词云图

0 引言

词语图，也叫文字云，是对文本出现频率较高的“关键词”予以视觉化的展现，词云图过滤掉大量的低频低质的文本信息，使得浏览者一下就可以知道文章的主旨。

1 模块准备

```
import jieba # 分词模块
import matplotlib.pyplot as plt # 画图模块
from wordcloud import WordCloud # 文字云模块
from scipy.misc import imread # 处理图像的函数，用于读取并处理背景图片
```

3 实现的思路

准备一份需要分析的文本材料，这里选用的是 19 年两会政府工作报告，首先用 jieba 模块对文本材料进行分词处理（即识别出一个个有意义的词语），然后对处理后的材料使用 WordCloud 文字云模块生成相应的词云图片即可

of course，你也可以选择一张背景图片，以此为背景生成特定的云图

4 代码实现

```
def wordcloud():
    """
    背景图片为自定义的一个矩阵
    :return: 词云图
    """
    # 读取词源文件 二进制的形式
    with open("./govreport.txt", "rb") as f:
        t = f.read()  # 保存为 str 类型
    ls = jieba.lcut(t)  # 进行分词
    txt = " ".join(ls)  # 把分词用空格连起来
    # 设置词云的参数
    w = WordCloud(
        font_path="msyh.ttc",  # 设置字体
        width=1000,  # 设置输出的图片宽度
        height=700,  # 设置输出的图片的高度
        background_color="white",)  # 设置输出图片的背景色
    w.generate(txt)  # 生成词云
    w.to_file("./wordColud.png")  # 将图片保存
    return None
```

```
def wordcloud2():
    """
    用指定的图片生成词云图
    :return: 词云图
```

```
    """
    # 词源的文本文件
    wf = "./govreport.txt"
    word_content = open(wf, "r", encoding="utf-8").read().replace("\n", "")
    # 设置背景图片
    img_file = "./map.jpg"

    # 解析背景图片
    mask_img = imread(img_file)

    # 进行分词
    word_cut = jieba.lcut(word_content)
    # 把分词用空格连起来
    word_cut_join = " ".join(word_cut)

    # 设置词云参数
    wc = WordCloud(
        font_path="SIMYOU.TTF",  # 设置字体
        max_words=2000,  # 允许最大的词汇量
        max_font_size=90,  # 设置最大号字体的大小
        mask=mask_img,  # 设置使用的背景图片，这个参数不为空时，width 和 height 会被忽略
        background_color="white",  # 设置输出的图片背景色
    )
    # 生成词云
    wc.generate(word_cut_join)
    # 用于显示图片，需要配合 plt.show()一起使用
    plt.imshow(wc)
    plt.axis("off")  # 去掉坐标轴
    plt.savefig("./wordcloudWithMap.png") # 保存词云图
    plt.show()
    return None
```

5 效果展示

左边的是不带背景图片的词云图，右边是带有中国地图的词云图