Python抓取动态网页内容

学习Python web编程的时候，就涉及一个Python的requests模块，这个模块可以轻松读取页面上面的静态信息，但随着时代的发展，越来越多的网页中更多的使用javascript、jQuery、PHP等语言动态生成页面信息.因此用urllib再去抓取页面HTML就不足以达到我们想要的效果；有一个思路最为简单可以实现动态解析页面信息，urllib不可以解析动态信息，但是浏览器可以．在浏览器上展现处理的信息其实是处理好的HTML文档，这为我们抓取动态页面信息提供了很好的思路，在Python中有一个很有名的图形库——PyQt，PyQt中有个QtWebkit工具，可以实现模拟浏览器．所以我们可以通过QtWebKit把页面中的信息读取加载到HTML文档中，再解析HTML文档．从HTML文档中提取我们想用得信息.

```
Step1.安装QT
url: http1://nchc.dl.sourceforge.net/project/pyqt/PyQt4/PyQt-4.9.5/PyQt-x11-gpl-4.9.5.tar.gz
./configure
make (时间很久)
make install

以上三个步骤都需要较长时间进行：QT安装的目录在/usr/local/Trolltech/Qt-4.8.x，需要将qmake所在路径添加到Linux搜索路径所定义的路径中.
此时在终端直接输入qmake -v很有可能找不到命令，那么需要把qmake添加到path中。
ln -s /usr/local/Trolltech/Qt-4.8.x/bin/qmake /usr/local/bin

Step2.安装PyQt SIP
下载PyQt
url: http1://nchc.dl.sourceforge.net/project/pyqt/sip/sip-4.14.1/sip-4.14.1.tar.gz
python configure.py
make
sudo make install

Step3.安装PYQT
url:http1://nchc.dl.sourceforge.net/project/pyqt/PyQt4/PyQt-4.9.5/PyQt-x11-gpl-4.9.5.tar.gz
python configure.py
make
sudo make install

Step4.安装Spynner
spynner是一个QtWebKit的客户端，它可以模拟浏览器，完成加载页面、引发事件、填写表单等操作.
pip install spynner -i http://mirrors.aliyun.com/pypi/simple/ --trusted-host mirrors.aliyun.com
```

示例代码：

In [ ]:
```
import spynner
browser = spynner.Browser()
browser.hide()
browser.load("http://www.toutiao.com")
print(browser.html.encode("utf-8"))
#browser 类中有一个成员是html，是页面进过处理后的源码的字符串.
browser.close()
```