# Crawler1.0说明文档

## 1.准备工作

解压crawler-1.0.rar到当前目录得到文件夹crawler-1.0

目录结构如下：

.

├── build

│   └── main

│   ├── main.exe.manifest

│   ├── out00-Analysis.toc

│   ├── out00-EXE.toc

│   ├── out00-PKG.pkg

│   ├── out00-PKG.toc

│   ├── out00-PYZ.pyz

│   ├── out00-PYZ.toc

│   ├── out00-Tree.toc

│   ├── out01-Tree.toc

│   └── warnmain.txt

├── dist

│   ├── conf

│   │   ├── Chinese\_Simp.conf

│   │   ├── Chinese\_Trad.conf

│   │   ├── Czech.conf

│   │   ├── Danish.conf

│   │   ├── Dutch.conf

│   │   ├── Estonian.conf

│   │   ├── Finnish.conf

│   │   ├── Hungarian.conf

│   │   ├── Japanese.conf

│   │   ├── Kazakh.conf

│   │   ├── Korean\_Hangul.conf

│   │   ├── Latvian.conf

│   │   ├── Lithuanian.conf

│   │   ├── Norwegian.conf

│   │   ├── Polish.conf

│   │   ├── Portuguese.conf

│   │   ├── pt\_Portuguese.conf

│   │   ├── Russian.conf

│   │   ├── Spanish.conf

│   │   ├── Swedish.conf

│   │   ├── Turkish.conf

│   │   ├── Ukrainian.conf

│   │   ├── uzn\_Uzbek.conf

│   │   └── uzs\_Uzbek.conf

│   ├── crawler

│   │   ├── logging.conf

│   │   └── parameter.conf

│   └── main.exe

└── main.spec

打开下图中的conf文件夹

├── dist

│   ├── conf

保留自己需要获取的语言的文件，其余的文件移动到其他目录

## 2.执行数据清洗工作

打开dist文件夹

双击main.exe文件（会弹出对话框）

按照对话框提示操作

程序执行完毕在这个文件夹（dist）下找到output文件夹，查看里面对应的国家的数据的大小

达到要求则打包

## 3.可能遇到的情况

如果数据大小不够，然而网站的信息会达到预期的大小

程序运行后会生成一些文件，进入文件logs

查看生唱的24.log文件，在日志中会找到类似这样的信息

2017-04-28 09:45:16,667 - 文件内容的语言ja和想获取的文章的语言sv不符合！

2017-04-28 09:45:22,089 - <http://telefonforsaljare.nu>

如果确认下面的网址是所要获取的语言，在文件夹crawler中打开文件parameter.conf

添加一条数据格式如下zh=Chinese\_Simp

把文件内容的语言ja写在等号左边，语言（eg:Chinese\_Simp）写在等号右边

├── dist

│   ├── crawler

重新运行main.exe