**CNKI爬虫对接方案**

最后修改：8/31/2023 4:07 PM Ver.0.1

1. **通用说明**

* **所有接口参数、返回值等定义编程实现时全部小写字母。(某些语言中大小写敏感)**
* **RESTful 服务通讯全部采用JSON**
* **按照面向服务领域概念，爬虫是服务提供者，客户端是服务请求者。**

1. **接口设计**

|  |  |
| --- | --- |
| 爬虫任务启动 | /cnkiSpider/newtask |
| 爬虫任务状态查询 | /cnkiSpider/getState |
| 爬取结果数据获取 | /cnkiSpider/getResult |

* /cnkiSpider/newtask

**描述：**根据检索词，启动爬虫任务

**HTTP请求方式：**POST

**请求参数：**

|  |  |  |
| --- | --- | --- |
| 参数 | 类型 | 说明 |
| access\_token | string | 服务密钥，每个客户端唯一一个密钥，由管理员统一分配。 |
| keywords | string | 关键词检索定义字符串  检索词，多个关键词用空格或逗号分隔 |
| requester | string | 服务请求者标识GUID |
| task\_type | string | 任务类型：realtime/batch  realtime：立即启动爬取（紧急立刻）  batch:加入抓取队列，非实时获取 |
| maxpage | int | 爬取页面数量，默认50，0表示取默认最大值（爬虫自定） |

**返回值：**

|  |  |  |
| --- | --- | --- |
| 参数 | 类型 | 说明 |
| code | string | 状态码  操作成功：100  系统内部错误：200 |
| task\_guid | string | 后台返回的任务GUID  服务请求者通过这个id可以和爬虫服务提供者通信，一个任务有唯一一个guid |
| msg | string | 返回信息（操作成功/操作失败+报错信息） |
| status | string | 添加任务成功且立即执行：0  添加成功已有爬虫任务未启动：1（实时队列）  批处理任务添加成功：2 |

* /cnkiSpider/getState

**描述：**获取爬虫任务的当前状态

**HTTP请求方式：**POST

**请求参数：**

|  |  |  |
| --- | --- | --- |
| 参数 | 类型 | 说明 |
| access\_token | string | 服务密钥 |
| task\_guid | string | 查询的任务GUID |

**返回值：**

|  |  |  |
| --- | --- | --- |
| 参数 | 类型 | 说明 |
| code | string | 状态码  操作成功：100  系统内部错误：200  任务GUID不存在：300 |
| msg | string | 返回信息 |
| data | string | 爬虫任务状态（运行完成100/正在爬取200/正在等待300、已成功爬取数据量、爬取失败数据量、计划爬取总数据量） |

* /cnkiSpider/getResult

**描述：**获取爬虫任务结果数据

**HTTP请求方式：**POST

**请求参数：**

|  |  |  |
| --- | --- | --- |
| 参数 | 类型 | 说明 |
| access\_token | string | 服务密钥 |
| task\_guid | string | 任务GUID |

**返回值：**

|  |  |  |
| --- | --- | --- |
| 参数 | 类型 | 说明 |
| code | string | 状态码 |
| msg | string | 返回信息 |
| data | string | 爬取数据结果JSON数组/列表，每个元素是一个文献结果。  一个文献的结果参考CNKI的检索结果页定义 |

1. **数据设计**
2. **爬虫程序自身任务存储设计**
3. **一个任务，用任务GUID建立一个目录，里面包括：**
4. 任务json（要记录任务执行状态信息）：spider.json
5. 检索结果列表json（cnki的检索列表页，包括爬虫抓取时间等信息），results.json
6. 文献题录信息json（根据检索列表页，一个一个的抓取的文献信息，相当于数据库）:BiB.json
7. 通讯记录json：log.json ，记录请求者查询状态、获取结果、结果返回内容的日志信息（任务启动100/状态查询200/结果获取300/异常报错400/某文献爬取失败500）
8. **所有爬虫任务记录task.json**

Task.json相当于爬虫本地的任务的总表，记录：

* 任务信息：任务GUID、请求者guid、请求者access\_token、任务类型、最大数量等。
* 执行信息：首次开始执行时间，任务完成时间、任务当前状态等。注：任务可能出错多次执行，执行信息记入任务自己的task.json和log.json

1. **Access\_token.json**

包括：token\_guid、客户名称、客户guid、是否禁用。

此json为客户端配置，已经配置且未禁用的access\_token才可以访问。

注意和newtask对应，实现客户端验证

1. **Requester.json**

记录每个客户端的统计信息：客户guid、最近一次任务提交时间、累计提交任务次数、累计成功任务次数、累计失败任务次数、累计获取结果次数、累计查询状态次数等。