# 功能概述

1. 使用多进程对上交所技术官网所有第一层超链接资源实现无重复并行下载。
2. 支持在远程机器部署、执行。

# 处理流程

主要分为remoteScript、localScript和main。remoteScript为远程机器部署、执行脚本，localScript为本机多进程执行脚本，main为单进程爬虫程序。

## remoteScript流程



## localScript流程



## main流程



# 包设计

## pathlib

包含一些与path操作相关的函数。

### 函数接口

|  |  |
| --- | --- |
| 函数接口 | 描述 |
| ConvertBackslashToSlash | 将“\\”转换为“/” |
| GetParentPath | 获取上级路径 |
| GetModulePath | 获取当前模块路径（工作目录需要在该模块中） |
| GetURLPath | 将url转换为路径，例如："https://a/b" => "a/b" |

## remotessh

基于ssh提供与远程机器进行命令行交互的功能。

### 结构设计



### 函数接口

|  |  |
| --- | --- |
| 函数接口 | 描述 |
| NewInteractor | 获取与ssh服务器交互的Interator |

## remotesftp

基于sftp提供与远程机器进行文件传输的功能。

### 结构设计



### 函数接口

|  |  |
| --- | --- |
| 函数接口 | 描述 |
| NewInteractor | 获取与服务器进行sftp交互的Interator |

## osutils

封装与os交互的功能，启动一个或多个进程、监听系统信号。

### 结构设计



### 函数接口

|  |  |
| --- | --- |
| 函数接口 | 描述 |
| NewCommand | 创建Command对象 |
| NewMultiProcCmd | 创建多进程执行的MultiProcCmd对象 |
| ListenSystemSignalsWithCtx | 利用上下文监听系统信号 |

### 状态图



## taskutils

任务调度执行工具。每个任务会有前置任务，所有前置任务均已完成，当前任务才能被执行。任务池根据任务状态进行任务的执行。

### 结构设计



### 函数接口

|  |  |
| --- | --- |
| 函数接口 | 描述 |
| NewTask | 创建新的Task |
| NewTaskPool | 创建新的TaskPool |

### 状态图

1. Task状态



1. TaskPool状态



### 流程图



## etcd

封装与etcd服务器的交互功能。

### 结构设计



### 函数接口

|  |  |
| --- | --- |
| 函数接口 | 描述 |
| GetClientConfig | 读取json格式的etcd配置文件 |
| NewInteractorWithEmbed | 获取与嵌入etcd交互的Interactor |
| NewInteractor | 获取与etcd交互的Interactor |

## crawler

包含与html处理相关的函数

### 函数接口

|  |  |
| --- | --- |
| 函数接口 | 描述 |
| GetElementNodesFromURL | 从html中获取指定类型的标签节点 |
| GetElementAttributeValue | 从标签节点中获取指定属性的值 |
| FilterURL | 根据规则对获取到的url进行过滤 |
| DownloadURL | 下载url内容至指定位置 |
| WriteToFile | 将内容写入指定文件中 |
| CreateFile | 创建指定文件 |