Nature数据导入工具使用说明文档

1. 前言

Nature数据导入工具用于提取Nature XML文章数据持久化到数据库，同时将PDF文章内容写入到指定文件系统，过程包含数据查重、日志记录等任务。

## 解压工具

解压程序ZIP包到任意支持JDK8+环境的操作系统。

## 目录说明

解压后获得目录：

|  |
| --- |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 名称 | 用途 | 备注 |
| 1 | tutor.txt | 简要使用步骤 |  |
| 2 | nature-data-import-fat.jar | 可执行程序jar |  |
| 3 | config.properties | 配置文件 |  |
| 4 | bin | 执行脚本目录 |  |

## 配置参数

使用任意编辑器打开配置文件config.properties，按照参数说明，指定必要参数：

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 名称 | 用途 | 备注 |
| 1 | dbHost | Nature数据库SqlServer实例所在主机的地址 | 格式IP:PORT  例如：192.168.2.51:1433 |
| 2 | dbName | Nature数据库名称 |  |
| 3 | username | Nature数据库登录用户名 | 要求：用户的角色拥有查询、删除、插入、创建PROCEDURE的权限。 |
| 4 | password | Nature数据库登录口令 |  |
| 5 | zipFileDir | Nature数据供应商提供的ZIP文件所在文件系统的绝对路径 | 应和该程序同属于一个操作系统。 |
| 6 | pdfBaseDir | Nature文章PDF写入的基目录 | 例如：  Windows OS(注意两个反斜杠):  Z:\\qikan-fulltext\\nature  Linux OS：  /mnt/data/nature |

## 运行

按照要求指定必要参数之后，进入bin目录，双击nature-data-import-start.bat

即刻，操作系统会打开控制台窗体，观察控制台程序执行过程和同步日志输出。

几个重要的日志输出：

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 名称 | 用途 | 备注 |
| 1 | ===Begin=== | 开始 |  |
| 2 | {TotalCount=2, SuccessCount=2, FailCount=0} | Nature资源ZIP文件所含文章数、本次导入成功数、本次导入失败数。 |  |
| 3 | ===Success=== | 本次运行完全成功 |  |
| 4 | ===End==== | 结束 |  |
| 5 | ERROR | 运行出现错误 | 错误也可以到运行开始后生成的logs-error目录查看。若出现错误，请根据错误输出，排查和解决后再次重新执行，直到成功为止。 |

## 日志

程序在执行后自动生成以下日志目录：

|  |
| --- |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 名称 | 用途 | 备注 |
| 1 | logs | 所有日志输出 |  |
| 2 | logs-error | 错误日志输出 | **重要级别：**若出现错误，请根据错误输出，排查和解决后再次重新执行，直到成功为止。 |
| 3 | logs-warn | 警告日志输出 | 一般不影响 |
| 4 | logs-debug | 程序开发者观察日志 | 用于排查程序逻辑 |

## 若出现错误，排查解决错误，再次运行(数据完整性不受影响)，直到成功。