



Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

# Stata 、 cURL 交互与网络爬虫

## 小花经济学术

彭文威

MPhil in Social Science  
Division of Social Science, HKUST

温州

2017 年 8 月 19-20 日



[www.uone-tech.cn](http://www.uone-tech.cn)

友万科技



# 目录

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

## ① 从爬虫说起

## ② cURL 是什么?

## ③ cURL 爬虫

cURL 与微博 API

实战之环保部空气质量日数据

实战之中国土地网

## ④ cURL has limits



[www.uone-tech.cn](http://www.uone-tech.cn)

友万科技



# Table of Contents

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

## ① 从爬虫说起

## ② cURL 是什么?

## ③ cURL 爬虫

cURL 与微博 API

实战之环保部空气质量日数据

实战之中国土地网

## ④ cURL has limits



友万科技

[www.uone-tech.cn](http://www.uone-tech.cn)



# 网络爬虫

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

爬虫，是网络时代的产物，互联网上数据量呈现指数级增长。大量有用的数据可能是非结构化的，如果用人工收集会耗费大量的人力和时间。而程序工具的出现，为实现这种工作量巨大、内容繁琐重复的工作提供了解决办法。最初始的也是最常见的爬虫就是搜索引擎（google、baidu、bing...）

爬虫一般可以分为

- API 等规则化数据收集：这种爬虫非常简单，利用网站提供的 API 接口，可返回干净、结构化的数据。缺点在于，你只能获得他想要你获得的数据，并且出于保护自身利益和服务器成本的原因，API 往往有返回数据量限制。
- 非结构化数据收集（占比最大，难度高）。



www.uone-tech.cn

友万科技



# Http 协议

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

超文本传输协议（英文：HyperText Transfer Protocol，缩写：HTTP）是一种用于分布式、协作式和超媒体信息系统的\*\*应用层协议\*\*。HTTP 是万维网的数据通信的基础。

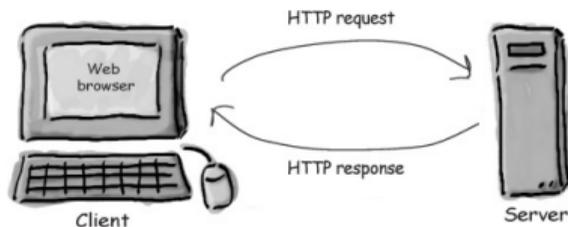


Figure: HTTP 协议

(Photo downloaded from: <http://www.ruanyifeng.com/blog/2016/08/http.html>)

HTTP 是一个客户端终端（用户）和服务器端（网站）请求和应答的标准（TCP）。通过使用网页浏览器、网络爬虫或者其它的工具，客户端发起一个 HTTP 请求到服务器上指定端口（默认端口为 80）。我们称这个客户端为用户代理程序（user agent）。应答的服务器上存储着一些资源，比如 HTML 文件和图像。我们称这个应答服务器为源服务器（origin server）。



# HTTP 请求方法

HTTP/1.1 协议中共定义了八种方法（也叫“动作”）来以不同方式操作指定的资源：

- **GET**：向指定的资源发出“显示”请求。使用 **GET** 方法应该只用在读取数据，而不应当被用于产生“副作用”的操作中，例如在 **Web Application** 中。其中一个原因是 **GET** 可能会被网络爬虫等随意访问。
- **POST**：向指定资源提交数据，请求服务器进行处理（例如提交表单或者上传文件）。数据被包含在请求本文中。这个请求可能会创建新的资源或修改现有资源，或二者皆有。
- **HEAD**：与 **GET** 方法一样，都是向服务器发出指定资源的请求。只不过服务器将不传回资源的本文部分。它的好处在于，使用这个方法可以在不必传输全部内容的前提下，就可以获取其中“关于该资源的信息”（元信息或称元数据）
- **PUT**：向指定资源位置上传其最新内容。
- **DELETE**：请求服务器删除 **Request-URI** 所标识的资源
- **TRACE**：回显服务器收到的请求，主要用于测试或诊断。
- .....

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么？

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits



- GET

- 当使用 GET 方法请求返回的数据是 Document/TEXT 时, Stata 的 copy 命令可以轻松解决。

- POST

- 当请求方法为 POST 时, copy 无能为力, 一个较为简单的解决办法就是 Stata 与 cURL 交互。





# Table of Contents

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

① 从爬虫说起

② cURL 是什么?

③ cURL 爬虫

cURL 与微博 API

实战之环保部空气质量日数据

实战之中国土地网

④ cURL has limits



[www.uone-tech.cn](http://www.uone-tech.cn)

友万科技



- cURL 是一个利用 URL 语法在命令行下工作的文件传输工具，1997 年首次发行。它支持文件上传和下载，所以是综合传输工具，但按传统，习惯称 cURL 为下载工具。cURL 还包含了用于程序开发的 libcurl。
- cURL 支持的通信协议有 FTP、FTPS、HTTP、HTTPS、TFTP、SFTP、Gopher、SCP、Telnet、DICT、FILE、LDAP、LDAPS、IMAP、POP3、SMTP 和 RTSP。





# cURL 语法

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

- 下载单个文件，默认将输出打印到标准输出中 (STDOUT) 中：curl www.stata.com

```

Microsoft Windows [版本 6.1.7601]
版权所有 (c) 2009 Microsoft Corporation。保留所有权利。

C:\Users\HP>curl www.stata.com
<!DOCTYPE html>
<html>
<head>
<title>
  Data Analysis and Statistical Software | Stata
</title>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
<meta name="viewport" content="width=device-width, initial-scale=1" />
<meta http-equiv="X-UA-Compatible" content="IE=11; IE=10; IE=9; IE=8; IE=EDGE" />
<meta name="msvalidate.01" content="789287E5ED5EC4153774BAD86B603235" />
<meta property="fb:page_id" content="151026434935843" />
<meta name="description" content="Data Analysis and Statistical Software for Prof
<meta name="keywords" content="statistics, statistical, software, statistical sof
<meta name="verify-v1" content="DsHleKDzVKG5hX6UJQIB+0BphiZyFzrG35ZQ026dNtus=" />
<meta name="google-site-verification" content="RMpgeXKA9weiGQd8pq_KPJm0IikGIwChd
  
```

- curl www.stata.com -o stata.txt

```

c:\Users\HP\Desktop>curl www.stata.com -o stata.txt
  % Total    % Received % Xferd  Average Speed   Time    Time     Current
                                 Dload  Upload   Total   Spent    Left   Speed
 100 23489    0 23489    0    0  33460    0 --:--:-- --:--:-- --:--:-- 33460
  
```

- 了解更多：curl -help

```

C:\Users\HP>curl --help
Usage: curl [options...] <url>
  
```





# Table of Contents

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

① 从爬虫说起

② cURL 是什么?

③ cURL 爬虫

cURL 与微博 API

实战之环保部空气质量日数据

实战之中国土地网

④ cURL has limits



[www.uone-tech.cn](http://www.uone-tech.cn)

友万科技



# cURL 与微博 API

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

为何以微博为例？API 端口，参数尽知。而在实战的爬虫中，参数获取需要分析、经验与琢磨。

原理：

- 使用新浪微博接口 update. 授权方式:Basic Authentication
- 如果你真的想测试下这个效果，那么你需要去新浪开放平台 (<http://open.weibo.com/>) 创建一个应用. 并取得 APP Key (我们需要一个 APP key 作为传送的参数)。你创建的应用完全不必去审核 (未审核的应用允许最多 15 个测试账号访问接口)
- 使用的接口地址：<https://api.weibo.com/2/statuses/update.json>



www.uone-tech.cn

友万科技



# 微博 API 文档查看 request method

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

open.weibo.com

微博·开放平台

微连接

微服务

文档

支持

推广

我的应用

登录



## API接口

微博开放平台开放了包括微博、评论、用户及关系在内的二十余类接口，通过Oauth2.0用户授权后即可在任意开发环境下使用。丰富齐全的功能，可以满足各种类型的产品需求。

立即查看

## 微博组件

微博开放平台封装了可直接部署在任意网站上的微



友万科技



原创



转发

www.uone-tech.cn



# 微博 API 文档查看 request method

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

微博·开放平台

微连接

微服务

文档

支持

推广

我的应用

登录

## 开发文档

open.weibo.com/wiki

跳转到: [导航](#), [搜索](#)

### statuses/update

发布一条新微博

### URL

<https://api.weibo.com/2/statuses/update.json>

### 支持格式

JSON

### HTTP请求方式

POST

首页

平台公告

平台概述

政策规范

视频直播

头条文章

微服务

轻应用

商业接口

评价此API



友万科技

www.uone-tech.cn



# 微博 API 文档查看 POST 参数

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

open.weibo.com/wiki/2/statuses/update

微博·开放平台

微连接

微服务

文档

支持

推广

我的应用



Wenwei\_P...

微博登录

移动应用

网站接入

微博API

资源下载

联系我们

常见问题

## 访问授权限制

访问级别：普通接口

频次限制：是

关于频次限制，参见 [接口访问权限说明](#)

## 请求参数

	必选	类型及范围	说明
<b>access_token</b>	true	string	采用OAuth授权方式为必填参数，OAuth授权后获得。
<b>status</b>	true	string	要发布的微博文本内容，必须做URLencode，内容不超过140个汉字。
<b>visible</b>	false	int	微博的可见性，0：所有人能看，1：仅自己可见，2：密友可见，3：指定分组可见，默认为0。
<b>list_id</b>	false	string	微博的保护投递指定分组ID，只有当visible参数为3时生效且必选。
<b>lat</b>	false	float	纬度，有效范围：-90.0到+90.0，+表示北纬，默认为0.0。
<b>long</b>	false	float	经度，有效范围：-180.0到+180.0，+表示东经，默认为0.0。
<b>annotations</b>	false	string	元数据，主要是为了方便第三方应用记录一些适合于自己使用的信息，每条微博可以包含一个或者多个元数据，必须以json字符串的形式提交，字符串长度不超过512个字符，具体内容可以自定义。
<b>rip</b>	false	string	开发者上报的操作用户真实IP，形如：211.156.0.1

## 注意事项



www.uone-tech.cn

友万科技

评价此API



# stata 透过 cURL 发微博

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

- 使用的接口参数: access\_token : 授权登录获得的参数; status: 发布的微博内容。(改版之前参数为: source = APP KEY;status= 发布的微博内容)
- 使用的 curl 的参数:-u 发送登陆请求格式是 -u username:password ;-d 发送 post 数据
- curl -u 微博账号: 微博密码 -d “source= 应用的 APP key&status=Test” https://api.weibo.com/2/statuses/update.json
- Stata : !curl -u 微博账号: 微博密码 -d “source= 应用的 APP key&status=Test ” https://api.weibo.com/2/statuses/update.json



www.uone-tech.cn

友万科技



# stata 透过 cURL 转发微博

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

- 参数: source:APP key; id: 你想转发的微博 ID; status : 转发评论语。
- Stata code: `!curl -u 18813795596:123456 -d "source=3186431957&id=3975371793325993&status= 期待各位优秀的大三学生申请我们的首届夏令营!" https://api.weibo.com/2/statuses/repost.json`



Wenwei\_PENG 🏠

1分钟前 来自 未通过审核应用

期待各位优秀的大三学生申请我们的首届夏令营！

@暨南大学经济与社会研究院 ✓

发布了头条文章：《暑期夏令营，送你两个大写加粗的“学术彩蛋”》学霸与学神的差距在哪里？

IESR暑期夏令营。1Russell Cooper教授送上三天免费课程。2诺贝尔经济学奖得主Heckman

等顶级学者，邀请首届硕士生参会！还不来戳报名：[www.uone-tech.cn](http://www.uone-tech.cn)

个大写加粗的“学术彩蛋”

友万科技



Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

# 实战之环保部空气质量日数据



友万科技

[www.uone-tech.cn](http://www.uone-tech.cn)



# 实战之环保部空气质量日数据

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量日数据

实战之中国土地网

cURL has limits



中华人民共和国环境保护部 数据中心

Ministry of Environmental Protection of the People's Republic of China

快速搜索

搜索你需要的



首页



政务信息



环境质量



污染防治



友万科技



环境影响评价



环保法律法规



自然生态

www.uone-tech.cn



# 空气质量日数据

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量日数据

实战之中国土地网

cURL has limits

datacenter.mep.gov.cn/index

## 全国城市空气质量日报

2017年08月17日,共367个城市

[更多...](#)

术语和定义: 空气质量指数 air quality index(AQI), 定量描述空气质量状况的无量纲指数。

0-50 优 级别: 一级

城市	AQI	级别	首要污染物
北京市	90	良	臭氧8小时
天津市	118	轻度污染	臭氧8小时
石家庄市	80	良	PM2.5
唐山市	107	轻度污染	臭氧8小时
秦皇岛市	65	良	臭氧8小时
邯郸市	78	良	PM2.5
邢台市	80	良	PM2.5
保定市	121	轻度污染	臭氧8小时
承德市	47	优	
沧州市	122	轻度污染	臭氧8小时
廊坊市	95	良	臭氧8小时



www.uone-tech.cn

友万科技



# 寻找 POST 参数

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

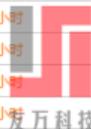
datacenter.mep.gov.cn:8099/ths-report/report/list.action?xmlName=146

城市  开始时间

截止时间

序号	城市	AQI指数	首要污染物	日期	空气质量级别
1	北京市	90	臭氧8小时	2017-08-17	良
2	天津市	118	臭氧8小时	2017-08-17	轻度污染
3	石家庄市	80	PM2.5	2017-08-17	良
4	唐山市	107	臭氧8小时	2017-08-17	轻度污染
5	秦皇岛市	65	臭氧8小时	2017-08-17	良
6	邯郸市	78	PM2.5	2017-08-17	良
7	邢台市	80	PM2.5	2017-08-17	良
8	保定市	121	臭氧8小时	2017-08-17	轻度污染
9	承德市	47		2017-08-17	优
10	沧州市	122	臭氧8小时	2017-08-17	轻度污染
11	廊坊市	95	臭氧8小时	2017-08-17	良
12	衡水市	68	臭氧8小时	2017-08-17	良
13	张家口市	85	臭氧8小时	2017-08-17	良
14	太原市	117	臭氧8小时	2017-08-17	轻度污染
15	大同市	95	臭氧8小时	2017-08-17	良
16	阳泉市	87	PM2.5	2017-08-17	良
17	长治市	113	臭氧8小时	2017-08-17	轻度污染
18	晋城市	136	臭氧8小时	2017-08-17	轻度污染
19	朔州市	102	臭氧8小时	2017-08-17	轻度污染
20	晋中市	111	臭氧8小时	2017-08-17	轻度污染

www.uone-tech.cn





# 寻找 POST 参数

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

datacenter.mep.gov.cn:8099/thz-report/reportlist.action

城市  开始时间

截止时间

城市	AQI指数	首要污染物	日期	空气质量级别
温州市	107	臭氧8小时	2017-08-17	轻度污染
温州市	35		2017-08-01	优
温州市	70	臭氧8小时	2017-08-15	良
温州市	98	臭氧8小时	2017-08-14	良
温州市	92	臭氧8小时	2017-08-13	良
温州市	94	臭氧8小时	2017-08-12	良

Elements Network

Filter   Regex  Hide data URLs

All XHR JS CSS Img Media Font Doc WS Manife

2000 ms 4000 ms 6000 ms

Name x Headers Preview Response Cookies >>

re... M": "30", "ID": "56F6ED50EC4FEDB9E05000003EE", "AQI": "102", "CITY": "鄂尔多斯市", "STATUS": "轻度污染", "MAIN\_POLLUTANT": "臭氧8小时", "CITYCODE": "150600", "CDATE": "2017-08-17"}]

isdesignpatterns: false

CITY: 温州市

V\_DATE: 2017-08-01

DATE: 2017-08-17

33 req...

www.uone-tech.cn

友万科技



# curl 模拟浏览器提交 POST 参数

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

```
global Begin = "2017-08-01"
global End = "2017-08-19"

# delimit ;
!curl "http://datacenter.mep.gov.cn:8099/ths-report/report%21list.action"
-H "User-Agent: Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/59.0.3071.115 Safari/537.36"
-H "Content-Type: application/x-www-form-urlencoded" -H "Cache-Control: no-cache"
-d "page.pageNo=1&xmlname=1462259560614&queryflag=open&isdesignpatterns=
false&V_DATE=$Begin&_DATE=$End"
--compressed -o temp.txt
;
#delimit cr
cls
type temp.txt
```



www.uone-tech.cn

友万科技



# curl 下载网页源代码

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

```

Stata/MP 14.2
File Edit Data Graphics Statistics User Window Help
Review T F X
Filter command
# Comma...
1 do "C:\... 1.
2 do "C:\...
3 type te...
4 do "C:\...
5 do "C:\...
6 do "C:\...
7 percent...
8 do "C:\...
9 ds
1 do "C:\...

> : break-word; display:none ">56F6ED50EC34EDB9E050007F010003EE</td>
<td rowid="3" mergecol="-1" mergerow="-1" colid="2" rowspan="1" colspan="1" style="text-align:center;WORD-WRAP
> : break-word;"><span style='color:#FFD91C;'>石家庄市</span></td>
<td rowid="3" mergecol="-1" mergerow="-1" colid="3" rowspan="1" colspan="1" style="text-align:center;WORD-WRAP
> : break-word;"><span style='color:#FFD91C;'>80</span></td>
<td rowid="3" mergecol="-1" mergerow="-1" colid="4" rowspan="1" colspan="1" style="text-align:center;WORD-WRAP
> : break-word;"><span style='color:#FFD91C;'>PM2.5</span></td>
<td rowid="3" mergecol="-1" mergerow="-1" colid="5" rowspan="1" colspan="1" style="text-align:center;WORD-WRAP
> : break-word; display:none ">二级</td>
<td rowid="3" mergecol="-1" mergerow="-1" colid="6" rowspan="1" colspan="1" style="text-align:center;WORD-WRAP
> : break-word;"><span style='color:#FFD91C;'>2017-08-17</span></td>
<td rowid="3" mergecol="-1" mergerow="-1" colid="7" rowspan="1" colspan="1" style="text-align:center;WORD-WRAP
> : break-word; display:none ">130100</td>
<td rowid="3" mergecol="-1" mergerow="-1" colid="8" rowspan="1" colspan="1" style="text-align:center;WORD-WRAP
> : break-word;"><span style='color:#FFD91C;'>良</span></td>
</tr>
<tr style="" onmouseover="this.style.backgroundColor='#FFDD7'" onmouseout="this.style.backgroundColor=''><td
> rowid="4" mergecol="-1" mergerow="-1" colid="0" rowspan="1" colspan="1" style="text-align:center; width:5%;W
> ORD-WRAP: break-word;"><span style='color:#FF7E00;'>4</span></td>
<td rowid="4" mergecol="-1" mergerow="-1" colid="1" rowspan="1" colspan="1" style="text-align:center;WORD-WRAP
> : break-word; display:none ">56F6ED50EC35EDB9E050007F010003EE</td>
Command

```



www.uone-tech.cn

友万科技

友万科技



# 清洗源代码

最后一步，利用 Stata 读入数据、清洗数据的强大功能对网页源代码进行整理。

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与 微博 API

实战之 环保部 空气质量 日数据

实战之 中国土地网

cURL has limits

```
. list , clean
```

	Grade	Rownum	ID	AQI	CITY	STATUS	MAIN_PO-T	CITYCODE	OPER_DATE
1.	二级	1	56F6ED50EC32EDB9E050007F010003EE	90	北京市	良	臭氧8小时	110000	2017-08-17
2.	三级	2	56F6ED50EC33EDB9E050007F010003EE	118	天津市	轻度污染	臭氧8小时	120000	2017-08-17
3.	二级	3	56F6ED50EC34EDB9E050007F010003EE	80	石家庄市	良	PM2.5	130100	2017-08-17
4.	三级	4	56F6ED50EC35EDB9E050007F010003EE	107	唐山市	轻度污染	臭氧8小时	130200	2017-08-17
5.	二级	5	56F6ED50EC36EDB9E050007F010003EE	65	秦皇岛市	良	臭氧8小时	130300	2017-08-17
6.	二级	6	56F6ED50EC37EDB9E050007F010003EE	78	邯郸市	良	PM2.5	130400	2017-08-17
7.	二级	7	56F6ED50EC38EDB9E050007F010003EE	80	邢台市	良	PM2.5	130500	2017-08-17
8.	三级	8	56F6ED50EC39EDB9E050007F010003EE	121	保定市	轻度污染	臭氧8小时	130600	2017-08-17
9.	一级	9	56F6ED50EC3AEDB9E050007F010003EE	47	承德市	优		130800	2017-08-17
10.	三级	10	56F6ED50EC3BEDB9E050007F010003EE	122	沧州市	轻度污染	臭氧8小时	130900	2017-08-17
11.	二级	11	56F6ED50EC3CEDB9E050007F010003EE	95	廊坊市	良	臭氧8小时	131000	2017-08-17
12.	二级	12	56F6ED50EC3DEDB9E050007F010003EE	68	衡水市	良	臭氧8小时	131100	2017-08-17
13.	二级	13	56F6ED50EC3EEDB9E050007F010003EE	85	张家口市	良	臭氧8小时	131200	2017-08-17
14.	三级	14	56F6ED50EC3FEDB9E050007F010003EE	117	太原市	轻度污染	臭氧8小时	140100	2017-08-17
15.	二级	15	56F6ED50EC40EDB9E050007F010003EE	95	大同市	良	臭氧8小时	140200	2017-08-17
16.	二级	16	56F6ED50EC41EDB9E050007F010003EE	87	阳泉市	良	PM2.5	140300	2017-08-17
17.	三级	17	56F6ED50EC42EDB9E050007F010003EE	113	长治市	轻度污染	臭氧8小时	140400	2017-08-17
18.	二级	18	56F6ED50EC43EDB9E050007F010003EE	136	晋城市	轻度污染	臭氧8小时	140500	2017-08-17
19.	三级	19	56F6ED50EC44EDB9E050007F010003EE	102	朔州市	轻度污染	臭氧8小时	140600	2017-08-17
20.	三级	20	56F6ED50EC45EDB9E050007F010003EE	111	晋中市	轻度污染	臭氧8小时	140700	2017-08-17
21.	三级	21	56F6ED50EC46EDB9E050007F010003EE	139	运城市	轻度污染	臭氧8小时	140800	2017-08-17
22.	三级	22	56F6ED50EC47EDB9E050007F010003EE	110	忻州市	轻度污染	臭氧8小时	140900	2017-08-17
23.	三级	23	56F6ED50EC48EDB9E050007F010003EE	126	临汾市	轻度污染	臭氧8小时	141000	2017-08-17
24.	三级	24	56F6ED50EC49EDB9E050007F010003EE	122	吕梁市	轻度污染	臭氧8小时	141100	2017-08-17
25.	二级	25	56F6ED50EC4AEDB9E050007F010003EE	95	呼和浩特市	良	臭氧8小时	150100	2017-08-17
26.	二级	26	56F6ED50EC4BEDB9E050007F010003EE	75	包头市	良	臭氧8小时	150200	2017-08-17
27.	二级	27	56F6ED50EC4CEDB9E050007F010003EE	93	乌海市	良	臭氧8小时	150300	2017-08-17
28.	一级	28	56F6ED50EC4DEDB9E050007F010003EE	28	赤峰市	优		150400	2017-08-17
29.	一级	29	56F6ED50EC4EEDB9E050007F010003EE	27	通辽市	优		150500	2017-08-17



Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

# 实战之中国土地网



友万科技

[www.uone-tech.cn](http://www.uone-tech.cn)



## 结果公告 landchina.com

### 查询条件

签订日期 : 2017-8-1 到 2017-8-16

行政区 : 北京市

土地用途 :

供应方式 :

电子监管号 :

土地坐落 :

总面积 : >

请选择排序方式:  签订日期[降]  创建时间[降]

查询

序号	行政区	土地坐落	总面积	土地用途	供应方式	签定日期
1.	朝阳区	朝阳区大屯路36号	0.435872	科教用地	划拨	2017/8/10
2.	密云县	密云生态商务区A1地块	1.390900	街巷用地	划拨	2017/8/9
3.	房山区	房山区长阳镇	0.640000	科教用地	划拨	2017/8/8
4.	北京市本级	北京市丰台区长辛店镇辛庄村(一...	10.212050	其他普通商品住房用地	挂牌出让	2017/8/7



# 找到 POST 参数

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

```
<input id="TAB_QuerySubmitConditionData" name="TAB_QuerySubmitConditionData" type="hidden"
value="9f2c3acd-0256-4da2-a659-6949c4671a2a 2017-8-1 2017-8-16 42ad98ae-c46a-40aa-aacc-
c0884036eeaf 11北京市"/><input id="TAB_QuerySubmitOrderData" name="TAB_QuerySubmitOrderData"
type="hidden" value="282:False"/><script type="text/javascript">QueryAction.formId='mainForm';
</script></TD>
```

Figure: 中国土地网源代码

(Photo cropped from: <http://www.landchina.com/default.aspx?tabid=263&ComName=default>)



www.uone-tech.cn

友万科技



# Download

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

```
# delimiter ;
| curl "http://www.landchina.com/default.aspx?tabid=263&ComName=default"
-H "User-Agent: Opera/9.80 (Android 2.3.4; Linux; Opera Mobi/build-1107180945; U; en-GB) Presto/2.8.149 Ver
-H "Content-Type: application/x-www-form-urlencoded"
-H "Cache-Control: no-cache"
-d "__EVENTVALIDATION=/wEWAgKg+ojgCwLN3cj/BEjpHh2IVeSnH7YAAyO/TP/eDYTr/Sg8kIkYr1hhTyNZ
&hidComName=default
&TAB_QueryConditionItem=9f2c3acd-0256-4da2-a659-6949c4671a2a
&TAB_QueryConditionItem=42ad98ae-c46a-40aa-aacc-c0884036eeaf
&TAB_QuerySortItemList=282:False
&TAB_QuerySubmitConditionData=9f2c3acd-0256-4da2-a659-6949c4671a2a:
2017-8-1~2017-8-16|42ad98ae-c46a-40aa-aacc-c0884036eeaf 11北京市
&TAB_QuerySubmitOrderData=282:False

&TAB_QuerySubmitPagerData=1"

-x 123.103.93.38:80 -o temp.txt
;
#delimiter cr
```



www.uone-tech.cn

友万科技



# 爬虫最后一步：数据清洗

Stata 出色的数据管理能力使我们在使用 copy 或 cURL 下载网络源代码以后，能轻松的整理成格式化的数据。

```
cap unicode analyze "temp.txt"
cap unicode encoding set gb18030
cap unicode translate "temp.txt", transutf8
cap unicode erasebackups, badidea
qui set obs 1
qui gen v = fileread("temp.txt")
qui split v,p("gridTdNumber")
qui drop v v1
qui sxpose ,clear
qui split _var1,p("</td><td class="queryCellBordy">")
qui gen num = ustrregexs(1) if ustrregexm(_var1,"^">([0-9]+)\.")
qui gen city = ustrregexs(1) if ustrregexm(_var12,"<span title="(.)">")
qui replace _var12 = city if city~=""
qui drop city
qui ren _var12 city
qui gen url = ustrregexs(1) if ustrregexm(_var13,"href="(.)")
qui gen address = ustrregexs(1) if ustrregexm(_var13,"target="_blank">(.)</a>$")
qui gen date = ustrregexs(1) if ustrregexm(_var17,"^"([0-9]+)/([0-9]+)/([0-9]+)</td>")
```

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么？

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits



www.uone-tech.cn

友万科技



# 爬虫最后一步：数据清洗

Stata Group Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

```
. list num city area type way address date url in 1
```

num	city	area	type	way	address	date
1	朝阳区	0.435872	科教用地	划拨	朝阳区大屯路36号	2017/8/10

```
url
default.aspx?tabid=386&comname=default&wmguid=75c72564-ffd9-426a-954b-8ac2df0903b7&recorderguid=350bd..
```

```
. list num city area type way address date url in 2
```

num	city	area	type	way	address	date
2	密云县	1.390900	街巷用地	划拨	密云生态商务区A1地块	2017/8/9

```
url
default.aspx?tabid=386&comname=default&wmguid=75c72564-ffd9-426a-954b-8ac2df0903b7&recorderguid=14961..
```

Figure: 北京市土地结果公告



www.uone-tech.cn

友万科技



# cURL 的本领

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

- POST
- Headers
- Cookies
- 上传、下载、重定向....



[www.uone-tech.cn](http://www.uone-tech.cn)

友万科技



# Table of Contents

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

① 从爬虫说起

② cURL 是什么?

③ cURL 爬虫

cURL 与微博 API

实战之环保部空气质量日数据

实战之中国土地网

④ cURL has limits



[www.uone-tech.cn](http://www.uone-tech.cn)

友万科技



# cURL has its' limits

Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

cURL 不知道何为 Javascript

cURL 无法解决验证码



[www.uone-tech.cn](http://www.uone-tech.cn)

友万科技



Stata Group  
Meeting

彭文威

从爬虫说起

cURL 是什么?

cURL 爬虫

cURL 与微博 API

实战之环保部空气质量  
日数据

实战之中国土地网

cURL has limits

Thank you!



友万科技

[www.uone-tech.cn](http://www.uone-tech.cn)