实验报告

环境配置

- python 3
- requests ， beautifulsoup， lxml第三方库（用于爬虫）
- jieba 第三方库 （用于分词）
- thulac 第三方库 （用于对比分词结果）
- coding: utf-8

关键代码

- 获取网页

```
url = 'http://www.xinhuanet.com/politics/leaders/2019-
09/21/c_1125023359.htm' #网址
headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/77.0.3865.90 Safari/537.36'
                } # 访问头部
response = requests.get(url, headers=headers)
response.encoding = 'utf-8' # 编码
# 调用beautifulsoup库
soup = BeautifulSoup(response.content, 'html.parser')
text = soup.find_all(text=True)
```

- 中文分词

```
f_read = open('web.txt', 'r', encoding='utf-8')
words = f_read.read()     # 从文本读取

jieba.add_word('史竞男')  # 加入默认词汇

paragraph = jieba.cut(words, cut_all=False) # 利用jieba库函数分词
with open('text.txt', 'w', encoding='utf-8')as f_write:
    f_write.write('Default Mode:')
    for i in paragraph:  # 空格和换行不加上‘|’
        if i == ' ' or i == '\n':
            f_write.write(i)
        else:
            f_write.write(i+'|')
```

所遇难题

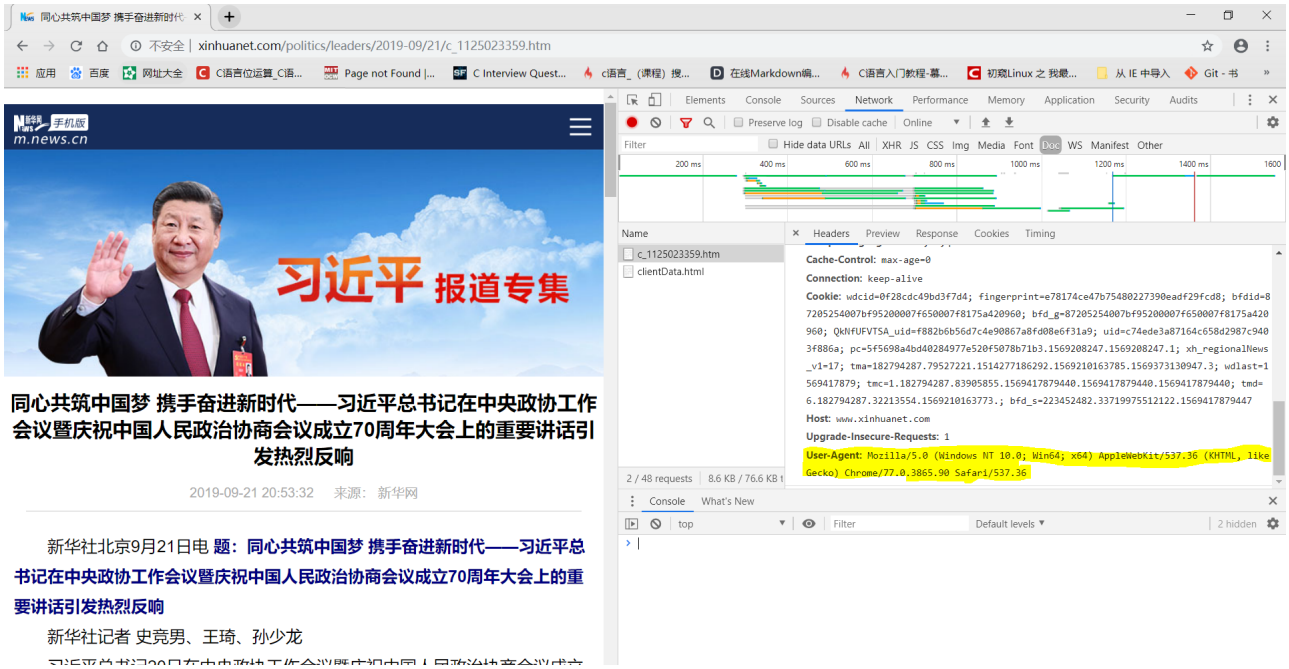
1. 在安装第三方库的时候requests，beautifulsoup ，lxml时，使用

```
pip install xxx
```

安装时出现Could not install packages due to an EnvironmentError报错。
经过一番搜索后，发现似乎是权限的原因。只要在前面加上--user即可。

```
pip install --user xxx
```

2. 在获取指定网页文本时，使用requests的get方法获取网页，得到的总是403forbidden。 同样进行一番学习后，找到了原因。在get方法后加上header，获取自该网页按F12后得到的浏览器头部。
如下图黄色处：



3. 在获取到整个网页的文本后，需要选择出所需的部分，如标题以及文章主题。
在看了网上的blog之后学会了如何查看获取到的text的各个部分。使用代码如下：

```
print(set([t.parent.name for t in text]))
```

然后将所需的部分经过挑选后写入web.txt

```
with open('web.txt', 'w', encoding='utf-8')as f_write:
for i in text:
    if i.parent.name == 'p'or i.parent.name == 'title':
        f_write.write(i.strip(' '))
```

4. 使用第三方库jieba的cut函数，默认为精确模式，进行中文分词。
但是一开始查看结果的时候，不是很满意。许多空格以及换行之间都出现了'|'。

而且记者的名字**史竞男**还被逐字分开。
后来，我到网上了解了jieba的更多方法函数，可以用add_word添加默认单词。

```
jieba.add_word('史竞男')
```

加上之后就不会出现逐字分开的情况了。
此外，还有用户词典用于批量导入。

5. 关于上一点提到的对空格以及换行之间都出现了'|'。我在原有的代码进行了修改。由

```
f_write.write('Default Mpode:'+'|'.join(paragraph))
```

改为

```
for i in paragraph:  # 空格和换行不加上‘|’
        if i == ' ' or i == '\n':
            f_write.write(i)
        else:
            f_write.write(i+'|')
```

但是仍然在每一段的开头都出现了空格+|，查看文本ASCII码才发现那两个空格不是'\n'而是'\u3000',只需在上述代码加上判断即可。
然而，加上后看起来并不变得美观，所以我没有选择加上。

6. 之后我又对一些有歧义的文字进行分词查看结果如下：

```
欢迎|武汉市|长江大桥|莅临指导|
|广州|市长|隆|马戏|欢迎您|
```

我对此感到一丝奇怪，后来想想应该jieba优先对热词进行划分，而忽略生词。

7. 之后，在同学的推荐下，我又下载了其他中文分词工具，如thulac来自清华的中文分词工具，对第6点中的文字进行分词，得到结果如下：

```
import thulac
thu1 = thulac.thulac()  # 默认模式
text = thu1.cut("欢迎武汉市长江大桥莅临指导", text=True)
```

```
欢迎_v 武汉市_ns 长江_ns 大桥_n 莅临_v 指导_v
广州市_ns 长隆_ns 马戏_n 欢迎_v 您_r
```

与jieba的分词结果不一，我想这应该是训练集不同所致，thulac模型是训练于人民日报分词语料库，大概是人民日报中长隆，马戏出现得比较多吧。

8. 关于jieba的cut方法的实现，我也进行了一些了解。大致如下：
   首先使用概率无向图，获得最大概率路径。概率无向图（DAG）的构建完全依赖于字典，最大概率路径求解也是依赖字典中的词频。 最后使用HMM模型来解决未登录词(Out Of Vocabulary) ，所以在整个过程如果没有模型也是可以的，只要你有一个很好的词典。