# spider设计文档

# 一、spider简介

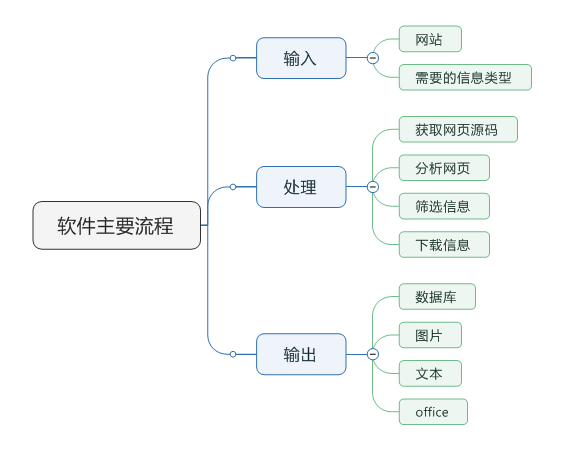
描述特点

封装了网络信息获取方法，数据库的操作，office的操作。轻量，多线程。

主要语言

python、java

### 功能



### 特点

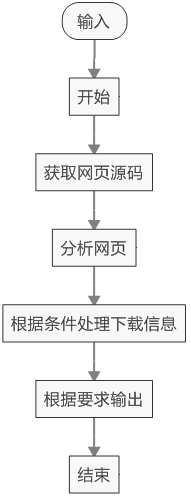
1、耗时短，采用多线程

2、灵活性强，对于不同的防爬虫有相应的措施

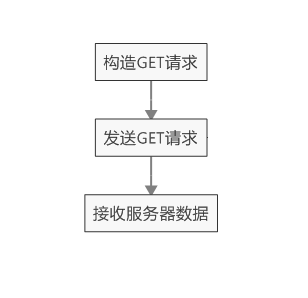
3、输出方式多变，支持图表，文本，数据库。

# 流程设计

## 软件主要流程



## 1、获取网页模块流程



### 应注意的事项与解决方法

1.服务器拒绝IP访问 ——— 创建代理IP

2.服务器拒绝爬虫 ——— 伪造浏览器头

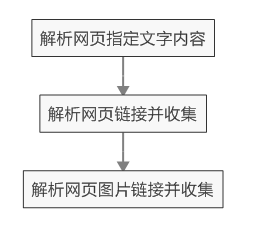
3.服务器拒绝多次访问 ——— 降低频率

4.网页需要载入js ——— 启用类浏览器selenium

5.网页需要cookie验证信息 ——伪造请求头，储存cookie

6.网页需要验证码 ——— 启用图像文字识别Tesseract或者验证码识别平台

## 2、分析网页模块流程



### 应注意的事项与解决方法

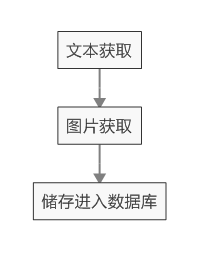
1.网页编码错误 ——— 启用json解析，保证输入输出编码相同

2.文字内容出错，重复 ——— 使用bs4/regax筛选

3.网页链接不存在，重复——— 处理异常，跳过并记录

4.图片过多，logo等一些不必要的图片 ——— 筛选

## 3、根据条件处理下载并储存信息流程



### 应注意的事项与解决方法

1、图片下载缓慢 ——— 启用多线程

2、文字乱码，重复 ——— 筛选，保证输入输出编码相同

3、数据库写入中断 ——— 增加互斥锁，增加线程间隔时间

4、python库的引入

5、文件编码的一致性

# 文档要求

## 1、异常的处理

1、界面错误，传值

2、数据库错误，

3、主程序错误, 网页的解析

4、常见错误，AttributeError：属性错误，赋值失败时会引发属性错误  
NameError：试图访问的变量名不存在  
SyntaxError：语法错误，代码形式错误  
Exception：所有异常的基类，因为所有python异常类都是基类Exception的其中一员，异常都是从基类Exception继承的，并且都在exceptions模块中定义。  
IOError：一般常见于打开不存在文件时会引发IOError错误  
KeyError：使用了映射中不存在的关键字（键）时引发的关键字错误  
IndexError：索引错误，使用的索引不存在，常索引超出序列范围，什么是索引  
TypeError：类型错误，函数应于在了错误类型的对象时会引发类型错误  
ZeroDivisonError：除数为0，在用除法操作时，第二个参数为0时引发了该错误  
ValueError：值错误，传给对象的参数类型不正确，像是给int()函数传入了字符串数据类型的参数。

## 2、注释的要求

1、模块的注释；

2、各个函数功能，输入值的注释

3、对错误判断的注释

## 3、测试要求

1、UI的测试。

1. 性能的测试
2. 移植性的测试
3. 输出结果的测试
4. 软件单元测试