Grab Data About China Administrative Division Code Using Shell Script

忽然想知道中國大陸有多少個城市、多少個鄉鎮、多少個村莊，記得以前在國家統計局官網上看到過。覈實後，想將數據抓取下來，一爲瞭解神州大地、二爲提升自己分析能力、編程能力、數據庫能力、三爲以後的計劃作資源儲備。

從項目立項，到最終實現，花了整一天半時間。

已經上傳至 GitHub

Table Of Content

Prerequisites

Geographical Division Introduction

- ROC 中華民國 Republic of China
- PRC 中華人民共和國 People's Republic of China

數據來自 中華人民共和國國家統計局 官網，具體頁面

- 行政區劃代碼

- 統計用區劃和城鄉劃分代碼
- 統計用區劃代碼和城鄉劃分代碼編制規則

中國大陸官方資料：目前有`34`個省級行政區，包括`23`個省、`5`個自治區、`4`個直轄市（包含臺灣省、香港特別行政區、澳門特別行政區）。但區劃代碼和城鄉劃分代碼不包括（臺灣省、香港特別行政區、澳門特別行政區）。

中國大陸六大地理分區

Geographical Division	Provices
`1` 華北	北京市、天津市、河北省、山西省、內蒙古自治區
`2` 東北	遼寧省、吉林省、黑龍江省
`3` 華東	上海市、江蘇省、浙江省、安徽省、福建省、江西省、山東省
`4` 中南	河南省、河北省、湖北省、廣東省、廣西壯族自治區、海南省
`5` 西南	重慶市、四川省、貴州省、雲南省、西藏自治區
`6` 西北	陝西省、甘肅省、青海省、寧夏回族自治區、新疆維吾爾族自治區

Administrative Division Code

行政區劃代碼

> 統計用區劃代碼基本長度爲`12`位，省、地、縣、鄉四級代碼不足12位用`0`補足。

行政區劃共分爲5級，由12位數字代碼表示。

Division Level	Code Position	English	Explanation
省級	1~2	Province	包含省、直轄市、自治區
地級	3~4	City	各省、自治區的地級市，各直轄市的市轄區
縣級	5 ~ 6	Country	各直轄市的市轄區，各地級市所轄的區、市、縣
鄉級	7 ~ 9(3位)	Town	各街道辦事處或鄉鎮
村級	10 ~ 12(3位)	Village	各社區居委會 、各村委會

- 縣以上行政區劃代碼由`1～6`位代碼組成；
- 縣以下區劃代碼由`7～12`位代碼組成，包括 *鄉級代碼* 和 *村級代碼* 兩部分;
    - 鄉級代碼編碼(7~9位)
        - `001～099`: 街道
        - `100～199`: 鎮
        - `200～399`: 鄉
        - `400～599`: 類似鄉級單位(民政部門未確認的開發區、工礦區、農場等)
    - 村級代碼編碼(10~12位)
        - `001~199`: 居民委員會
        - `200~399`: 村民委員會
        - `400~499`: 類似居民委員會(不含498代碼)
        - `500~599`: 類似村民委員會(不含598代碼)
        - `498`: 虛擬社區 (街道、鎮以及類似鄉級單位的開發區、科技園區、工業園區、工礦區、高校園區、科研機構園區等)
        - `598`: 虛擬生活區 (鄉以及類似鄉級單位的農、林、牧、漁場和其它農業活動區域)

Urban and rural classification code

城鄉分類代碼

通常由`3`位數字構成，首位是`1`表示 **城鎮**，首位是`2`表示 **鄉村**。

Code	Explanation
`111`	主城區
`112`	城鄉結合區
`121`	鎮中心區域
`122`	鎮鄉結合區
`123`	特殊區域
`210`	鄉中心區
`220`	村莊

Web Crawler Methods

直接使用`curl`抓取頁面，返回的數據出現中文亂碼，考慮是編碼問題。使用`curl -s URL | grep charset`返回的數據中有`charset=gb2312`，即頁面採用`GB2312`編碼。

解決方案是使用`iconv`命令進行編碼轉換，格式`iconv -f fromcode -t tocode [file ...]`。

使用命令`iconv -l | grep -Ei '^(gb|utf)'`，返回的編碼集中有`GB2312`、`GBK`、`UTF8`、`UTF-8`。經過測試源編碼可使用`GB2312`或`GBK`，目標編碼可使用`UTF8`或`UTF-8`。

即使用`curl -s URL | iconv -f GBK -t UTF-8`。

> -s quiet靜默模式 –retry 重試次數 –retry-delay 間隔時間 -x 代理 -o保存路徑 ipaddr 代理IP port 代理端口

使用代理進行抓取操作，如`curl -s --retry 5 --retry-delay 5 -x ipaddr:port | iconv -f GBK -t UTF-8`

使用`sed -r 's@<[^>]*>@@g;'`去除HTML頁面中標籤。

Official Page Analysis

對 2014年統計用區劃代碼和城鄉劃分代碼 相關頁面進行分析。

Specific Datas

- 省級: `<tr class='provincetr'></tr>`
- 地級: `<tr class='citytr'>`
- 縣級: `<tr class='countytr'>`
- 鄉級: `<tr class='towntr'>`
- 村級: `<tr class='villagetr'>`

URL Analysis

- 首頁地址: `http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/index.html`
- 省級列表地址: `http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/13.html`，13是省級區劃代碼，如河北省`13`，對應替換爲其它省份代碼
- 地級列表地址: `http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/13/1306.html`，13是河北省，`1306`是保定市，格式是`2位省份代碼/地級市代碼`
- 縣級列表地址:

`http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/13/06/130637.html`，13是河北省，06在河北省下代表保定，`130637`是河北省保定市博野縣，格式是`2位省份代碼/2位所屬地級市代碼/縣級代碼`

- 鄉級列表地址: `http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/13/06/37/130637100.html`，13是河北省，06在河北省下代表保定，37在保定下代表博野縣，`130637100`河北省保定市博野縣博野鎮，顯示具體村莊列表信息，格式是`2位省份代碼/2位所屬地級市代碼/2位縣級代碼/鄉級代碼`

Crawler Codes

- 首頁地址: `index.html`

```
curl -s http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/index.html | iconv -f GBK -t UTF-8 | tr -d "'" | grep "class=provincetr" | sed -n '/<tr class=provincetr>/,/<\/tr>/ p' | sed -r 's@<br/>@\n@g;s@</?(a|tr|td)>@@g;s@<tr class=provincetr>@@g;s@(<a |href=|.html)@@g;s@>@ @g;'
```

- 省級列表地址: `13.html`

```
curl -s http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/13.html | iconv -f GBK -t UTF-8 | tr -d "'" | grep "class=citytr" | sed -r 's@ class=citytr@@g' |sed -n '/<tr>/,/<\/tr>/ p' | sed -r 's@<tr>@\n@g;s@>@>@g;s@<[^>]*>@@g;'
```

- 地級列表地址: `13/1306.html`

```
curl -s http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/13/1306.html | iconv -f GBK -t UTF-8 | tr -d "'" | grep "class=countytr" | sed -r 's@ class=countytr@@g' |sed -n '/<tr>/,/<\/tr>/ p' | sed -r 's@<tr>@\n@g;s@>@> @g;s@<[^>]*>@@g;'
```

- 縣級列表地址: `13/06/130637.html`

```
curl -s http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/13/06/130637.html | iconv -f GBK -t UTF-8 | tr -d "'" | grep "class=towntr" | sed -r 's@ class=towntr@@g' |sed -n '/<tr>/,/<\/tr>/ p' | sed -r 's@<tr>@\n@g;s@>@> @g;s@<[^>]*>@@g;'
```

- 鄉級列表地址: `13/06/37/130637100.html`

```
curl -s http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/13/06/37/130637100.html | iconv -f GBK -t UTF-8 | tr -d "'" | grep "class=villagetr" | sed -r 's@ class=villagetr@@g' |sed -n '/<tr>/,/<\/tr>/ p' | sed -r 's@<tr>@\n@g;s@>@> @g;s@<[^>]*>@@g;'
```

Code Design Architecture

代碼架構設計

按照 `省級` –> `地級` –> `縣級` –> `鄉級` –> `村級` 層級,一級一級操作。

數據抓取通過代理IP進行，數據提取使用`tr`, `grep`, `sed`。數據插入使用多條數據同時插入形式，以提高入庫效率。

共四個文件

```
[flying@lemp chinaReginCode]$ tree
.
├── crawler.sh
├── impotProxyIP.sh
├── proxyList.txt
└── SQLDesign.sql

0 directories, 4 files
[flying@lemp chinaReginCode]$
```

- `crawler.sh`: 數據抓取腳本
- `impotProxyIP.sh`: 導入代理腳本
- `proxyList.txt`: 代理IP列表
- `SQLDesign.sql`: 數據表設計

SQL Design

數據表設計，總共7張表，關聯表之間建有外鍵約束

```
#lemp-馬雪東
#https://lempstacker.com
#2016.03.13 12:30 Tue

#explain https://mariadb.com/kb/en/mariadb/explain/

#Database chinaCode
#table
#    proxy
#    province
#    city
#    country
#    town
#    urbanruralCode
#    village

#數據庫名 chinaCode
drop database if exists chinaCode;

create database if not exists chinaCode character set=utf8 collate=utf8
_general_ci;

use chinaCode

#Proxy代理
#drop table if exists proxy;
create table if not exists proxy (
    id int(10) unsigned not null auto_increment primary key comment '代
理IP列表自增id',
    ipaddr char(18) not null comment '代理IP地址',
    port smallint(5) not null comment '代理IP端口',
    protocol char(10) comment 'HTTP類型',
    anonymity varchar(20) comment '匿名等級',
    country varchar(20) comment '代理IP所屬國家',
    province varchar(20) comment '代理IP所屬省份地區',
    city varchar(20) comment '代理IP所屬城市',
    uptime varchar(10) comment '代理IP正常運行時間',
    createTime timestamp default current_timestamp comment '數據入庫時間
YYYY-MM-DD HH:MM:SS',
    updateTime timestamp null on update current_timestamp comment '數據
更新時間 YYYY-MM-DD HH:MM:SS',
    key indProxy_id_ipaddr (id,ipaddr,port),
    unique key indProxy_ipaddr_port (ipaddr,port)
)engine=innodb default charset=utf8 collate=utf8_general_ci comment='代
理IP列表';

#alter table proxy add index index_id_ipaddr (id,ipaddr,port);
#alter table proxy add unique index index_ipaddr_port (ipaddr,port);
#drop index index_id_ipaddr on proxy;
```

```
#drop index index_ipaddr_port on proxy;


#province省、直轄市、自治區
#drop table if exists province;
create table if not exists province (
    id int(10) unsigned not null auto_increment primary key comment '省級列表自增id',
    name char(24) not null comment '省份名稱',
    name_roc char(24) comment '省份名稱(正體)',
    code bigint(12) unsigned not null comment '行政區劃代碼，共12位，由前2位指定',
    region enum('華北','東北','華東','中南','西南','西北') not null comment '省所屬區域:1華北、2東北、3華東、4中南(華中,華南)、5西南、6西北',
    abbr char(30) comment '省份簡稱或別稱',
    abbr_roc char(30) comment '省份簡稱(正體)',
    createTime timestamp default current_timestamp comment '數據入庫時間 YYYY-MM-DD HH:MM:SS',
    updateTime timestamp null on update current_timestamp comment '數據更新時間 YYYY-MM-DD HH:MM:SS',
    key indProvince_name_abbr (id,name,abbr),
    key indProvince_roc_name_abbr (id,name_roc,abbr_roc),
    unique key ind_code (code)
)engine=innodb default charset=utf8 collate=utf8_general_ci comment='中國大陸省份列表';

#alter table province add index index_name_abbr (name,abbr);
#alter table province add index index_roc_name_abbr (name_roc,abbr_roc);
#alter table province add index index_id_code (id,code);


# city地市級
#drop table if exists city;
create table if not exists city (
    id int(10) unsigned not null auto_increment primary key comment '地級市列表自增id',
    provinceId int(10) unsigned not null comment '省級id，對應表province，外鍵約束',
    name char(60) not null comment '地級市名稱',
    name_roc char(60) comment '地級市名稱(正體)',
    code bigint(12) unsigned not null comment '行政區劃代碼，共12位，由前4位指定',
    abbr char(30) comment '地級市簡稱或別稱',
    abbr_roc char(30) comment '地級市簡稱(正體)',
    createTime timestamp default current_timestamp comment '數據入庫時間 YYYY-MM-DD HH:MM:SS',
    updateTime timestamp null on update current_timestamp comment '數據更新時間 YYYY-MM-DD HH:MM:SS',
    key indCity_name_abbr (id,name(12),abbr),
```

```
    key indCity_roc_name_abbr (id,name_roc(12),abbr_roc),
    key incCity_fkey_province (provinceId),
    unique key indCity_code (code),
    constraint fkey_province_city foreign key (provinceId) references p
rovince(id) on update cascade
)engine=innodb default charset=utf8 collate=utf8_general_ci comment='中
國大陸地級市列表';

#alter table city add constraint fkey_province_city foreign key(provinc
eId) references province(id) on update cascade;


# country縣級
#drop table if exists country;
create table if not exists country (
    id int(10) unsigned not null auto_increment primary key comment '縣
級列表自增id',
    cityId int(10) unsigned not null comment '地級市id，對應表city，外鍵約
束',
    name char(60) not null comment '縣級市名稱',
    name_roc char(60) comment '縣級市名稱(正體)',
    code bigint(12) unsigned not null comment '行政區劃代碼，共12位，由前6位
指定',
    abbr char(30) comment '縣級市簡稱或別稱',
    abbr_roc char(30) comment '縣級市簡稱(正體)',
    createTime timestamp default current_timestamp comment '數據入庫時間
YYYY-MM-DD HH:MM:SS',
    updateTime timestamp null on update current_timestamp comment '數據
更新時間 YYYY-MM-DD HH:MM:SS',
    key indCoun_name_abbr (id,name(12),abbr),
    key indCoun_roc_name_abbr (id,name_roc(12),abbr_roc),
    key indCoun_fkey_city (cityId),
    unique key indCoun_code (code),
    constraint fkey_city_coun foreign key (cityId) references city(id)
on update cascade
)engine=innodb default charset=utf8 collate=utf8_general_ci comment='中
國大陸縣級列表';


# town鄉級
#drop table if exists town;
create table if not exists town (
    id int(10) unsigned not null auto_increment primary key comment '鄉
鎮列表自增id',
    countryId int(10) unsigned not null comment '縣級市id，對應表country，
外鍵約束',
    name char(60) not null comment '鄉鎮名稱',
    name_roc char(60) comment '鄉鎮名稱(正體)',
    code bigint(12) unsigned not null comment '行政區劃代碼，共12位，由前9位
指定',
```

```
    abbr char(30) comment '鄉鎮簡稱或別稱',
    abbr_roc char(30) comment '鄉鎮簡稱(正體)',
    createTime timestamp default current_timestamp comment '數據入庫時間 YYYY-MM-DD HH:MM:SS',
    updateTime timestamp null on update current_timestamp comment '數據更新時間 YYYY-MM-DD HH:MM:SS',
    key indTown_name_abbr (id,name(12),abbr),
    key indTown_roc_name_abbr (id,name_roc(12),abbr_roc),
    key indTown_fkey_coun (countryId),
    unique key indTown_code (code),
    constraint fkey_coun_town foreign key (countryId) references country(id) on update cascade
)engine=innodb default charset=utf8 collate=utf8_general_ci comment='中國大陸鄉鎮列表';


# Urban and rural classification code 城鄉分類代碼
#drop table if exists urbanruralCode;
create table if not exists urbanruralCode (
    id int(10) unsigned not null auto_increment primary key comment '城鄉分類代碼自增id',
    code tinyint(3) unsigned not null comment '城鄉分類代碼，共3位',
    name char(60) not null comment '城鄉分類代碼名稱',
    name_roc char(60) not null comment '城鄉分類代碼名稱(正體)',
    createTime timestamp default current_timestamp comment '數據入庫時間 YYYY-MM-DD HH:MM:SS',
    updateTime timestamp null on update current_timestamp comment '數據更新時間 YYYY-MM-DD HH:MM:SS',
    key indurc_name_abbr (id,code,name),
    unique key indurc_code (code)
)engine=innodb default charset=utf8 collate=utf8_general_ci comment='中國大陸城鄉分類代碼表';

insert into urbanruralCode (code,name,name_roc) values (111,'主城区','主城區'),(112,'城乡结合区','城鄉結合區'),(121,'镇中心区','鎮中心區域'),(122,'镇乡结合区','鎮鄉結合區'),(123,'特殊区域','特殊區域'),(210,'乡中心区','鄉中心區'),(220,'村庄','村莊');


# village村級
#drop table if exists village村級;
create table if not exists village (
    id int(10) unsigned not null auto_increment primary key comment '村級列表自增id',
    townId int(10) unsigned not null comment '鄉鎮id，對應表town，外鍵約束',
    name char(60) not null comment '村莊名稱',
    name_roc char(60) comment '村莊名稱(正體)',
    urbanruralCode tinyint(3) unsigned not null comment '城鄉分類代碼，對應表urbanruralCode中code，不設外鍵約束',
```

```
    code bigint(12) unsigned not null comment '行政區劃代碼，共12位，後3位是
村莊代碼',
    abbr char(30) comment '村莊簡稱或別稱',
    abbr_roc char(30) comment '村莊簡稱(正體)',
    createTime timestamp default current_timestamp comment '數據入庫時間
YYYY-MM-DD HH:MM:SS',
    updateTime timestamp null on update current_timestamp comment '數據
更新時間 YYYY-MM-DD HH:MM:SS',
    key indVillage_name_abbr (id,name(12),abbr),
    key indVillage_roc_name_abbr (id,name_roc(12),abbr_roc),
    key indVillage_urcode (urbanruralCode),
    key indVillage_fkey_town (townId),
    unique key indVillage_code (code),
    constraint fkey_town_village foreign key (townId) references town(i
d) on update cascade
)engine=innodb default charset=utf8 collate=utf8_general_ci comment='中
國大陸村级列表';
```

impotProxyIP.sh

將同目錄下`proxyList.txt`文件中的數據導入數據庫

```
#!/bin/bash
#lemp-馬雪東
#https://lempstacker.com
#2016.03.12 23:30 Sat
#獲取代理IP列表並寫入數據庫

# http://www.freeproxylists.net/
dbname='chinaCode'
file='./proxyList.txt';

sed -i '/^$/d;s@High Anonymous@HighAnonymous@g' $file
#createTime設置屬性timestamp default current_timestamp，無需手動指定入庫時間

mysql -e "truncate table $dbname.proxy;"

awk '{print $1,$2,$3,$4,$5,$8,$9,$10}' $file | while read line;do
    # now=`date +'%Y-%m-%d %H:%M:%S'`
    # now=`date +'%F %T'`
    #存入數組
    arr=(${line})
    ipaddr=${arr[0]}
    port=${arr[1]}
    protocol=${arr[2]}
    anonymity=${arr[3]}
    country=${arr[4]}
    region=${arr[5]}
    city=${arr[6]}
    uptime=${arr[7]}
    mysql -e "insert into $dbname.proxy set ipaddr='$ipaddr',port='$por
t',protocol='$protocol',anonymity='$anonymity',country='$country',provi
nce='$region',city='$city',uptime='$uptime';"

    # mysql -e "insert into $dbname.proxy set ipaddr='$ipaddr',port='$p
ort',protocol='$protocol',anonymity='$anonymity',country='$country',pro
vince='$region',city='$city',uptime='$uptime',createTime='$now';"

done

mysql -e "select count(*) from $dbname.proxy;"
```

crawler.sh

數據抓取腳本

```
#!/bin/bash
#lemp-馬雪東
#https://lempstacker.com
#2016.03.13 23:30 Sun
#抓取數據入庫

dbname='chinaCode'
originUrl='http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2014/'

#curl參數 retry重試次數 retryDelay時間間隔
declare i retry=5
declare i retryDelay=5
# curl -s --retry $retry --retry-delay $retryDelay -x ipaddr:port

#獲取代理IP
function getProxyIP () {
    temp=`mysql -Bse "select ipaddr,port from $dbname.proxy order by ra
nd() limit 1;"`
    arr=($temp)
    ipaddr=${arr[0]}
    port=${arr[1]}
    echo $ipaddr:$port
}
#調用自定義函數getProxyIP
curltool="curl -s --retry $retry --retry-delay $retryDelay -x "`getProx
yIP`

# set foreign_key_checks=0


##抓取首頁省份列表
##查詢數據庫，表province中是否有數據，如果沒有則抓取網頁提取數據入庫

provinceCount=`mysql -Bse "select count(*) from $dbname.province;"`
if [[ $provinceCount -eq 0 ]]; then
    mainPage=$originUrl'index.html'     #拼接網頁
    tempfile=`mktemp -t tempXXXXX.txt`
    $curltool $mainPage | iconv -f GBK -t UTF-8 | tr -d "'" | grep "cla
ss=provincetr" | sed -n '/<tr class=provincetr>/,/<\/tr>/ p' | sed -r
's@<br/>@\n@g;s@</?(a|tr|td)>@@g;s@<tr class=provincetr>@@g;s@(<a |href
=|.html)@@g;s@>@ @g;' | tr -s "\r\n" "\n" > $tempfile

    str=''  #定義變量str，用於拼接字符串入庫
    while read line; do
        arr=($line)
        pcode=${arr[0]}
        pregion=${pcode:0:1}
        pname=${arr[1]}
        # mysql -Bse "insert into $dbname.province set name='"$pname"',
```

```
code=$pcode,region=$pregion;" &> /dev/null
        #將插入數據拼接成字符串，提高插入效率
        str=$str"('"$pname"',$pcode,$pregion),"
        unset arr
        unset pcode
        unset pregion
        unset pname
    done < $tempfile
    rm -f $tempfile
    str=${str%,*}   #刪除字符串尾部逗號，
    mysql -Bse "insert into $dbname.province (name,code,region) values
$str;" &> /dev/null
    unset str
    unset mainPage
fi
unset provinceCount



##抓取各省地級市列表
##地級表city中查看字段provinceId是否含有表province中的數據，如果爲空開值抓取對應
地級市數據

cityCount=`mysql -Bse "select count(a.id) from $dbname.city a join $dbn
ame.province b on a.provinceId=b.id;"`
if [[ $cityCount -eq 0 ]]; then
    tempfile=`mktemp -t tempXXXXX.txt`
    mysql -Bse "select id,code,name from $dbname.province;" > $tempfile
#id用於外鍵查選和入庫，code用於拼接網頁URL
    while read line; do
        arr=($line)
        pid=${arr[0]}
        pcode=${arr[1]}
        pname=${arr[2]}
        num=`mysql -Bse "select count(id) from $dbname.city where provi
nceId=$pid"`
        if [[ $num -eq 0 ]]; then
            tempfilecity=`mktemp -t tempXXXXX.txt`
            provincePage=$originUrl$pcode'.html'
            $curltool $provincePage | iconv -f GBK -t UTF-8 | tr -d "'"
| grep "class=citytr" | sed -r 's@ class=citytr@@g' |sed -n '/<tr>/,/<
\/tr>/ p' | sed -r 's@<tr>@\n@g;s@>@> @g;s@<[^>]*>@@g;' | tr -s "\r\n"
"\n" > $tempfilecity
            sed -i '/^$/d' $tempfilecity
            unset provincePage

            cstr=''
            while read cityline; do
                cityarr=($cityline)
                ccode=${cityarr[0]}
```

```
                cname=${cityarr[1]}
                if [[ $cname == '县' ]]; then
                    cname=$pname' 县'
                fi
                cstr=$cstr"('"$cname"',$ccode,$pid),"
                unset cityarr
                unset ccode
                unset cname
            done < $tempfilecity
            rm -f $tempfilecity

            cstr=${cstr%,*}    #删除字符串尾部逗號，
            mysql -Bse "insert into $dbname.city (name,code,provinceId)
values $cstr;" &> /dev/null
            unset cstr
        fi
        unset arr
        unset pid
        unset pcode
        unset pname

    done < $tempfile
    rm -f $tempfile
fi
unset cityCount


##抓取各地級市縣列表
##地級表city中查看字段cityId是否含有表city中的數據，如果爲空開值抓取對應縣級數據

countyCount=`mysql -Bse "select count(a.id) from $dbname.country a join
$dbname.city b on a.cityId=b.id;"`
if [[ $countyCount -eq 0 ]]; then
    tempfile=`mktemp -t tempXXXXX.txt`
    mysql -Bse "select id,left(code,2),left(code,4),name from $dbname.c
ity;" > $tempfile  #id用於外鍵查選和入庫，code用於拼接網頁URL 截取前4爲 省2位、
地級2位
    while read line; do
        arr=($line)
        cityid=${arr[0]}
        provinceCode=${arr[1]}
        cityCode=${arr[2]}
        cityName=${arr[3]}
        num=`mysql -Bse "select count(id) from $dbname.country where ci
tyId=$cityid"`
        if [[ $num -eq 0 ]]; then
            tempfilecountry=`mktemp -t tempXXXXX.txt`
            cityPage=$originUrl$provinceCode'/'$cityCode'.html'
            $curltool $cityPage | iconv -f GBK -t UTF-8 | tr -d "'" | g
rep "class=countytr" | sed -r 's@ class=countytr@@g' |sed -n '/<tr>/,/<
```

```
\/tr>/ p' | sed -r 's@<tr>@\n@g;s@>@> @g;s@<[^>]*>@@g;' > $tempfilecoun
try

            sed -i '/^$/d' $tempfilecountry
            unset cityPage
            countryStr=''
            while read countryline; do
                countryarr=($countryline)
                countryCode=${countryarr[0]}
                countryName=${countryarr[1]}
                if [[ $countryName == '市辖区' ]]; then
                    countryName=$cityName' 市辖区'
                fi
                # echo 'countryName is '$countryName
                countryStr=$countryStr"('"$countryName"',$countryCod
e,$cityid),"
                unset countryarr
                unset countryCode
                unset countryName
            done < $tempfilecountry
            rm -f $tempfilecountry

            countryStr=${countryStr%,*}    #删除字符串尾部逗號，
            mysql -Bse "insert into $dbname.country (name,code,cityId)
values $countryStr;" &> /dev/null
            unset countryStr
        fi
        unset arr
        unset cityid
        unset provinceCode
        unset cityCode
        unset cityName
    done < $tempfile
    rm -f $tempfile

fi
unset countyCount


##抓取各地縣的鄉鎮列表
##鄉級表country中查看字段countryId是否含有表Country中的數據， 如果爲空開值抓取對
應縣級數據

townCount=`mysql -Bse "select count(a.id) from $dbname.town a join $dbn
ame.country b on a.countryId=b.id;"`
if [[ $townCount -eq 0 ]]; then
    tempfile=`mktemp -t tempXXXXX.txt`
    mysql -Bse "select id,left(code,2),substring(code,3,2),left(code,
6),name from $dbname.country;" > $tempfile  #id用於外鍵查選和入庫，code用於
拼接網頁URL 截取前4爲 省2位、地級2位
```

```
    while read line; do
        arr=($line)
        countryid=${arr[0]}
        provinceCode=${arr[1]}
        cityCode=${arr[2]}
        countryCode=${arr[3]}
        countryName=${arr[4]}
        num=`mysql -Bse "select count(id) from $dbname.town where count
ryId=$countryid"`
        if [[ $num -eq 0 ]]; then
            tempfileTown=`mktemp -t tempXXXXX.txt`
            countryPage=$originUrl$provinceCode'/'$cityCode'/'$countryC
ode'.html'
            $curltool $countryPage | iconv -f GBK -t UTF-8 | tr -d "'"
| grep "class=towntr" | sed -r 's@ class=towntr@@g' |sed -n '/<tr>/,/<
\/tr>/ p' | sed -r 's@<tr>@\n@g;s@>@> @g;s@<[^>]*>@@g;' > $tempfileTown

            sed -i '/^$/d' $tempfileTown
            unset countryPage
            townStr=''
            while read townline; do
                townarr=($townline)
                townCode=${townarr[0]}
                townName=${townarr[1]}
                townStr=$townStr"('"$townName"',$townCode,$countryid),"
                unset townarr
                unset townCode
                unset townName
            done < $tempfileTown
            rm -f $tempfileTown

            townStr=${townStr%,*}    #删除字符串尾部逗號，
            # echo $townStr
            mysql -Bse "insert into $dbname.town (name,code,countryId)
values $townStr;" &> /dev/null
            unset townStr
        fi
        unset arr
        unset countryid
        unset provinceCode
        unset cityCode
        unset countryCode
        unset countryName
    done < $tempfile
    rm -f $tempfile

fi
unset townCount
```

```
##抓取各鄉鎮下的村列表
##村級表village中查看字段townId是否含有表town中的數據， 如果爲空開值抓取對應鄉級數據

villageCount=`mysql -Bse "select count(a.id) from $dbname.village a join $dbname.town b on a.townId=b.id;"`
if [[ $villageCount -eq 0 ]]; then
    tempfile=`mktemp -t tempXXXXX.txt`
    mysql -Bse "select id,left(code,2),substring(code,3,2),substring(code,5,2),left(code,9),name from $dbname.town;" > $tempfile  #id用於外鍵查選和入庫，code用於拼接網頁URL 截取前6位 省2位、地級2位 縣級2位
    while read line; do
        arr=($line)
        townid=${arr[0]}
        provinceCode=${arr[1]}
        cityCode=${arr[2]}
        countryCode=${arr[3]}
        townCode=${arr[4]}
        townName=${arr[5]}
        num=`mysql -Bse "select count(id) from $dbname.village where townId=$townid"`
        if [[ $num -eq 0 ]]; then
            tempfileVillage=`mktemp -t tempXXXXXX.txt`
            townPage=$originUrl$provinceCode'/'$cityCode'/'$countryCode'/'$townCode'.html'

            $curltool $townPage | iconv -f GBK -t UTF-8 | tr -d "'" | grep "class=villagetr" | sed -r 's@ class=villagetr@@g' |sed -n '/<tr>/,/<\/tr>/ p' | sed -r 's@<tr>@\n@g;s@>@> @g;s@<[^>]*>@@g;' > $tempfileVillage

            sed -i '/^$/d' $tempfileVillage
            unset townPage
            villageStr=''
            while read villageline; do
                villagearr=($villageline)
                villageCode=${villagearr[0]}
                urbanruralCode=${villagearr[1]}
                villageName=${villagearr[2]}
                villageStr=$villageStr"('"$villageName"',$villageCode,$urbanruralCode,$townid),"
                unset villagearr
                unset villageCode
                unset villageName
            done < $tempfileVillage
            rm -f $tempfileVillage

            villageStr=${villageStr%,*}    #删除字符串尾部逗號，
            #
            # echo "insert into $dbname.village (name,code,urbanruralCo
```

```
de,townId) values $villageStr;"
            mysql -Bse "insert into $dbname.village (name,code,urbanrur
alCode,townId) values $villageStr;" &> /dev/null
            unset villageStr
        fi
        unset arr
        unset townid
        unset provinceCode
        unset cityCode
        unset countryCode
        unset townCode
        unset townName
    done < $tempfile
    rm -f $tempfile

fi
unset villageCount
```

Data Presentation

數據展示，數據就不展示了

Problems

- 文末`^M`出現，使用`tr -s "\r\n" "\n"`去除
- Shell腳本中自定義函數必須先定義，而後才能調用，位置有先後。將函數返回值賦值給變量，可使用'`function_name`'實現，反引號包裹函數名
- MySQL中字符串截取，下標從1開始 `substring(string,pos,len)`

Change Log

- 2016.03.14 00:01 Mon
    - 初稿完成

- Note Time: 2016.03.14 00:01 Mon
- Note Location: Beijing
- Writer: lemp-馬雪東
- Blog: https://lempstacker.com