In [ ]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
nltk.download('stopwords')

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk.stem.porter import PorterStemmer

import string
import re
import textblob
from textblob import TextBlob
import os

from wordcloud import WordCloud, STOPWORDS
from wordcloud import ImageColorGenerator
import warnings
%matplotlib inline
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [ ]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

In [ ]:

```
tweets_df = pd.read_csv(r'/content/drive/MyDrive/dataset_sma/Tweets.csv')
```

In [ ]:

```
tweets_df.head(5)
```

Out [54]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gol
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN

In [ ]:

```
tweets_df.shape
```

Out [55]:

```
(14640, 15)
```

In [ ]:

```
tweets_df.head()
```

Out [56]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gol
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gol
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN

In [ ]:

```
tweets_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   tweet_id                      14640 non-null  int64  
 1   airline_sentiment             14640 non-null  object 
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object 
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object 
 6   airline_sentiment_gold        40 non-null     object 
 7   name                          14640 non-null  object 
 8   negativereason_gold           32 non-null     object 
 9   retweet_count                 14640 non-null  int64  
 10  text                          14640 non-null  object 
 11  tweet_coord                   1019 non-null   object 
 12  tweet_created                 14640 non-null  object 
 13  tweet_location                9907 non-null   object 
 14  user_timezone                 9820 non-null   object 
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
```

In [ ]:

```
tweets_df.value_counts(tweets_df['airline'])
```

Out [58]:

```
airline
United            3822
US Airways        2913
American          2759
Southwest         2420
Delta             2222
Virgin America     504
Name: count, dtype: int64
```

In [ ]:

```
tweets_df.value_counts(tweets_df['airline_sentiment_gold'])
```

Out [59]:

```
airline_sentiment_gold
negative    32
positive     5
neutral      3
Name: count, dtype: int64
```

In [ ]:

```
tweets_df['airline_sentiment_gold'].isnull().sum()
```

Out [60]:

```
14600
```

In [ ]:

```
tweets_df.value_counts()
```

Out [61]:

```
tweet_id            airline_sentiment  airline_sentiment_confidence  negativereason    negativereason_confidence  airline
airline_sentiment_gold  name              negativereason_gold  retweet_count  text
tweet_coord                  tweet_created              tweet_location  user_timezone
567778009013178368  negative           1.0000                        Cancelled Flight  1.0000                     United      negative
realmikesmith    Cancelled Flight     0             @united So what do you offer now that my flight was Cancelled Flighted and I'm
stranded away from home and work?  [26.37852293, -81.78472152]  2015-02-17 12:10:00 -0800  Chicago         Eastern Time (US & Canada)
1
569887533267611648  negative           0.8563                        Late Flight       0.5938                     US Airways  negative
ConstanceSCHERE  Late Flight          0             @USAirways Seriously doubt that as I am still sitting inside at the gate.
[39.8805621, -75.23893393]  2015-02-23 07:52:30 -0800  Boston, MA      Atlantic Time (Canada)        1
Name: count, dtype: int64
```

In [ ]:

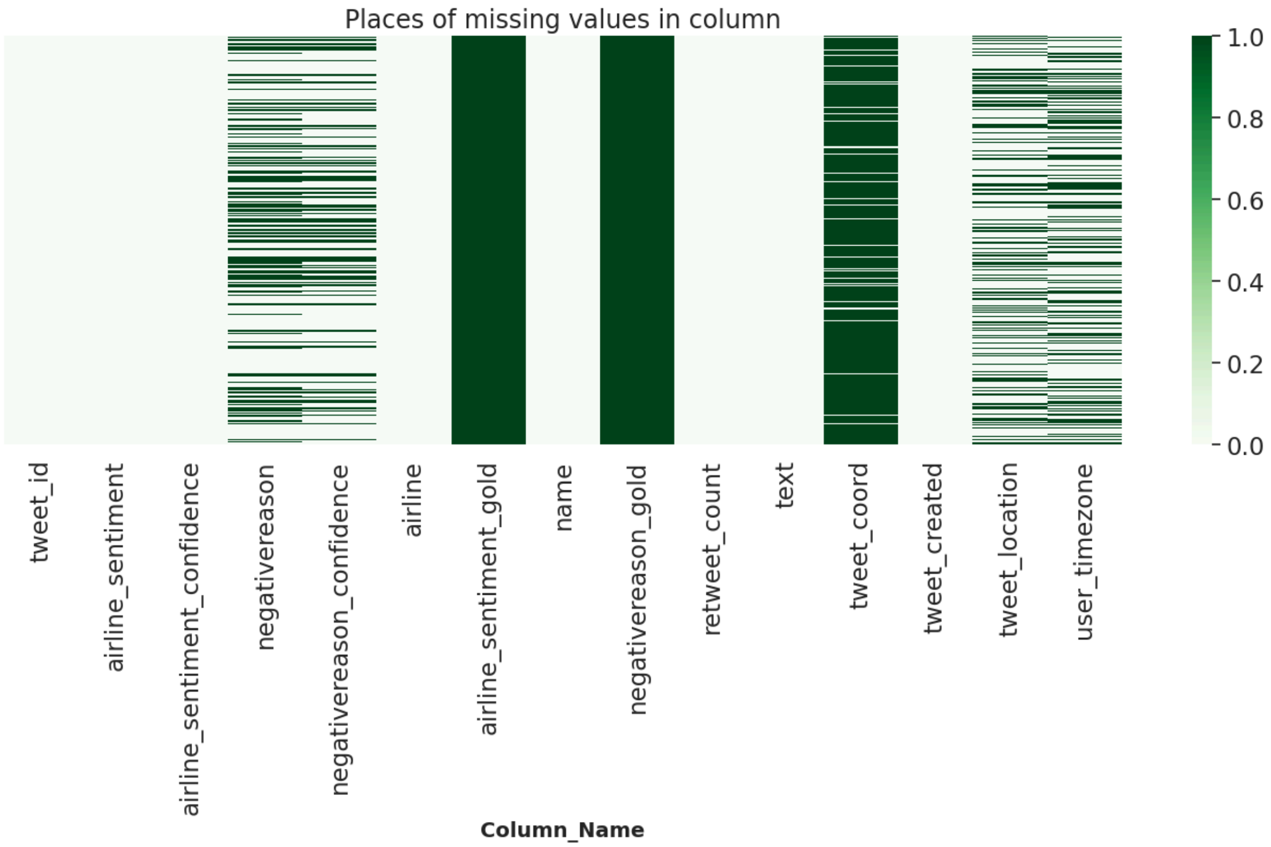
```
tweets_df.isnull().sum()
```

Out [62]:

```
tweet_id                            0
airline_sentiment                   0
airline_sentiment_confidence        0
negativereason                   5462
negativereason_confidence        4118
airline                             0
airline_sentiment_gold          14600
name                                0
negativereason_gold             14608
retweet_count                       0
text                                0
tweet_coord                     13621
tweet_created                       0
tweet_location                   4733
user_timezone                    4820
dtype: int64
```

In [ ]:

```
plt.figure(figsize=(17, 5))
sns.heatmap(tweets_df.isnull(), cbar=True, yticklabels=False,cmap="Greens")
plt.xlabel("Column_Name", size=14, weight="bold")
plt.title("Places of missing values in column",size=17)
plt.show()
```

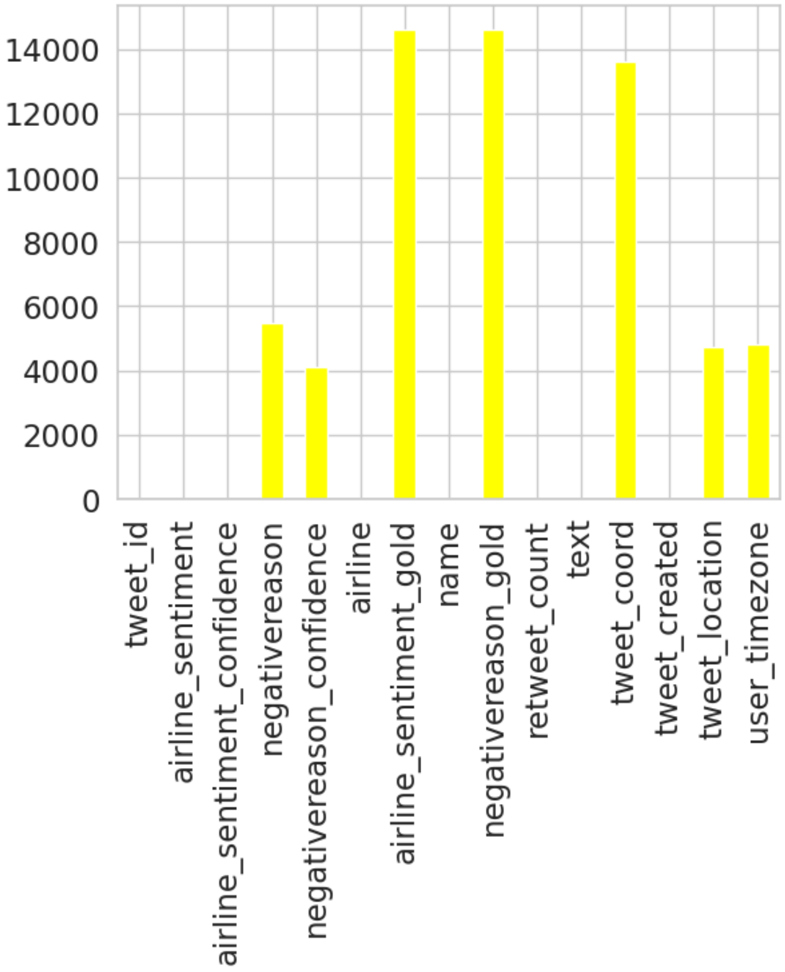


In [ ]:

```
tweets_df.isnull().sum().plot(kind="bar",color="yellow")
```

Out [64]:

```
<Axes: >
```



In [ ]:

```
import plotly.graph_objects as go
Top_Location_Of_tweet= tweets_df['airline'].value_counts().head (10)
```

In [ ]:

```
print(Top_Location_Of_tweet)
```

```
airline
United            3822
US Airways        2913
American          2759
Southwest         2420
Delta             2222
```

```
Virgin America     504
Name: count, dtype: int64
```

In [ ]:

```
from nltk. corpus import stopwords
stop = stopwords.words('english')
tweets_df['text'].apply(lambda x: [item for item in x if item not in stop])
tweets_df.shape
```

Out [67]:

```
(14640, 15)
```

In [ ]:

```
tweets_df['text'].head(10)
```

Out [68]:

```
0                  @VirginAmerica What @dhepburn said.
1    @VirginAmerica plus you've added commercials t...
2    @VirginAmerica I didn't today... Must mean I n...
3    @VirginAmerica it's really aggressive to blast...
4    @VirginAmerica and it's a really big bad thing...
5    @VirginAmerica seriously would pay $30 a fligh...
6    @VirginAmerica yes, nearly every time I fly VX...
7    @VirginAmerica Really missed a prime opportuni...
8      @virginamerica Well, I didn't…but NOW I DO! :-D
9    @VirginAmerica it was amazing, and arrived an ...
Name: text, dtype: object
```

In [ ]:

```
!pip install tweet-preprocessor
```

```
Requirement already satisfied: tweet-preprocessor in /usr/local/lib/python3.10/dist-packages (0.6.0)
```

In [ ]:

```
punct  =  ['%','/',':','\\','&amp','&',';','?']

def remove_punctuations(text):
  for punctuation in punct:
    text = text.replace(punctuation,'')
  return text
```

In [ ]:

```
tweets_df['text'] = tweets_df['text'].apply(lambda x: remove_punctuations(x))
```

In [ ]:

```
tweets_df['text'].isnull().sum()
```

Out [72]:

```
0
```

In [ ]:

```
tweets_df['text'].replace( '', np.nan, inplace=True)
tweets_df.dropna(subset=["text"],inplace=True)
len(tweets_df)
```

Out [73]:

```
14640
```

In [ ]:

```
tweets_df = tweets_df.reset_index(drop=True)
tweets_df.head()
```

Out [74]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gol
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN

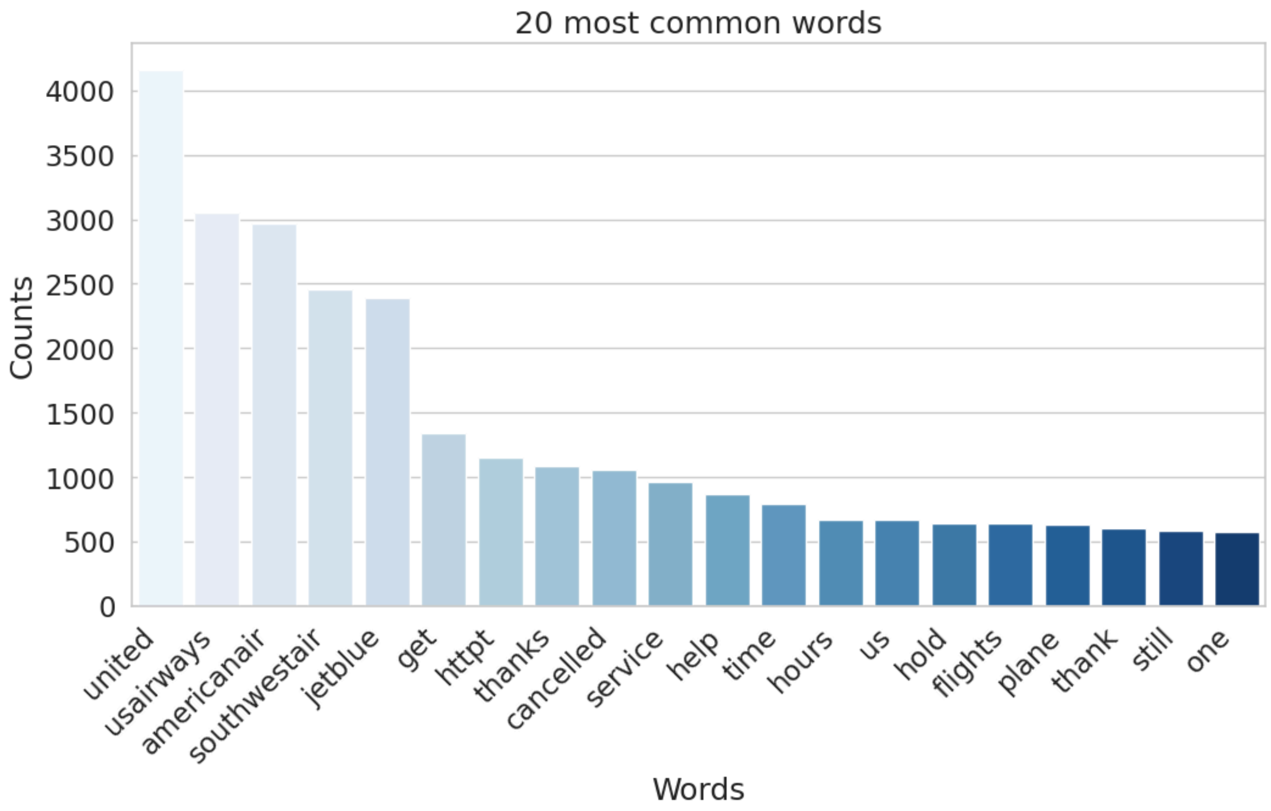
In [ ]:

```
from sklearn.feature_extraction. text import TfidfVectorizer, CountVectorizer
```

In [ ]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer

sns.set_style('whitegrid')
%matplotlib inline

stop = stop + ['Virgin America', 'San Francisco', 'Boston', 'New York', 'customer', 'flight', 'airline', 'San

def plot_20_most_common_words(count_data, count_vectorizer):
    words = count_vectorizer.get_feature_names_out()
    total_counts = np.zeros(len(words))

    for t in count_data:
        total_counts += t.toarray()[0]

    count_dict = dict(zip(words, total_counts))
    count_dict = sorted(count_dict.items(), key=lambda x: x[1], reverse=True)[:20]

    words = [w[0] for w in count_dict]
    counts = [w[1] for w in count_dict]

    x_pos = np.arange(len(words))

    plt.figure(figsize=(12, 6))
    sns.set_context('notebook', font_scale=1.5)
    sns.barplot(x=x_pos, y=counts, palette='Blues')
    plt.title('20 most common words')
    plt.xticks(x_pos, words, rotation=45, ha='right')
    plt.xlabel('Words')
    plt.ylabel('Counts')
    plt.show()


count_vectorizer = CountVectorizer(stop_words=stop)
count_data = count_vectorizer.fit_transform(tweets_df['text'])
plot_20_most_common_words(count_data, count_vectorizer)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/feature_extraction/text.py:409: UserWarning:

Your stop_words may be inconsistent with your preprocessing. Tokenizing the stop words generated tokens ['america', 'boston',
'california', 'diego', 'francisco', 'new', 'oakland', 'san', 'virgin', 'york'] not in stop_words.

<ipython-input-76-5eac74ad3f22>:28: FutureWarning:



Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.
```



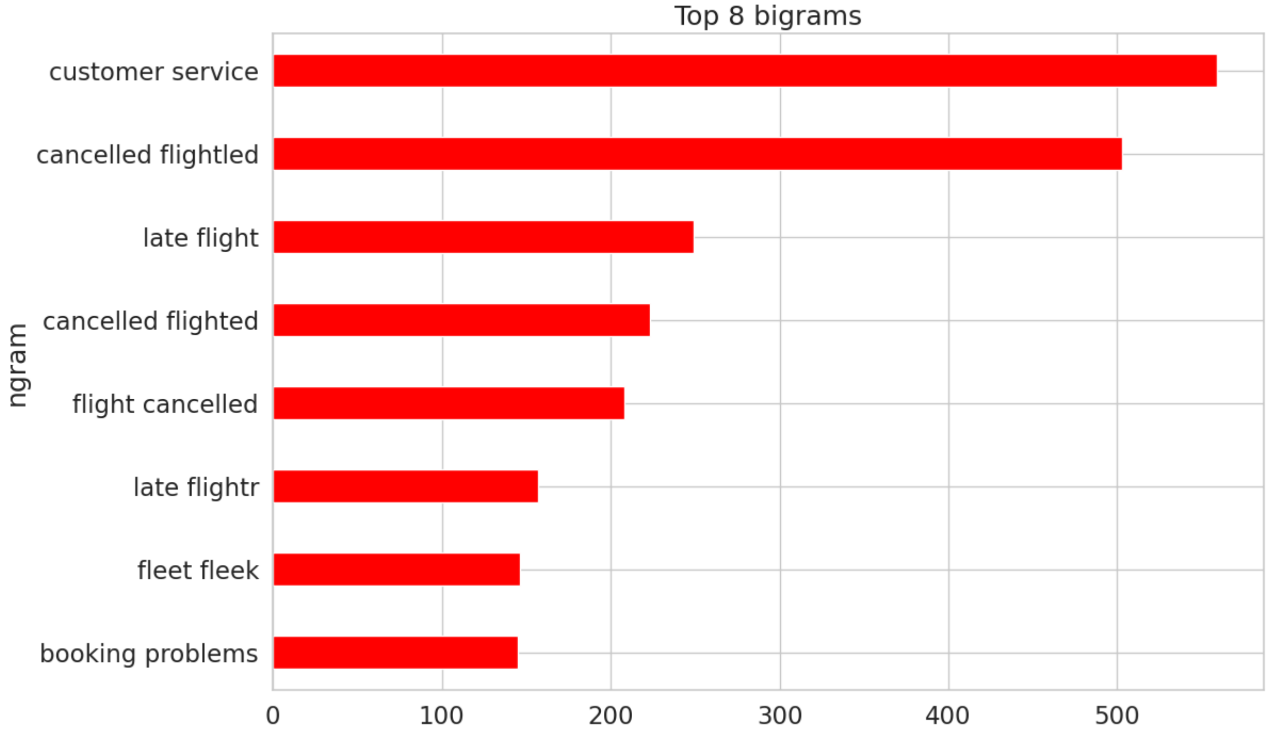
In [ ]:

```
import cufflinks as cf
cf.go_offline()
cf.set_config_file(offline=False, world_readable=True)

def get_top_n_bigram(corpus, n=None) :
  vec = CountVectorizer(ngram_range=(2, 4), stop_words="english").fit(corpus)
  bag_of_words = vec.transform(corpus)
  sum_words = bag_of_words.sum(axis=0)
  words_freq =[(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items()]
  words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
  return words_freq[:n]
```

```
common_words = get_top_n_bigram(tweets_df['text'] , 8)
mydict={}
for word, freq in common_words:
  bigram_df = pd.DataFrame(common_words,columns = ['ngram', 'count'])

bigram_df.groupby( 'ngram' ).sum()['count'].sort_values(ascending=False).sort_values().plot.barh(title = 'Top
```

Out [77]:

```
<Axes: title={'center': 'Top 8 bigrams'}, ylabel='ngram'>
```



APSIT REVIEW DATASET

In [ ]:

```
tweets_df1 = pd.read_csv(r'/content/drive/MyDrive/dataset_sma/google (1).csv')
```

In [ ]:

```
tweets_df1.head(5)
```

Out [79]:

	Link	Username	Rank	Timeline	Review	Response
0	https://lh3.googleusercontent.com/a-/ALV-UjWHb...	Saurabh Kanade	Local Guide · 37 reviews · 20 photos	6 months ago	Amezing ClassRooms.\nAir Conditioner & Fan Bot...	Like
1	https://lh3.googleusercontent.com/a/ACg8ocKaRp...	Rasika Pujare	Local Guide · 18 reviews · 170 photos	2 years ago	I visited the institute as it was my examinati...	NaN
2	https://lh3.googleusercontent.com/a-/ALV-UjV4l...	V S	Local Guide · 22 reviews · 7 photos	2 months ago	This was my CAT exam center. Though exam didn'...	Share
3	https://lh3.googleusercontent.com/a-/ALV-UjUV4...	Sneha Yadav	Local Guide · 7 reviews · 61 photos	4 years ago	It's the best college in Thane, growing rapidl...	NaN
4	https://lh3.googleusercontent.com/a-/ALV-UjWll...	AJP Travel	Local Guide · 112 reviews · 3,469 photos	2 years ago	It is beside highway so keep check on take in ...	NaN

In [ ]:

```
tweets_df1.shape
```

Out [80]:

```
(190, 6)
```

In [ ]:

```
tweets_df1.head()
```

Out [81]:

	Link	Username	Rank	Timeline	Review	Response
0	https://lh3.googleusercontent.com/a-/ALV-UjWHb...	Saurabh Kanade	Local Guide · 37 reviews · 20 photos	6 months ago	Amezing ClassRooms.\nAir Conditioner & Fan Bot...	Like
1	https://lh3.googleusercontent.com/a/ACg8ocKaRp...	Rasika Pujare	Local Guide · 18 reviews · 170 photos	2 years ago	I visited the institute as it was my examinati...	NaN
2	https://lh3.googleusercontent.com/a-/ALV-UjV4l...	V S	Local Guide · 22 reviews · 7 photos	2 months ago	This was my CAT exam center. Though exam didn'...	Share

	Link	Username	Rank	Timeline	Review	Response
3	https://lh3.googleusercontent.com/a-/ALV-UjUV4...	Sneha Yadav	Local Guide · 7 reviews · 61 photos	4 years ago	It's the best college in Thane, growing rapidl...	NaN
4	https://lh3.googleusercontent.com/a-/ALV-UjWll...	AJP Travel	Local Guide · 112 reviews · 3,469 photos	2 years ago	It is beside highway so keep check on take in ...	NaN

```
In [ ]: tweets_df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 190 entries, 0 to 189
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Link      190 non-null    object
 1   Username  190 non-null    object
 2   Rank      190 non-null    object
 3   Timeline  190 non-null    object
 4   Review    190 non-null    object
 5   Response  168 non-null    object
dtypes: object(6)
memory usage: 9.0+ KB
```

```
In [ ]: tweets_df1.value_counts(tweets_df1['Review'])
```

```
Out [83]: Review
The college is very good and road touch.\n\nThe faculty is nice and good infrastructure. …
2
2019-20 batch have not yet received their convocation wheras 2020-21 batch have received??
1
One of the best engineering college in Mumbai.\nCollege has Nice infrastructure, AC Classrooms and Advanced Labs.\nMany Courses
conducted by college free of cost other than curriculum.\nCollege having excellent  TPO cell as well.
1
One of the BEST College lf MUMBAI UNIVERSITY to study at and a Great Organization to work with. Perfectly connected to Central and
Western region and exactly in front of upcoming Metro station. Amazing amenities and environment to live, learn and work. Setting
benchmarks in terms of education and recruitements.    1
One of the best Engineering colleges in thane with best infrastructure including all labs & classrooms Air Conditioned. …
1

..
Excellent engineering college in Mumbai university.\nLocation is road touch.\nAll classes are equipped with smart boards, projector and
A.C. …
1
Excellent infrastructure and expert faculties in every domain offer students the opportunity to be industry ready. Hands-on training and
recent software courses are being offered to students
1
Excellent infrastructure, Air conditioning class rooms and laboratories. Great initiative's like Project Based Learning, attendance
rewards, incubation centre.\nOverall, I would like to give 5 star rating to APSIT.
1
Faculty ,management staff is great . Infrastructure is great . Inshort one of the best college in Thane site.
1
very nice college. friendly cooperative and helpful teaching and non teaching staff.\nAmazing infrastructure both civil and IT. …
1
Name: count, Length: 189, dtype: int64
```

```
In [ ]: tweets_df1.value_counts(tweets_df1['Response'])
```

```
Out [84]: Response
Share    165
Like       3
Name: count, dtype: int64
```

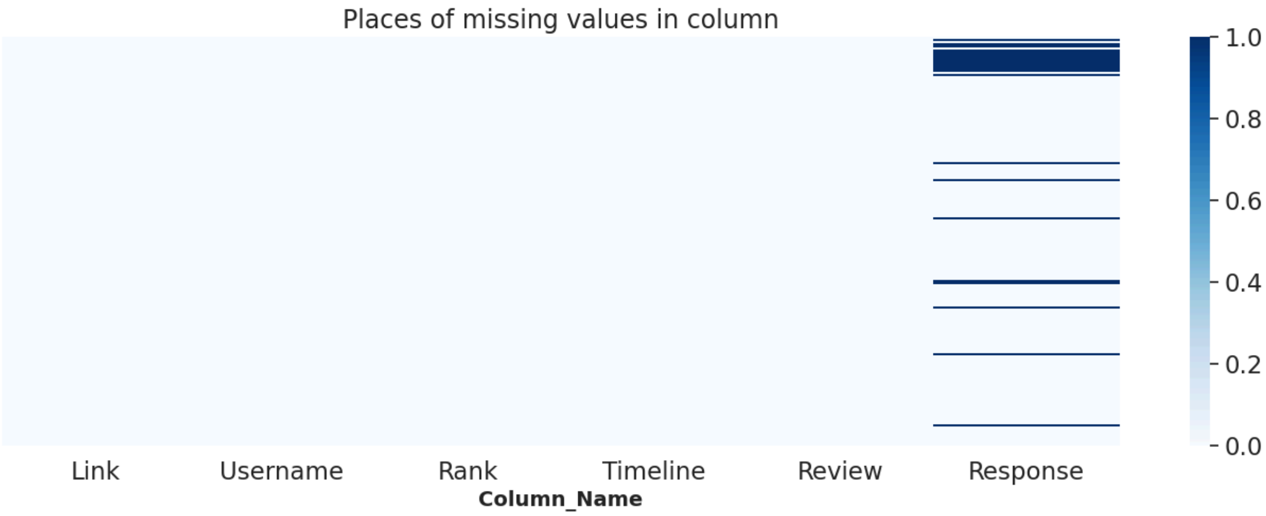
```
In [ ]: tweets_df1['Response'].isnull().sum()
```

```
Out [85]: 22
```

```
In [ ]: tweets_df1.isnull().sum()
```

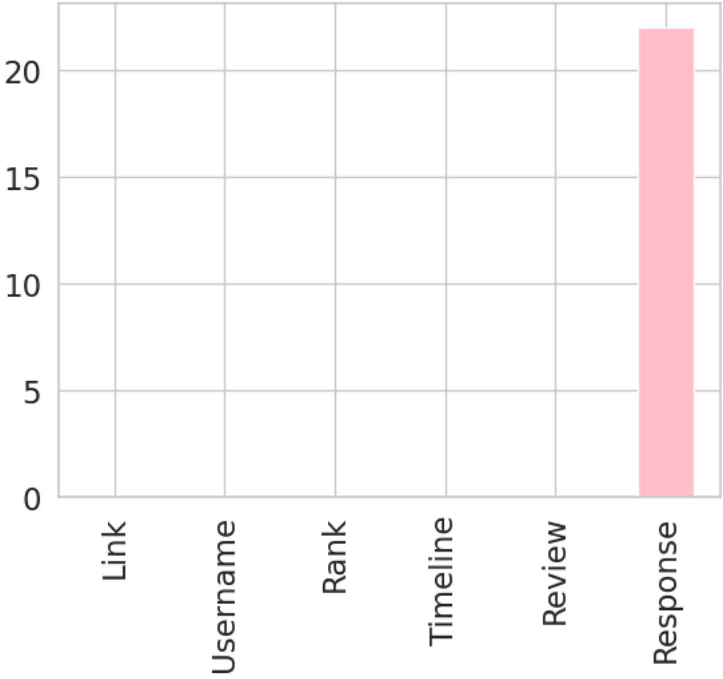
```
Out [86]: Link         0
Username     0
Rank         0
Timeline     0
Review       0
Response    22
dtype: int64
```

```
In [ ]: plt.figure(figsize=(17, 5))
        sns.heatmap(tweets_df1.isnull(), cbar=True, yticklabels=False,cmap="Blues")
        plt.xlabel("Column_Name", size=14, weight="bold")
        plt.title("Places of missing values in column",size=17)
        plt.show()
```



In [ ]:
```
tweets_df1.isnull().sum().plot(kind="bar",color="pink")
```

Out [88]: <Axes: >



In [ ]:
```
import plotly.graph_objects as go
Top_Location_Of_tweet= tweets_df1['Review'].value_counts().head (10)
```

In [ ]:
```
print(Top_Location_Of_tweet)
```

```
Review
The college is very good and road touch.\n\nThe faculty is nice and good infrastructure. …
Amezing ClassRooms.\nAir Conditioner & Fan Both are Available In Each Class.\n5 Floors Building. …
Impossible for you to reach without google ma.\nNot even a single auto rickshawala or bus conductor knows where the college is.\nReally difficult to
Apsit is the best education center in Thane with the best learning exposure.College infrastructure is very good.The college provides workshops and c
Best College Ever.\nBest Management and Awesome Teachers.\nIT department is the best and most preferred department. …
I have visited this to give exam of railway, everything was well managed and hygienic.
A very good center for online exams.. Nice campus with proficient staff
Very bad placement cell for civi engineering this college has no tie ups only good for it and software engg if you want to go for good college don't
The infrastructure is really nice but it lacks in campus area the fest done by the student council is very up to the mark
Surely one of the best Institute for Engineering in Mumbai Thane region..\n\nAwesome Teaching - Learning Environment.. …
Name: count, dtype: int64
```

In [ ]:
```
from nltk. corpus import stopwords
stop = stopwords.words('english')
tweets_df1['Review'].apply(lambda x: [item for item in x if item not in stop])
tweets_df1.shape
```

Out [91]: (190, 6)

In [ ]:
```
tweets_df1['Review'].head(10)
```

Out [92]:
```
0    Amezing ClassRooms.\nAir Conditioner & Fan Bot...
1    I visited the institute as it was my examinati...
2    This was my CAT exam center. Though exam didn'...
3    It's the best college in Thane, growing rapidl...
4    It is beside highway so keep check on take in ...
5    Best college for engineering colleges if in fr...
6    AP SHAH  INSTITUTE OF                     TEC...
7    A.P. Shah Institute of Technology, one of the ...
8    This is one of the best engineering college in...
9    The Infrastructure of this college is excellen...
Name: Review, dtype: object
```

In [ ]:
```
tweets_df1['Review'] = tweets_df1['Review'].apply(lambda x: remove_punctuations(x))
```

In [ ]:
```
tweets_df1['Review'].head(10)
```

Out [94]:
```
0    Amezing ClassRooms.\nAir Conditioner  Fan Both...
1    I visited the institute as it was my examinati...
2    This was my CAT exam center. Though exam didn'...
3    It's the best college in Thane, growing rapidl...
4    It is beside highway so keep check on take in ...
5    Best college for engineering colleges if in fr...
6    AP SHAH  INSTITUTE OF                      TEC...
7    A.P. Shah Institute of Technology, one of the ...
8    This is one of the best engineering college in...
9    The Infrastructure of this college is excellen...
Name: Review, dtype: object
```

In [ ]:
```
tweets_df1['Review'].isnull().sum()
```

Out [95]:
```
0
```

In [ ]:
```
tweets_df1['Review'].replace( '', np.nan, inplace=True)
tweets_df1.dropna(subset=["Review"],inplace=True)
len(tweets_df1)
```

Out [96]:
```
190
```

In [ ]:
```
tweets_df1 = tweets_df1.reset_index(drop=True)
tweets_df1.head()
```

Out [97]:

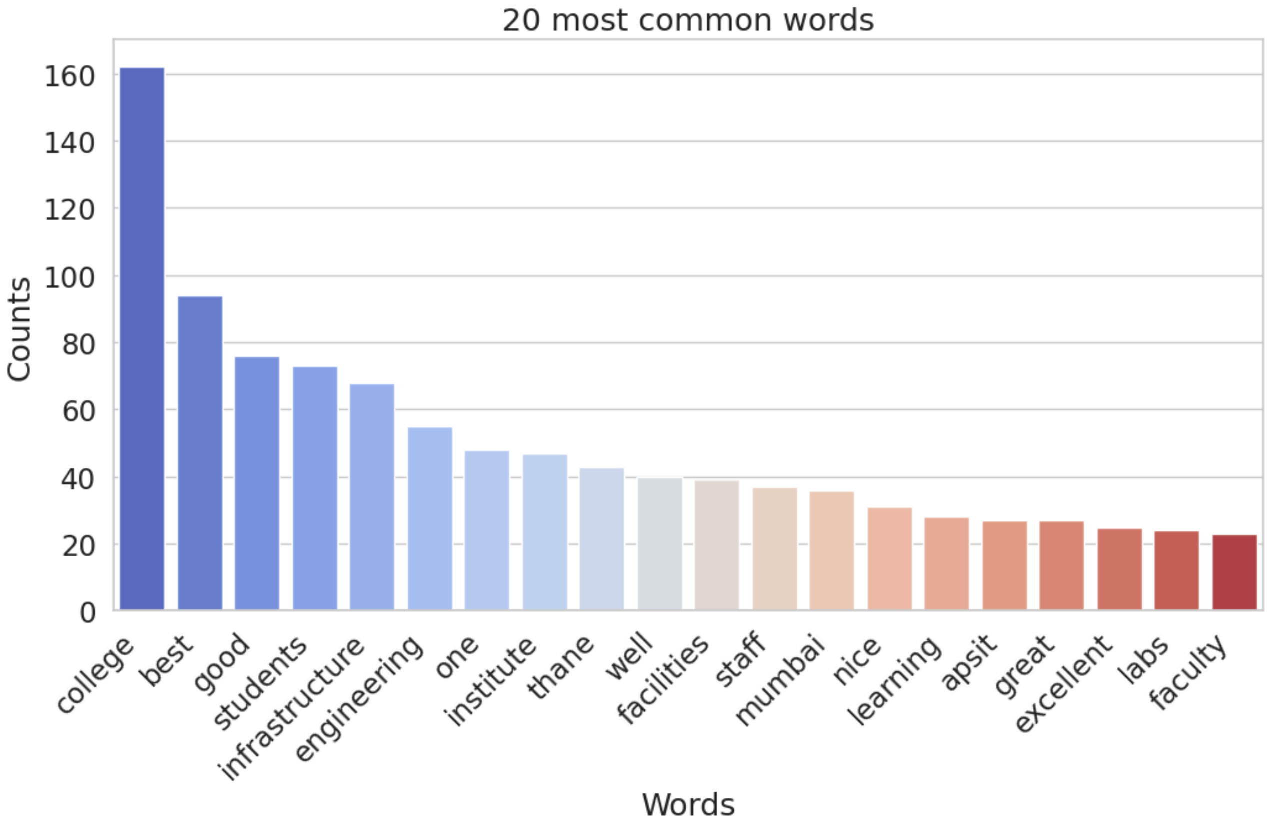
	Link	Username	Rank	Timeline	Review	Response
0	https://lh3.googleusercontent.com/a-/ALV-UjWHb...	Saurabh Kanade	Local Guide · 37 reviews · 20 photos	6 months ago	Amezing ClassRooms.\nAir Conditioner Fan Both...	Like
1	https://lh3.googleusercontent.com/a/ACg8ocKaRp...	Rasika Pujare	Local Guide · 18 reviews · 170 photos	2 years ago	I visited the institute as it was my examinati...	NaN
2	https://lh3.googleusercontent.com/a-/ALV-UjV4l...	V S	Local Guide · 22 reviews · 7 photos	2 months ago	This was my CAT exam center. Though exam didn'...	Share
3	https://lh3.googleusercontent.com/a-/ALV-UjUV4...	Sneha Yadav	Local Guide · 7 reviews · 61 photos	4 years ago	It's the best college in Thane, growing rapidl...	NaN
4	https://lh3.googleusercontent.com/a-/ALV-UjWIl...	AJP Travel	Local Guide · 112 reviews · 3,469 photos	2 years ago	It is beside highway so keep check on take in ...	NaN

In [ ]:
```
from sklearn.feature_extraction. text import TfidfVectorizer, CountVectorizer
```

In [ ]:
```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer

sns.set_style('whitegrid')
%matplotlib inline

stop = stop + ['Institute', 'APSIT', 'AP', 'Shah', 'Technology']

def plot_20_most_common_words(count_data, count_vectorizer):
    words = count_vectorizer.get_feature_names_out()
    total_counts = np.zeros(len(words))

    for t in count_data:
        total_counts += t.toarray()[0]

    count_dict = dict(zip(words, total_counts))
    count_dict = sorted(count_dict.items(), key=lambda x: x[1], reverse=True)[:20]

    words = [w[0] for w in count_dict]
    counts = [w[1] for w in count_dict]

    x_pos = np.arange(len(words))

    plt.figure(figsize=(12, 6))
    sns.set_context('notebook', font_scale=1.5)
    sns.barplot(x=x_pos, y=counts, palette='coolwarm')
    plt.title('20 most common words')
    plt.xticks(x_pos, words, rotation=45, ha='right')
    plt.xlabel('Words')
    plt.ylabel('Counts')
    plt.show()
```

```
count_vectorizer = CountVectorizer(stop_words=stop)
count_data = count_vectorizer.fit_transform(tweets_df1['Review'])
plot_20_most_common_words(count_data, count_vectorizer)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/feature_extraction/text.py:409: UserWarning:

Your stop_words may be inconsistent with your preprocessing. Tokenizing the stop words generated tokens ['ap', 'apsit', 'institute',
'shah', 'technology'] not in stop_words.

<ipython-input-99-3aee185b1fee>:28: FutureWarning:



Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.
```

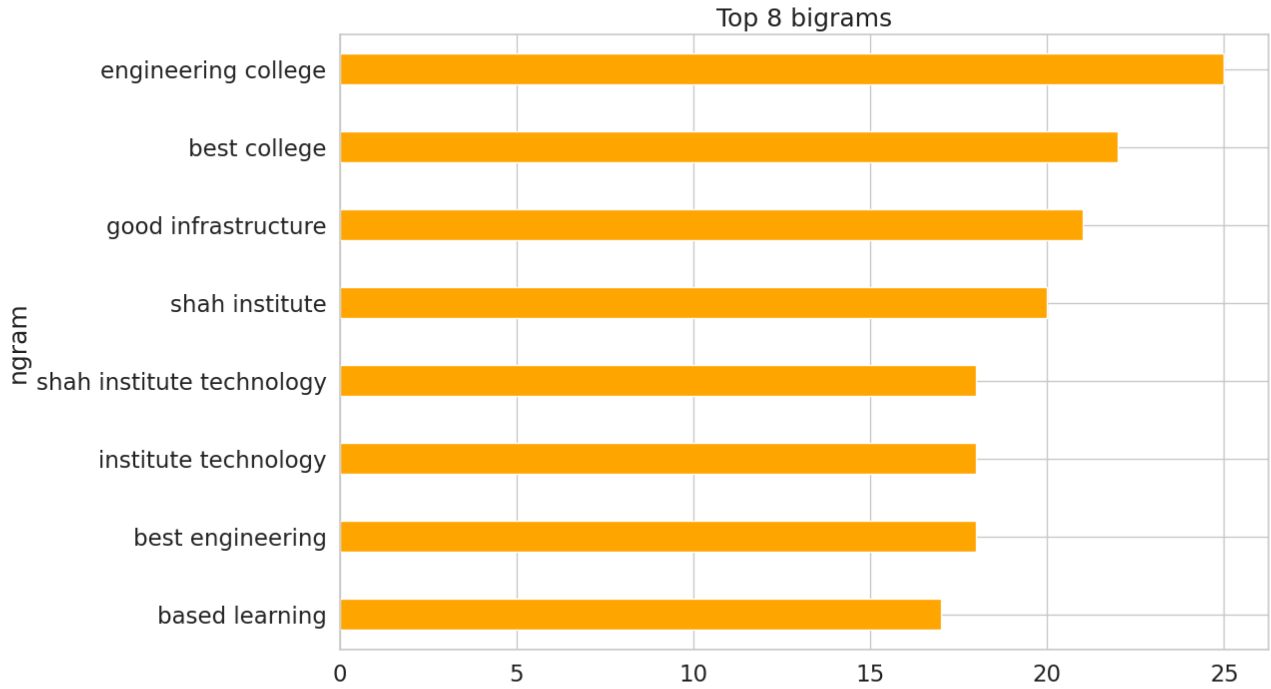


In [ ]:

```
import cufflinks as cf
cf.go_offline()
cf.set_config_file(offline=False, world_readable=True)

def get_top_n_bigram(corpus, n=None) :
  vec = CountVectorizer(ngram_range=(2, 4), stop_words="english").fit(corpus)
  bag_of_words = vec.transform(corpus)
  sum_words = bag_of_words.sum(axis=0)
  words_freq =[(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items()]
  words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
  return words_freq[:n]

common_words = get_top_n_bigram(tweets_df1['Review'] , 8)
mydict={}
for word, freq in common_words:
  bigram_df = pd.DataFrame(common_words,columns = ['ngram', 'count'])

bigram_df.groupby( 'ngram' ).sum()['count'].sort_values(ascending=False).sort_values().plot.barh(title = 'Top
```

Out [100]:

```
<Axes: title={'center': 'Top 8 bigrams'}, ylabel='ngram'>
```



```
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
In [ ]:
```

In [ ]:

In [ ]:

In [ ]: