

数据处理的部分

首先数据处理是采用python语言去处理的，所使用的工具为anaconda，该工具是专门用于数据科学使用的，是为广大数据分析师，数据挖掘师，最火的工具。因为代码众多，无法一一解释每一行代码的含义，不过对于每一块的内容，我都写在对应的代码里面了，可以通过该注释大概明白每一块，代码块所代表的意义是什么，这是总的一个介绍的，如果有什么不齐的地方，到时候你再微信过来问我，在服务范围之内，我看到都会帮你解答的。

共为以下几个步骤：

1. 首先删除重复项

```
content = pd.concat([content1, content2, content3, content4])
content.drop_duplicates(keep='first', inplace=True)
content
```

2. 然后筛选中文长度大于100的

```
: df = df[df['正文长度'] >= 100]
df
```

3. 然后把不包含三亚或者海南的文本去除

```
] : df['内容'] = df['内容'].astype(str)
def sjqx(c):
    c = c.replace('\n', '').replace('\r', '')
    c = c.replace(' ', '').replace('前言', '')
    c = c.replace(r"([\uD800-\uDBFF][\uDC00-\uDFFF])", '')
    c = c.strip(" ")
    if '海南' in c or '三亚' in c:
        return c
    else:
        return np.nan
```

4. 然后再对文中内容进行情感分析，因为paddlehub是百度开源的情感分析库，准确率是目前国内最高的

所以无需做图片情感分类器训练语料这一块内容

```
] : import paddlehub as hub
#这里使用了百度开源的成熟NLP模型来预测情感倾向
senta = hub.Module(name="senta_bilstm")
texts = df['内容'].tolist()
input_data = {'text': texts}
res = senta.sentiment_classify(data=input_data)
df['情感分值'] = [x['positive_probs'] for x in res]
df.head()
```

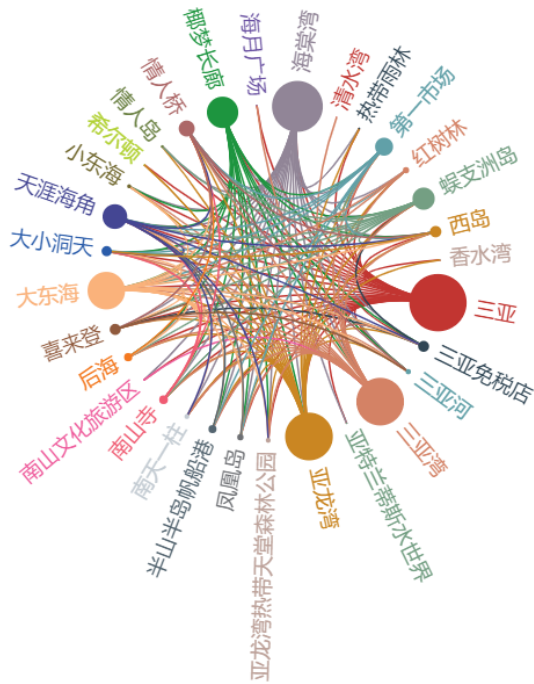
5. 具体介绍在这里

senta_bilstm模型的介绍

官方文档：https://www.paddlepaddle.org.cn/hubdetail?name=senta_bilstm&en_category=SentimentAnalysis

游客行为规律分析

首先这是效果图



具体做法:

先把路线用txt保存下来, 然后把去哪儿和携程的路径保存下来

```
fp = open('./data/路线.txt', 'r', encoding='utf8')
for line in fp:
    fq = open('./data/路线1.txt', 'a', encoding='utf8') #这里用追加模式
    fq.write(line)
fp.close()
fq.close()
```

```
with open('./data/路线1.txt', 'w', encoding='utf-8') as f:
    for d in data2['作者去了这些地方']:
        f.write(str(d)+'\n')
```

然后根据总的线路先去划分它们, 然后去计算它们的权重

计算权重的方法

```
#循环遍历关键词所在位置 设置word_vector计数
i = 0
j = 0
while i<len(nums):          #ABCD共现 AB AC AD BC BD CD加1
    j = i + 1
    w1 = nums[i]             #第一个单词
    while j<len(nums):
        w2 = nums[j]        #第二个单词
        #从word数组中找到单词对应的下标
        k = 0
        n1 = 0
        while k<len(word):
            if w1==word[k]:
                n1 = k
                break
            k = k + 1
        #寻找第二个关键字位置
        k = 0
        n2 = 0
        while k<len(word):
            if w2==word[k]:
                n2 = k
                break
            k = k + 1
        #重点: 词频矩阵赋值 只计算上三角
```

对应的矩阵样式如下：

首先创建一个共现矩阵

$$\begin{bmatrix} - & A & B & C & D \\ A & 0 & 1 & 2 & 1 \\ B & 1 & 0 & 1 & 1 \\ C & 2 & 1 & 0 & 1 \\ D & 1 & 1 & 1 & 0 \end{bmatrix}$$

然后通过共现矩阵分别获取两两关系及权重，再写入CSV或Excel文件中

$$\begin{bmatrix} Source & Target & Weight \\ A & B & 1 \\ A & C & 2 \\ A & D & 1 \\ B & C & 1 \\ B & D & 1 \\ C & D & 1 \end{bmatrix}$$

最后的样式如下

29]:

	Word1	Word2	Weight
0	三亚	海棠湾	64
1	三亚	椰梦长廊	34
2	三亚	亚龙湾	57
3	三亚	后海村	4
4	三亚	亚特兰蒂斯水世界	7
...
2944	琼海	莲花墩	1
2945	爱心大世界	兰花世界	1
2946	乐会古城	蔡家宅	1
2947	乐会古城	莲花墩	1
2948	蔡家宅	莲花墩	1

然后再去统计每一个站点的权重有多少

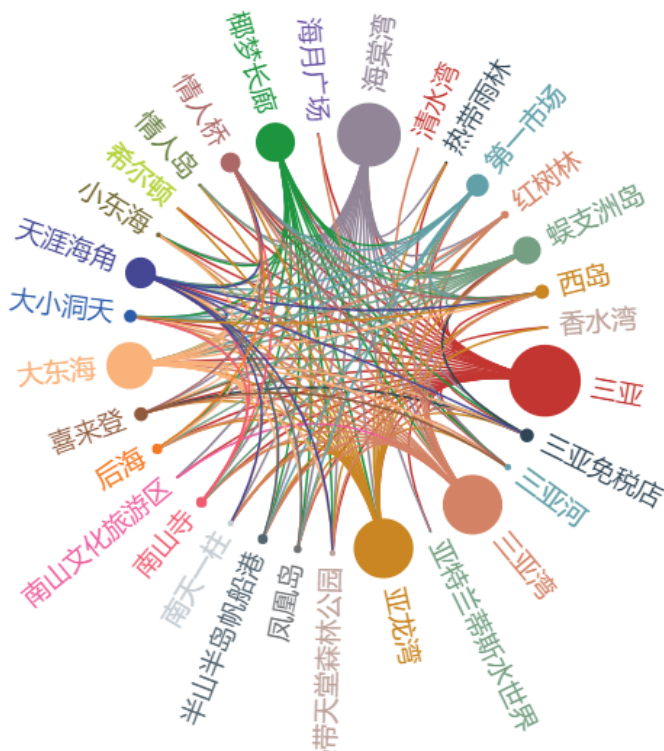
这里因为数量太多，并没什么意义，所以做了一个筛选，word1-word2权重小于5的全部删除，然后再去统计剩余每一个站点它的权重，然后去画出对应的图形

计算的权重如下：

33]:

Word	Weight
三亚	514
三亚免税店	98
三亚河	46
三亚湾	427
亚特兰蒂斯水世界	12
亚龙湾	428
亚龙湾热带天堂森林公园	42
凤凰岛	56
半山半岛帆船港	71
南天一柱	42
南山寺	77

最后的图形就是刚刚的那副图形



出游动机是怎么划分的

出游动机是首先根据它的内容

然后去判断里面是否存在某个词是符合出游类型的类型

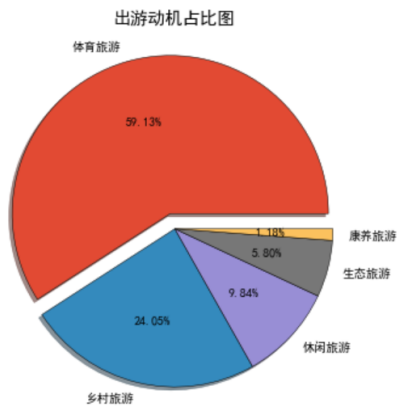
具体方法如下：

这一块是计算它们的出游动机，根据一些关键词的存在，判断出它们的出游动机属于哪一列

```
] def cydj(x):  
    x = str(x)  
    if '徒步' in x or '骑' in x or '越野' in x or '运动' in x or '登山' in x:  
        return '体育旅游'  
    elif '游居' in x or '野' in x or '村' in x or '乡' in x:  
        return '乡村旅游'  
    elif '生态' in x or '污染' in x or '自然' in x:  
        return '生态旅游'  
    elif '人与自然' in x or '心智' in x or '精神' in x or '身体' in x or '心态' in x or '和谐共处' in x or '修身' in x or '养性' in x:  
        return '康养旅游'  
    else:  
        return '休闲旅游'
```

然后再根据它的统计的数量去做成一个饼图，统计每个类型的数量及占比

```
plt.style.use('ggplot')
plt.figure(figsize=(12,6))
explo = [0.1,0.0,0.0]
plt.pie(list_values, labels=list_key, explode=explode, shadow=True, startangle=0, autopct='%1.2f%%', wedgeprops={'edgecolor':'black'})
plt.title('出游动机占比图')
plt.savefig('出游动机占比图.jpg')
plt.show()
```



关于滞留时间

是根据他们的天数

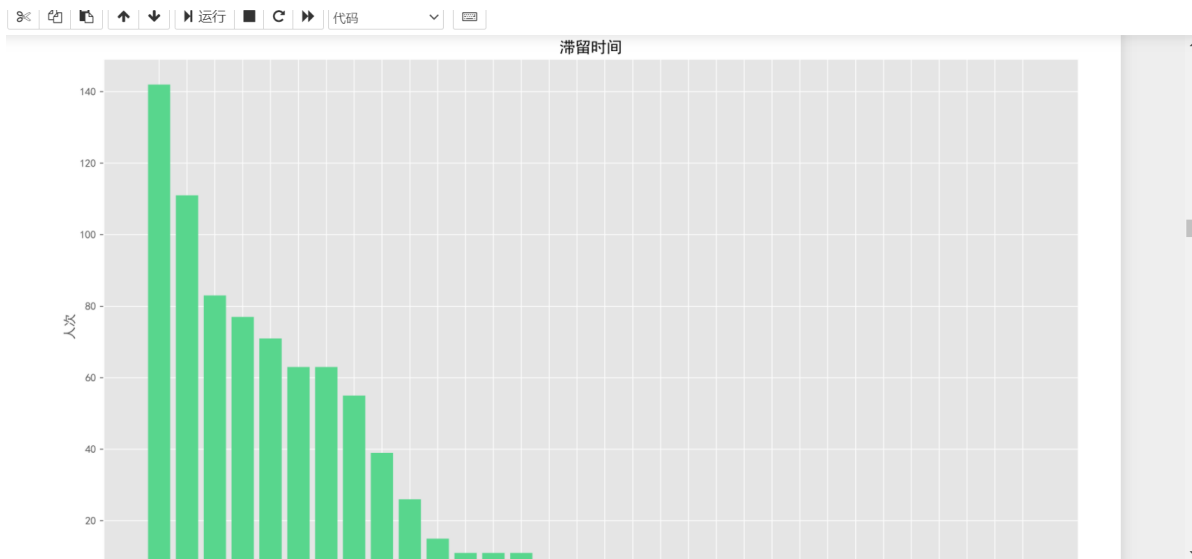
0]:

	页面网址	发表时间	标题	作者	城市	天数	旅游时间	人均	和谁	作者去了这些地方	游记目录	正文
0	https://you.ctrip.com/travels/sanya61/4014427....	2021-06-01 20:32	三亚美食攻略 住进奢华亚特兰蒂斯,赏三亚各种美食	vivi慢生活	三亚	6天	5月	15000元	和朋友	三亚-海棠湾-椰梦长廊-亚龙湾-后海村-亚特兰蒂斯水世界-南山文化旅游景区-三亚国际免税城-三...	NaN	✎写在前面 三亚这座城市总是有着独特的魅力,特别是在无法出国的日子里,想享受...
1	https://you.ctrip.com/travels/sanya61/3997496....	2021-03-01 00:49	三亚怎么玩?这份吃喝玩乐宝藏攻略,你可得收好了	小飞侠 Finn	三亚	4天	3月	2000元	一个人	第一市场-三亚-亚龙湾-海棠湾-热带雨林-热带雨林	NaN	林姐香味海鲜(第一市场总店) 一家开了23年的本地特色老店,好吃且不贵,来讨过...

或者逗留时间啥的,统计

然后它们存在天的,把天这个去掉,只剩数字方便我们统计,然后把它们对应每个人逗留天数做一个统计然后用图表的方式展示

```
def tsjs(x):
    x = str(x)
    x = x.replace('天','').strip(" ")
    if x == np.nan:
        return 0
    else:
        return x
data1['逗留天数'] = data1['逗留天数'].apply(tsjs)
data2['天数'] = data2['天数'].apply(tsjs)
data3['出行天数'] = data3['出行天数'].apply(tsjs)
```



关于认知形象

首先是它的词频统计

这里是采用停用词把一些无意义的词去掉，然后再去统计这些词的长度是不是大于2，单个词不算词语，然后统计好之后再把这个词存入csv文件，名为高频词.csv

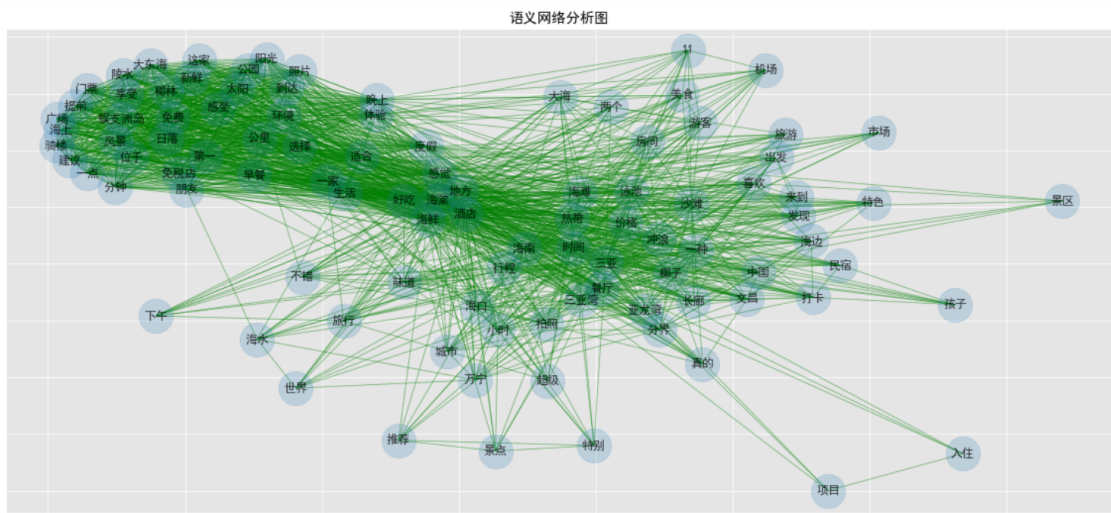
上面是 对文本进行词频统计的代码

```
# 定义分词函数
def get_cut_words(content_series):
    # 读入停用词表
    stop_words = []

    with open("./data/stopwords_cn.txt", 'r', encoding='utf-8') as f:
        lines = f.readlines()
        for line in lines:
            stop_words.append(line.strip())

    # 分词
    word_num = jieba.lcut(content_series.str.cat(sep='。'), cut_all=False)

    # 条件筛选
    word_num_selected = [i for i in word_num if i not in stop_words and len(i)>=2]
    for word in word_num_selected:
        for w in word:
            list_word.append(w)
    return list_word
```

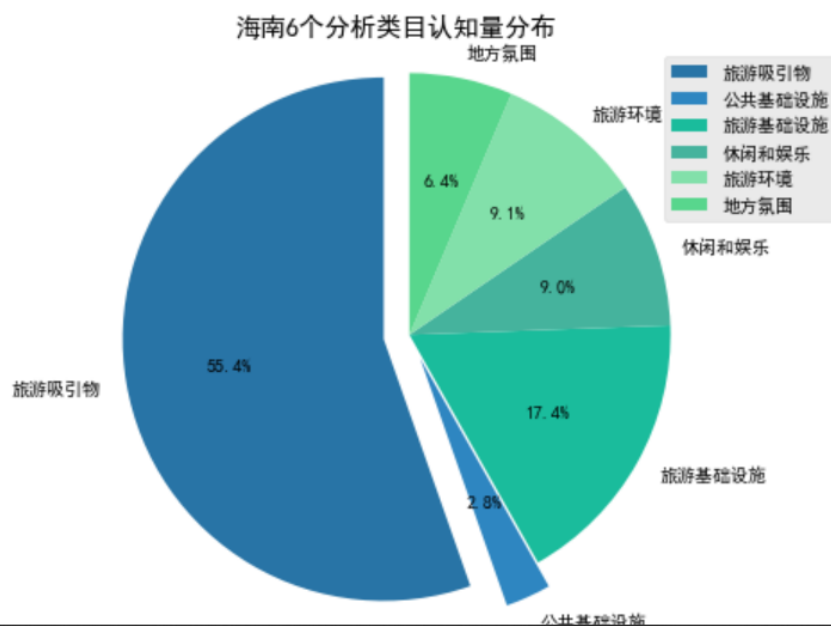



关于旅游吸引物、旅游交通、旅游住宿、旅游饮食、旅游娱乐、旅游商品6类 描述游客的认知形象

维度		高频词 (频次)	词汇总数	词汇比例
旅游吸引物	自然类	亚龙湾 (4986), 沙滩 (4657), 三亚湾 (2805), 海棠 (2864), 蜈支洲岛 (1615), 大东海 (1508), 陵水 (1487), 椰林 (1472), 日落 (1448), 大港 (2480), 海边 (3555)	28877	55.40%
	人文类	老街 (1316), 骑楼 (1380), 椰子 (5420), 海鲜 (6071), 免税店 (1410), 美食 (2181), 码头 (1305), 森林公园 (1265), 鹿回头 (1253), 网红 (986), 亚特兰蒂斯 (923), 文昌鸡 (875), 天涯海角 (823)	25208	
公共基础设施		地铁 (43), 火车 (404), 公交车 (323), 公共交通 (77), 火车站 (251), 车站 (121), 汽车站 (23), 飞机 (1140), 海口机场 (111), 三亚机场 (105), 出租车 (143)	2741	2.90%
旅游基础设施		酒店 (14222), 房间 (2287), 饭店 (311), 宾馆 (143)	16963	17.40%
休闲和娱乐		步行 (641), 旅游 (1691), 旅行 (2135), 拍照 (3111), 游览 (593), 参观 (471), 游船 (105)	8747	9.00%
旅游环境	自然环境	方便快捷 (8), 干净 (796), 卫生 (113), 美丽 (1071)	1988	9.10%
	社会环境	经济 (173), 人文 (211), 安全感 (36), 治安 (4)	424	
	政治环境	昂贵 (1431), 价格 (2077), 免费 (1660), 便宜 (1296)	6464	
地方氛围		自然 (1058), 热闹 (496), 漂亮 (526), 著名 (425), 热情 (528), 美好 (850), 浪漫 (831), 开心 (765), 舒服 (753)	6232	6.40%

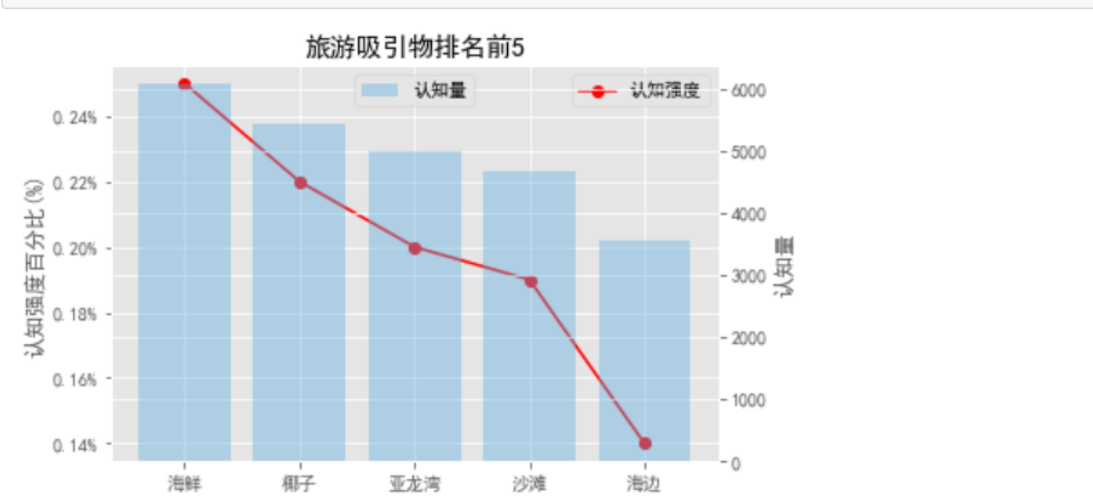
这个它的总的统计，因为数量太多，所以一部分作为参考，选词是从高频词.csv文件这里选取的

并且把相应的内容做成饼图



分别画出“旅游吸引物、旅游交通、旅游住宿、旅游饮食、旅游娱乐、旅游商品”6方面排名前5的认知关键词图表

如图所示，该前五主要是从上面的表格选取最高的前5个词，如果数量不够，那么就是去高频词.csv这个文件里面找，最高的词，这样的图有六个，图片太多没必要截，到时候拿到代码查看即可

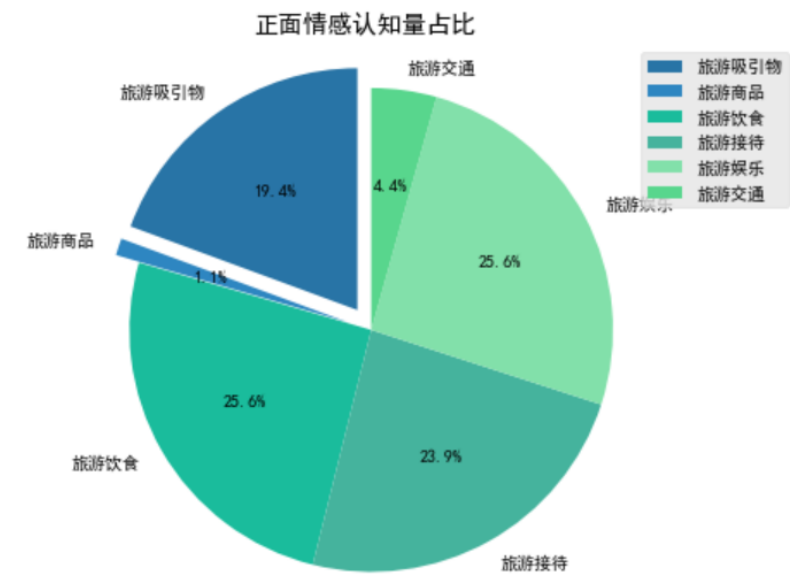


情感形象

分析类目	正面认知量	正面认知比例(%)
旅游吸引物	4595	19.40%
旅游商品	269	1.10%
旅游饮食	6071	25.60%
旅游接待	5666	23.90%
旅游娱乐	6079	25.60%
旅游交通	1033	4.40%

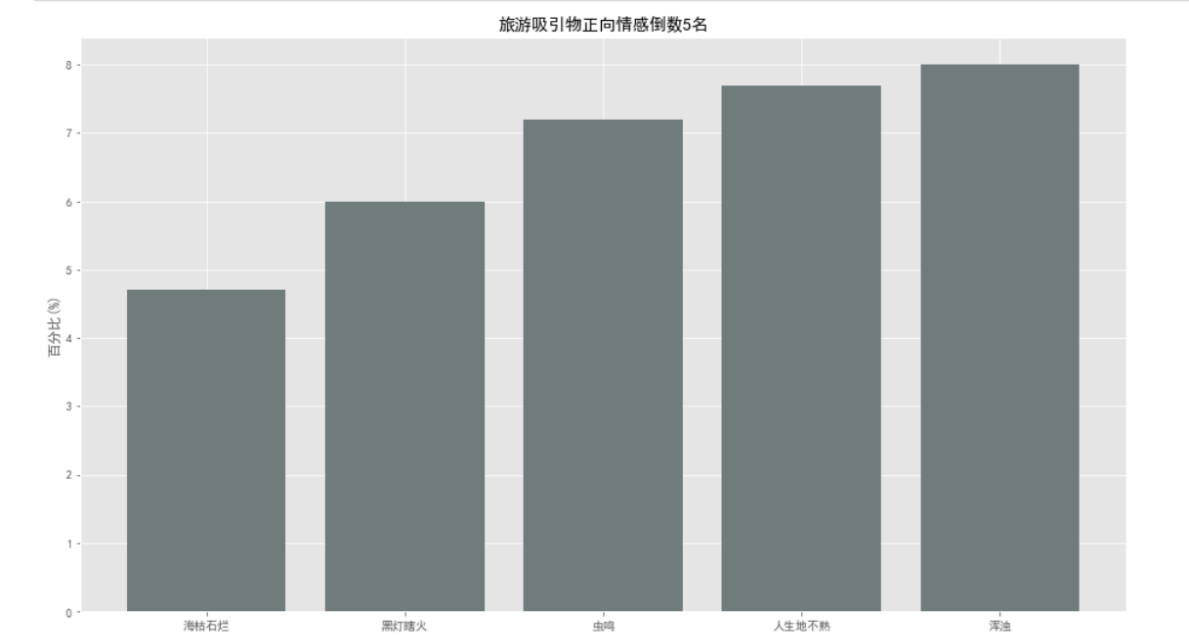
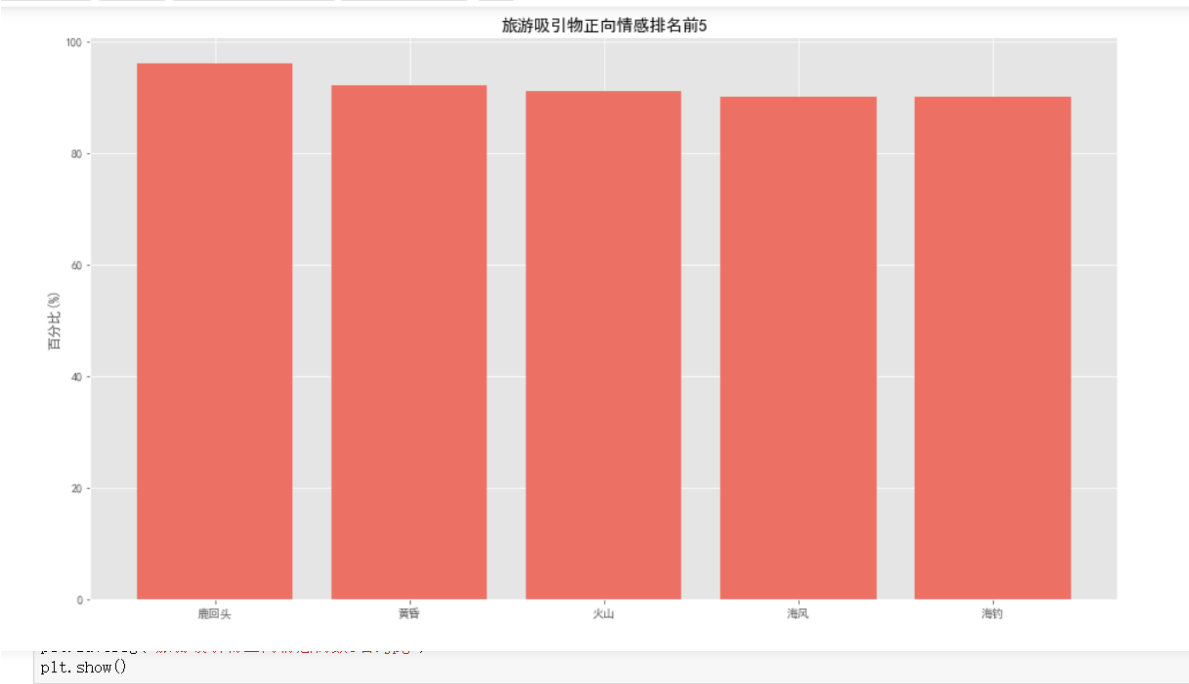
该表格主要从正面情感TOP100高频词这个文件里面，对词进行一个归类，后统计的认知量及认知比例等

并且根据该表格做成一个饼图





分别画出“旅游吸引物、旅游交通、旅游住宿、旅游饮食、旅游娱乐、旅游商品”6方面正向情感排名前5（排名最低的5个也要）的认知关键词图表

这样的图一共有12个，正向的6个，反向的也有6个



是根据这两个文件，对相对应的词进行分类，归纳，选择倒数5个和前5个

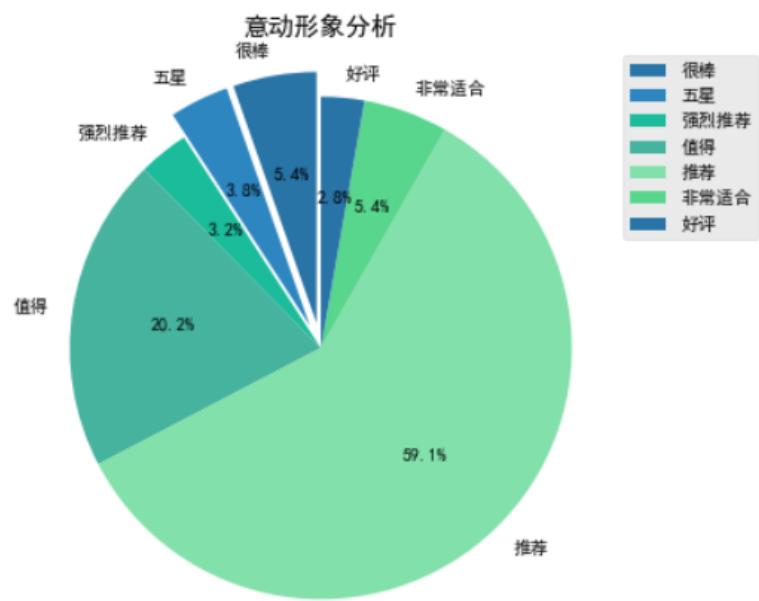
 正面情感最低的高频词.xlsx	2021/12/20 14:22	Microsoft Excel 工...	26 KB
 正面情感最高的高频词.xlsx	2021/12/20 14:21	Microsoft Excel 工...	12 KB

意动形象分析

意动形象分析		
关键词	词汇频次	词汇比例
很棒	241	5.40%
五星	171	3.80%
强烈推荐	145	3.20%
值得	908	20.20%
推荐	2652	59.10%
非常适合	244	5.40%
好评	125	2.80%

该表主要结合意动形象的相关词去<正面情感最高的高频词>这个文件里面查找对应词的词频，并且做出对应的统计

这是对应的饼图



认知形象、情感形象、意动形象

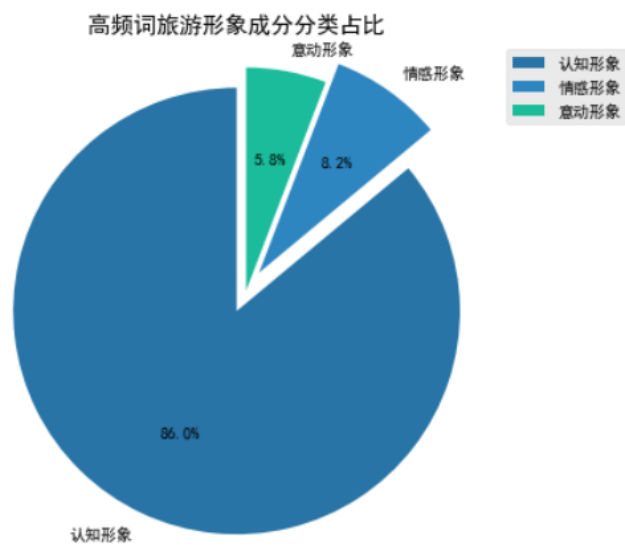
类别		高频词	词汇比例
认知形象	景点	三亚湾（2805），亚龙湾（4986），三亚（15970），蜈支洲岛（1615），大东海（1508），免税店（1410），森林公园（1265），鹿回头（1253），石梅湾（1162），博鳌（1102）	86.00%
	内容	环境（1448），酒店（14222），风景（1421），建筑（1140），飞机（1140），海岛（1104），海口（4540），海边（3555），大海（2480），海滩（2449）	
情感形象		放心（206），满意（211），喜欢（3295），愉悦（118），贴心（333），快乐（387），温馨（245），开心（765），舒服（753），	8.20%
意动形象		很棒（241），五星（171），强烈推荐（145），值得（908），推荐（2652），非常适合（244），好评（125）	5.80%

因为词频数量太多，所以我们对应认知形象去筛选前10个该类型的高频词去统计，去从高频词.csv这个文件去进行筛选的，

关于情感形象是根据《正面情感TOP100高频词》这个文件去进行统计的

关于意动形象是直接采用上面归类好的进行统计

最后这个它的饼图



这个则是根据高频词对词进行归类，然后找寻对应最高的前10的词