

匿名墙帖子爬取

（潘再峰、游伟 中国人民大学）

在本教程中，将演示如何使用提供的python脚本爬取匿名墙上的帖子内容。

本实验一共提供了两个python脚本，分别为：

- create_tbl.py: 用于创建数据库中的表
- crawler.py: 用于帖子的爬取

环境配置

在Kali虚拟机中，本实验需要的环境依赖全部自带，因此不需要额外安装与配置。

如果想手动进行配置，可以参考以下步骤，在终端里完成配置：

1. 更新apt软件列表

```
sudo apt-get update
```

2. 下载python3

```
sudo apt-get install python3
```

3. 下载pip（python的包管理工具）

```
sudo apt-get install python3-pip
```

4. 通过pip下载python的requests包

```
pip install requests
```

5. 下载sqlite的可视化工具sqlitebrowser

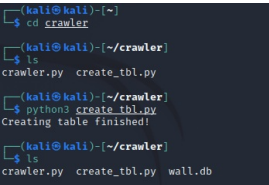
```
sudo apt-get install sqlitebrowser
```

创建数据库中的表

在命令行里通过cd进入代码所在的目录，通过以下命令运行脚本：

```
python3 create_tbl.py
```

如图所示，运行脚本后看到 Creating table finished! 的信息说明创建成功。此时可以发现目录下多了一个wall.db的数据库文件。注意此脚本只需要运行一次。

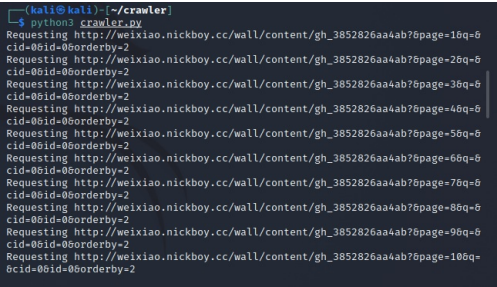


爬取帖子

在命令行里通过运行以下命令运行脚本来爬取帖子：

```
python3 crawler.py
```

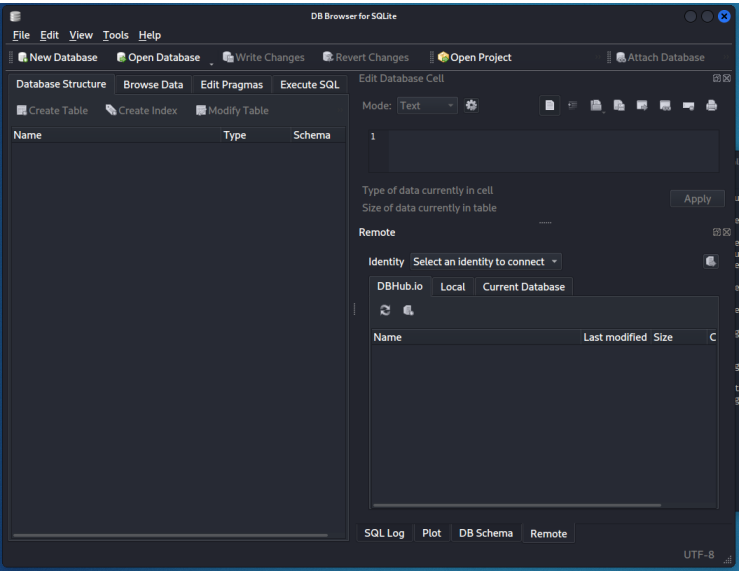
看到图中输出说明正在爬取url里的内容：



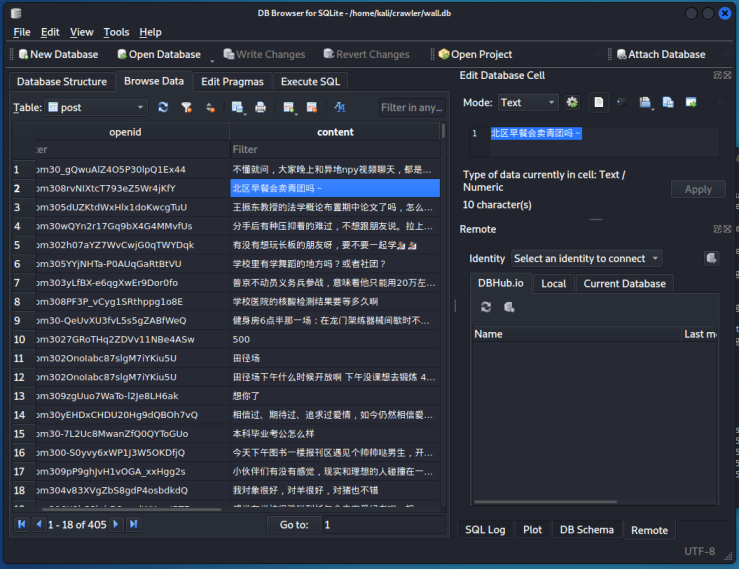
第一次爬取的时间会比较久，请耐心等待。目前的脚本没有断点继续的功能，因此如果中断运行后再运行，将会重新爬取所有内容。

查看爬取的帖子内容

爬取脚本运行完后，帖子的内容都被加入到了wall.db的数据库当中。可以使用sqlitebrowser可视化工具来进行内容的查看。在命令行里输入sqlitebrowser即可打开可视化界面，如下图所示。



点击 Open Database 选项，选择 wall.db 打开数据库。然后选择 Browse Data 标签，在 Table 框里下拉选择 post 表，即可看到表中记录的帖子的内容。如下图所示。



进阶实验