

**实 习 周 报**

**作者：林晓丹**

**日期：2017.07.29**

**广州探迹科技有限公司 版权所有**

1 本周工作

|  |  |
| --- | --- |
| 任务项 | 完成进度 |
| KSW\_0001：慧聪网关键词爬取 | 100% |
| KSW\_0003：中国交易制造网关键词爬取 | 100% |
| KSW\_0005：中国供应商关键词爬取 | 100% |
| ZC\_0002：爬虫\_搜狗企业微信公众号信息 | 100% |

# 2 学习计划

|  |  |
| --- | --- |
| 计划项 | 进度/计划时间 |
| python | 0%/2周 |

# 3 心得体会

本周完成了三个快塑网的关键词爬取企业信息的爬虫，其中在慧聪网中耗时比较长，主要在查找其中优质商家的标识是根据什么做的，之后发现在一开始请求的页面中只有该页面显示的部分产品信息，然后在另外的请求中发现了部分产品的信息，并且和优质商家中的内容相符合，翻页之后除了主页面外会请求前面相同的链接，并显示前面优质商家内容的部分在前一页没有显示的内容，根据这个可以判断到优质商家的产品更加像是打广告的内容，另外添加请求就可以了。

之后主要精力放在搜狗企业微信公众号的爬虫上，整个请求过程比较简单，但是搜狗那边应该是有一批黑名单，所以在使用不稳定代理时基本会遇到验证码的问题，这里要感谢大佬支撑破解验证码，验证码的通过率目前为1/3，但是在更改了cookies和headers之后还没有进行批量的测试， 具体要看在服务器上面的效果。