**《LQAB爬虫微服务 1.0 技术手册与说明文档》**

**一、技术特点:**

1.局域内网的微服。遵循简单可靠的http协议，在内网网络节点间架起微服务。主程序和爬虫引擎无需再部署在同一台物理主机，可以做到跨操作系统和初级的分布式负载均衡。

2.当作为远端的云爬虫引擎使用时，除了上述优点，提供了成熟高效的对称加密功能，保证了外网通讯中爬取结果的安全与私密性。因为做到了跨网段部署。可以承担一部分代理云主机的角色，例如:可以部署在国外，获得国外IP和网络带宽，对一些特殊网站可以很好的采集。

3.无论部署在内网还是外网，都支持基于关键词的语义爬取。语义爬取内容主要是输入关键词从搜索引擎采集， 因此可以用于传统的元搜索，搜索小偷和智能问答机器人等高端数据采集需求。

**二、环境搭建**

1.python 3.5及以上版本 需要安装 flask,tornado,pymysql及若干小型专业库。

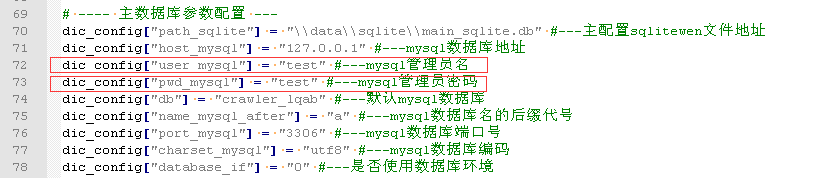
2.数据库环境：

（1）执行数据库脚本

data/sql/basedata\_00.sql #建立标准示例数据库及基本表

data/sql/crawler\_key.sql #用于更新默认对称加密的密钥

（2）修改根目录下config.py文件中相关参数为本机数据库环境参数



建议：单独建立一个 用户名test 密码 test 权限比较大的数据库账户，用于本地调试，当正式部署时，一次性修改config.py配置文件。

**三、调用方式：**

(1)flask框架

运行 python web\_flask.py

(2)tornado框架

运行 python web\_tornado.py

默认支持自动识别是网址（普通爬虫）还是关键词（语义模式）

(3)打开浏览器(默认)

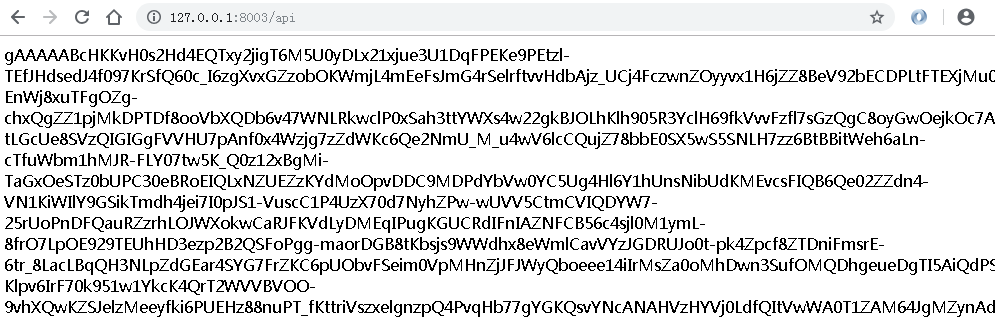
<<tornado框架下>>

本地调用测试：

web：http://127.0.0.1:8003



获得加密的网页内容



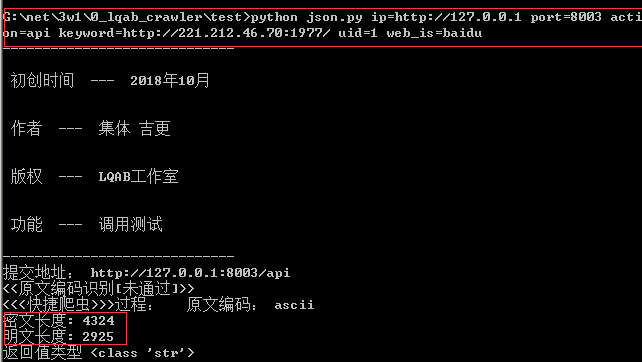
获得加密的语义内容：



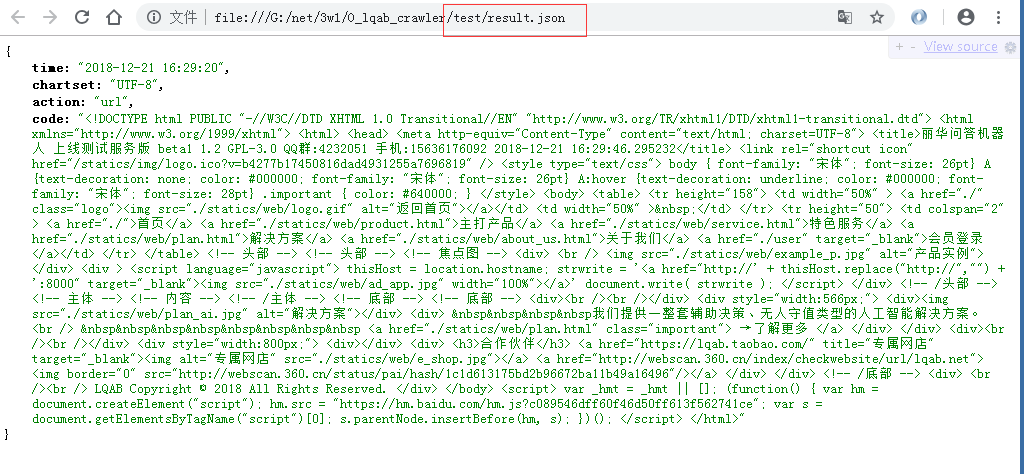


test目录下：

python json.py ip=http://127.0.0.1 port=8003 action=api keyword=http://221.212.46.70:1977/ uid=1 web\_is=baidu



生成result.json的结果文件 本地的生产调用可参考此模块

（根目录下生成test\_result.txt文件查看,可以设置crawler\_key表某id的密钥值为"n/a",则返回结果不会被加密）

<<flask 框架下>>

本地调用测试：

web：http://127.0.0.1:8004





API爬取测试：http://127.0.0.1:8004/statics/test.html





**特别说明：**

1） 对于爬虫类微服务最重要的问题就是编码问题，一定要确保运行环境的编码固定，建议统一为utf-8编码，无论是windows，lunix或mac。

比如我们在Windows上进行开发时，Python工程及代码文件都使用的是默认的GBK编码，也就是说Python代码文件是被转换成GBK格式的字节码保存到磁盘中的。

了兼容Python2和Python3，在代码头部声明字符编码：-\*- coding:utf-8 -\*-

同时最好做一次检测：

>>> # Python3

>>> import sys

>>> sys.getdefaultencoding()

'utf-8'

2）友情资助与联系方式：

如果您觉得效果不错，请扫码资助我们的开源项目《LQAB爬虫微服务》



QQ群：88786426



联系电话：15636176092（微信同号，早9:00 -- 晚16:00）

客服微信二维码：



演示网址：http://lqab.net:8003 (api)

http://lqab.net:8004 (web)

出品：LQAB工作室（http://www.lqab.net）  
主创：王滕辉 吉更