# Soqi爬虫说明文档v1.0

**安装说明：**

先安装好PYTHON 2.7.3版本

再双击安装PyQt-Py2.7-x86-gpl-4.9.4-1.exe （界面支持库）

再双击安装MySQL-python-1.2.3.win32-py2.7.exe （MYSQL连接库）

再解压缩pyexcelerator-0.6.4.1.zip，打开pyexcelerator-0.6.4.1 目录，用命令行运行：

python setup.py install (注意，python要配置好环境变量)

最后再解压chardet-1.0.1.tar.gz 打开chardet-1.0.1目录，用命令行运行：

python setup.py install (注意，python要配置好环境变量)

最后运行根目录的main.py即可打开窗口

**主界面：**

运行根目录的main.py



为减少磁盘I/O，excel文件将在窗口析构后生成，mysql数据将在窗口析构后才commit录入。txt文件是及时生成、及时更新的。

所以推荐每次搜索的范围是一个省。

例如：

起始ID ：110000　北京市

结束ID： 119999 及在天津市120000之前

**起始ID和结束ID的说明：**

起始ID与结束ID均为**6位数字。**

详细ID列表详见下文，是以中国 **市级**行政区为划分，若ID为空会自动跳过。

例如：

设置起始ID 为000001 结束ID为999999（此时结束ID会自动取到最大值820000）

这样就能搜索全国公司。

**关键词选择：**

推荐选择公司和厂即可。

**配置文件：**

Spider.exe.config

其中需要配置的参数完全可以在窗口界面配置。所以这个配置文件可以忽略

**详细ID表：**

ID号 　对应名称

110000　北京市

110100　　市辖区

110101　　　东城区

110102　　　西城区

110103　　　崇文区

110104　　　宣武区

110105　　　朝阳区

110106　　　丰台区

110107　　　石景山区

110108　　　海淀区

110109　　　门头沟区

110111　　　房山区

110112　　　通州区

110113　　　顺义区

110114　　　昌平区

110115　　　大兴区

110116　　　怀柔区

110117　　　平谷区

110200　　县

110228　　　密云县

110229　　　延庆县

120000　天津市

120100　　市辖区

120101　　　和平区

120102　　　河东区

120103　　　河西区

120104　　　南开区

120105　　　河北区

120106　　　红桥区

120107　　　塘沽区

120108　　　汉沽区

120109　　　大港区

120110　　　东丽区

120111　　　西青区

120112　　　津南区

120113　　　北辰区

120114　　　武清区

120115　　　宝坻区

120200　　县

120221　　　宁河县

120223　　　静海县

120225　　　蓟　县

130000　河北省

130100　　石家庄市

130200　　唐山市

130300　　秦皇岛市

130400　　邯郸市

130500　　邢台市

130600　　保定市

130700　　张家口市

130800　　承德市

130900　　沧州市

131000　　廊坊市

131100　　衡水市

140000　山西省

140100　　太原市

140200　　大同市

140300　　阳泉市

140400　　长治市

140500　　晋城市

140600　　朔州市

140700　　晋中市

140800　　运城市

140900　　忻州市

141000　　临汾市

141100　　吕梁市

150000　内蒙古自治区

150100　　呼和浩特市

150200　　包头市

150300　　乌海市

150400　　赤峰市

150500　　通辽市

150600　　鄂尔多斯市

150700　　呼伦贝尔市

150800　　巴彦淖尔市

150900　　乌兰察布市

152200　　兴安盟

152500　　锡林郭勒盟

152900　　阿拉善盟

210000　辽宁省

210100　　沈阳市

210200　　大连市

210300　　鞍山市

210400　　抚顺市

210500　　本溪市

210600　　丹东市

210700　　锦州市

210800　　营口市

210900　　阜新市

211000　　辽阳市

211100　　盘锦市

211200　　铁岭市

211300　　朝阳市

211400　　葫芦岛市

220000　吉林省

220100　　长春市

220200　　吉林市

220300　　四平市

220400　　辽源市

220500　　通化市

220600　　白山市

220700　　松原市

220800　　白城市

222400　　延边朝鲜族自治州

230000　黑龙江省

230100　　哈尔滨市

230200　　齐齐哈尔市

230300　　鸡西市

230400　　鹤岗市

230500　　双鸭山市

230600　　大庆市

230700　　伊春市

230800　　佳木斯市

230900　　七台河市

231000　　牡丹江市

231100　　黑河市

231200　　绥化市

232700　　大兴安岭地区

310000　上海市

310100　　市辖区

310101　　　黄浦区

310103　　　卢湾区

310104　　　徐汇区

310105　　　长宁区

310106　　　静安区

310107　　　普陀区

310108　　　闸北区

310109　　　虹口区

310110　　　杨浦区

310112　　　闵行区

310113　　　宝山区

310114　　　嘉定区

310115　　　浦东新区

310116　　　金山区

310117　　　松江区

310118　　　青浦区

310119　　　南汇区

310120　　　奉贤区

310200　　县

310230　　　崇明县

320000　江苏省

320100　　南京市

320200　　无锡市

320300　　徐州市

320400　　常州市

320500　　苏州市

320600　　南通市

320700　　连云港市

320800　　淮安市

320900　　盐城市

321000　　扬州市

321100　　镇江市

321200　　泰州市

321300　　宿迁市

330000　浙江省

330100　　杭州市

330200　　宁波市

330300　　温州市

330400　　嘉兴市

330500　　湖州市

330600　　绍兴市

330700　　金华市

330800　　衢州市

330900　　舟山市

331000　　台州市

331100　　丽水市

340000　安徽省

340100　　合肥市

340200　　芜湖市

340300　　蚌埠市

340400　　淮南市

340500　　马鞍山市

340600　　淮北市

340700　　铜陵市

340800　　安庆市

341000　　黄山市

341100　　滁州市

341200　　阜阳市

341300　　宿州市

341400　　巢湖市

341500　　六安市

341600　　亳州市

341700　　池州市

341800　　宣城市

350000　福建省

350100　　福州市

350200　　厦门市

350300　　莆田市

350400　　三明市

350500　　泉州市

350600　　漳州市

350700　　南平市

350800　　龙岩市

350900　　宁德市

360000　江西省

360100　　南昌市

360200　　景德镇市

360300　　萍乡市

360400　　九江市

360500　　新余市

360600　　鹰潭市

360700　　赣州市

360800　　吉安市

360900　　宜春市

361000　　抚州市

361100　　上饶市

370000　山东省

370100　　济南市

370200　　青岛市

370300　　淄博市

370400　　枣庄市

370500　　东营市

370600　　烟台市

370700　　潍坊市

370800　　济宁市

370900　　泰安市

371000　　威海市

371100　　日照市

371200　　莱芜市

371300　　临沂市

371400　　德州市

371500　　聊城市

371600　　滨州市

371700　　菏泽市

410000　河南省

410100　　郑州市

410200　　开封市

410300　　洛阳市

410400　　平顶山市

410500　　安阳市

410600　　鹤壁市

410700　　新乡市

410800　　焦作市

410900　　濮阳市

411000　　许昌市

411100　　漯河市

411200　　三门峡市

411300　　南阳市

411400　　商丘市

411500　　信阳市

411600　　周口市

411700　　驻马店市

420000　湖北省

420100　　武汉市

420200　　黄石市

420300　　十堰市

420500　　宜昌市

420600　　襄樊市

420700　　鄂州市

420800　　荆门市

420900　　孝感市

421000　　荆州市

421100　　黄冈市

421200　　咸宁市

421300　　随州市

422800　　恩施土家族苗族自治州

429000　　省直辖行政单位

429004　　　仙桃市

429005　　　潜江市

429006　　　天门市

429021　　　神农架林区

430000　湖南省

430100　　长沙市

430200　　株洲市

430300　　湘潭市

430400　　衡阳市

430500　　邵阳市

430600　　岳阳市

430700　　常德市

430800　　张家界市

430900　　益阳市

431000　　郴州市

431100　　永州市

431200　　怀化市

431300　　娄底市

433100　　湘西土家族苗族自治州

440000　广东省

440100　　广州市

440200　　韶关市

440300　　深圳市

440400　　珠海市

440500　　汕头市

440600　　佛山市

440700　　江门市

440800　　湛江市

440900　　茂名市

441200　　肇庆市

441300　　惠州市

441400　　梅州市

441500　　汕尾市

441600　　河源市

441700　　阳江市

441800　　清远市

441900　　东莞市

442000　　中山市

445100　　潮州市

445200　　揭阳市

445300　　云浮市

500000　重庆市

500100　　市辖区

500101　　　万州区

500102　　　涪陵区

500103　　　渝中区

500104　　　大渡口区

500105　　　江北区

500106　　　沙坪坝区

500107　　　九龙坡区

500108　　　南岸区

500109　　　北碚区

500110　　　万盛区

500111　　　双桥区

500112　　　渝北区

500113　　　巴南区

500114　　　黔江区

500115　　　长寿区

500116　　　江津区

500117　　　合川区

500118　　　永川区

500119　　　南川区

500200　　县

500222　　　綦江县

500223　　　潼南县

500224　　　铜梁县

500225　　　大足县

500226　　　荣昌县

500227　　　璧山县

500228　　　梁平县

500229　　　城口县

500230　　　丰都县

500231　　　垫江县

500232　　　武隆县

500233　　　忠　县

500234　　　开　县

500235　　　云阳县

500236　　　奉节县

500237　　　巫山县

500238　　　巫溪县

500240　　　石柱土家族自治县

500241　　　秀山土家族苗族自治县

500242　　　酉阳土家族苗族自治县

500243　　　彭水苗族土家族自治县

450000　广西壮族自治区

450100　　南宁市

450200　　柳州市

450300　　桂林市

450400　　梧州市

450500　　北海市

450600　　防城港市

450700　　钦州市

450800　　贵港市

450900　　玉林市

451000　　百色市

451100　　贺州市

451200　　河池市

451300　　来宾市

451400　　崇左市

460000　海南省

460100　　海口市

460200　　三亚市

510000　四川省

510100　　成都市

510300　　自贡市

510400　　攀枝花市

510500　　泸州市

510600　　德阳市

510700　　绵阳市

510800　　广元市

510900　　遂宁市

511000　　内江市

511100　　乐山市

511300　　南充市

511400　　眉山市

511500　　宜宾市

511600　　广安市

511700　　达州市

511800　　雅安市

511900　　巴中市

512000　　资阳市

513200　　阿坝藏族羌族自治州

513300　　甘孜藏族自治州

513400　　凉山彝族自治州

520000　贵州省

520100　　贵阳市

520200　　六盘水市

520300　　遵义市

520400　　安顺市

522200　　铜仁地区

522300　　黔西南布依族苗族自治州

522400　　毕节地区

522600　　黔东南苗族侗族自治州

522700　　黔南布依族苗族自治州

530000　云南省

530100　　昆明市

530300　　曲靖市

530400　　玉溪市

530500　　保山市

530600　　昭通市

530700　　丽江市

530800　　思茅市

530900　　临沧市

532300　　楚雄彝族自治州

532500　　红河哈尼族彝族自治州

532600　　文山壮族苗族自治州

532800　　西双版纳傣族自治州

532900　　大理白族自治州

533100　　德宏傣族景颇族自治州

533300　　怒江傈僳族自治州

533400　　迪庆藏族自治州

540000　西藏自治区

540100　　拉萨市

542100　　昌都地区

542200　　山南地区

542300　　日喀则地区

542400　　那曲地区

542500　　阿里地区

542600　　林芝地区

610000　陕西省

610100　　西安市

610200　　铜川市

610300　　宝鸡市

610400　　咸阳市

610500　　渭南市

610600　　延安市

610700　　汉中市

610800　　榆林市

610900　　安康市

611000　　商洛市

620000　甘肃省

620100　　兰州市

620200　　嘉峪关市

620300　　金昌市

620400　　白银市

620500　　天水市

620600　　武威市

620700　　张掖市

620800　　平凉市

620900　　酒泉市

621000　　庆阳市

621100　　定西市

621200　　陇南市

622900　　临夏回族自治州

623000　　甘南藏族自治州

630000　青海省

630100　　西宁市

632100　　海东地区

632200　　海北藏族自治州

632300　　黄南藏族自治州

632500　　海南藏族自治州

632600　　果洛藏族自治州

632700　　玉树藏族自治州

632800　　海西蒙古族藏族自治州

640000　宁夏回族自治区

640100　　银川市

640200　　石嘴山市

640300　　吴忠市

640400　　固原市

640500　　中卫市

650000　新疆维吾尔自治区

650100　　乌鲁木齐市

650200　　克拉玛依市

652100　　吐鲁番地区

652200　　哈密地区

652300　　昌吉回族自治州

652700　　博尔塔拉蒙古自治州

652800　　巴音郭楞蒙古自治州

652900　　阿克苏地区

653000　　克孜勒苏柯尔克孜自治州

653100　　喀什地区

653200　　和田地区

654000　　伊犁哈萨克自治州

654200　　塔城地区

654300　　阿勒泰地区

659000　　省直辖行政单位

710000　台湾省

810000　香港特别行政区

820000　澳门特别行政区

**输出文件：**

在该项目的根目录下：

按照时间生成的excel文件和txt文件：

companyinformation 2012-07-23 21 30 37.xls

2012-07-23 21 30 37.txt

如果mysql配置无误，会自动把数据插入mysql中，并且会自动去重。

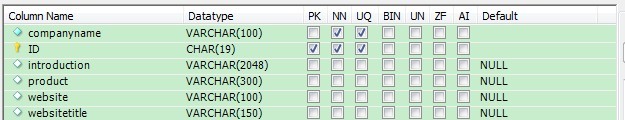
**MYSQL数据库配置：**

MYSQL数据库默认 host地址="localhost"，用户名='root'，密码='123456'，数据库名='companyinformation'，插入表名='companyinformation'

**数据库结构图：**

数据库名companyinformation

表名 companyinformation

****

**数据库建表SQL语句：**

CREATE DATABASE `CompanyInformation` DEFAULT CHARACTER SET utf8 ;

CREATE TABLE `companyinformation`.`CompanyName` (

`idCompanyName` CHAR(19) NOT NULL ,

PRIMARY KEY (`idCompanyName`) ,

UNIQUE INDEX `idCompanyName\_UNIQUE` (`idCompanyName` ASC) )

DEFAULT CHARACTER SET = utf8;

ALTER TABLE `companyinformation`.`companyname` ADD COLUMN `companyname` VARCHAR(100) NOT NULL AFTER `ID` , ADD COLUMN `introduction` VARCHAR(2048) NULL AFTER `companyname` , ADD COLUMN `product` VARCHAR(300) NULL AFTER `introduction` , ADD COLUMN `website` VARCHAR(100) NULL AFTER `product` , ADD COLUMN `websitetitle` VARCHAR(150) NULL AFTER `website` , CHANGE COLUMN `idCompanyName` `ID` CHAR(19) NOT NULL

, DROP PRIMARY KEY

, ADD PRIMARY KEY (`ID`)

, DROP INDEX `idCompanyName\_UNIQUE`

, ADD UNIQUE INDEX `idCompanyName\_UNIQUE` (`ID` ASC)

, ADD UNIQUE INDEX `companyname\_UNIQUE` (`companyname` ASC) , RENAME TO `companyinformation`.`companyinformation` ;

ALTER TABLE `companyinformation`.`companyinformation` CHANGE COLUMN `introduction` `introduction` VARCHAR(4096) NULL DEFAULT NULL , CHANGE COLUMN `product` `product` VARCHAR(600) NULL DEFAULT NULL ;