python爬虫微博24小时热搜_GitHub - Writeup007/weibo_Hot_Search: 微博爬虫：每天定时爬取微博热搜榜的内容，留下互联网人的记忆。...

原创

梧芗 于 2021-02-10 18:47:02 发布 111 收藏
文章标签： python爬虫微博24小时热搜

本文链接：https://blog.csdn.net/weixin_34352414/article/details/113984617

Weibo_Hot_Search

都说互联网人的记忆只有七秒钟，可我却想记录下这七秒钟的记忆。

项目已部署在服务器，会在每天的上午 11 点和晚上11 点定时爬取微博的热搜榜内容，保存为 Markdown 文件格式，然后上传备份到 GitHub 你可以随意下载查看。

不要问我为什么选择 11 这两个时间点，因为个人总感觉这两个时间点左右会有大事件发生。

不管微博热搜上是家事国事天下事，亦或是娱乐八卦是非事，我只是想忠实的记录下来...

运行环境

Python 3.0 +

pip install requests

pip install lxml

pip install bs4

或者执行

pip install -r requirements.txt

进行安装运行所需的环境

运行

请确保你已准备好所需的运行环境

运行方法(任选一种)

在仓库目录下运行 weibo_Hot_Search_bs4.py(新增) 或 weibo_Hot_Search.py

在cmd中执行 python weibo_Hot_Search_bs4.py(新增) 或 python weibo_Hot_Search.py

自动运行：利用 Windows 或 Linux 的任务计划程序实现即可

scrapy版本运行

项目的结构如下

> ├── hotweibo

│ ├── __init__.py

```
│   ├── items.py
│   ├── middlewares.py
│   ├── pipelines.py
│   ├── __pycache__
│   │   ├── __init__.cpython-38.pyc
│   │   ├── items.cpython-38.pyc
│   │   ├── pipelines.cpython-38.pyc
│   │   └── settings.cpython-38.pyc
│   ├── settings.py
│   ├── spiders
│   │   ├── hot.py
│   │   ├── __init__.py
│   │   └── __pycache__
│   │   ├── hot.cpython-38.pyc
│   │   └── __init__.cpython-38.pyc
│   └── TimedTask.py # 可以运行此文件直接启动爬虫
└── scrapy.cfg
```

请确保准备好 MongoDB 环境和 Scrapy 环境

推荐使用 Docker 安装 MongoDB

数据库和集合不需要预先创建

TimedTask.py 用于执行定时爬取,默认为每分钟爬取一次

在linux下可以在TimedTask脚本所在目录执行

nohup python Timer.py >/dev/null 2>&1 &

具体用法可参考这里

生成文件

运行结束后会在当前文件夹下生成以时间命名的文件夹，如下：

2019年11月08日

并且会生成以具体小时为单位的具体时间命名的 Markdown 文件，如下：

2019年11月08日23点.md

接口来源

更新日志

2020年08月08日：

1.将原有保存的 Markdown 文件数据进行整理，保存至新开仓库 weibo_Hot_Search_Data 此仓库以后用作代码更新及保存，不再在此存放数据内容。

声明

本项目的所有数据来源均来自 新浪微博 数据内容及其解释权归新浪微博所有。

License