# Twitter爬虫说明

## 整体思路

首先确定一些关键字根据关键字搜索发布过该关键词相关的用户名，twitter用户名是用户的唯一性标识符，之后最后，爬取这些用户发布的动态，获取用户发布的动态。

## 运行步骤

### 前提条件

已安装Python环境

已安装必要的依赖库

拥有有效的Cookie信息

参考库：https://github.com/Altimis/Scweet

### 步骤 1: 确定关键字

确定关键字

确定你想要搜索的关键词列表，例如：“区块链”、“人工智能”、“大数据”等。

步骤 2: 确定关键词推特用户唯一性标识符

搜索并存储用户名

使用推特的搜索功能或API，根据关键词搜索发布过相关内容的用户名。

将输出的用户名存储在 username.json 文件中。

### 步骤 3: 根据推特用户唯一性标识符来爬取用户信息

运行 scraping.py 脚本

scraping.py 中的 scrap() 函数将读取 username.json 文件中的用户名。

运行脚本后，浏览器将自动弹出并访问推特页面，你需要输入自己的推特账户和密码。

网页会自动执行对用户主页的访问并收集用户主页信息。

继续中断的爬取

使用 scraping.py 中的 current\_indexes1 = [number1, number2] 参数来控制要爬取的用户在 username.json 中的索引。

如果因网络问题中断自动爬取，可以通过这个索引继续从中断的位置恢复爬取。

输出文件

每个用户的主页动态信息将以 JSON 文件格式存储在 UserPosts 文件夹中。

输出文件的格式为：userPostsInformation0-{UID}.json，例如 userPostsInformation0-2014254514.json。