

In [17]:

```
import pandas as pd
import numpy as np
```

In [18]:

```
data1 = pd.read_excel('./data/去哪儿游记.xlsx')
content1 = data1['正文']
```

In [19]:

```
data2 = pd.read_excel('./data/携程游记.xlsx')
content2 = data2['正文']
```

In [20]:

```
data3 = pd.read_excel('./data/马蜂窝游记.xlsx')
content3 = data3['正文']
```

In [21]:


```
data4 = pd.read_excel('./data/驴妈妈游记.xlsx').iloc[:11]
data4['正文'] = data4['游记内容']
content4 = data4['正文']
```

In [22]:

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

Out [22]:

0 这是一篇搁置了两个月的游记，虽然对三亚行程记忆犹新，只是不知道该如何用文字描述这种悸动，终于...

1 第一次来写  游记，这次跟以往每一次出去玩都信誓旦旦的说“我要记录下来”不一样！这次我真的...

2 浓浓的秋意由西至东、从北到南地拂遍祖国大地，向往温暖阳光的候鸟们即将蠢蠢欲动，准备启程远航……

3 +1\r\n+1\r\n《请到天涯海角来》 “请到天涯海角来，这里四季春常在；海南岛上春风暖，...

4 久违了！\r\n\r\n阔别250多天后，才出发\r\n\r\n原本今年2月就打算去韩国转一...

		...		
6	前言\n	\n	\n	...
7	前言\n	\n	\n	...
8	前言\n	\n	\n	...
9	前言\n	\n	\n	...
10	前言\n	\n	\n	...

Name: 正文, Length: 1254, dtype: object

把全部的内容合并在一起，并且删除重复行，上面的代码主要的做的是这个事情

下面这一块是处理计算中文字数，大于100的内容才保留下来

In [23]:

```
import string
def str_count(str):
    count_en = count_dg = count_sp = count_zh = count_pu = 0
    for s in str:
        if s in string.ascii_letters:
            count_en += 1
        elif s.isdigit():
            count_dg += 1
        elif s.isspace():
            count_sp += 1
        elif s.isalpha():
            count_zh += 1
        else:
            count_pu += 1
    return count_zh
```

In [24]:

```
content = content.astype(str)
content_cd = content.apply(str_count)
```

In [25]:

```
comment = content.values
cd = content_cd.values
```

In [26]:

```
df = pd.DataFrame()
df['内容'] = comment
df['正文长度'] = cd
df.head()
```

Out[26]:

	内容	正文长度
0	这是一篇搁置了两个月的游记，虽然对三亚行程记忆犹新，只是不知道该如何用文字描述这种悸动，终于...	18592
1	第一次来写这篇游记，这次跟以往每一次出去玩都信誓旦旦的说"我要记录下来"不一样！这次我真的...	12460
2	浓浓秋意由西至东、从北到南地拂遍祖国大地，向往温暖阳光的候鸟们即将蠢蠢欲动，准备启程远航.....	9256
3	+1\r\n+1\r\n《请到天涯海角来》“请到天涯海角来，这里四季春常在；海南岛上春风暖，...	17151
4	久违了！\r\n\r\n阔别250多天后，才出发\r\n\r\n原本今年2月就打算去韩国转一...	11980

In [27]:

```
df = df[df['正文长度'] >=100]
df
```

Out[27]:

	内容	正文长度
0	这是一篇搁置了两个月的游记，虽然对三亚行程记忆犹新，只是不知道该如何用文字描述这种悸动，终于...	18592
1	第一次来写这篇游记，这次跟以往每一次出去玩都信誓旦旦的说"我要记录下来"不一样！这次我真的...	12460
2	浓浓秋意由西至东、从北到南地拂遍祖国大地，向往温暖阳光的候鸟们即将蠢蠢欲动，准备启程远航.....	9256
3	+1\n+1\n《请到天涯海角来》“请到天涯海角来，这里四季春常在；海南岛上春风暖，...	17151
4	久违了！\n\n\n阔别250多天后，才出发\n\n\n原本今年2月就打算去韩国转一...	11980
...	...	...
1249	前言\n\n\n...	2508
1250	前言\n\n\n...	1758
1251	前言\n\n\n...	894
1252	前言\n\n\n...	2389
1253	前言\n\n\n...	2321

1226 rows × 2 columns

这一块是清洗数据，删除一些多余的字符和空格等无关紧要的内容，并且如果没有包含三亚或者海南的就去掉

In [28]:

```
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
```

In [29]:

```
df['内容'] = df['内容'].apply(sj qx)  
df['内容']
```

Out[29]:

0 这是一篇搁置了两个月的游记，虽然对三亚行程记忆犹新，只是不知道该如何用文字描述这种悸动，终于...

1 第一次来写 游记，这次跟以往每一次出去玩都信誓旦旦的说“我要记录下来”不一样！这次我真的...

2 浓浓秋意由西至东、从北到南地拂遍祖国大地，向往温暖阳光的候鸟们即将蠢蠢欲动，准备启程远航.....

3 +1+1《请到天涯海角来》“请到天涯海角来，这里四季春常在；海南岛上春风暖，好花叫你喜心怀”...

4 久违了！阔别250多天后，才出发原本今年2月就打算去韩国转一圈，各种装备、攻略、签证都搞定了...

1249 【 引 】如果你心中自带星辰和大海，那你眼里的世界，一定光彩熠熠！【 开篇 】你会不会有特别...

1250 【出发前的小笔记】为了庆祝老妈正式迈入退休阶段，原本计划今年要带她出去旅行好好玩玩的。结果，...

1251 在这个六月，想与大海来次亲密接触，在这个季节，你一定不能错过神州半岛喜来。有多少人说要遍海...

1252 【】打包好行李又要出发了，小半年没有出门旅行，感觉自己都要变成原始人了。这次的目的地 海南 ...

1253 【疫后出行小贴士】今年因为疫情的缘故，待在家里大半年都没有出门。现在国内疫情终于好转，大部分...

Name: 内容, Length: 1226, dtype: object

这一块是删除一些空值的内容

In [30]:

```
df.dropna(how='any', inplace=True)
df
```

Out[30]:

	内容	正文长度
0	这是一篇搁置了两个月的游记，虽然对三亚行程记忆犹新，只是不知道该如何用文字描述这种悸动，终于...	18592
1	第一次来写这篇游记，这次跟以往每一次出去玩都信誓旦旦的说"我要记录下来"不一样！这次我真的...	12460
2	浓浓秋意由西至东、从北到南地拂遍祖国大地，向往温暖阳光的候鸟们即将蠢蠢欲动，准备启程远航.....	9256
3	+1+1《请到天涯海角来》“请到天涯海角来，这里四季春常在；海南岛上春风暖，好花叫你喜心怀”...	17151
4	久违了！阔别250多天后，才出发原本今年2月就打算去韩国转一圈，各种装备、攻略、签证都搞定了...	11980
...	...	...
1249	【引】如果你心中自带星辰和大海，那你眼里的世界，一定光彩熠熠！【开篇】你会不会有特别...	2508
1250	【出发前的小笔记】为了庆祝老妈正式迈入退休阶段，原本计划今年要带她出去旅行好好玩的。结果，...	1758
1251	在这个六月，想与大海来次亲密接触，在这个季节，你一定不能错过神州半岛喜来。有多少人说要耍遍海...	894
1252	【】打包好行李又要出发了，小半年没有出门旅行，感觉自己都要变成原始人了。这次的目的地 海南 ...	2389
1253	【疫后出行小贴士】今年因为疫情的缘故，待在家里大半年都没有出门。现在国内疫情终于好转，大部分...	2321

1189 rows × 2 columns

这一块是删除表情包和标点符号的内容

In [31]:

```
import re
def clear_characters(text):
    return re.sub('\W', '', text)
df['内容'] = df['内容'].apply(clear_characters)
df['内容']
```

Out[31]:

```
0      这是一篇搁置了两个月的游记虽然对三亚行程记忆犹新只是不知道该如何用文字描述这种悸动终于有时间...
1      第一次来写游记这次跟以往每一次出去玩都信誓旦旦的说我要记录下来不一样这次我真的来写了原本计划...
2      浓浓秋意由西至东从北到南地拂遍祖国大地向往温暖阳光的候鸟们即将蠢蠢欲动准备启程远航它们心驰神...
3      11请到天涯海角来请到天涯海角来这里四季春常在海南岛上春风暖好花叫你喜心怀经历过这不平凡的2...
4      久违了阔别250多天后才出发原本今年2月就打算去韩国转一圈各种装备攻略签证都搞定了然而五一那...

...
1249   引如果你心中自带星辰和大海那你眼里的世界一定光彩熠熠开篇你会不会有特别想念阳光的时候每当我特...
1250   出发前的小笔记为了庆祝老妈正式迈入退休阶段原本计划今年要带她出去旅行好好玩玩的结果计划赶不上...
1251   在这个六月想与大海来次亲密接触在这个季节你一定不能错过神州半岛喜来有多少人说要耍遍海南很难做...
1252   打包好行李又要出发了小半年没有出门旅行感觉自己都要变成原始人了这次的目的地海南三亚一个只听名...
1253   疫后出行小贴士今年因为疫情的缘故待在家里大半年都没有出门现在国内疫情终于好转大部分城市也开始...
Name: 内容, Length: 1189, dtype: object
```

这一块是采用百度的开源NLP，这个是一个成熟的框架，且准确率高，不需要去计算其他什么的内容，该框架就会自动帮你去分析判断这里面的情感倾向

In [32]:

```
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()
```

```
[2021-12-19 13:41:06,247] [ INFO] - Installing senta_bilstm module
[2021-12-19 13:41:06,256] [ INFO] - Module senta_bilstm already installed in C:\Users\96075\paddlehub\modules\senta_bilstm
[2021-12-19 13:41:08,998] [ INFO] - Installing lac module
[2021-12-19 13:41:09,004] [ INFO] - Module lac already installed in C:\Users\96075\paddlehub\modules\lac
```

Out[32]:

	内容	正文长度	情感分值
0	这是一篇搁置了两个月的游记虽然对三亚行程记忆犹新只是不知道该如何用文字描述这种悸动终于有时间...	18592	0.9954
1	第一次来写游记这次跟以往每一次出去玩都信誓旦旦的说我要记录下来不一样这次我真的来写了原本计划...	12460	0.9886
2	浓浓秋意由西至东从北到南地拂遍祖国大地向往温暖阳光的候鸟们即将蠢蠢欲动准备启程远航它们心驰神...	9256	0.3345
3	11请到天涯海角来请到天涯海角来这里四季春常在海南岛上春风暖好花叫你喜心怀经历过这不平凡的2...	17151	0.9966
4	久违了阔别250多天后才出发原本今年2月就打算去韩国转一圈各种装备攻略签证都搞定了然而五一那...	11980	0.9959

senta\_bilstm模型的介绍  
官方文档:[https://www.paddlepaddle.org.cn/hubdetail?name=senta\\_bilstm&en\\_category=SentimentAnalysis](https://www.paddlepaddle.org.cn/hubdetail?name=senta_bilstm&en_category=SentimentAnalysis)

In [33]:

```
清洗数据保存为EXCEL
df.to_excel("./data/clean_data.xlsx", index=False)
```

In [72]:

```
df1 = pd.read_csv('./data/高频词.csv', encoding='gbk')
```

上面的计算内容的感情倾向 下面这个是计算所有高频词的情感倾向

In [73]:

```
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 = df1['word'].tolist()
input_data = {'text':texts}
res = senta.sentiment_classify(data=input_data)
df1['情感分值'] = [x['positive_probs'] for x in res]
df1.head()
```

```
[2021-12-20 11:43:52,921] [ INFO] - Installing senta_bilstm module
[2021-12-20 11:43:52,925] [ INFO] - Module senta_bilstm already installed in C:\Users\96075\paddlehub\modules\senta_bilstm
[2021-12-20 11:43:55,471] [ INFO] - Installing lac module
[2021-12-20 11:43:55,475] [ INFO] - Module lac already installed in C:\Users\96075\paddlehub\modules\lac
```

Out[73]:

Unnamed: 0	word	count	情感分值
0	0 三亚	15970	0.8208
1	1 酒店	14222	0.5330
2	2 海南	9519	0.7336
3	3 海鲜	6071	0.6311
4	4 椰子	5420	0.5367

统计情感倾向大于0.9的，并且频次高于200的，获取正面情感最高的那100个词频内容，并且保存为xlsx文件

In [82]:

```
data1 = df1[df1['情感分值'] >=0.9]
data1 = data1[data1['count'] >=200]
data1.to_excel("./data/正面情感TOP100高频词.xlsx", index=False)
```

统计情感倾向大于0.9的，并且频次高于100的，获取正面情感词频的主要内容内容，并且保存为xlsx文件

In [85]:

```
data2 = df1[df1['count'] >=100]
data2 = data2[data2['情感分值'] >=0.9]
data2.to_excel("./data/正面情感最高的高频词.xlsx", index=False)
```

统计情感倾向大于0.2的，并且频次高于10的，获取正面情感最低词频的主要内容内容，并且保存为xlsx文件

In [86]:

```
data3 = df1[df1['count'] >=10]
data3 = data3[data3['情感分值'] <=0.2]
data3.to_excel("./data/正面情感最低的高频词.xlsx", index=False)
```



In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: