# 爬虫-----构建新闻分类数据集

一个深度学习 AI 项目是从收集和整理数据集开始的。正如我们开始写一本书或开始最准备做一件事情，总是从收集素材收集资料开始，完成一个 AI 项目也不例外。但一般的 TensorFlow 教程都是使用之前别人已经做好的数据集，本项目中提供的新闻分类数据集信息有限，可以找到的新闻分类数据集多为前几年的数据集，因此有必要构建一个集合较大，主流新闻的数据集。本项目基于python及相关库文件完成。

## 构建数据集前的准备

为了方便获取新闻分类数据集，在收集数据集前，应调研目前互联网中主流新闻网站的新闻分类情况，分析可获得的新闻量和网页元素排列。选取较为统一规范的新闻网站，以便构建新闻分类数据集。本次调研的主流新闻网站有：新浪新闻网、腾讯新闻网、环球新闻网、网易新闻网、人民新闻网、17173网络游戏门户站、澎湃新闻网。

### 分析并获得新闻分类label

在新闻网站首页获取新闻url很简单，但获取的新闻url为未知标签且很难从新闻页面获取分类。根据对主流新闻网站的调研后，发现主流新闻网站的首页有清晰的分类栏（图1），根据分类栏的连接，可以观察到不同分类标签的url不同。



图1 新浪首页分类栏

例如：新浪体育的域名为：sports.sina.com.cn，

新浪娱乐的域名为：ent.sina.com.cn，

环球网军事滚动API 域名为：mil.huanqiu.com，

环球网汽车滚动API 域名为：auto.huanqiu.com。

利用这一特点，在获取的新闻url中解析域名或下属网址，可以很方便的获取每个新闻url的分类label。

### 分析新闻的获得方式及可获得的新闻量

对于深度学习AI 项目而言，数据集的大小对模型的准确率极为重要。因此，在选择主流新闻网站时，应该考虑该网站的新闻数据获取方式和可获得的新闻量。这里有两个获取新闻url的方法，其一，分析网页中的元素，提取新闻url；其二，加载网站时抓包，在Network里面抓取到json新闻url列表。下面以新浪新闻网为例：

#### （1）直接提取网页元素

在新浪首页（https://news.sina.com.cn/）中可以直接提取网页元素，得到新闻url和对应的label。但是一页新闻量很少，可以在url中添加信息，查询新浪历史首页的新闻。

例如：2021年5月11日上午：<https://news.sina.com.cn/head/news20210511am.shtml>

2021年5月13日下午：[https://news.sina.com.cn/head/news20210513pm.shtml](https://news.sina.com.cn/head/news20210511am.shtml)

这里就可根据预设定的信息爬取前几年的首页新闻，即可获得较多新闻量。

#### （2）获取滚动json新闻

在后续爬取新闻之后，发现军事、汽车、房产、游戏等label的新闻量较少，因此需要特定标签的细纹，我们又分析了新浪首页的军事url、汽车url、房产url、游戏url入口。爬取对应label滚动API的json新闻。这里获取的API可直接得到json文件，里面含有新闻url列表。（这个方法是在Network找到的）

例如：军事API：<http://mil.news.sina.com.cn/roll/index.d.html?cid=57918&page=3>

汽车API：<https://interface.sina.cn/auto/news/getWapNewsNe>wBycID.d.json?

cid=78590&page=1&limit=20&tagid=0

到此，以及获得了新浪网大量的新闻url并且有了对应的label。

### 分析网页元素排列

在确定网站之前一定要分析各label网页元素的排列，一旦选择了该网站，新闻详细页面的元素排列应大致统一规律，这为后面多线程爬取解析新闻详细内容创造条件。根据HTML和CSS的特点，定位新闻主体内容（title和contents）对应的元素。

例如：新浪网大部分新闻页面的标题都是<h3>标签，大部分新闻内容都是<p>标签。

## 获取新闻数据集详解

创建新闻数据集主要分为两个步骤，第一步是获取新闻url和对应的label列表保存到txt文件。从网站中获取新闻url和label列表，存入对应网站的txt文件中，用于判别下次爬取的新闻url是否与之重复；第二步是多线程获取新闻的详细信息并保存到json文件。从第一步对应网站的txt文件中加载url列表，通过多线程的方式提取新闻的标题和正文内容，最后存入对应网站的json文件。其主要框架如图2所示。

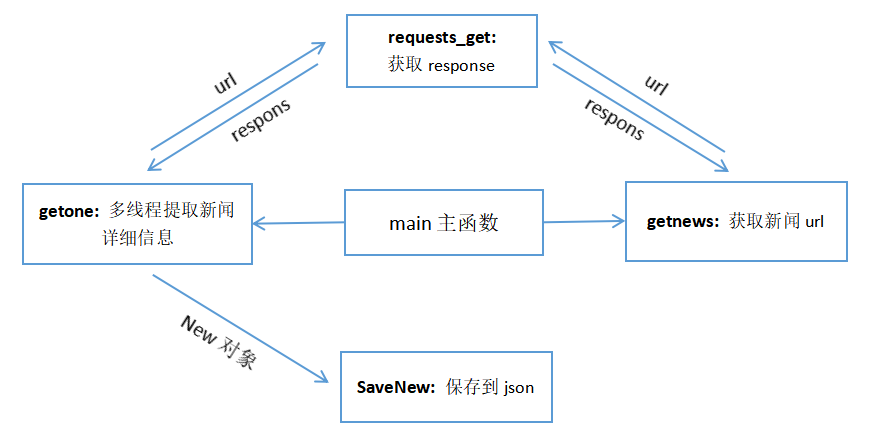


图2 爬虫--获取新闻分类数据集总框架

本程序使用模块化的设计，有利于规划项目的进程，也有利于后期程序的改进和维护。在介绍本程序框架前需要说明，爬取不同网站都遵循了统一框架准则，但是不同网站在getnews和getone模块有所区别。下面来介绍一下统一的模块：

**main函数**主要调用于两个模块（getnews和getone），是程序的核心。

**requests\_get模块**主要作用是获取url的响应response对象，是为了减少代码的赘余。这里使用了python中的requests库并对其进行了改进。

**SaveNew模块**用于保存新闻详细数据到json文件，传递的参数主要有New对象和文件对象。使用了python中的threading库，用于添加线程锁，防止多线程写入文件错乱。

因为不同的网站获取信息的方式不同，所以在getnews和getone模块有一定的区别，但功能和目的都是一样的。详细步骤如下：

### 获取新闻url列表和对应label（getnews）

为了更好的获取新闻url列表和改善项目，将获取新闻url模块单独分出来很有必要。因为新闻url获取条件和数量很受限，后续更新数据集时，数据可能会重复。单独设立一个模块，可以很方便地更新新闻url，为后期多线程爬取新闻详情打下基础，操作较简单。本模块主要利用了第一点第2小点中的两个方法，其主要框架如图3。

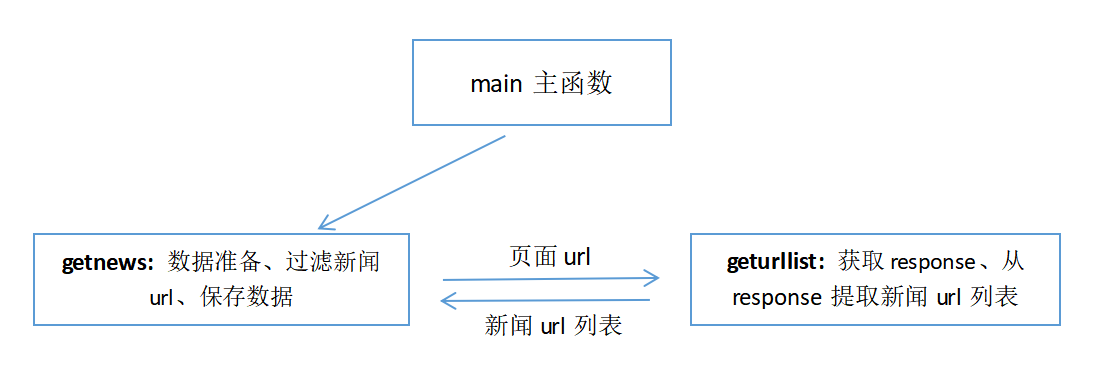


图3 获取新闻url 框架

本模块主要的流程包括数据准备、获取response、提取新闻url列表、过滤新闻url、保存数据。

**数据准备：**根据调研发现新闻网站分类情况，准备各分类新闻页面对应的url，以便获取新闻url列表（可参考第一点的1、2小点）。

**获取response：**直接调用requests\_get模块，传入参数页面url，返回response对象。

**提取新闻url列表：**使用python中的BeautifulSoup库或re库，准确定位并提取response.text中新闻url内容。

**过滤新闻url：**首先排除非新闻url，然后读取对应网站的txt文件，得到以前获取的新闻url列表，将新的新闻url与其对比，去除重复项，得到新的新闻url。

**保存数据：**将新的新闻url列表以追加的方式写入对应网站的txt文件。

为了获取更多的新闻数据，对于爬取网站首页的新闻url（因为一次爬取的新闻量很少少）需要定期爬取，直到获得满意的数据量。至此我们可以获得大量的新闻url和对应的label，为后面多线程获取新闻详细数据提供了便利。

### 多线程提取新闻详细信息（getone）

新闻url保存在了对应的txt文件后，随时可以爬取新闻url提取新闻详细信息并保存在json文件中，建立新闻分类数据集就大致完成了。多线程提取新闻详细信息框架如图4。

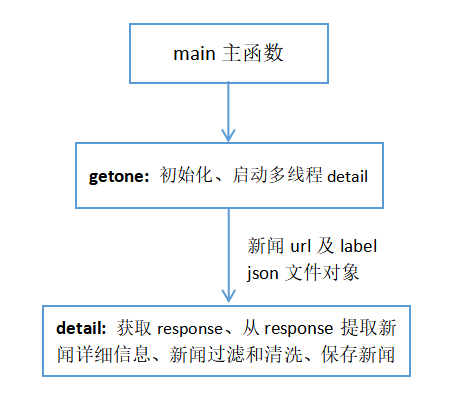


图4 多线程提取新闻详细信息框架

如图4，多线程提取新闻详细信息模块的主要流程有：初始化、启动多线程detail，

最主要工作在detail函数内完成，将新闻详细信息爬取下来并保存。

**初始化：**主要是提取对应网站txt中的新闻信息（包括新闻url及对应的label），创建相应json对象（新闻详情保存的位置），并把所有新闻说传入的detail函数创建一个多线程列表。

**启动多线程：**启动多线程时应控制线程数量，因为硬件性能不够。当进程中的线程超过预定值，后续线程应处于等待状态。

**detail函数：**用于提取新闻标题和内容。首先调用requests\_get模块返回response对象，使用BeautifulSoup库或re库相关方法准确提取response.text中新闻标题和内容。过滤和清理新闻内容，最后调用SaveNew模块保存新闻。

这个模块基本上完成对数据的获取，可以快速完成对数据的抓取。但要得到一个较好的数据集，还要经过完善的数据过滤和清洗，防止“脏数据”对模型精确度的影响。

## 新闻的过滤和清理

在getone模块中提到了新闻数据的过滤和清理，这也是构建数据集的一个重要步骤。因为即使是一个网站内的新闻，页面不规范，爬取下来的信息就会出现乱码或缺失。

#### 网站数据缺失

网站数据缺失即新闻url无效，此时找到的内容较少（非新闻内容）或内容不存在，所以需要淘汰这部分数据。在多线程爬取之前，分析了14632条新闻内容字数的分布，如图5。

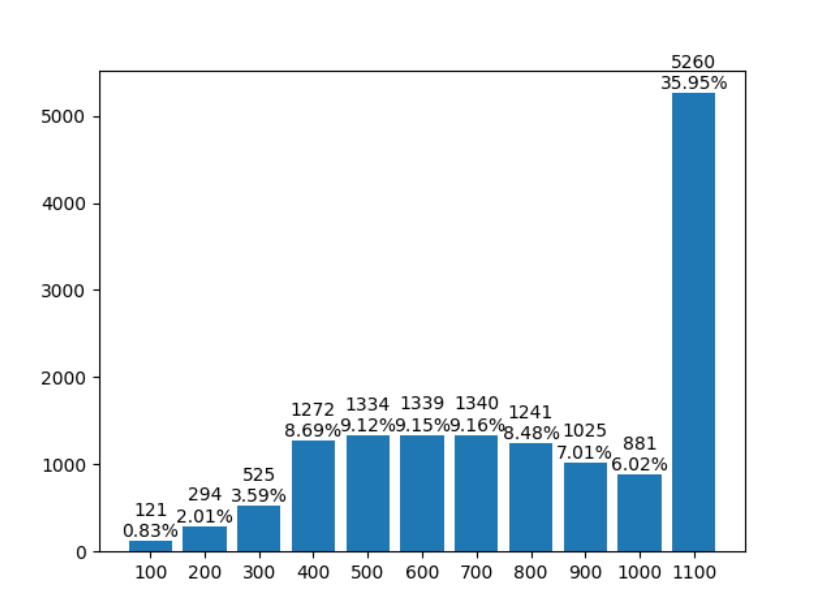


图5 新闻字数分布柱状图

由柱状图得知，新闻字数在100以内的新闻占比很少，对数据集的影响不大，因此可以设定一个字数限制，当新闻的字数少于45字时，舍弃该新闻。由于模型对新闻的字数也有限制，因此我们设置了一个最大字数限制32000。此时可以舍弃大部分失效链接和少量无效内容。

#### 数据出现乱码

在爬取数据时，经常会遇到编码问题，常见的编码有utf-8，GBK，对于爬取少量数据时可以查找编码，但寻找编码过程较缓慢，会给大量数据爬取带来压力，大多数新闻网站的页面编码为utf-8，因此我们只需要预先分析网站的编码，并设定编码即可解决问题。如果有网页不符合这个编码，即可查找编码付给response对象的encoding属性。

#### 数据出现不相关的广告等内容



图5 新闻详细页

如图5，新闻详细页可能会出现广告连接等非新闻主体信息，可以分布在文章前，中，后。因此我们要对每一新闻都要进行筛选。但是发现这种隐蔽性的文字比较困难，这个过程非常费时间，因此我们采用xlsx查看寻找相同项的方法。我主要的清理流程为：初步清理、在xlsx文件中观察搜索相同信息，在代码中添加需要清理匹配内容。重复第二步和第三步。最终得到较为“干净”的数据。

经过清理后的数据集基本上可以直接使用了。