中，苏宁商品列表页提取商品名称、商品id、店铺名称、店铺id，商品名称和店铺名称存到数据库，商品id和店铺id进行商品价格和商品评论的json请求拼接后放入特殊url队列中，淘宝商品列表页提取商品名称、商品id、店铺名称、店铺id、价格，商品名称、店铺名称、价格存入数据库，商品id和店铺id进行商品评论的json请求拼接后放入特殊url队列。



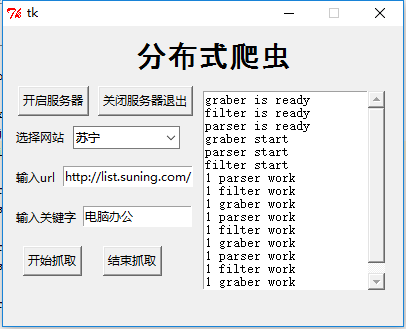
# 四、测试结果

## 主界面

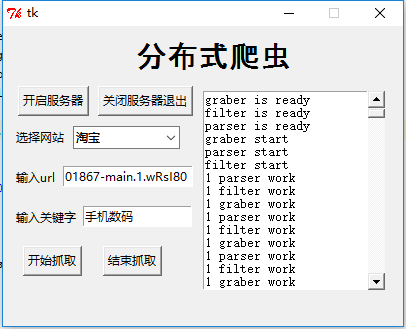
京东



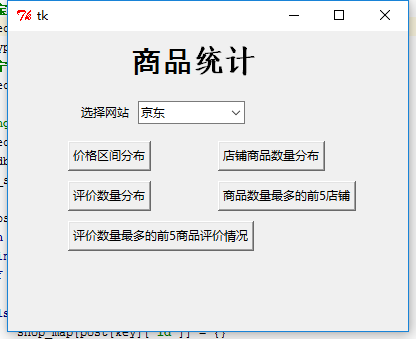
苏宁



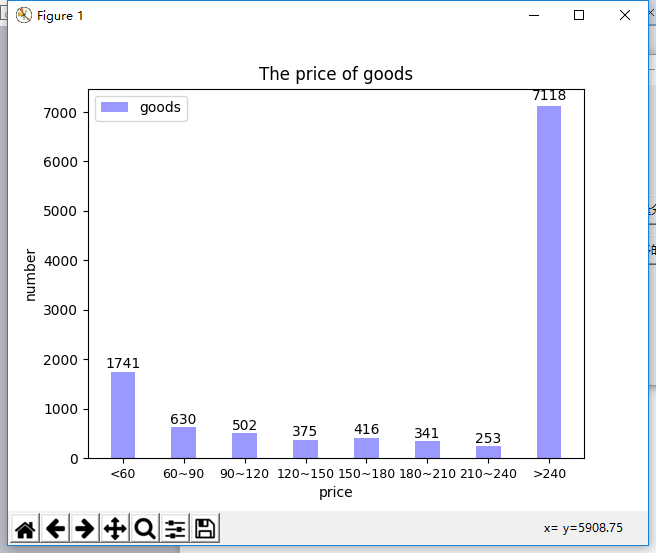
淘宝



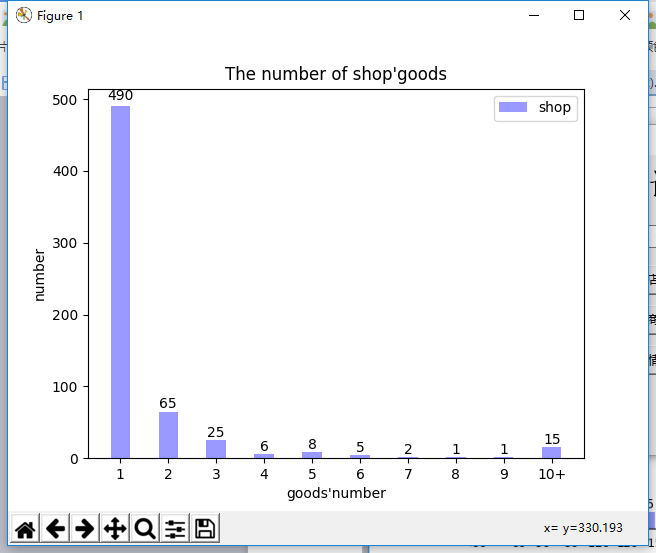
## 抓取数据的可视化

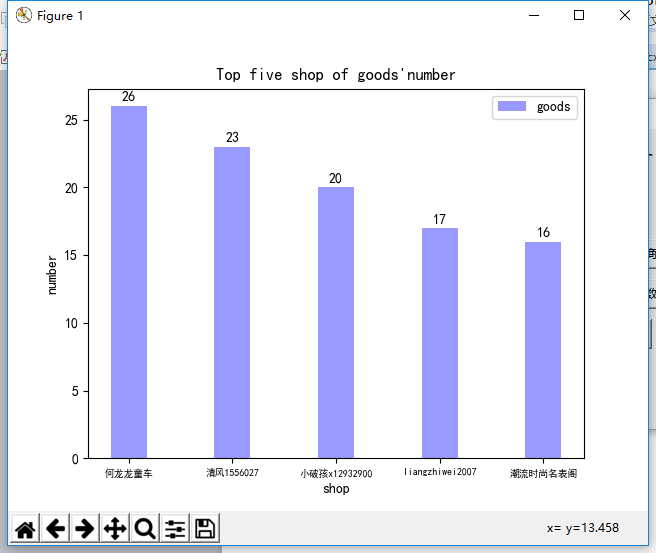


京东价格区间



淘宝商品数量





苏宁商品评论

