In [51]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import wordcloud
from wordcloud import WordCloud, STOPWORDS
import re
import string
import math
```

In [52]:
```
import seaborn as sns
```

In [53]:
```
data = pd.read_csv("HotelReview_Clean_v1.1.csv")
```

In [54]:
```
data.head()
```

Out[54]:

	Unnamed: 0	hotel_name	overall_rating	date_of_review	user_name	review_rating	review_title	revie
0	0	Innside New York NoMad	45	22-Mar	Kim C	5.0	Great hotel!	We in a loft v c
1	1	Innside New York NoMad	45	21-Jun	Peggy M	5.0	The Innside Front Desk Staff	I ju tc mor s w
2	2	Innside New York NoMad	45	4-Mar	Imran	5.0	NYC gem	I go a l hand
3	3	Innside New York NoMad	45	2-Mar	Jay B	5.0	Great Experience!	revie bec
4	4	Innside New York NoMad	45	22-Feb	Jeweliana159	5.0	Amazing Hospitality Team - Kudos to Christina,...	I've h nu oc b

5 rows × 21 columns

In [55]:
```
del data["Unnamed: 0"]
```

In [55]:
```
data.head(2)
```

Out[55]:

	hotel_name	overall_rating	date_of_review	user_name	review_rating	review_title	review_text	date_
0	Innside New York NoMad	45	22-Mar	Kim C	5.0	Great hotel!	We stayed in a family loft which is two connec...	Mar
1	Innside New York NoMad	45	21-Jun	Peggy M	5.0	The Innside Front Desk Staff	I just want to take a moment to say how wonder...	M

In [56]:
```
data.shape
```

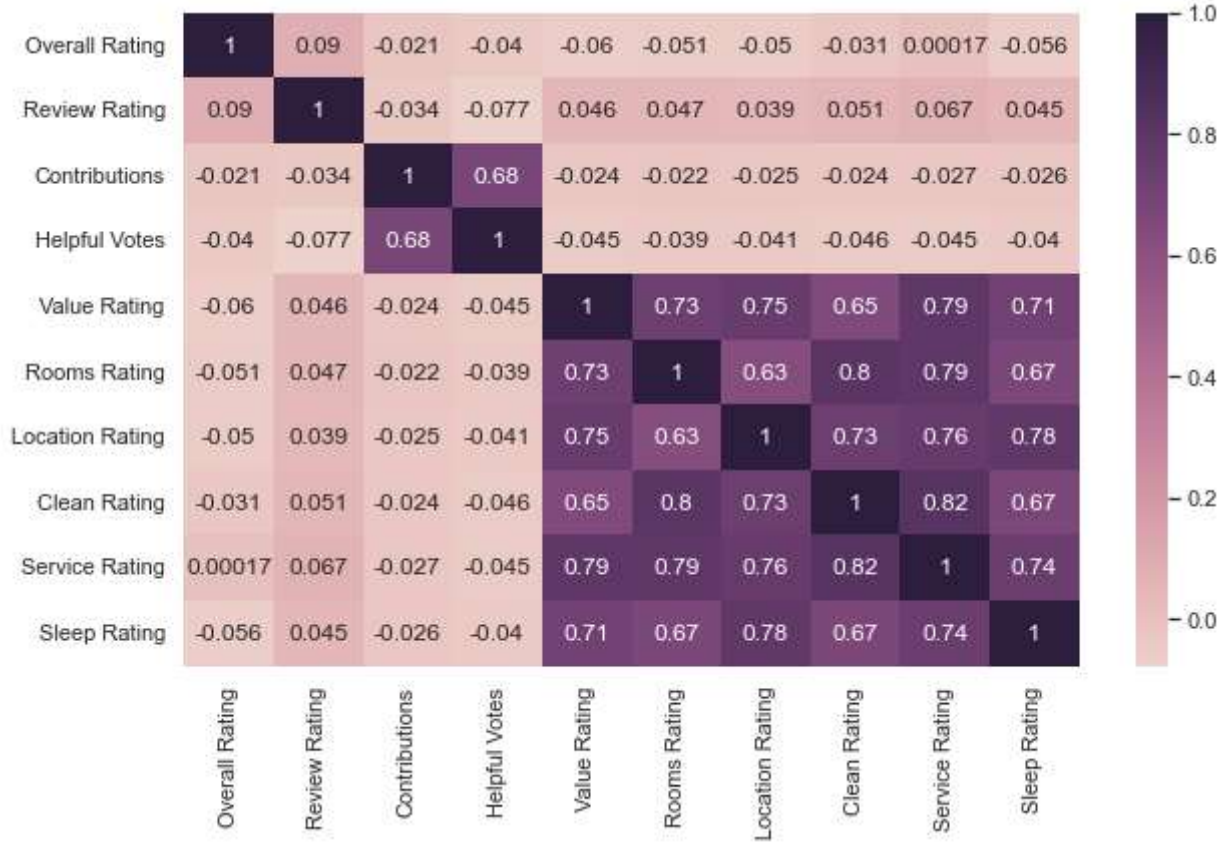
Out[56]:
```
(1115, 20)
```

In [57]:
```
data.corr()
```

Out[57]:

	overall_rating	review_rating	contributions	helpful_votes	value_rating	rooms_rating	lo
overall_rating	1.000000	0.090441	-0.020950	-0.040424	-0.059563	-0.050714	
review_rating	0.090441	1.000000	-0.034467	-0.077291	0.046352	0.046782	
contributions	-0.020950	-0.034467	1.000000	0.682356	-0.023740	-0.022342	
helpful_votes	-0.040424	-0.077291	0.682356	1.000000	-0.045075	-0.038896	
value_rating	-0.059563	0.046352	-0.023740	-0.045075	1.000000	0.728445	
rooms_rating	-0.050714	0.046782	-0.022342	-0.038896	0.728445	1.000000	
location_rating	-0.050160	0.038905	-0.024937	-0.041359	0.750696	0.628023	
clean_rating	-0.030687	0.051462	-0.024443	-0.045748	0.649471	0.802283	
service_rating	0.000168	0.067265	-0.027270	-0.045258	0.787675	0.788806	
sleep_rating	-0.055971	0.045119	-0.025708	-0.039639	0.713376	0.669804	

In [58]:
```
h_labels = [x.replace('_', ' ').title() for x in list(data.select_dtypes(include=['numb
fig, ax = plt.subplots(figsize=(10,6))
_ = sns.heatmap(data.corr(), annot=True, xticklabels=h_labels, yticklabels=h_labels, cm
```



In [59]:
```
data.hotel_name.unique()
```

Out[59]:
```
array(['Innside New York NoMad', 'Motto by Hilton New York City Chelsea',
       'Pod Times Square',
       'Residence Inn New York Downtown Manhattan/world Trade Center Area',
       'Homewood Suites by Hilton New York/Midtown Manhattan Times Square-South, NY',
       'Hotel Riu Plaza New York Times Square'], dtype=object)
```

In [60]:
```
# #how to name curve with corresponding hotel name??
# dist=data['review_rating'].groupby(data["hotel_name"])
# dist.plot(kind='kde', figsize=(10,6), title='Distribution of Review Rating')
```

In [61]:
```
data['review_rating'].value_counts()
```

Out[61]:
```
5.0    774
4.0    166
3.0     71
1.0     57
2.0     47
Name: review_rating, dtype: int64
```

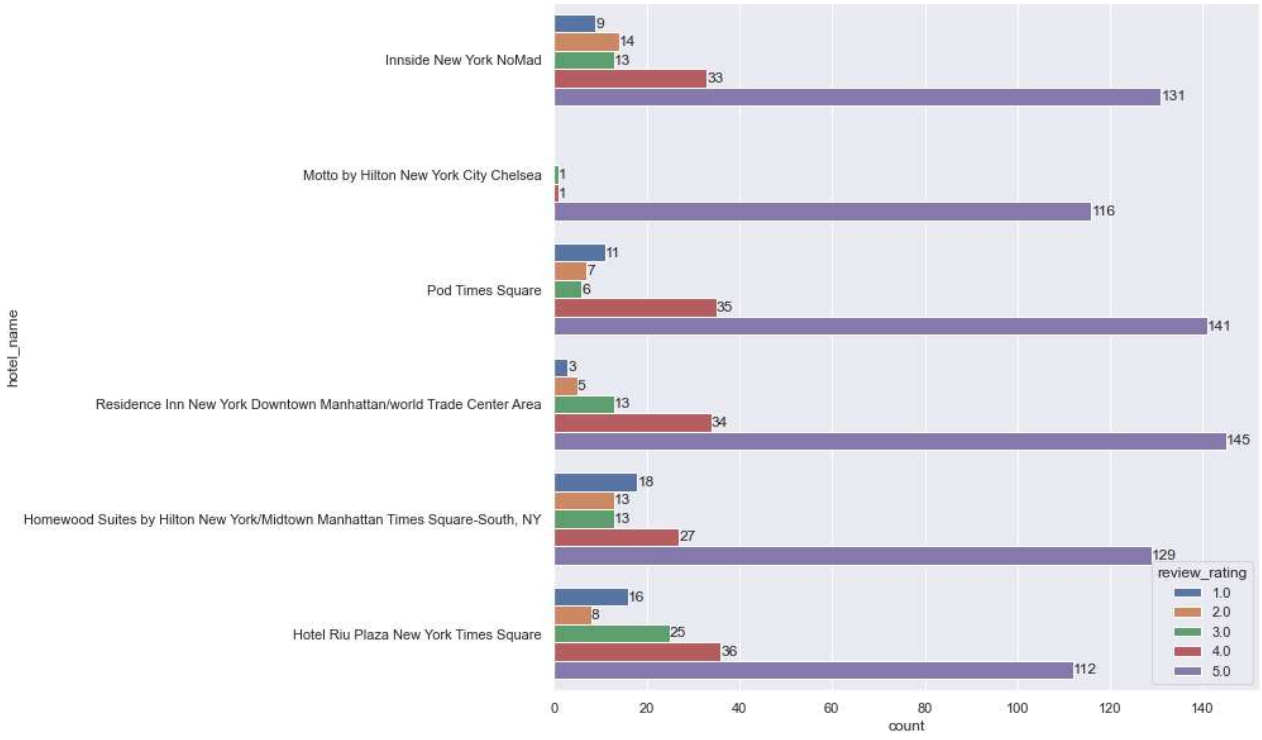
In [62]:
```
data.describe()
```

Out[62]:

	overall_rating	review_rating	contributions	helpful_votes	value_rating	rooms_rating	location_ra
count	1115.000000	1115.000000	1115.000000	1115.000000	1115.000000	1115.000000	1115.00(
mean	44.632287	4.392825	42.759641	17.713901	0.079821	0.077130	0.07

	overall_rating	review_rating	contributions	helpful_votes	value_rating	rooms_rating	location_ra
std	2.645961	1.110559	188.474361	43.073064	0.611012	0.591203	0.62
min	40.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.00
25%	45.000000	4.000000	0.000000	0.000000	0.000000	0.000000	0.00
50%	45.000000	5.000000	2.000000	1.000000	0.000000	0.000000	0.00
75%	45.000000	5.000000	18.000000	11.000000	0.000000	0.000000	0.00
max	50.000000	5.000000	3915.000000	455.000000	5.000000	5.000000	5.00

In [63]:
```
ax1=sns.countplot(y="hotel_name", hue="review_rating", data=data)
for container in ax1.containers:
    ax1.bar_label(container)
sns.set(rc={'figure.figsize':(10,10)})
```



In [64]:
```
#if we consider users who have provided a review 5 times before
criteria = data[data['contributions'] >= 5.0]

criteria
```

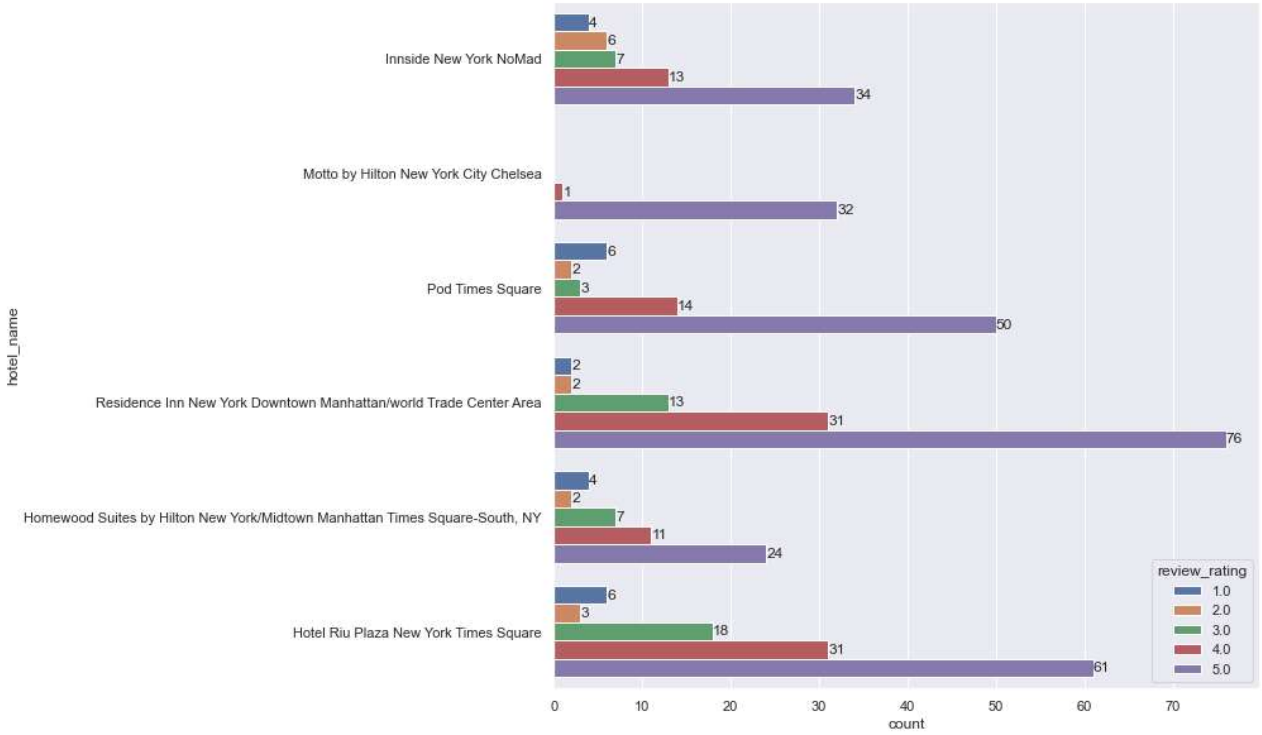
Out[64]:

	hotel_name	overall_rating	date_of_review	user_name	review_rating	review_title	review_
1	Innside New York NoMad	45	21-Jun	Peggy M	5.0	The Innside Front Desk Staff	I just want to a moment to how wonc

	hotel_name	overall_rating	date_of_review	user_name	review_rating	review_title	review_
4	Innside New York NoMad	45	22-Feb	Jeweliana159	5.0	Amazing Hospitality Team - Kudos to Christina,...	I've stayed I on numer occasions be
7	Innside New York NoMad	45	22-Jan	Jay C	5.0	Great Weekend Stay in NYC	Spent a Satu night in cit join frier
11	Innside New York NoMad	45	22-Jan	Mark R	4.0	Ideally situated	We stayed at Innside wher visited wi
13	Innside New York NoMad	45	22-Jan	etremat	5.0	NYE in New York	Stayed at beautiful hot celebrate
...	...	...	...	...	...	...	
1103	Hotel Riu Plaza New York Times Square	45	21-Jul	Denny820	5.0	Perfect	Perfect locat Hotel is new modern. I
1105	Hotel Riu Plaza New York Times Square	45	21-Jul	Tachita	4.0	Best location, price, convenience I found for ...	Quick chec friendly s Wake up soor
1106	Hotel Riu Plaza New York Times Square	45	21-Jul	faubel16	5.0	Cleanest Hotel In New York & Nicest Staff	All staff was friendly accommodati
1109	Hotel Riu Plaza New York Times Square	45	21-May	Robert W	4.0	Nice hotel located near Times Square	Spent weekend at for our first
1110	Hotel Riu Plaza New York Times Square	45	21-May	Trina Samara	5.0	Great Mother's Day weekend!	When we arr at the RIU cl in was fas

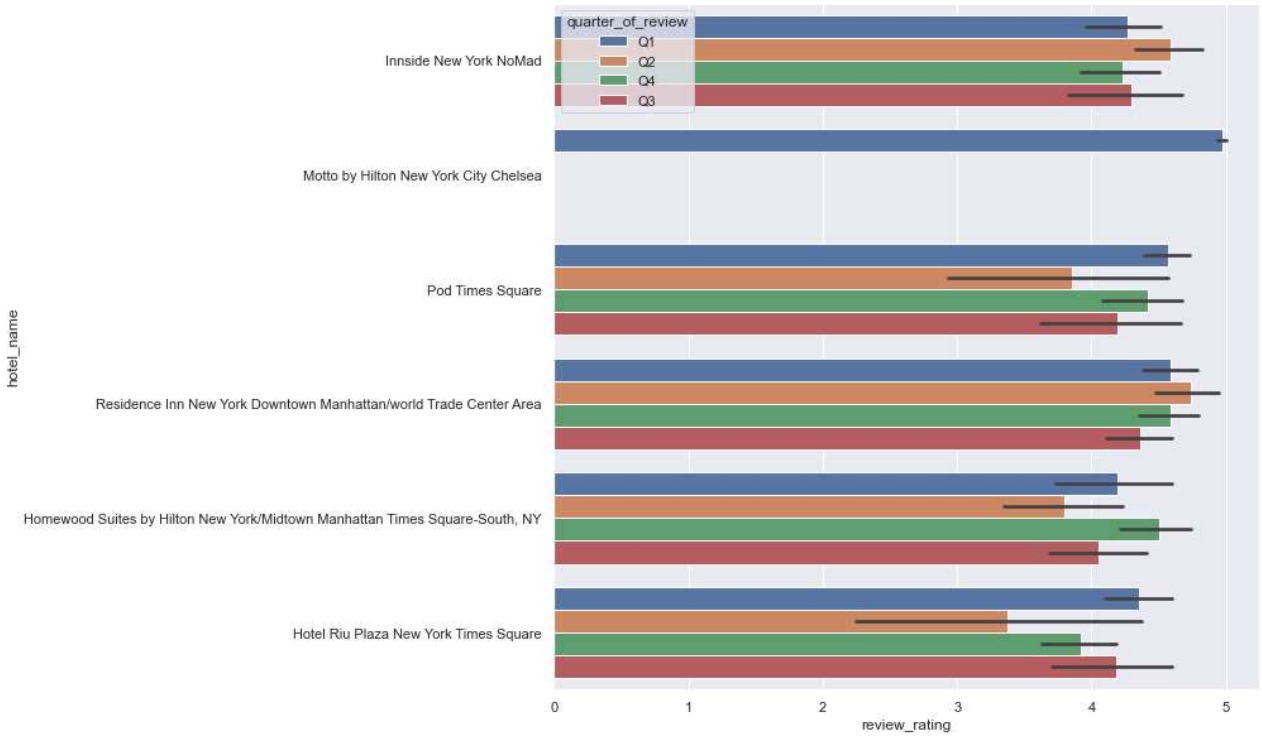
463 rows × 20 columns

In [65]:
```
#overall review rating for users who have significant contributions in reviews on Triva
ax2=sns.countplot(y="hotel_name", hue="review_rating", data=criteria)
for container in ax2.containers:
    ax2.bar_label(container)
sns.set(rc={'figure.figsize':(10,10)})
```



In [66]:

```
# Quarter wise distribution of review rating
sns.barplot(y="hotel_name",x="review_rating", hue="quarter_of_review", data=data)
sns.set(rc={'figure.figsize':(10,10)})
```



In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

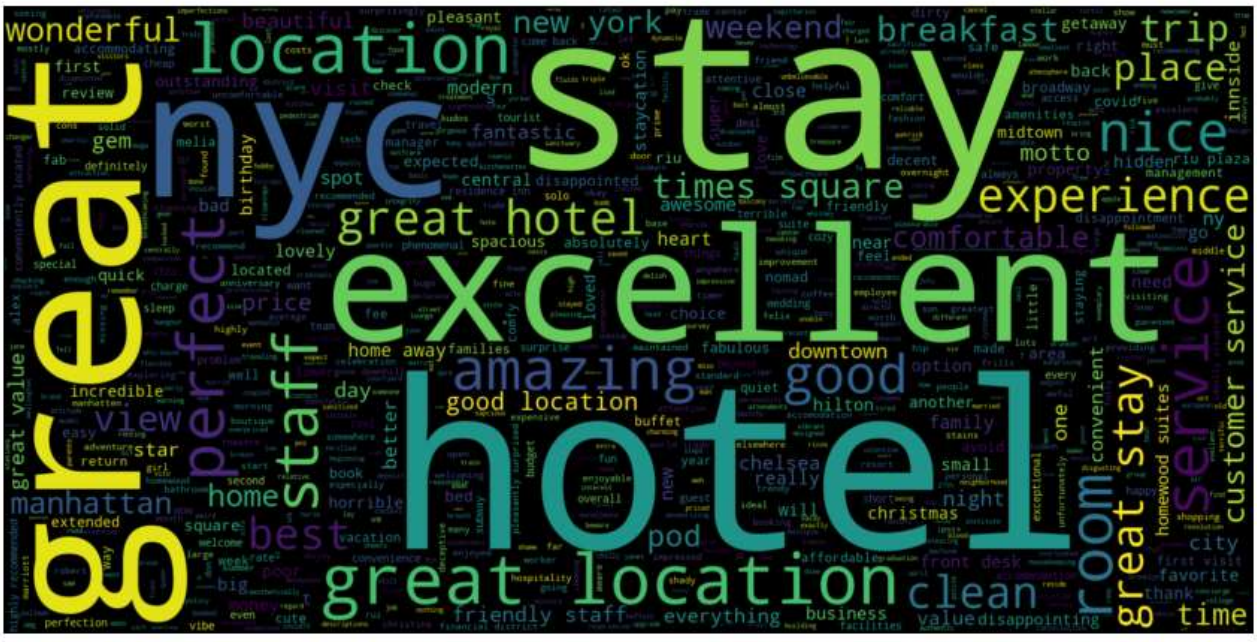
In [ ]:

NLP

In [67]:
```
import textblob
from textblob import TextBlob
```

In [68]:
```
word_string=" ".join(data['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 300,  width=1600,
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[68]: (-0.5, 1599.5, 799.5, -0.5)



In [69]:
```
data['cleaned_review_text']=data['review_text'].apply(lambda x: x.lower())
```

In [70]:
```
# Remove digits and words containing digits
```

```
data['cleaned_review_text']=data['cleaned_review_text'].apply(lambda x: re.sub('\w*\d\w
# Remove Punctuations
data['cleaned_review_text']=data['cleaned_review_text'].apply(lambda x: re.sub('[%s]' %
# Removing extra spaces
data['cleaned_review_text']=data['cleaned_review_text'].apply(lambda x: re.sub(' +',' '
```

In [71]:

```
# Importing spacy
import spacy
import en_core_web_sm
# Loading model
nlp = spacy.load('en_core_web_sm',disable=['parser', 'ner'])
# Lemmatization with stopwords removal
data['lemmatized']=data['cleaned_review_text'].apply(lambda x: ' '.join([token.lemma_ f
```

In [72]:

```
data.head(3)
```

Out[72]:

	hotel_name	overall_rating	date_of_review	user_name	review_rating	review_title	review_text	date_
0	Innside New York NoMad	45	22-Mar	Kim C	5.0	Great hotel!	We stayed in a family loft which is two connec...	Mar
1	Innside New York NoMad	45	21-Jun	Peggy M	5.0	The Innside Front Desk Staff	I just want to take a moment to say how wonder...	M
2	Innside New York NoMad	45	4-Mar	Imran	5.0	NYC gem	I go to NYC a lot, and this is hands down the ...	Ju

3 rows × 22 columns

In [73]:

```
import textblob
from textblob import TextBlob

# link : https://www.analyticsvidhya.com/blog/2020/04/beginners-guide-exploratory-data-
#need to check how polarity is calculated

data['polarity']=data['lemmatized'].apply(lambda x:TextBlob(x).sentiment.polarity)
```

In [74]:

```
print("3 Random Reviews with Highest Polarity:")
for index,review in enumerate(data.iloc[data['polarity'].sort_values(ascending=False)[:
    print('Review {}:\n'.format(index+1),review)
```

```
3 Random Reviews with Highest Polarity:
Review 1:
```

```
 Excellent place to stay for a show and to experience all the city has to offer! Great l
ocation for access to transportation and city sites! I will be back again and again! Exc
ellent staff and service!
Review 2:
 Came to visit a few friends in the tri state area and chose this hotel as my place of s
tay. I was very impressed with not only the ambience but the hotel staff from front desk
greeters to house keeping was incredibly awesome. Beautiful hotel and it is kept very be
autiful by its staff. Keep up the good work , Motto!!
Review 3:
 Stayed at the Residence Inn for a family vacation and it was great.  Location was great
- very close to 9/11 museum (one block) and Battery Park (5-8 min walk).  Jane at the fr
ont desk was wonderful.  About a 15-20 min ride on subway to Times Square.  Breakfast in
the morning.  Would stay again.
```

In [75]:
```
data.head(3)
```

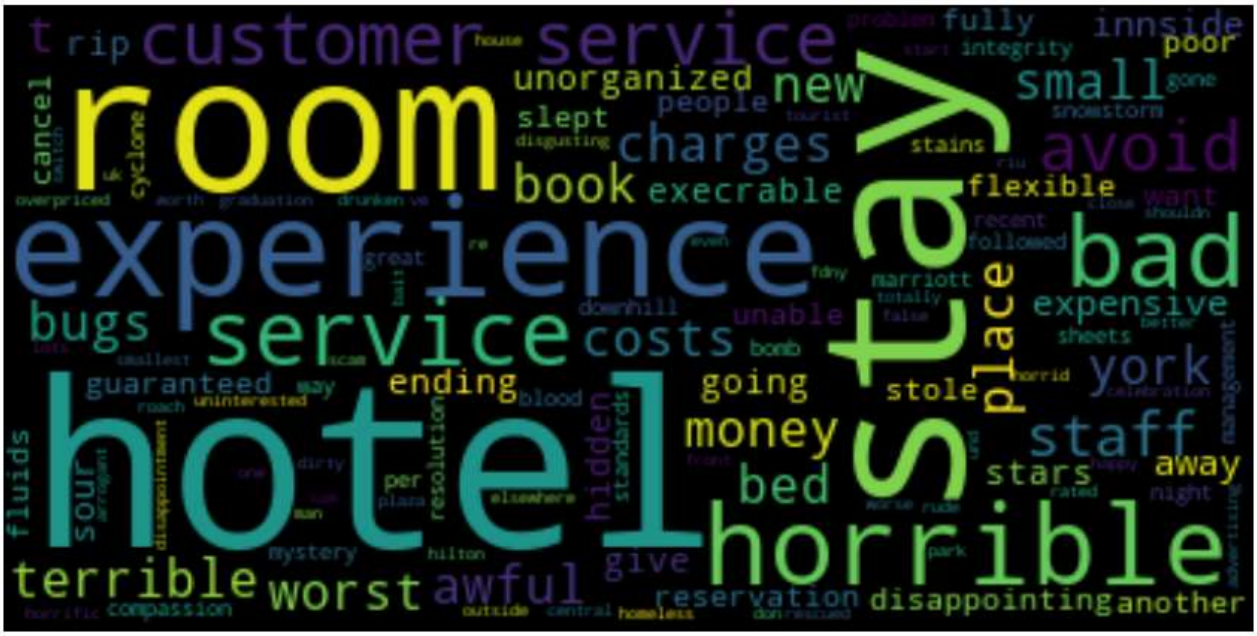
Out[75]:

	hotel_name	overall_rating	date_of_review	user_name	review_rating	review_title	review_text	date_
0	Innside New York NoMad	45	22-Mar	Kim C	5.0	Great hotel!	We stayed in a family loft which is two connec...	Mar
1	Innside New York NoMad	45	21-Jun	Peggy M	5.0	The Innside Front Desk Staff	I just want to take a moment to say how wonder...	M
2	Innside New York NoMad	45	4-Mar	Imran	5.0	NYC gem	I go to NYC a lot, and this is hands down the ...	Ju

3 rows × 23 columns

In [76]:
```
#user that have given rating == 1
filtered_df = data[data.review_rating.isin([1.0])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[76]:
```
(-0.5, 399.5, 199.5, -0.5)
```

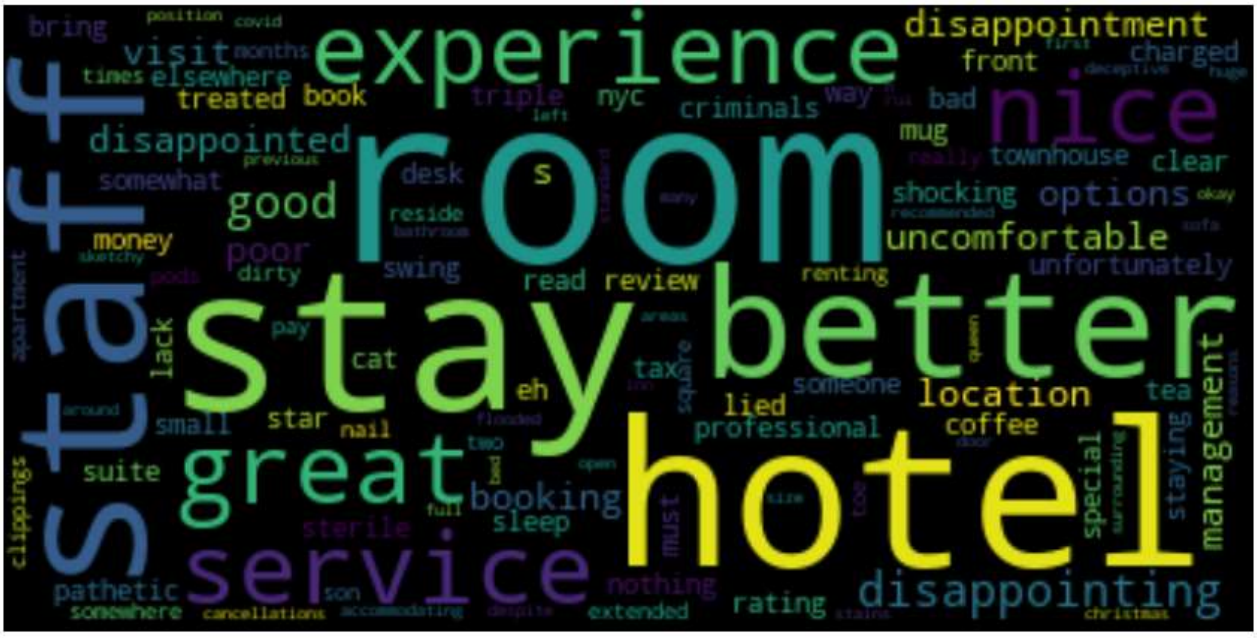


In [77]:

```
#user that have given rating == 2
filtered_df = data[data.review_rating.isin([2.0])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[77]:

```
(-0.5, 399.5, 199.5, -0.5)
```

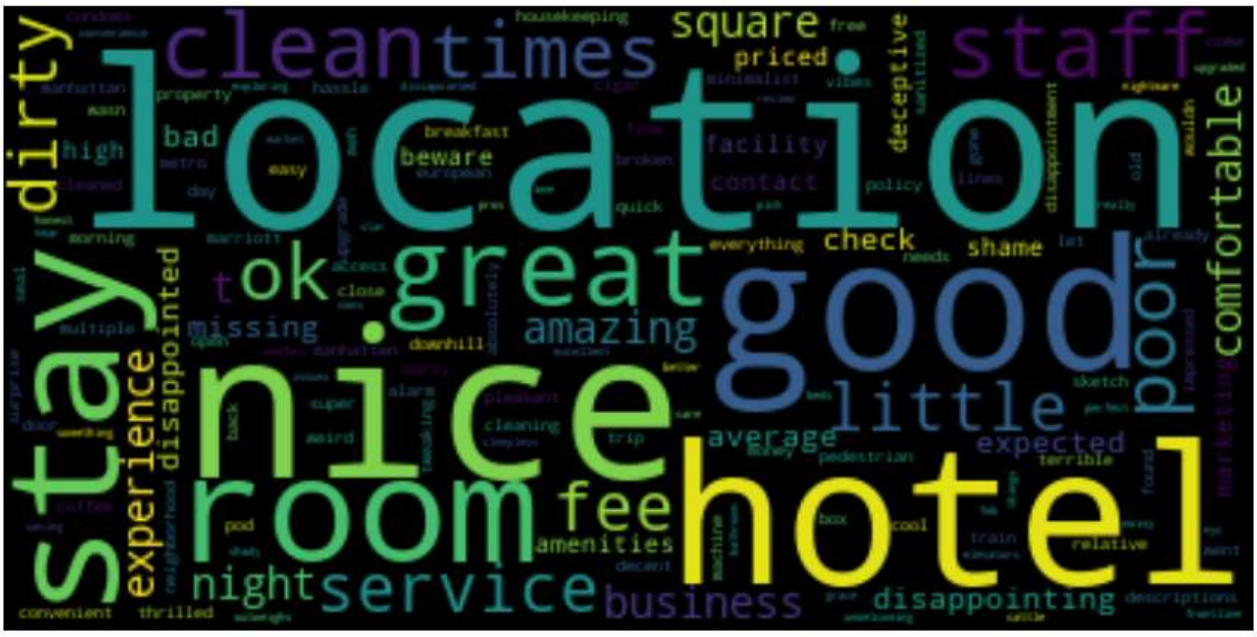


In [78]:

```
#user that have given rating == 3
filtered_df = data[data.review_rating.isin([3.0])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000,max_font_size= 75,width=400, heig
wc.generate(word_string)
```

```
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[78]: (-0.5, 399.5, 199.5, -0.5)



In [79]:
```
#user that have given rating == 4
filtered_df = data[data.review_rating.isin([4.0])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000,max_font_size= 75,width=400, heig
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[79]: (-0.5, 399.5, 199.5, -0.5)

In [80]:
```
#user that have given rating == 5
```

```
filtered_df = data[data.review_rating.isin([5.0])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

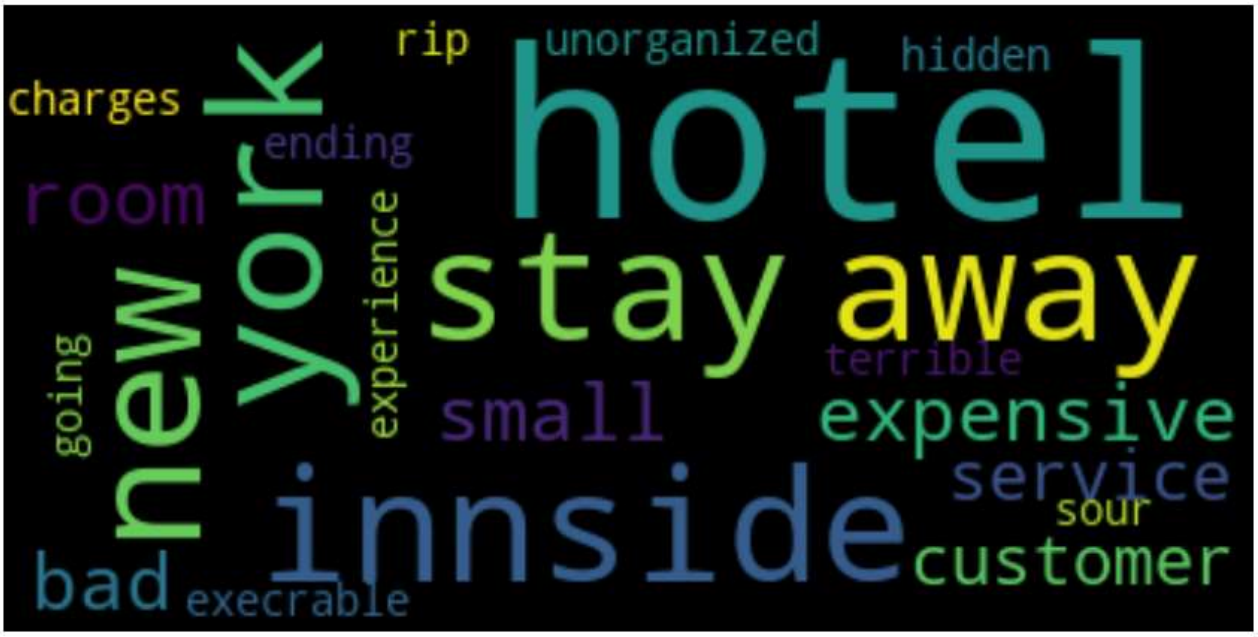
Out[80]: (-0.5, 399.5, 199.5, -0.5)

In [81]:
```
data['hotel_name'].unique()
```

Out[81]:
```
array(['Innside New York NoMad', 'Motto by Hilton New York City Chelsea',
       'Pod Times Square',
       'Residence Inn New York Downtown Manhattan/world Trade Center Area',
       'Homewood Suites by Hilton New York/Midtown Manhattan Times Square-South, NY',
       'Hotel Riu Plaza New York Times Square'], dtype=object)
```

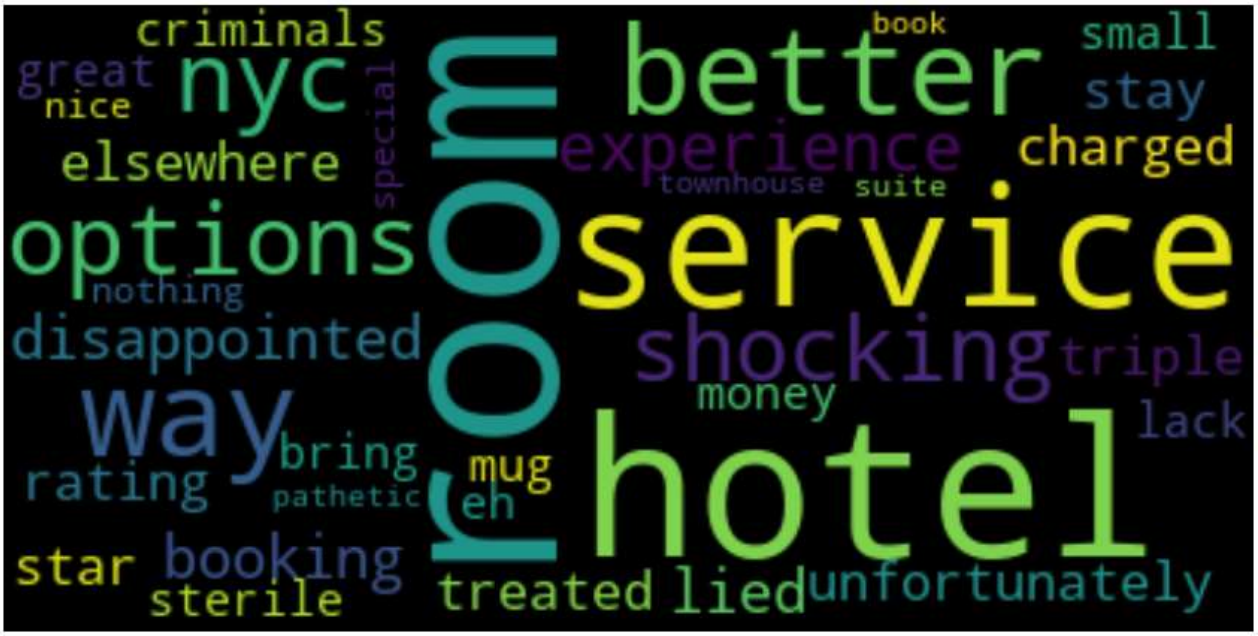
In [82]:
```
# For reviews for hotel "Innside New York NoMad" where review rating ==1
filtered_df = data[data.review_rating.isin([1.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Innside New York NoMad'])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[82]: (-0.5, 399.5, 199.5, -0.5)



In [83]:
```
# For reviews for hotel "Innside New York NoMad" where review rating ==2
filtered_df = data[data.review_rating.isin([2.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Innside New York NoMad'])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

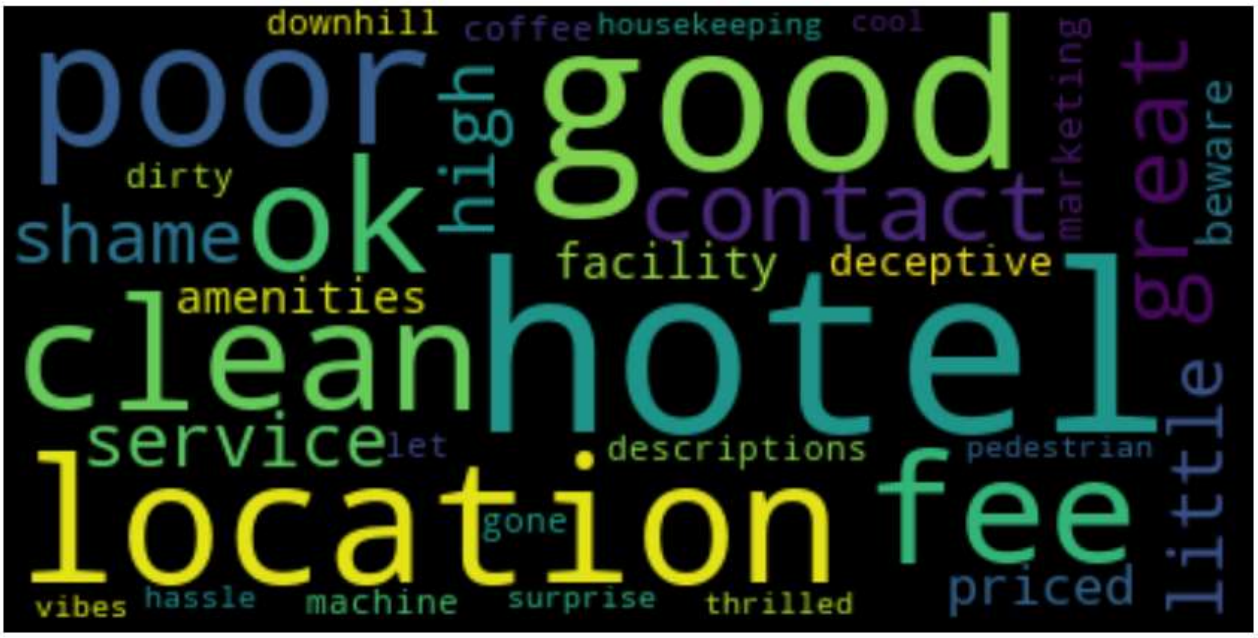
Out[83]: (-0.5, 399.5, 199.5, -0.5)



In [84]:
```
# For reviews for hotel "Innside New York NoMad" where review rating ==3
filtered_df = data[data.review_rating.isin([3.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Innside New York NoMad'])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
```

```
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[84]: (-0.5, 399.5, 199.5, -0.5)



In [85]:

```
# For reviews for hotel "Innside New York NoMad" where review rating ==4
filtered_df = data[data.review_rating.isin([4.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Innside New York NoMad'])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[85]: (-0.5, 399.5, 199.5, -0.5)

In [86]:

```
# For reviews for hotel "Innside New York NoMad" where review rating ==5
filtered_df = data[data.review_rating.isin([5.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Innside New York NoMad'])]
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[86]: (-0.5, 399.5, 199.5, -0.5)

In [87]:

```
# For reviews for hotel "Hotel Riu Plaza New York Times Square" where review rating ==1
filtered_df = data[data.review_rating.isin([1.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Hotel Riu Plaza New York Times Sq
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
```

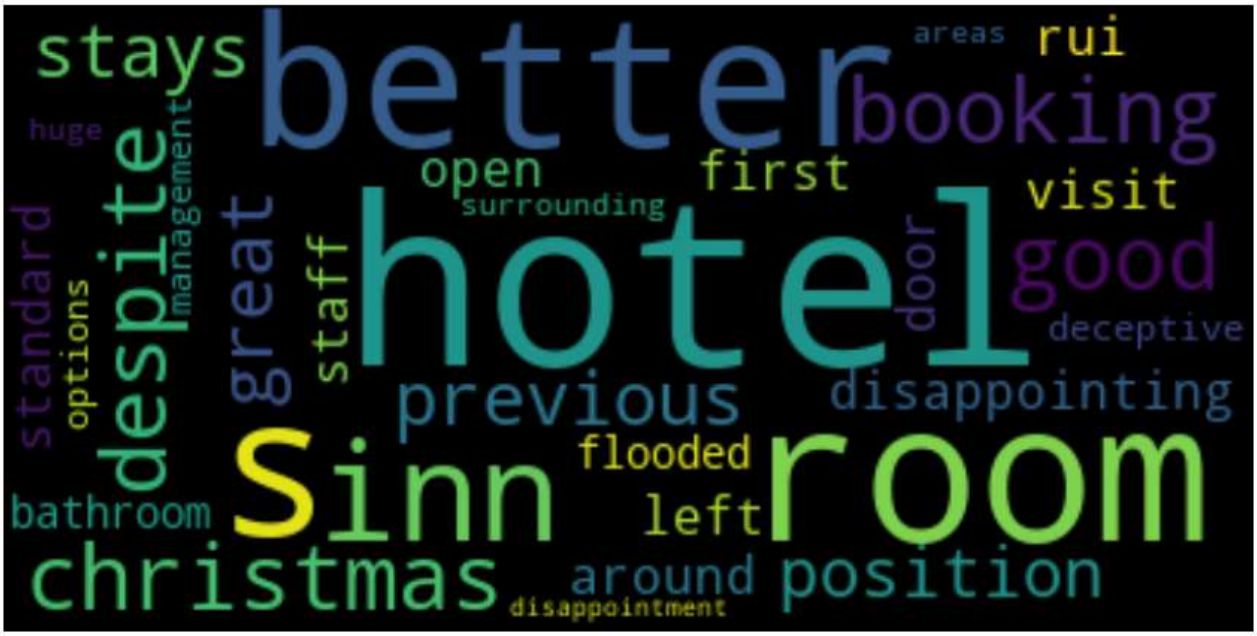
```
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[87]: (-0.5, 399.5, 199.5, -0.5)

In [88]:

```
# For reviews for hotel "Hotel Riu Plaza New York Times Square" where review rating ==2
filtered_df = data[data.review_rating.isin([2.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Hotel Riu Plaza New York Times Sq
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[88]: (-0.5, 399.5, 199.5, -0.5)



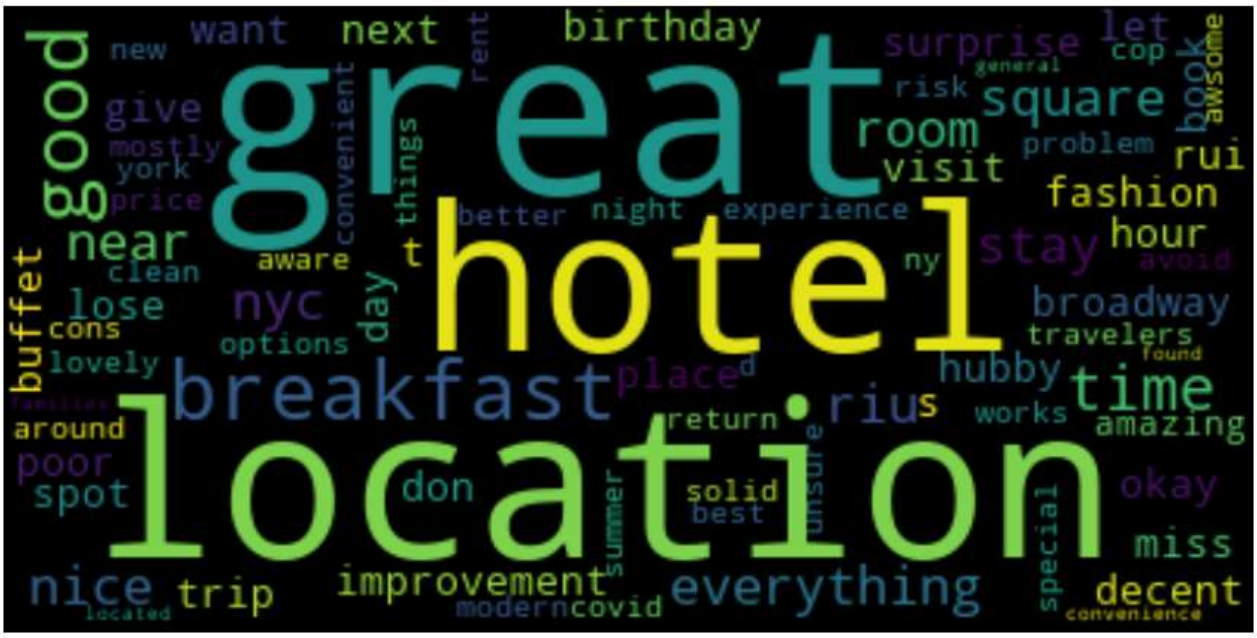
In [89]:
```
# For reviews for hotel "Hotel Riu Plaza New York Times Square" where review rating ==3
filtered_df = data[data.review_rating.isin([3.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Hotel Riu Plaza New York Times Sq
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[89]:
```
(-0.5, 399.5, 199.5, -0.5)
```

In [90]:
```
# For reviews for hotel "Hotel Riu Plaza New York Times Square" where review rating ==4
filtered_df = data[data.review_rating.isin([4.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Hotel Riu Plaza New York Times Sq
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
```

```
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[90]: (-0.5, 399.5, 199.5, -0.5)



In [91]:

```
# For reviews for hotel "Hotel Riu Plaza New York Times Square" where review rating ==5
filtered_df = data[data.review_rating.isin([5.0])]
filtered_df=filtered_df[filtered_df.hotel_name.isin(['Hotel Riu Plaza New York Times Sq
word_string=" ".join(filtered_df['review_title'].str.lower())
plt.figure(figsize=(15,15))
wc = WordCloud( stopwords = STOPWORDS, max_words=2000, max_font_size= 75,width=400, hei
wc.generate(word_string)
plt.imshow(wc.recolor( colormap= 'viridis' , random_state=17), interpolation="bilinear"
plt.axis('off')
```

Out[91]: (-0.5, 399.5, 199.5, -0.5)


hotel
perfect
location
stay
great
excellent
comfortable
beautiful
back
time
recommend
visit
trip
riu
gem
yanet
highly
room
close
john
overlooked
fabulous
view
new york
nyc
square
nice
subway
best
short
place
enjoyed
fantastic
customer
christmas
first
service
amazing
staff
value
clean
posh
wonderful
mega
really
night
riu plaza
experience
go
absolutely
day
times square
breakfast
shopping
good
broadway
alternative

In [ ]:

In [ ]: