**分布式爬虫系统作品简介**

**与参赛感言**



1. **作品简介**

**1.1 作品名称**

分布式爬虫系统

**1.2 作品作者**

电子科技大学iMake小组，成员：郑毅，王骞，祝含颀；

指导教师：吴晓华教授。

**1.3 作品功能**

对电商网站如淘宝，天猫，京东，网易严选中海量商品数据的爬取和数据分析。实现反爬虫机制，网页自动结构化，爬取速度快而稳定。

**1.4 作品技术**

1.4.1反爬虫机制

1. 伪造请求头
2. 设置代理

1.4.2通讯机制

1. 浏览器和后台用websocket通讯，后台使用tornado接收数据
2. 服务器之间用socket通讯（含阻塞型和非阻塞型），使用了socketserver，可以每秒钟处理上千次请求。

1.4.3爬虫模块

1. 用phantomjs模拟浏览器采集动态网页的内容。
2. 用requests采集静态网页内容
3. 用面向对象方法封装一系列操作

1.4.4 UI

1. Echarts动态实时刷新图表
2. 基于websocket接收数据
3. 使用js脚本动态刷新数据

1.4.5 分布式技术

1. redis分布式数据库
2. 分布式去重
3. 监控爬虫子节点的内存占用和cpu使用率
4. 主--从任务分配

**1.5 作品创新和实现的功能**

1.5.1 创新点

（1）自己研发的后台展示系统。

（2）自己研发爬虫模块。

（3）自制爬虫框架。

（4）自制通讯模块。

（5）实现了分布式的节点调度和分布式的数据存储，分节点监听。

1.5.2 实现的功能

1. 爬取数据页面展示
2. 反爬虫机制
3. 爬取速度快
4. 网页自动结构化，URL去重调度。
5. **参赛感言**

队员一:

一开始我们对软件比赛完全没有概念，我们小组也是在一门专业课上组建的，老师鼓励我们自主寻找课题，后来组内的一个同学提议“中国软件杯”中的课题，这才开始逐渐了解比赛的各个方面。所以我们主要还是抱着学习的心态来参加比赛的，希望更好地加强自己的实践能力和团队合作能力。

队员二：

准备比赛过程中我最大的感受，就是只有亲自动手，才能体会到理论是如何指导实践的。平时所学看似简单的理论，实践起来并不是那么轻松，一个小小的代码都会导致整个系统的崩溃，更何况还有很多知识都是平时未曾涉及的。

队员三：

我觉得在比赛过程中从其他参赛队伍中学到了很多。无论获奖与否，各个队伍对同一问题的不同解决方式让我大开眼界。当然最关键的还是指导老师从始至终的鼓励，最大的收获还是培养了团队合作精神。

此次比赛，我们队最看重的是动手实践，对于这样开源性的赛题，有很多开源的框架，但是我们并不愿意去用，因为有很多技术并不是止步在现有的技术，我们需要自己去探索，自己去发现现有技术的不足，而不是做一些没有用的外包技术，简单的重叠代码。