[interesting-python](https://github.com/Alfred1984/interesting-python)/[CaiXuKun](https://github.com/Alfred1984/interesting-python/tree/master/CaiXuKun)/**CaiXuKun**/

[@Alfred1984](https://github.com/Alfred1984)

[**Alfred1984**](https://github.com/Alfred1984/interesting-python/commits?author=Alfred1984) [Update settings.py](https://github.com/Alfred1984/interesting-python/commit/6403935da6032bbd5b0b51e41e3decc8b7685508)

Latest commit [6403935](https://github.com/Alfred1984/interesting-python/commit/6403935da6032bbd5b0b51e41e3decc8b7685508) 14 days ago

| **Type** | **Name** | **Latest commit message** | **Commit time** |
| --- | --- | --- | --- |
|  | [**..**](https://github.com/Alfred1984/interesting-python/tree/master/CaiXuKun) |  |  |
|  | [spiders](https://github.com/Alfred1984/interesting-python/tree/master/CaiXuKun/CaiXuKun/spiders) | [Update caixukun.py](https://github.com/Alfred1984/interesting-python/commit/fa22bfbbc641cf23788e09d7e3ea7c12adaaeb8d) | 15 days ago |
|  | [\_\_init\_\_.py](https://github.com/Alfred1984/interesting-python/blob/master/CaiXuKun/CaiXuKun/__init__.py) | [Create \_\_init\_\_.py](https://github.com/Alfred1984/interesting-python/commit/344547eb80163f5d1ad3498755377748f38d574e) | 20 days ago |
|  | [items.py](https://github.com/Alfred1984/interesting-python/blob/master/CaiXuKun/CaiXuKun/items.py) | [Add files via upload](https://github.com/Alfred1984/interesting-python/commit/d7f29b747268ca2c620313b3e0289beeb6351448) | 20 days ago |
|  | [middlewares.py](https://github.com/Alfred1984/interesting-python/blob/master/CaiXuKun/CaiXuKun/middlewares.py) | [Add files via upload](https://github.com/Alfred1984/interesting-python/commit/d7f29b747268ca2c620313b3e0289beeb6351448) | 20 days ago |
|  | [pipelines.py](https://github.com/Alfred1984/interesting-python/blob/master/CaiXuKun/CaiXuKun/pipelines.py) | [Update pipelines.py](https://github.com/Alfred1984/interesting-python/commit/63c613731c6c078a20e375ba2eec088e853715e3) | 20 days ago |
|  | [readme.md](https://github.com/Alfred1984/interesting-python/blob/master/CaiXuKun/CaiXuKun/readme.md) | [Update readme.md](https://github.com/Alfred1984/interesting-python/commit/f3495972ef017d4ff71d058fca0ddf221a780838) | 15 days ago |
|  | [settings.py](https://github.com/Alfred1984/interesting-python/blob/master/CaiXuKun/CaiXuKun/settings.py) | [Update settings.py](https://github.com/Alfred1984/interesting-python/commit/6403935da6032bbd5b0b51e41e3decc8b7685508) | 14 days ago |

**readme.md**

**Scarpy爬虫项目文件**

**项目结构**

├── CaiXuKUN

│ ├── \_\_init\_\_.py # 初始文件

│ ├── items.py # 项目项定义文件

│ ├── middlewares.py # 项目中间件文件

│ ├── pipelines.py # 项目管道文件

│ ├── settings.py # 项目设置文件

│ └── spiders # 爬虫

│ ├── \_\_init\_\_.py # 爬虫初始文件

│ └── caixukun.py # 爬虫文件

└── scrapy.cfg # 部署配置文件

**需要注意的问题**

**1. 爬虫名**

* caixukun

**2. 启动爬虫**

* 切换进Scarpy爬虫项目文件目录，命令行输入 scrapy crawl caixukun

**3. 浏览器抓包**

* 爬取的链接：<https://m.weibo.cn/api/statuses/repostTimeline?id=4347741368557605&page=1> ，是通过浏览器抓包得来（微博移动端）
* 4347741368557605是每条微博对应的id，只要通过抓包得到这个id，便可以爬取你想要爬取的任何微博的转发数据
* page的最大参数随转发量不同而变化，需要自行测试

**4. 微博反爬的应对措施**

直接大批量抓取，很快会被微博反爬，很多page抓取不到数据，应对措施有：

* 在middlewares.py里面增加IP代理
* 在settings.py里面设置爬取间隔时间
* 由于我的代理IP质量不高，所以没有配置代理IP，而是在middlewares.py里增加了一个重试类TooManyRequestsRetryMiddleware，只要返回418状态码，整个爬虫暂停1分钟，并在setting.py里面设置了重试次数，当然如果你的代理IP质量好的话，可以直接上代理