# 1 通用网络爬虫：

### 1.1.应用背景

通用网络爬虫的主要功能是对录入的网站或网页进行相关性正文采集。通用网络爬虫能够实现以网站为初始起点，以个性化配置的方式实现自动数据采集，并支持分布式部署运行的方式，满足稳定性、并发性和易用性。

### 1.2.调用说明

**requirements.txt环境:**

beautifulsoup4==4.9.3  
boilerpipe==1.3.0.0  
Django==3.1.2  
Flask==1.1.2  
ipython==7.16.1  
loguru==0.5.3  
lxml==4.5.2  
numpy==1.19.2  
pandas==1.1.3  
pymongo==3.11.0  
redis==3.5.3  
requests==2.24.0  
retrying==1.3.3  
stanfordcorenlp==3.9.1.1  
tqdm==4.50.2  
urllib3==1.25.10

**文件目录结构**：

.

BIT\_PSPIDER

│ first\_extractor.py 网页内容抽取器

│ rowblock\_extractor.py 网页内容抽取器

│ input\_to\_redis.py 输入文件导入redis

│ main.py 启动主程序

│ mongodb\_to\_redis.py mongodb中未成功采集的url导入redis

│

├─configure

│ config.py 配置文件

│

├─envs

│ pspider.zip Anaconda虚拟环境：使用方法简单：复制到本地机器的anaconda下的envs目录，解压即可正常激活和退出

│ python\_Levenshtein-0.12.0-cp36-cp36m-win\_amd64.whl 一个安装包

│

├─input

│ input\_url.csv 输入URL文件

│ test.csv 测试URL文件

│

└─log 日志目录

**环境准备：**

1. **安装mongodb**
2. **安装redis**
3. **安装python依赖**

**方法一：**使用Anaconda虚拟环境pspider.zip

**方法二：**安装requirements.txt要求的依赖环境

**使用步骤：**

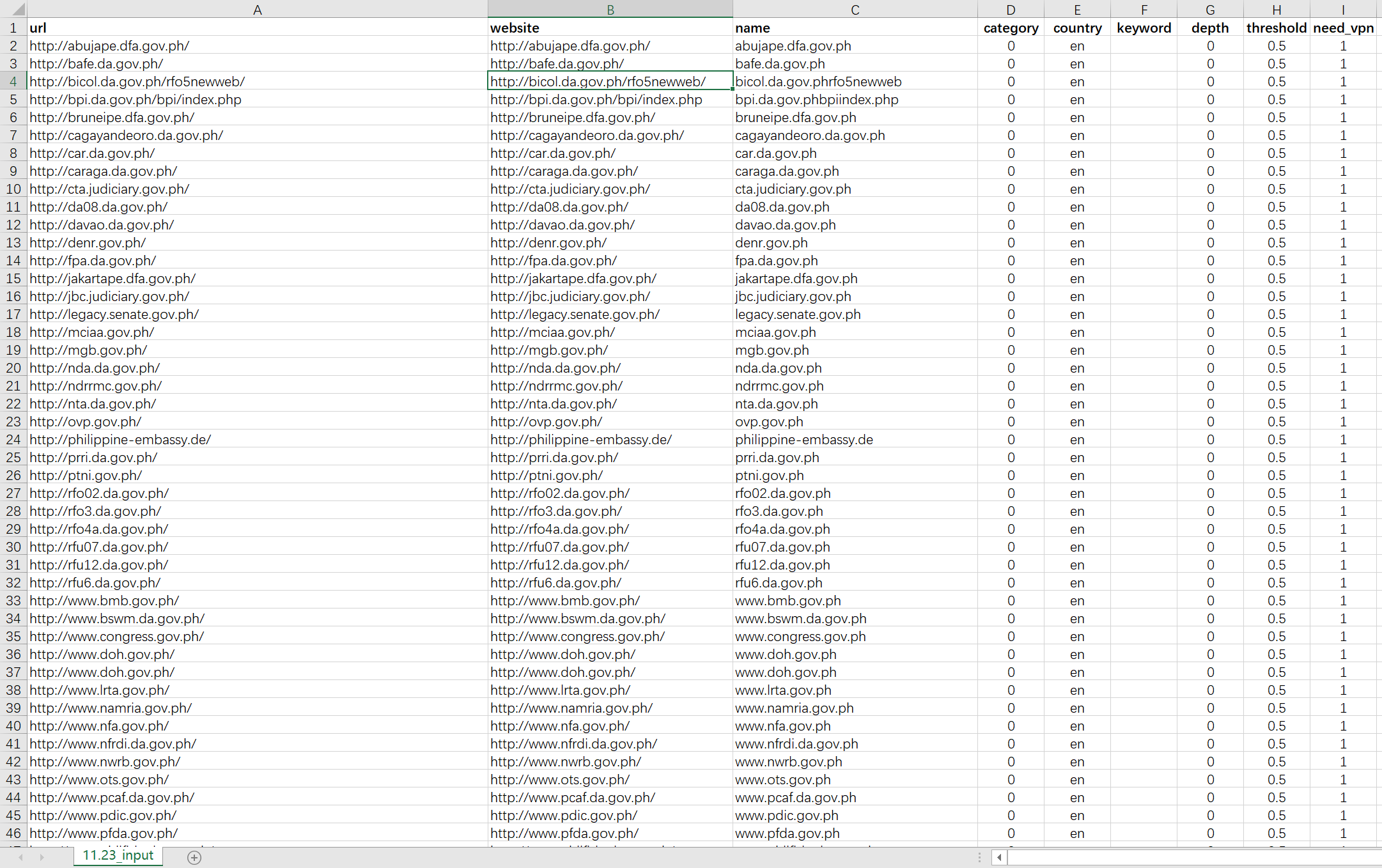
1. **启动数据库、编写输入文件：input/input\_url.csv**
2. **配置爬虫相关参数：configure/config.py**
3. **导入数据到redis**：python input\_to\_redis.py
4. **启动爬虫：python main.py**

#### 数据集准备

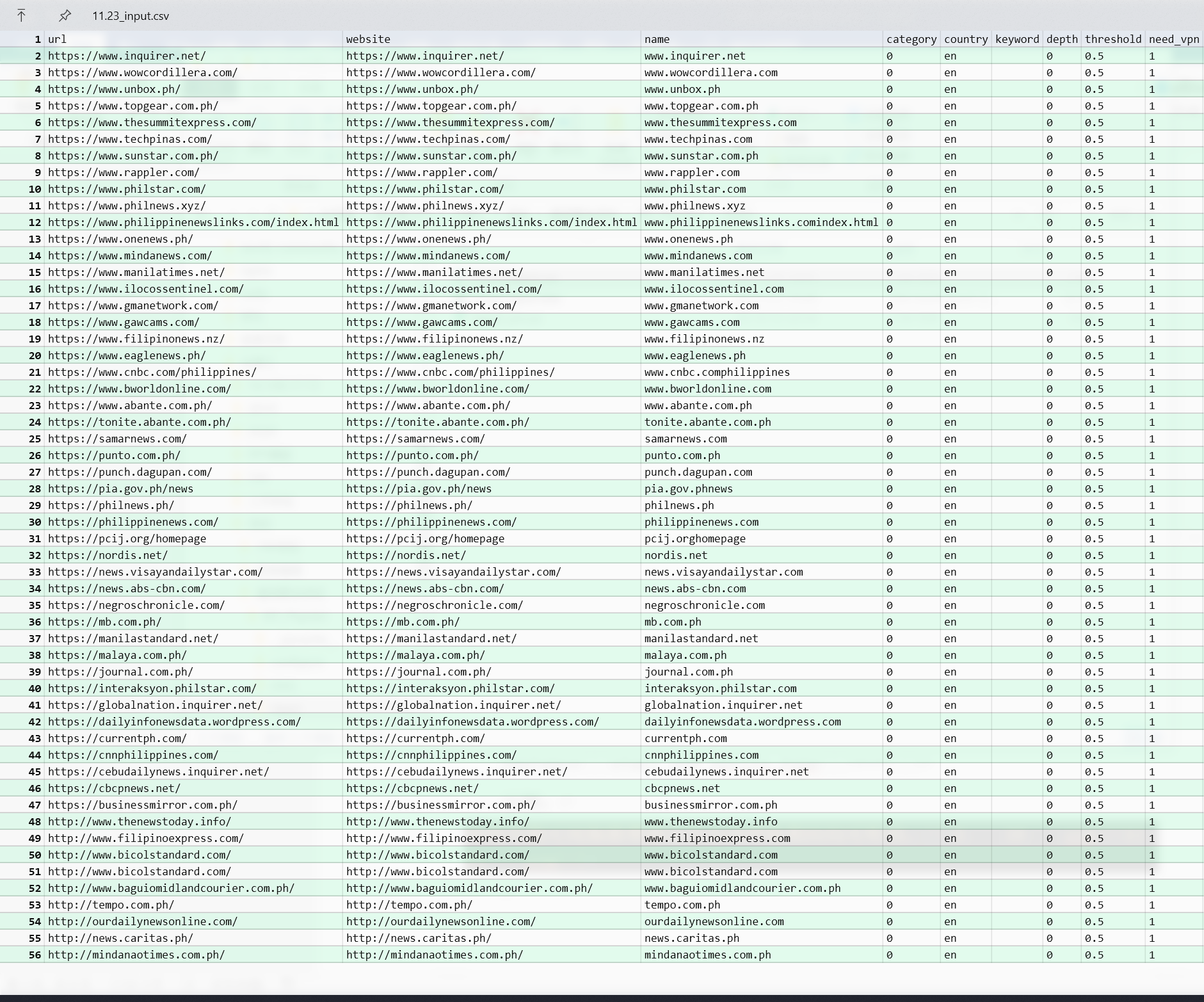
#### 在input目录，然后新建csv文件或者使用已有csv文件，将数据按照指定格式编写，并修改配置文件里的输入文件路径和文件名。

#### 数据集格式

Excel格式示例，需要导出为csv作为输入文件：



Csv格式示例：



输入数据文件必须是csv格式。

列头是['url', 'website', 'name', 'category', 'country', 'keyword', 'depth', 'threshold', 'need\_vpn']

url：网址

website：网址所属域，根地址

name：网址名称

category：网址类型

country：语种

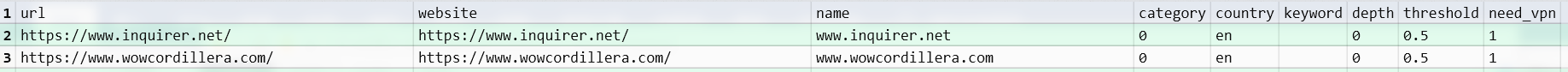
keyword：关键词表

depth：url在website的深度

threshold：相关性阈值

need\_vpn：是否需要翻墙访问

示例：



#### 运行测试

