实验五：三国演义词云图

“词云”就是对网络文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信息，使浏览网页者只要一眼扫过文本就可以领略文本的**主旨**。

实验目的：

1. 熟悉文件的读取
2. 熟悉中文分词及词频统计
3. 理解数据清洗
4. 熟练使用词云进行文本信息的展示
5. 理解文本处理、文本分析、文本可视化的基本思想

实验内容：

编写程序，找出《三国演义》这本小说的人物关键词，并对“关键词渲染”形成词云图。

实验要求：

在提供的文本中选择其中的一个文本，提取关键词，并对“关键词渲染”形成词云图。

实验步骤：

1．安装 WordCloud 词云

按照最常规的 pip install wordcloud 命令安装

2．熟悉 WordCloud 库（可以通过 help（WordCloud）查看其用法）

WordCloud(font_path=None,width=400,height=200,margin=2,ranks_only=None,prefer_horizontal=0.9, mask=None, scale=1, color_func=None, max_words=200, min_font_size=4, stopwords=None, random_state=None, background_color='black', max_font_size=None, font_step=1, mode='RGB', relative_scaling='auto', regexp=None, collocations=True, colormap=None, normalize_plurals=True, contour_width=0, contour_color='black', repeat=False)

参数	描述
width	指定词云对象生成图片的宽度，默认400像素
height	指定词云对象生成图片的高度，默认200像素
min_font_size	指定词云中字体的最小字号，默认4号
max_font_size	指定词云中字体的最大字号，根据高度自动调节
font_step	指定词云中字体字号的步进间隔，默认为1
font_path	指定字体文件的路径，默认None
max_words	指定词云显示的最大单词数量，默认200
stop_words	指定词云的排除词列表，即不显示的单词列表
mask	指定词云形状，默认为长方形，需要引用imread()函数
background_color	指定词云图片的背景颜色，默认为黑色

步骤 1：配置对象参数

步骤 2：加载词云文本

步骤 3：输出词云文件

WordCloud 库提供的常用的方法：

```
w = wordcloud.WordCloud()      #以 WordCloud 对象为基础进行操作
wc.generate (text)          #根据文本生成词云
to_file(filename)                    #输出文件
```

3.读取文本文件

```
txt = open("threekingdoms.txt", "r", encoding='utf-8').read()
bg_pic=imread(“pic.png”)     #读入背景图（如需按照特定图片样子生成词云图则使用，否则不使用）
```

4.对读取的文本进行分词，使用 jieba.lcut()

```
words    = jieba.lcut(txt)
```

5.词频统计

```
counts = {}
for word in words:
    if len(word) == 1:
        continue
    elif word == "诸葛亮" or word == "孔明曰":
        rword = "孔明"
```

```
        elif word == "关公" or word == "云长":
            rword = "关羽"
        elif word == "玄德" or word == "玄德曰":
            rword = "刘备"
        elif word == "孟德" or word == "丞相":
            rword = "曹操"
        else:
            rword = word
    counts[rword] = counts.get(rword,0) + 1
```

6.设置停用词，排除无效词

```
    excludes = {"将军","却说","荆州","二人","不可","不能","如此"}
    for word in excludes:
    del counts[word]
```

7.对统计的结果按照词频进行降序排序，打印前五十个词

```
    items = list(counts.items())
    items.sort(key=lambda x:x[1], reverse=True)
    for i in range(50):
        word, count = items[i]
        print ("{0:<10}{1:>5}".format(word, count))
    c.append(word)
```

8.生成词云图

```
    text=" ".join(c)
    w=WordCloud(width=600,height=400,font_path="msyh.ttc",mask=bg_pic,
    background_color="white",max_words=50)
    w.generate(text) #生成词云
    #image_colors=ImageColorGenerator(bg_pic)
```

9.显示词云图，并保存到本地

```
    plt.imshow(w)       #用词云图片
    plt.axis('off')    #不显示坐标
    plt.show()              #显示生成的词云图
```

w.to_file("threekingdoms.png")　#保存到本地

10.结果如下所示

曹操	1451
孔明	1383
刘备	1252
关羽	784
张飞	358
商议	344
如何	338
主公	331
军士	317
吕布	300
左右	294
军马	293
赵云	278
引兵	276
次日	271
大喜	268
孙权	264
天下	255
东吴	251
于是	250
今日	243
不敢	239
魏兵	233
陛下	223
一人	221
都督	221