**分布式爬虫系统软硬件环境说明**



一、概述：

分布式爬虫系统主要用于爬取电商网站，如京东，淘宝，天猫，网易严选，对海量的分散的电商网站数据进行采集的系统，并且展示爬取的商品信息，店铺信息和商品价格，以及实时展示爬虫系统的性能。

二、主要结构及工作原理：

1.主要结构

分布式爬虫系统由通讯模块，爬虫模块，前端模块，数据库模块组成。

2.工作原理

用户在前端web界面输入待提取的电商网页，点击开始提取，并可以在页面上看到相应电商网站商品的信息和爬虫性能数据。

三、主要性能参数：

1. 输入URL参数：

<http://www.taobao.com,>

<http://www.jd.com,>

<http://www.tianmao.com,>

http://you.163.com；

2、编写语言：html，python；

3、工作环境：windows;

四、操作指南：

1.安装Python3模块

https://jingyan.baidu.com/article/e73e26c0bad76224acb6a766.html

2.安装phantomjs模块

https://segmentfault.com/a/1190000009020535

3.安装依赖

pip install -r requirements.txt

1. 安装Chrome/Firefox浏览器
2. 开启code/main\_server/server\_main.py还有code/main\_server/spider\_main.py 。
3. 开启浏览器后打开离线web页面：login.html 输入url后点击提交即可开始采集 。
4. 页面信息隔段时间自动刷新一次。
5. 中间会有多次弹框，属正常现象。
6. 使用中略有报错，属正常现象，不影响使用。