网站镜像

简介

将一个网站的镜像到本地，主要用于学习和提升，涉及到python，数据库，多线程，锁，字符编码，http规范等知识点。目前仅在linux平台测试运行过。

功能：

1. 本程序目前支持断点下载，就是如果程序在运行中意外终止，重新运行就可以继续恢复到之前状态，不用重新再去运行程序。
2. 支持多配置文件，通过在程序运行时指定不同的配置文件，就可以通过运行多个不同的任务并行下载.
3. 通过自定义线程池，可达到在所有链接访问一遍之后，自动停止运行并推出。
4. 编码自适应，通过解析response响应头的数据和网页中的meta信息来筛选出最符合当前网页的编码
5. 不仅能够解析出html中的url，同时也支持解析css中的URL
6. 可指定运行目录，如果指定运行目录，在数据和日志就会输出到指定目录下

环境配置

1. 系统：Linux 或 Mac OS
2. 数据库: mysql
3. Python3, pip3, 开发环境是3.6.4

使用方式

1. 首先需要初始化环境，通过运行 bin/init.sh脚本来初始化环境, 目的是创建mysql数据库和表，初始化python3虚拟环境
2. 配置文件, 主要关注site::key, site::domain, site::start*urls, site::thread*cnt, log::path, mysql::*, 这几个配置具体含义在示例中都有说明。
3. 运行方式, bash bin/mirror.sh, 如果不加参数则使用 config/conf.ini配置.
    1. `bash bin/mirror.sh -c /tmp/conf.ini` 表示使用/tmp/conf.ini配置文件

2. `bash bin/mirror.sh -c /tmp/conf-dygang.ini`

3. `bash bin/mirror.sh -d /tmp/` 表示/tmp/作为执行目录，日志和下载的数据会存储到这个目录下，并且会优先从这个目录下 `/tmp/config` 寻找配置文件

4. `bash bin/mirror.sh -d /tmp/ -c /tmp1/conf-test.ini` 表示/tmp/作为执行目录，日志和下载的数据会存储到这个目录下，但是会优先使用 `/tmp1/conf-test.ini`

5. 但是指定本次程序中加载的配置文件 `*.ini` 其中如果配置日志和数据路径为绝对路径，则数据会存储到绝对路径下。