**基于nlp分词数据采集程序说明书（v1.5）**

## 1系统架构图

程序入口

关键词提取

**训练模型**

**BeautifulSoup采集引擎擎**

NLP 分词处理

关键词

MySQL

## 2目录结构

**Main.py**：程序入口、定时采集。

**Config**：配置信息模块

def\_config.py：数据库配置信息、程序配置信息；

jieba\_userdict.txt：分词处理特殊名词文件、可添加；

key\_word\_data.txt：关键词（违规词）、可添加；

def\_nlp\_jieba：npl 分词模块。

def\_houses：房产信息采集逻辑模块。

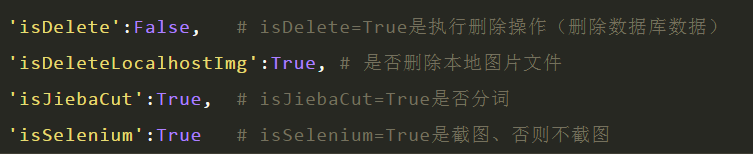
def\_commons：共用模块、包括采集模型。

Data\_file：程序生成数据文件。

img：程序生成图片文件。

## 2.1config模块详情说明

2.1.1Config.py文件重要配置信息



2.1.2提高分词模型准确率

Jieba（结巴分词）的准确率百分之八十。如果有特殊名称jieba分词不准确，例子:（**大学**），而文中出现的是**大学路12号（这是位置名称）**，不应该算是违规词。出现这种情况可以是分词准确率不高，需要我们自己**添加训练分词模型**jieba\_userdict.txt文件中添加模型（已经改为数据库添加）

2.1.3违规广告名词文件

key\_word\_data.txt文件是添加违规关键词，需要注意的是要用**、**分割。

如：**最、最好、最便宜**

## 3、项目环境

3.1语言环境

Python3.6

3.2依赖第三方模块

Pip、Jieba、pymysql、BeautifulSoup、selenium、geckodriver驱动

Requests、bs4、apscheduler

## 3、centos7环境安装

Python2.7和3.6共存：<https://www.cnblogs.com/chenyang13677/p/7827694.html>

安装pip:

<https://www.cnblogs.com/lilidun/p/6041198.html>

使用Python2.7：python2

使用Python3.6：python(已经默认使用3.6，除了特殊如yum外使用2.7)

我们使用pip安装一个python第三方库：**python -m pip install jieba**

安装selenium和版本：

<https://blog.csdn.net/u014283248/article/details/80631072>

<https://www.cnblogs.com/lgh344902118/p/8309114.html>





# <https://www.cnblogs.com/shaosks/p/9174329.html>（[CentOS7下命令安装火狐浏览器](https://www.cnblogs.com/shaosks/p/9174329.html)）

Install第三方模块：先进入pip安装目录（/root/pip/pip-8.0.2）

执行：python pip install 模块名

Pip 升级：python -m pip install -U pip

**运行主函数：**

**脚本执行：（推荐）**

**[root@host-10-5-0-59 ~]# ./start\_py\_houses.sh**

**任何目录脚本执行:bash /root/server/python/ start\_py\_houses.sh**

**任何目录Python执行：**

**python /root/server/python/sanya\_house\_py/py\_client/main.py**

权限：chmod 600 /root/server/python/start\_py\_houses.sh

chmod +x /root/server/python/start\_py\_houses.sh

定时任务（每天下午3点开始执行）：

[root@host-10-5-0-59 ~]# crontab -e

00 15 \* \* \* /root/server/python/start\_py\_houses.sh

查看定时任务：[root@host-10-5-0-59 ~]# crontab -l