《用 Python 玩转数据》项目—新闻标题内容挖掘

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：

- 权威解读：从零开始建设的雄安新区未来是啥样？ (04月21日 21:20)
- 雄安新区城乡空间布局：“一主五辅 多节点” (04月21日 21:17)
- 雄安国土空间格局：建设区按1万人/平方公里控制 (04月21日 21:17)

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
   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：

```
for word, flag in word_list:
    if not word in stop_words and flag == 'n':
        newslist.append(word)
```

5. 根据词频画出词云
   将所有的名词直接作为 WordCloud()函数的参数，默认 WordCloud 内部通过统计词频对词进行排序
   代码如下：

```
content = ' '.join(newslist)
```

wordcloud = WordCloud(font_path='simhei.ttf', background_color="grey", mask=mask_image, max_words=40).generate(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)
```

近期新闻标题热点词的云图如下所示：

三、安装

1. 安装结巴分词器

$ 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

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

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

def fetch_sina_news():
    PATTERN = re.compile('.shtml" target="_blank">(.*?)</a><span>(.*?)</span></li>')
    BASE_URL = "http://roll.news.sina.com.cn/news/gnxw/gdxw1/index_"
    MAX_PAGE_NUM = 2

    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.encoding='gb2312'
            data = r.text
            p = re.findall(PATTERN, data)
            for s in p:
                f.write(s[0])
            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
            word_list = pseg.cut(subject)
            for word, flag in word_list:
                if not word in stop_words and flag == 'n':
                    newslist.append(word)

    d = path.dirname(__file__)
    mask_image = imread(path.join(d, "mickey.png"))

    content = ' '.join(newslist)
    wordcloud = WordCloud(font_path='simhei.ttf', background_color="grey",
mask=mask_image, max_words=40).generate(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()
```