网页通知爬虫使用说明书

机笑

2024 年 10 月 27 日

摘要

本文为本人开发的山东大学网页通知爬虫使用说明书.

目录

1 项目简介 **1**

1.1 地址及依赖 1

1.2 功能 1

2 功能介绍 **1**

2.1 操作选项 1

2.2 爬取操作 1

2.3 功能操作 2

3 其他 **2**

3.1 输入与输出 2

3.1.1 输入 2

3.2 输出 2

3.3 定时自动爬取功能 2

3.3.1 爬取时间 2

3.3.2 操作 2

1 项目简介

1.1 地址及依赖

项目仓库地址为 https://github.com/Arshtyi/T4

仅依赖于 Python12.3 及以上版本.

1.2 功能

该脚本能够爬取以下三个网站发布的通知：

(1) 山大视点-山大要闻

(2) 本科生院-工作通知

(3) 计算机学院-本科教育

包括网站通知的来源，链接，发布时间，标题，简述.

2 功能介绍

2.1 操作选项

运行仓库内 T4_1_code.py 文件后, 将给出对应一定的提示信息供你选择：

(1) 山大视点-山大要闻

(2) 本科生院-工作通知

(3) 计算机学院-本科教育

(4) 在已定时间自动进行三个网站内容的爬取

(5) 立刻停止，同时格式化输出表格

(6) 删除原有表格，新建输出表格

2.2 爬取操作

选择 $1-3$ 后将提示选择期望的爬取页数并立即对指定网站进行爬取. 结果输出为当前目录下的 ret.xlsx 文件.

2.3 功能操作

4 操作使程序进入定时自动爬取，具体见下一节.

5 操作使程序对输出的表格做格式处理后立刻停止.

6 操作将会删除原有表格，新建一个输出表格.

3 其他

3.1 输入与输出

3.1.1 输入

输入只需要依据提示选择需要的功能即可，特别地，定时自动爬取中 Ctrl+C 会让程序在下一次爬取时终止.

3.2 输出

每次启动程序都会重置 ret.xlsx 文件，默认完成爬取后不会进行格式处理，只有进行 5 操作停止后才会对表格做格式处理.

输出的 ret.xlsx **文件在当前目录下，推荐使用** Excel **查看**.

3.3 定时自动爬取功能

3.3.1 爬取时间

自动爬取时间默认为 7、12、18、22 四个时间点.

3.3.2 操作

选择 4 后，将进入自动爬取状态，到达已定时间点后，程序将进行爬取，默认每个网站爬取一页通知.

此时可 Ctrl+C 在下一次爬取时退出.