**爬虫和某机构主页检索系统说明文档**

爬虫模块：

* 1. 爬虫模块介绍：

该模块用于爬取指定机构下的所有静态网页（html或htm后缀）并提取网页源码，然后将网页源码保存到指定的硬盘目录中。

* 1. 爬虫模块初始化：

要实现对指定机构下的所有静态网页进行爬取，首先要对爬虫模块初始化以创建一个爬虫类的实例来支持后续操作，可以通过如下的接口创建一个爬虫类实例：

**\_\_init\_\_(base\_url, output\_dir)**

|  |  |  |
| --- | --- | --- |
| 参数 | 参数类型 | 描述 |
| base\_url | string | 爬取的起始URL |
| output\_dir | string | 网页保存的本地目录 |

该接口无返回值。