InplusLab－爬虫任务

爬虫任务一（建议完成周期1周时间）

一.背景知识

在长时间的互联网快速发展的基础下，随着大数据挖掘技术的兴起。爬虫技术作为在最大数据载体－互联网中获取数据的工具也在最近几年非常火热。

目前学术界和工业界对爬虫技术都有比较大的需求。当然，现在也有多种多样的爬虫技术。选择合适的爬虫技术稳定并高效应用于现实面临的问题是爬虫技术成功实施的关键。

二.问题

1. 希望对已有的爬虫技术做一个技术综述［简要书写］。
2. 静态页面爬取［**P2P网络贷款黑名单,逾期名单**］：
   http://www.dailianmeng.com/p2pblacklist/index.html
   爬取信息规整化输出（类似数据库表的形势）。

上交资料：技术源码＋执行效果PPT＋技术综述pdf＋数据信息

爬虫任务二（建议完成周期2周时间）

一.背景知识

在长时间的互联网快速发展的基础下，随着大数据挖掘技术的兴起。金融P2P业务火爆发展。但是最近一年在面临转变期。如何利用爬虫技术做好用户风控控制也是面临的巨大挑战。

目前社会各界都有各类各样的黑名单数据。其中，法院公布的失信人员名单及法人是风控把控的重点对象。本次调研任务需要完成对法院执信名单的爬取。

二.问题

1. 动态页面爬去［**P2P中国执行信息网**］法院执行黑名单数据爬取
   http://shixin.court.gov.cn/
   爬取信息规整化输出（类似数据库表的形势）。

上交资料：技术源码＋执行效果PPT＋数据信息