### （1）爬虫相关原理：

爬虫基本原理是通过编写程序模拟人的行为，自动化地访问网页并提取所需的信息。具体的步骤如下：

发起HTTP请求：通过编写程序发送HTTP请求，模拟浏览器访问网页的过程。本采集程序使用的是requests工具发送网络请求。

获取网页内容：程序接收到网页的响应后，可以获取到网页提供的接口数据。使用json解析库解析到相关信息。

提取信息：根据需要，从网页的接口中提取出所需的信息。可以通过查找特定的标签、属性或者内容来定位所需的信息。

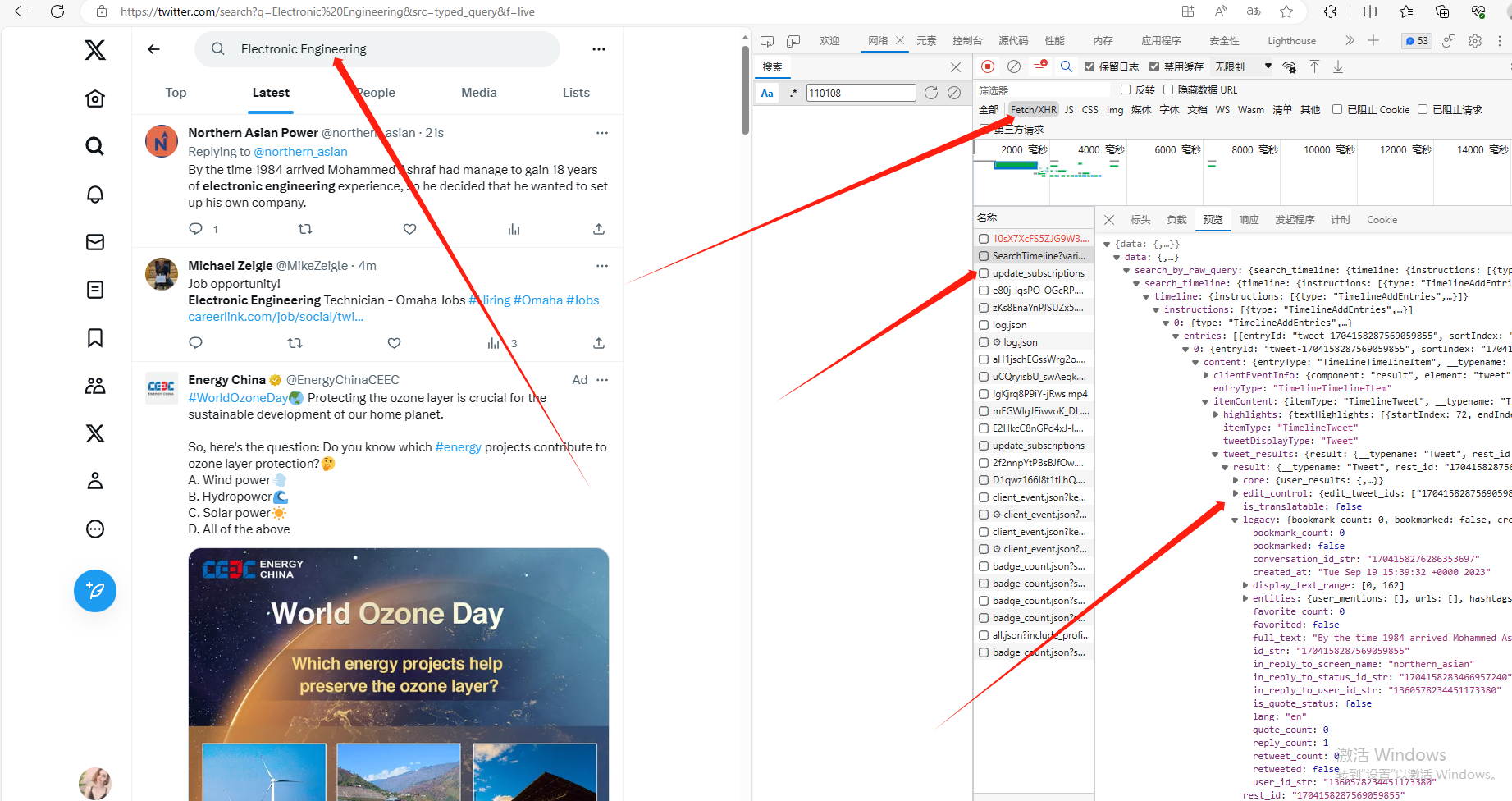
存储数据：将提取到的信息存储到本地文件或者数据库中，以供后续处理和分析。

遍历链接：如果需要爬取多个页码比如下一页，可以在程序中编写逻辑来遍历链接，自动发起HTTP请求并提取信息。

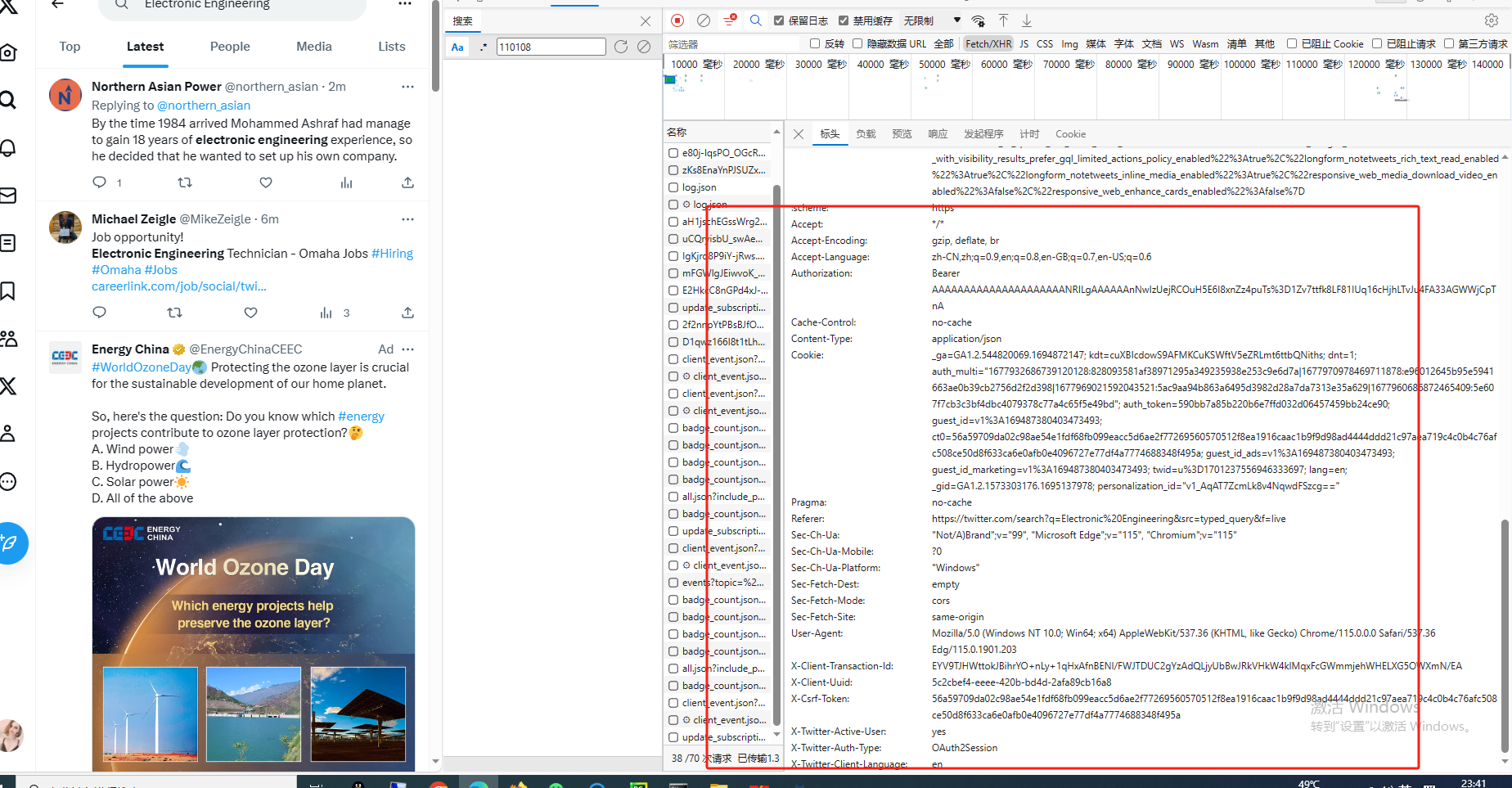
### 通过浏览器开发者工具，查看数据所在接口，

通过f12打开浏览器控制台，然后输入关键字，通过下拉页面，查看浏览器发送了什么包，大概这个包就是需要采集的接口信息，定位到网页接扣分析请求头和请求参数

1 . 查看数据所在接口操作如下：

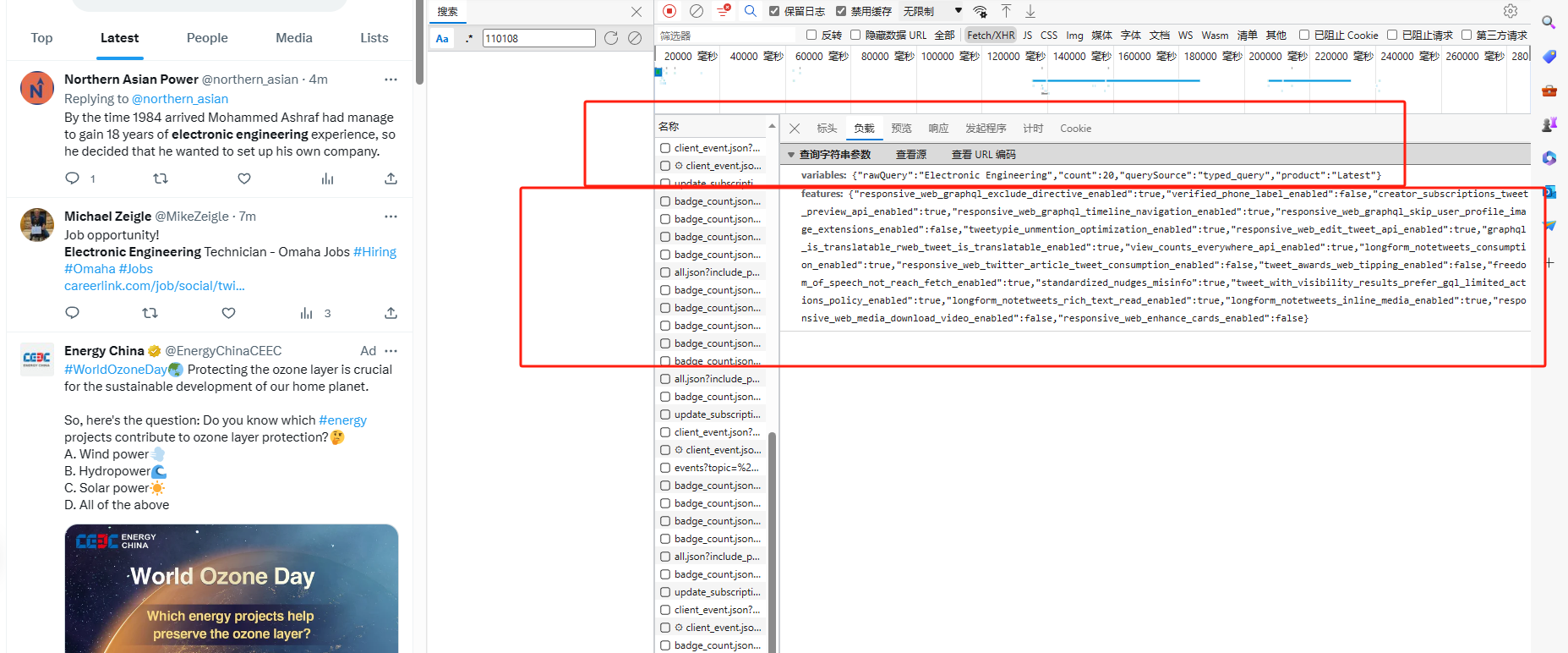


1. 查看请求头信息如下：这些请求头字段，都需要作为requests发包所需要的字段



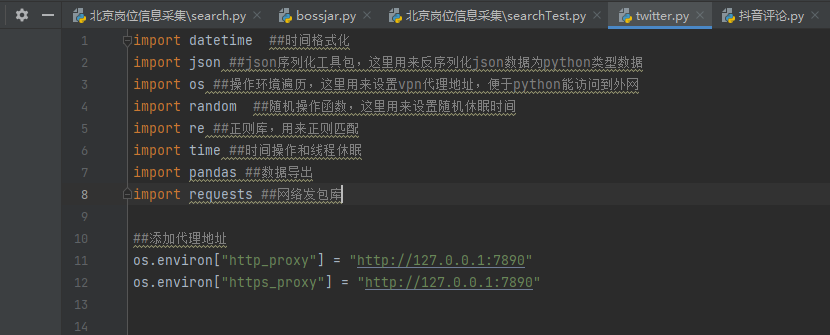
1. 查看请求参数，请求参数主要分为variables和feauters两部分

通过多次浏览器发包的规则可以看出，variables中rawQuery参数控制了查询条件，count为每页数量、query\_source为查询类型，cursor是下一页参数、product为分类

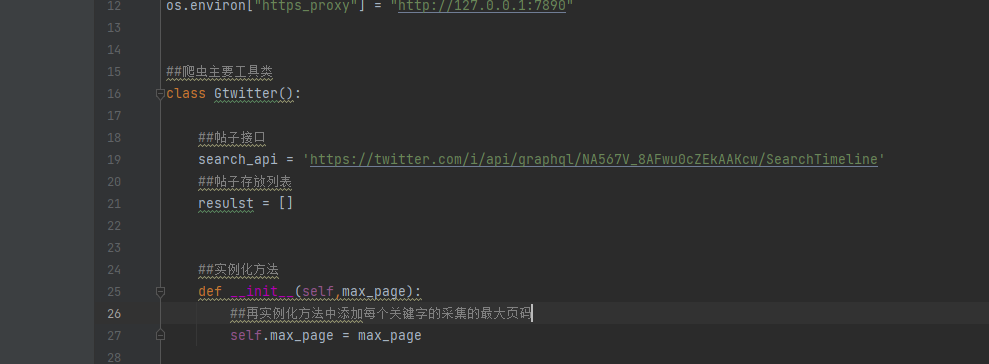


### 数据采集编码

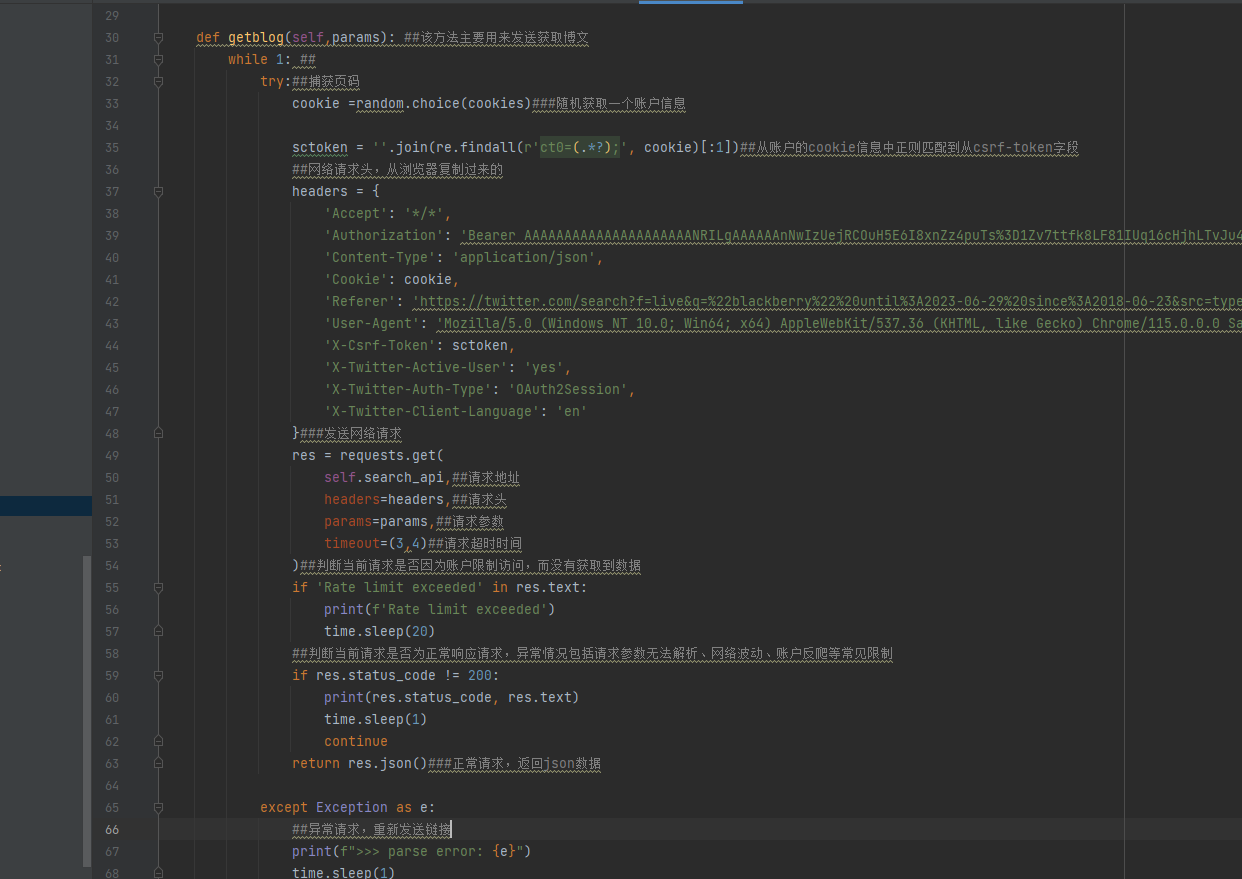
1. 导入相关库，库的作用和本爬虫作用说明如下



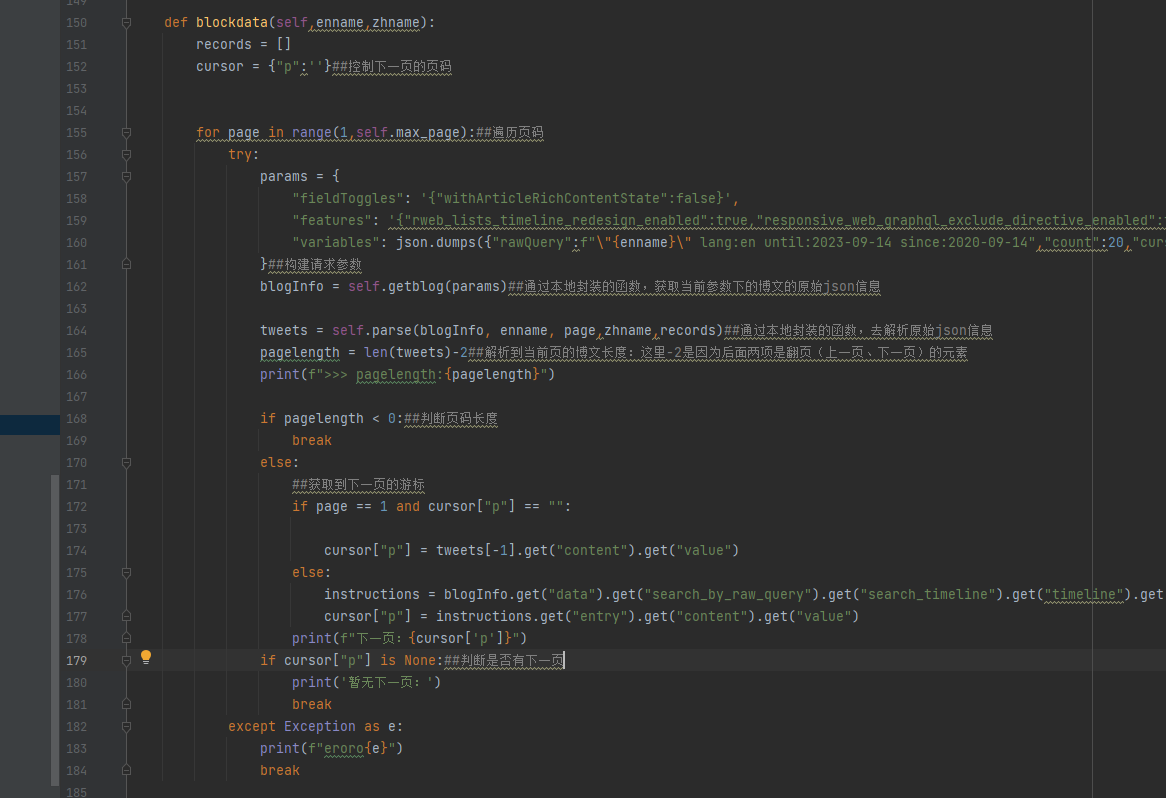
1. 定义爬虫类和实例化方法



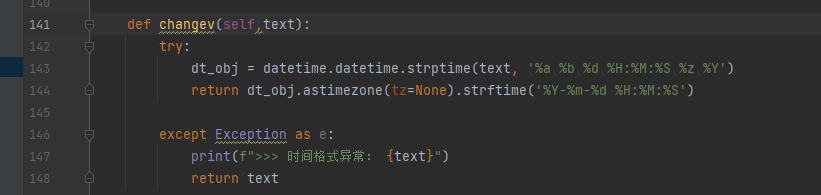
1. 构建一个能捕获异常的发包函数



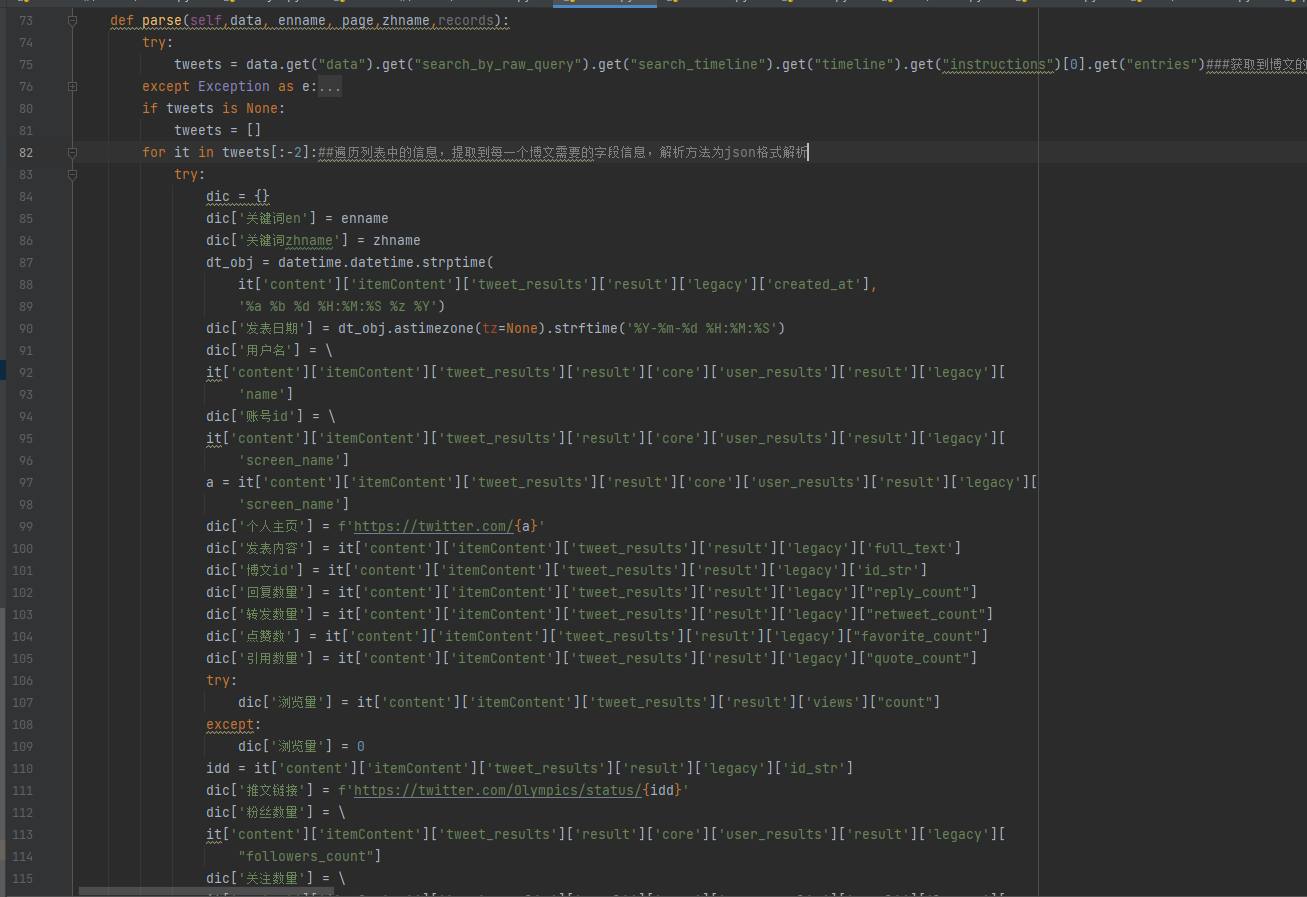
1. 定义一个方法，用来实现翻页操作



1. 定义一个函数，用来将twitter时间转换为标准的时间格式



1. 定义一个函数，用来解析网络原始接口中的博文信息



1. 调用类的方法，完成采集工作



### （4）相关反爬与解决方法：

1 . 相关反爬：常见的反爬有封ip，限流账号、接口保护如js加密等，再本网站中，网站对数据的保护措施为账号限流，即再一定的时间段，如十五分钟内，只能对网站进行45次下拉翻页操作，

1. 反爬解决措施，由于采集的数据量大，所以在限流的情况下无法完成数据采集工作，需要使用到大量的账号，通过购买第三方推特账号搭建账号池，以便能完整的采集数据，相关第三方twitter购买地址如下：https://www.accstw.com/