《用 Python 玩转数据》项目—动态新闻标题热点挖掘

Dazhuang@NJU

一、背景

新闻标题是新闻的主旨，从新闻标题中可以进行多种内容的挖掘，例如可以爬取一定时间段内的新闻进行分析获得热点词。新浪各地新闻中的新闻标题形式如下：

- 进博会门票网售1500元一张？上海警方：假的 别信 (11月06日 07:45)
- 赣州人大常委会原主任骆炳峰再获减刑八个月 (10月19日 12:27)
- 辽宁越狱事件调查：部分监狱管理人员非在编干警 (10月16日 21:25)

url：http://roll.news.sina.com.cn/news/gnxw/gdxw1/index_1.shtml

可以通过观察网页源代码，可以发现这些新闻标题和时间都有明显的特征：

```
<li><a href="http://news.sina.com.cn/o/2018-11-06/doc-ihmutuea7351575.shtml" target="_blank">进博会门票网售 1500 元一张？上海警方：假的 别信</a><span>(11 月 06 日 07:45)</span></li>
<li><a href="http://news.sina.com.cn/o/2018-10-19/doc-ifxeuwws5952620.shtml" target="_blank">赣州人大常委会原主任骆炳峰再获减刑八个月</a><span>(10 月 19 日 12:27)</span></li>
<li><a href="http://news.sina.com.cn/s/2018-10-16/doc-ihmhafis0742825.shtml" target="_blank">辽宁越狱事件调查：部分监狱管理人员非在编干警</a><span>(10 月 16 日 21:25)</span></li>
```

可以利用正则表达式方便的获取新闻标题和发布时间，同时观察网页也可以看到 url 是有规律的，所以可以方便地获取多条滚动新闻标题，可以基于这些标题中的热点词构建词云。

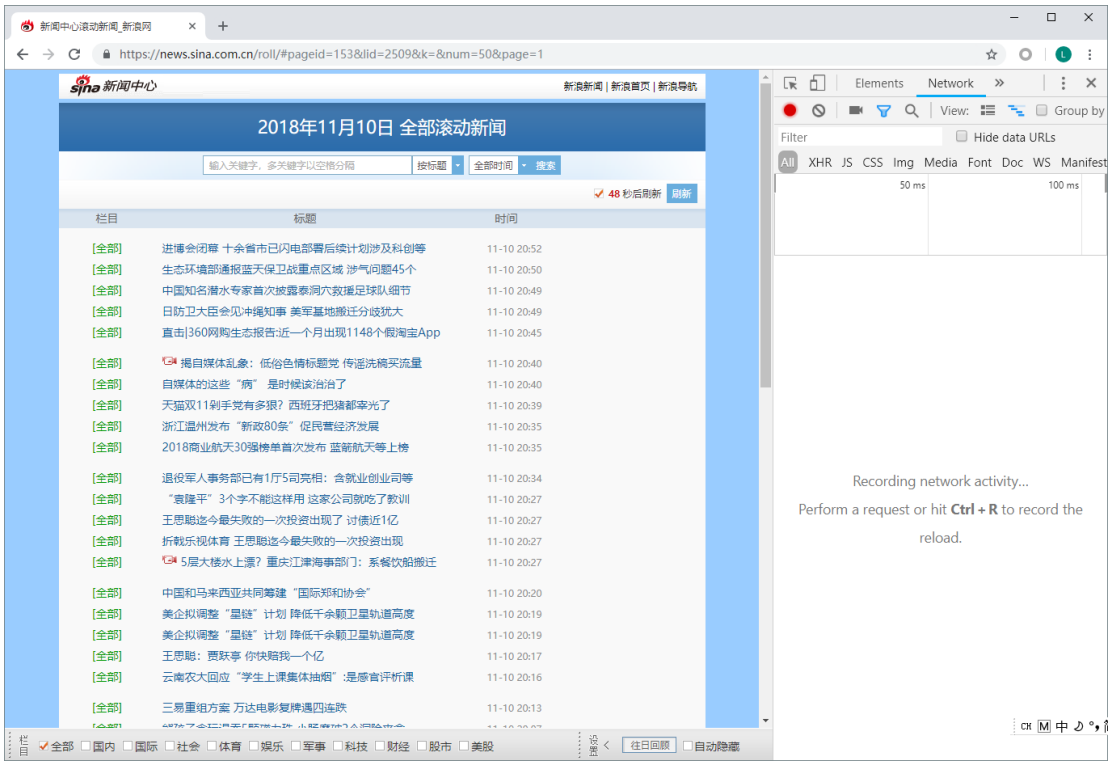
但是并非所有的网站数据都能在源代码中方便地解析到，有的网站由 Ajax 动态生成网页，其数据的获得需要不同的方法，例如新浪全部滚动新闻，页面部分内容示意如下：

栏目	标题	时间
[全部]	小米/诺基亚/三星/索尼/摩托8款机型获得TWRP支持	11-07 07:38
[全部]	腾讯游戏“站在高岗上”：在游戏和电竞上决心不会变	11-07 07:38
[全部]	国元证券出资60亿加入质押纾困队伍 5维度选帮扶对象	11-07 07:37

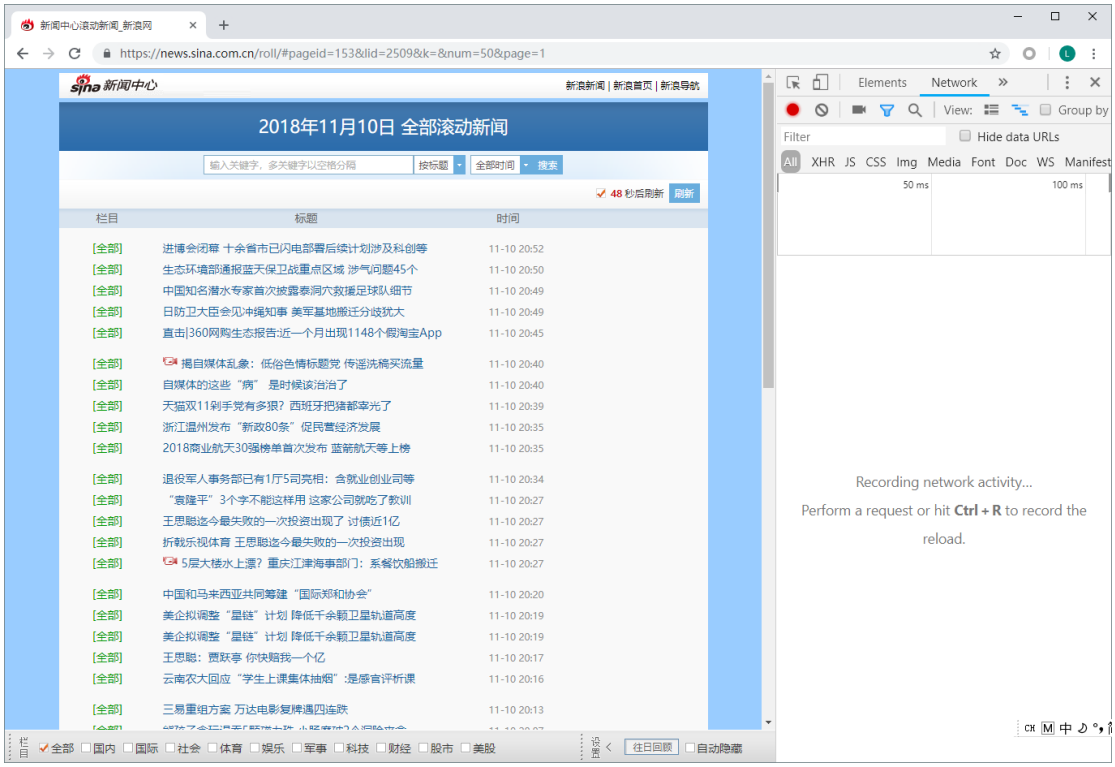
url: https://news.sina.com.cn/roll/#pageid=153&lid=2509&k=&num=50&page=1

页面右上角有一行文字“n 秒后刷新”和一个刷新按钮用于动态生成页面，查看该网页源代码也确实不能找到网页上显示的新闻标题，这种情况就需要用到浏览器的“开发者工具”来进行查看。

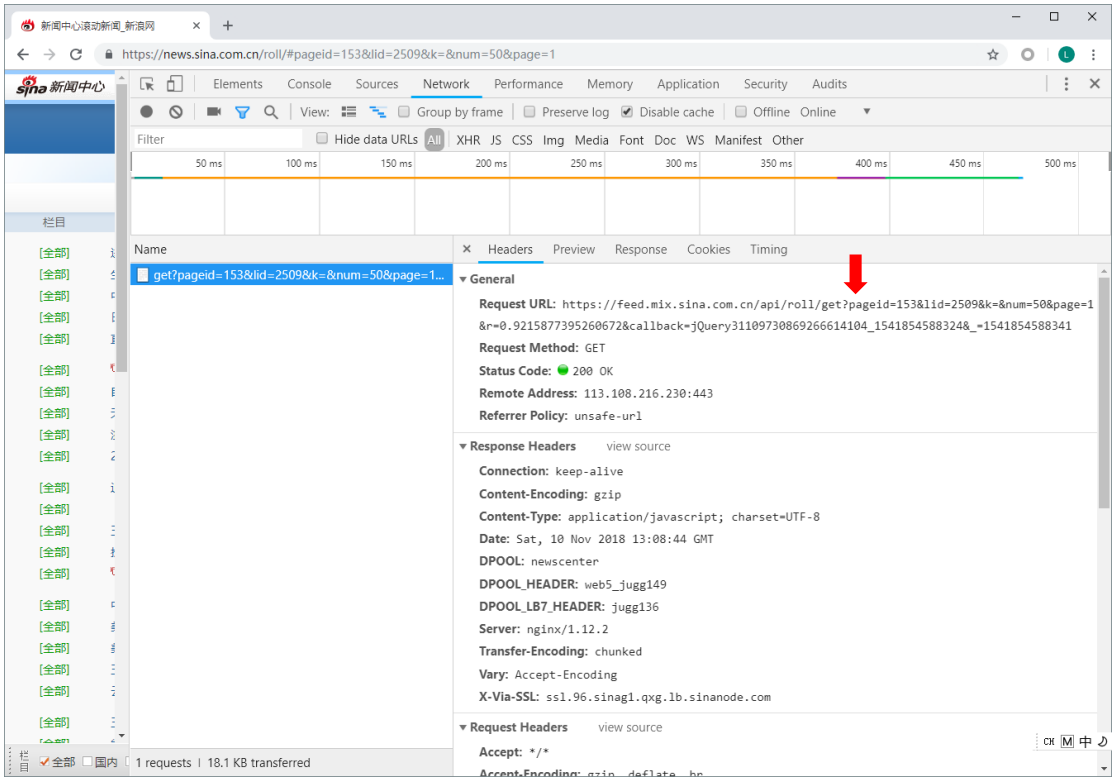
在浏览器中找到“开发者工具”命令，打开后页面如下图所示：



本页面可以等待其自动刷新或者点击“刷新”按钮记录网页数据日志，点击红色的“stop recording network log”按钮结束记录。



可以看到动态产生数据的页面 Name 是“get?pageid=153…&r=…&callback=…”(蓝色背景焦点部分)，观察该页面的 Headers 部分，可以看到 Request URL 字段，此 url 即为真正产生数据的页面，也可在页面下方的“Query String Parameters”中查看，其为一个字典，可以传给 get()或 post()等函数的 params 参数后获取需要的网页。



打开该页面可以看到新闻标题的 unicode 编码，对其做相应转换即可。注意一定要在遵循爬虫协议的前提下使用“开发者工具”。

二、算法

以获取一定时间段内新闻标题中的热点词并绘制词云为例，该算法的主要步骤如下：

1. 从新闻网站爬取若干新闻标题并进行解析
   1.1 利用 Requests 库的 get()函数爬取网页，动态网页需要深入查看
   1.2 找到其中的新闻标题模式
   1.3 利用 re 模块中的 findall()函数提取出标题，将它们存入文件；
2. 标题分词（Text Segmentation）
   要抓热点词首先要将新闻标题进行分词，可利用 Python 中著名的分词器 jieba（结巴分词）
   逐行用 jieba 分词，单行分词的代码如下：
   word_list = pseg.cut(subject)
3. 去除停用词
   很多如“的”和“我们”这样的功能词对于主题分析并无帮助，因此需要使用停用词表进行词的过滤
   代码如下：
   stop_words = set(line.strip() for line in open('stopwords.txt', encoding='utf-8'))
4. 选择名词
   jieba 中的词性标签使用了传统方式，例如"n"是名词，“a”是形容词，“v”是动词等。
   新闻标题中的名词更能代表热点，可以单独选择名词进行后续处理
   选择所有的名词放到一个列表中的代码如下：
   p = re.compile("n[a-z0-9]{0,2}") # n, nr, ns, ... 等都是名词标记

```
        for word, flag in word_list:
            if not word in stop_words and p.search(flag) != None:
                newslist.append(word)
```

5. 根据词频画出词云

手动计算词频：

```
content = {}
for item in newslist:
    content[item] = content.get(item, 0) + 1
```

利用 WordCloud()函数基于词创建词云，这里选择词频最高的 10 个词，代码如下：

```
wordcloud = WordCloud(font_path='simhei.ttf', background_color="grey",
mask=mask_image, max_words=10).generate_from_frequencies(content)
```

其中 simhei.ttf 为字体文件，用于程序运行后词云中词的字体显示。也可以基于一些图的轮廓来设置词云形状，代码如下：

```
d = path.dirname(__file__)
mask_image = imread(path.join(d, "mickey.png"))
plt.imshow(wordcloud)
```

获取当天一页新闻标题热点词的云图如下所示，若要获取多页则可以利用“开发者工具”获取其他产生数据页面的 url：

本词云基于 2019 年 4 月 29 日新浪滚动新闻生成

如果仅仅是统计词频不使用特殊模型，则生成词云也可简单地使用模块中的方法，例如：

```
wordcloud = WordCloud(font_path='simhei.ttf', background_color="grey",
        mask=mask_image, max_words=10)
wordcloud.fit_words(content)
```

三、安装

1. 安装结巴分词器（均在操作系统终端而非 Python 终端中安装）

$ pip install -i https://pypi.tuna.tsinghua.edu.cn/simple jieba

2. 安装词云包

$ pip install -i https://pypi.tuna.tsinghua.edu.cn/simple wordcloud

或

$ conda install -c conda-forge wordcloud

安装词云包 wordcloud 可能遇到编码问题的解决方法

（1） 修改 Python（Anaconda）安装目录下的 lib\site-packages\pip\compat__init__.py 文件，将 75 行附近的“return s.decode('utf-8')”修改成“return s.decode('gb2312')”

（2）在 Anaconda 的 Python 控制台中重启 kernel（单击控制台的齿轮形状的“Options”按钮，在打开的下拉菜单中选择“Restart kernel”命令）

四、参考资料

1. jieba 中文分词器

https://github.com/fxsjy/jieba/

其他相关资料

2. WordCloud 词云

https://amueller.github.io/word_cloud/

其他相关资料

五、参考代码

提示： fetch_sina_news()函数中与几条语句对应部分的注释语句为地方新闻抓取方式，供参考。

```
import jieba.posseg as pseg
import matplotlib.pyplot as plt
from os import path
import re
import requests
# import time
from scipy.misc import imread
from wordcloud import WordCloud

def fetch_sina_news():
    # PATTERN = re.compile('.shtml"
    target="_blank">(.*?)</a><span>(.*?)</span></li>')
    PATTERN = re.compile('"title":(.*?),')
    # BASE_URL = "http://roll.news.sina.com.cn/news/gnxw/gdxw1/index_"
```

```
    BASE_URL =
    'https://feed.mix.sina.com.cn/api/roll/get?pageid=153&lid=2509&k=&nu
    m=50&page=1&r=0.07257693576113322&callback=jQuery1112032872146402846
    9_1556541915945&_=1556541915947'

    # MAX_PAGE_NUM = 10
    with open('subjects.txt', 'w', encoding='utf-8') as f:
        # for i in range(1, MAX_PAGE_NUM):
            # print('Downloading page #{}'.format(i))
            # r = requests.get(BASE_URL + str(i)+'.shtml')
            r = requests.get(BASE_URL)
            # r.encoding='gb2312'
            # data = r.text

            # unicode to utf-8 code
            data = r.text.encode('utf-8').decode('unicode-escape')
            p = re.findall(PATTERN, data)
            for s in p:
                # f.write(s[0])
                f.write(s)
            # time.sleep(5)

def extract_words():
    with open('subjects.txt','r', encoding='utf-8') as f:
        news_subjects = f.readlines()

    stop_words = set(line.strip() for line in open('stopwords.txt',
    encoding='utf-8'))

    newslist = []

    for subject in news_subjects:
        if subject.isspace():
            continue

        # segment words line by line
        # n, nr, ns, ... are the flags of nouns
        p = re.compile("n[a-z0-9]{0,2}")
        word_list = pseg.cut(subject)
        for word, flag in word_list:
            if not word in stop_words and p.search(flag) != None:
                newslist.append(word)

    content = {}
```

```
    for item in newslist:
        content[item] = content.get(item, 0) + 1

    d = path.dirname(__file__)
    mask_image = imread(path.join(d, "mickey.png"))
    wordcloud = WordCloud(font_path='simhei.ttf',
    background_color="grey", mask=mask_image,
    max_words=10).generate_from_frequencies(content)
    # Display the generated image:
    plt.imshow(wordcloud)
    plt.axis("off")
    wordcloud.to_file('wordcloud.jpg')
    plt.show()

if __name__ == "__main__":
    fetch_sina_news()
    extract_words()
```