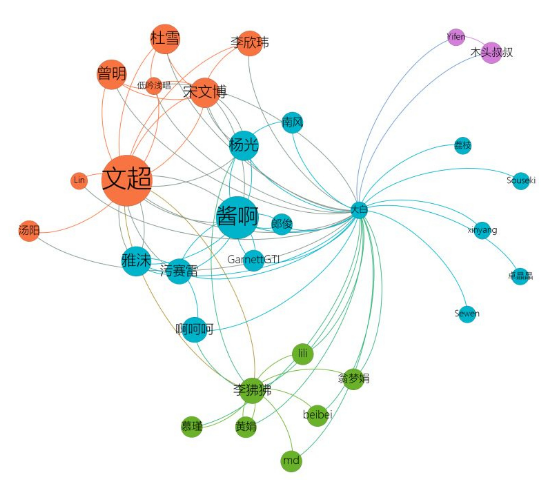
demo博客：<https://blog.csdn.net/bone_ace/article/details/50771839>

demo源码：<https://github.com/LiuXingMing/QQSpider>

功能：对某个qq空间的访客、说说、日志、个人四种信息进行爬取，后期可以通过Gephi生成社交热力图，如



技术栈：python2.7

数据库：mongodb

使用 requests 模块抓取，部分使用 BeautifulSoup 解析。

同一节点下的多个同级访客的多线程爬取使用 multiprocessing.dummy 。

登录qq空间需要抓取 Cookie， 使用 selenium 和 PhantomJS 。

判重使用 BitVector 。

采用pymongo访问mongodb数据库

代码构架

{

“代码结构”:

Controller

       spide\_controller

Service:”爬虫功能单元（下载页面、解析页面等方法）”

       blog\_spider

       friend\_spider

       information\_spider

       mood\_spider

Util

       public\_methods

Config

       init\_messages（爬虫的初始化配置）

       myQQ.txt:”存放爬虫初始爬取的源QQ”

log

       ghostdriver.log

       QQFailSpided.txt:”爬取失败”

       QQForSpider.txt:”待爬”

       QQHadSpided.txt:”已爬”

Plugin

       Phantomjs.exe:” 抓取 Cookie”

}

代码运行流程

1.reload(sys)设置程序统一编码格式

2. init\_messages读取本地的爬虫信息，并对爬虫进行参数初始化

3. spide\_controller生成多线程爬虫

4. spide\_controller运行爬虫

未解决启动警告

UserWarning: Selenium support for PhantomJS has been deprecated, please use headless versions of Chrome or Firefox instead

原因

新版本的Selenium不再支持PhantomJS了

解决方案

将Selenuim版本降级，通过pip show selenuim显示，默认安装版本为3.8.1。 将其卸载pip uninstall selenuim，重新安装并指定版本号pip install selenuim=2.48.0。

感谢[LiuXingMing](https://github.com/LiuXingMing)大神提供的学习资料