Lab Assignment 08

The objective of this lab assignment is to process and analyze a dataset of Yelp reviews classified by sentiment (positive or negative) (`data_lab_08.csv`).

Instructions:

Complete each task and question by filling in the blanks (`...`) with one or more lines of code or text. Each task is worth **0.5 points** and each question is worth **1 point** (out of **10 points**).

Submission:

This assignment is due **Friday, December 13, at 11:59PM (Central Time)**.

This assignment must be submitted on Gradescope as a **PDF file** containing the completed code for each task and the corresponding output. **No late submissions will be accepted for this assignment**.

This assignment is individual. Offering or receiving any kind of unauthorized or unacknowledged assistance is a violation of the University's academic integrity policies, will result in a grade of zero for the assignment, and will be subject to disciplinary action.

Installation instructions:

Install the `wordcloud` package by running the following command in the Anaconda prompt: `conda install -c conda-forge wordcloud`

In [53]:
```
# Install stopwords package
import nltk
nltk.download('stopwords')
```

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

Out[53]:
```
True
```

In [54]:
```
# Load libraries
import pandas as pd
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from wordcloud import WordCloud
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
import re
```

In [55]:
```
# Load the dataset and display the first five rows
data = pd.read_csv('data_lab_08.csv')
data.head()
```

Out[55]:

	text	class
0	Wow... Loved this place.	1
1	Crust is not good.	0
2	Not tasty and the texture was just nasty.	0
3	Stopped by during the late May bank holiday of...	1
4	The selection on the menu was great and so wer...	1

Task 01 (of 10): Convert the text to lowercase.

In [56]:
```
data['text'] = [i.lower() for i in data['text']]
data.head()
```

Out[56]:

	text	class
0	wow... loved this place.	1
1	crust is not good.	0
2	not tasty and the texture was just nasty.	0
3	stopped by during the late may bank holiday of...	1
4	the selection on the menu was great and so wer...	1

Task 02 (of 10): Remove punctuation and special characters (that is, every character that is not a letter, a digit, or a whitespace) from the text.
Hint: Use regular expressions and the `str.replace()` function.

In [57]:
```
data['text'] = [re.sub('[^A-Za-z0-9 ]', '', i) for i in data['text']]
data.head()
```

Out[57]:

	text	class
0	wow loved this place	1
1	crust is not good	0
2	not tasty and the texture was just nasty	0
3	stopped by during the late may bank holiday of...	1
4	the selection on the menu was great and so wer...	1

In [58]:
```
# Print list of stop words
stop_list = stopwords.words('english')
print(stop_list)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll",
"you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's",
'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themsel
ves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'ar
e', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doin
g', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by',
'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'abov
e', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'furthe
r', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'f
ew', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'tha
n', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now',
'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "did
n't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'm
a', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shou
ldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

In [59]:
```
# Remove stop words from text
data['text'] = data['text'].apply(lambda x: " ".join([word for word in x.split() if word not in stop_
list]))
data.head()
```

Out[59]:

	text	class
0	wow loved place	1
1	crust good	0
2	tasty texture nasty	0
3	stopped late may bank holiday rick steve recom...	1
4	selection menu great prices	1

Question 01 (of 05): What are the disadvantages of removing these stop words from the text?

Answer: We lose text relevance by removing stop words. For example, "we are happy" will lose "we" and "are" so that that sentence would only be "happy," which does not give any relevance to what was said. Who is happy?

Task 03 (of 10): Stem the text. *Hint:* Use the `stem()` function. For a similar lambda function, see the code for removing stop words above.

In [60]:
```
st = PorterStemmer()

data['text'] = data['text'].apply(lambda x: " ".join(
    [st.stem(word) for word in x.split()])
                                  )
data.head()
```

Out[60]:

	text	class
0	wow love place	1
1	crust good	0
2	tasti textur nasti	0
3	stop late may bank holiday rick steve recommen...	1
4	select menu great price	1

Task 04 (of 10): Split the text into tokens. *Hint:* Use the `split()` function.

In [66]:
```
data['tokens'] = data['text'].apply(lambda x: list(word for word in x.split()))
data.head()
```

Out[66]:

	text	class	tokens
0	wow love place	1	[wow, love, place]
1	crust good	0	[crust, good]
2	tasti textur nasti	0	[tasti, textur, nasti]
3	stop late may bank holiday rick steve recommen...	1	[stop, late, may, bank, holiday, rick, steve, ...
4	select menu great price	1	[select, menu, great, price]

In [71]:
```
# Create word cloud using text from positive reviews
list_words_positive = []
for index, row in data.iterrows():
    if row['class'] == 1:
        list_words_positive.extend(row['tokens'])
all_words = ' '.join(list_words_positive)
wordcloud = WordCloud(width = 800, height = 500).generate(all_words)
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

Task 05 (of 10): Create a word cloud using the text from negative reviews.

In [72]:
```
list_words_negative = []
for index, row in data.iterrows():
    if row['class'] == 0:
        list_words_negative.extend(row['tokens'])
all_words = ' '.join(list_words_negative)
wordcloud = WordCloud(width = 800, height = 500).generate(all_words)
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

Question 02 (of 05): What can you conclude from the word clouds?

Answer: words such as good, love, and great are all positive and attune themselves to the positive word cloud, which makes sense. For the negative word cloud, we can see that things more relevant to things you can expect to make you upset at a restaurant appear, such as the service, the wait, time, order, and bland.

Task 06 (of 10): Convert the text into a matrix of word counts. *Hint:* Use `sklearn`'s `CountVectorizer` object.

In [73]:
```
vectorizer = CountVectorizer()
count_matrix = vectorizer.fit_transform(data['text'])
```

Task 07 (of 10): Print all the words in the word count matrix.

In [122]:
```
print(len(vectorizer.vocabulary_))
print(vectorizer.vocabulary_)
# print(vectorizer.get_feature_names())
```

1628
{'wow': 1607, 'love': 840, 'place': 1074, 'crust': 350, 'good': 621, 'tasti': 1410, 'textur': 1422, 'nasti': 944, 'stop': 1359, 'late': 802, 'may': 873, 'bank': 114, 'holiday': 700, 'rick': 1193, 'steve': 1352, 'recommend': 1161, 'select': 1250, 'menu': 894, 'great': 634, 'price': 1109, 'get': 605, 'angri': 54, 'want': 1549, 'damn': 359, 'pho': 1062, 'honeslti': 703, 'didnt': 395, 'tast': 1408, 'fresh': 580, 'potato': 1100, 'like': 824, 'rubber': 1208, 'could': 322, 'tell': 1415, 'made': 852, 'ahead': 35, 'time': 1445, 'kept': 786, 'warmer': 1551, 'fri': 581, 'touch': 1465, 'servic': 1259, 'prompt': 1121, 'would': 1603, 'go': 615, 'back': 105, 'cashier': 238, 'care': 230, 'ever': 487, 'say': 1235, 'still': 1354, 'end': 471, 'wayyy': 1560, 'overpr': 1012, 'tri': 1474, 'cape': 226, 'cod': 285, 'ravoli': 1151, 'chickenwith': 262, 'cranberrymmmm': 335, 'disgust': 411, 'pretti': 1107, 'sure': 1394, 'human': 720, 'hair': 655, 'shock': 1269, 'sign': 1281, 'indic': 744, 'cash': 236, 'highli': 692, 'waitress': 1546, 'littl': 829, 'slow': 1295, 'worth': 1602, 'let': 816, 'alon': 43, 'vega': 1520, 'burritto': 206, 'blah': 156, 'food': 564, 'amaz': 47, 'also': 44, 'cute': 357, 'less': 815, 'interior': 759, 'beauti': 130, 'perform': 1054, 'that': 1425, 'rightth': 1196, 'red': 1162, 'velvet': 1526, 'cakeohhh': 217, 'stuff': 1372, 'hole': 699, 'wall': 1548, 'mexican': 898, 'street': 1364, 'taco': 1400, 'friendli': 584, 'staff': 1340, 'took': 1458, 'hour': 715, 'tabl': 1399, 'restaur': 1185, 'luke': 847, 'warm': 1550, 'sever': 1262, 'run': 1210, 'around': 76, 'total': 1464, 'overwhelm': 1013, 'worst': 1600, 'salmon': 1220, 'sashimi': 1228, 'combo': 292, 'burger': 204, 'beer': 134, '23': 10, 'decent': 366, 'deal': 365, 'final': 539, 'blow': 165, 'found': 573, 'accid': 23, 'happier': 667, 'seem': 1248, 'quick': 1138, 'grab': 628, 'bite': 154, 'familiar': 517, 'pub': 1126, 'favor': 526, 'look': 835, 'elsewher': 467, 'overal': 1008, 'lot': 838, 'redeem': 1163, 'qualiti': 1136, 'inexpens': 747, 'ampl': 52, 'portion': 1096, 'poor': 1092, 'waiter': 1545, 'feel': 529, 'stupid': 1373, 'everi': 488, 'came': 221, 'first': 546, 'visit': 1537, 'hiro': 697, 'delight': 381, 'suck': 1380, 'shrimp': 1278, 'tender': 1418, 'moist': 916, 'enough': 474, 'drag': 431, 'establish': 481, 'hard': 668, 'judg': 782, 'whether': 1576, 'side': 1280, 'gross': 644, 'melt': 891, 'styrofoam': 1375, 'eat': 451, 'fear': 528, 'sick': 1279, 'posit': 1097, 'note': 968, 'server': 1258, 'attent': 88, 'provid': 1125, 'frozen': 586, 'puck': 1129, 'peopl': 1049, 'behind': 136, 'regist': 1170, 'thing': 1431, 'prime': 1111, 'rib': 1190, 'dessert': 392, 'section': 1246, 'bad': 107, 'gener': 603, 'beef': 132, 'cook': 314, 'right': 1195, 'sandwich': 1226, 'firehous': 545, 'greek': 637, 'salad': 1219, 'dress': 436, 'pita': 1072, 'hummu': 722, 'refresh': 1167, 'order': 996, 'duck': 444, 'rare': 1145, 'pink': 1070, 'insid': 753, 'nice': 956, 'char': 248, 'outsid': 1004, 'us': 1511, 'realiz': 1155, 'husband': 726, 'left': 810, 'sunglass': 1391, 'chow': 270, 'mein': 889, 'horribl': 709, 'attitud': 89, 'toward': 1467, 'custom': 355, 'talk': 1404, 'one': 988, 'dont': 420, 'enjoy': 473, 'huge': 719, 'itfriendli': 766, 'wonder': 1590, 'imagin': 735, 'heart': 680, 'attack': 87, 'grill': 640, 'downtown': 430, 'absolut': 21, 'flatlin': 551, 'excus': 496, 'much': 930, 'seafood': 1241, 'string': 1367, 'pasta': 1036, 'bottom': 176, 'amount': 51, 'sauc': 1232, 'power': 1103, 'scallop': 1236, 'perfectli': 1053, 'rip': 1199, 'banana': 113, 'petrifi': 1058, 'tasteless': 1409, 'least': 807, 'think': 1432, 'refil': 1165, 'water': 1557, 'struggl': 1370, 'wave': 1558, '10': 0, 'minut': 909, 'receiv': 1159, 'star': 1343, 'appet': 68, 'cocktail': 283, 'handmad': 662, 'delici': 379, 'wed': 1563, 'definit': 375, 'glad': 610, 'give': 608, 'militari': 904, 'discount': 409, 'alway': 46, 'do': 416, 'gringo': 641, 'updatew': 1507, 'second': 1245, 'got': 625, 'appar': 66, 'never': 953, 'heard': 679, 'salt': 1222, 'batter': 124, 'fish': 547, 'chewi': 260, 'way': 1559, 'finish': 543, 'includ': 740, 'drink': 439, 'jeff': 772, 'went': 1571, 'beyond': 143, 'expect': 498, 'meh': 888, '30': 11, 'min': 907, 'milkshak': 906, 'not

h': 969, 'chocol': 267, 'milk': 905, 'guess': 649, 'known': 794, 'excalibur': 492, 'use': 1512, 'com
mon': 295, 'sens': 1253, 'dish': 412, 'quit': 1140, 'appal': 65, 'valu': 1517, 'well': 1570, 'swee
t': 1397, 'season': 1243, 'today': 1450, 'ive': 768, 'lunch': 849, 'buffet': 199, 'cheat': 253, 'was
t': 1555, 'opportun': 993, 'rice': 1192, 'compani': 296, 'come': 293, 'experienc': 501, 'underwhel
m': 1494, 'relationship': 1174, 'parti': 1032, 'cant': 225, 'wait': 1544, 'person': 1057, 'ask': 80,
'break': 184, 'walk': 1547, 'smell': 1300, 'old': 983, 'greas': 632, 'trap': 1472, 'other': 998, 'tu
rkey': 1484, 'roast': 1201, 'bland': 158, 'pan': 1024, 'cake': 216, 'everyon': 489, 'rave': 1150, 's
ugari': 1384, 'disast': 407, 'tailor': 1401, 'palat': 1021, 'six': 1291, 'year': 1615, 'spring': 133
8, 'roll': 1203, 'oh': 980, 'yummi': 1626, 'meat': 879, 'ratio': 1149, 'chicken': 261, 'unsatisfi':
1503, 'omelet': 986, 'die': 396, 'everyth': 490, 'summari': 1387, 'larg': 798, 'disappoint': 405, 'd
ine': 399, 'experi': 500, 'realli': 1156, 'sexi': 1264, 'mouth': 925, 'your': 1622, 'outrag': 1002,
'flirt': 556, 'hottest': 714, 'rock': 1202, 'casino': 239, 'step': 1351, 'forward': 572, 'best': 14
1, 'breakfast': 185, 'bye': 211, 'tip': 1447, 'ladi': 797, 'arriv': 78, 'quickli': 1139, 'cafe': 21
5, 'serv': 1257, 'fantast': 520, 'wife': 1581, 'garlic': 599, 'bone': 171, 'marrow': 870, 'ad': 31,
'extra': 505, 'meal': 877, 'anoth': 55, 'help': 687, 'bloddi': 162, 'mari': 867, 'town': 1468, 'cann
ot': 224, 'beat': 128, 'mussel': 936, 'wine': 1583, 'reduct': 1164, 'better': 142, 'tigerlilli': 144
4, 'afternoon': 33, 'bartend': 118, 'ambienc': 50, 'music': 935, 'play': 1081, 'next': 955, 'trip':
1477, 'sooooo': 1316, 'real': 1154, 'sushi': 1396, 'lover': 841, 'honest': 704, 'yama': 1612, '40mi
n': 16, 'pass': 1034, 'wasnt': 1554, 'busi': 207, 'thai': 1423, 'spici': 1332, 'check': 254, 'atmosp
her': 83, 'kind': 791, 'mess': 896, 'steak': 1348, 'although': 45, 'sound': 1320, 'actual': 30, 'bi
t': 152, 'know': 793, 'manag': 862, 'blandest': 159, 'eaten': 452, 'prepar': 1105, 'indian': 743, 'c
uisin': 353, 'boot': 173, 'worri': 1598, 'fine': 541, 'guy': 651, 'son': 1312, 'said': 1217, 'he': 6
76, 'thought': 1438, 'youd': 1620, 'ventur': 1528, 'away': 96, 'hit': 698, 'spot': 1336, 'night': 95
8, 'host': 711, 'lack': 796, 'word': 1593, 'bitch': 153, 'number': 971, 'reason': 1157, 'review': 11
87, 'ill': 733, 'leav': 809, 'phenomen': 1060, 'ambianc': 49, 'wouldnt': 1604, 'return': 1186, 'stri
p': 1368, 'pork': 1095, 'belli': 139, 'im': 734, 'mediocr': 883, 'penn': 1048, 'vodka': 1538, 'exce
l': 494, 'massiv': 872, 'meatloaf': 881, 'crispi': 344, 'wrap': 1608, 'delish': 382, 'tuna': 1483,
'rude': 1209, 'nyc': 974, 'bagel': 108, 'cream': 339, 'chees': 256, 'lox': 844, 'caper': 227, 'eve
n': 485, 'subway': 1378, 'fact': 510, 'meet': 886, 'serious': 1255, 'solid': 1306, 'bar': 115, 'extr
em': 507, 'mani': 865, 'weekend': 1566, 'empti': 470, 'suggest': 1385, 'ate': 82, 'curri': 354, 'bam
boo': 112, 'shoot': 1271, 'blanket': 160, 'moz': 927, 'top': 1459, 'done': 419, 'cover': 330, 'subpa
r': 1377, 'bathroom': 123, 'clean': 278, 'decor': 369, 'chang': 247, 'consid': 308, 'pace': 1018, 't
humb': 1443, 'watch': 1556, 'pay': 1042, 'ignor': 732, 'fianc': 534, 'middl': 901, 'day': 363, 'gree
t': 639, 'seat': 1244, 'mandalay': 863, 'bay': 125, 'forti': 571, 'five': 548, 'vain': 1515, 'crosti
ni': 345, 'stale': 1341, 'highlight': 693, 'nigiri': 959, 'joint': 779, 'differ': 397, 'cut': 356,
'piec': 1066, 'flavor': 552, 'voodoo': 1540, 'id': 730, 'sinc': 1287, 'gluten': 614, 'free': 577, 'a
go': 34, 'unfortun': 1496, 'must': 937, 'bakeri': 109, 'leftov': 811, 'reloc': 1177, 'impress': 738,
'immedi': 736, 'divers': 415, 'avoid': 94, 'cost': 319, 'full': 591, 'handsdown': 663, 'phoenix': 10
63, 'metro': 897, 'area': 72, 'treat': 1473, 'bacon': 106, 'hella': 685, 'salti': 1223, 'spinach': 1
334, 'avocado': 93, 'ingredi': 750, 'sad': 1213, 'liter': 828, 'zero': 1627, 'hand': 660, 'list': 82
7, 'lordi': 836, 'khao': 787, 'soi': 1305, 'miss': 911, 'terrif': 1421, 'thrill': 1440, 'accommod':
24, 'vegetarian': 1524, 'daughter': 362, 'perhap': 1055, 'caught': 241, 'inspir': 754, 'desir': 389,
'modern': 915, 'hip': 696, 'maintain': 858, 'cozi': 332, 'weekli': 1567, 'haunt': 673, 'sat': 1229,

'20': 8, 'take': 1402, 'overcook': 1009, 'charcoal': 249, 'decid': 367, 'send': 1252, 'verg': 1530, 'probabl': 1114, 'dirt': 402, 'someth': 1310, 'healthi': 678, 'quantiti': 1137, 'lemon': 814, 'raspberri': 1146, 'ice': 729, 'incred': 742, 'interest': 758, 'crepe': 342, 'station': 1346, 'hot': 713, 'bread': 183, 'butter': 209, 'home': 701, 'chip': 264, 'topveri': 1461, 'origin': 997, 'egg': 461, 'gyro': 652, 'wing': 1584, 'satisfi': 1231, 'joey': 777, 'vote': 1541, 'dog': 417, 'valley': 1516, 'reader': 1153, 'magazin': 855, 'bowl': 179, 'live': 830, 'friday': 582, 'insult': 757, 'felt': 533, 'disrespect': 414, 'drive': 441, 'exceed': 493, 'hope': 708, 'dream': 434, 'serivc': 1256, 'brunch': 194, 'invit': 760, '1979': 7, 'last': 800, 'foot': 567, 'mix': 913, 'mushroom': 934, 'yukon': 1624, 'gold': 618, 'pure': 1132, 'white': 1577, 'corn': 316, 'beateou': 129, 'bug': 200, 'show': 1276, 'given': 609, 'climb': 279, 'kitchen': 792, 'soon': 1314, 'friend': 583, 'tartar': 1407, 'wont': 1591, 'though': 1437, 'soggi': 1304, 'jamaican': 770, 'mojito': 917, 'small': 1296, 'shower': 1277, 'rins': 1198, 'unless': 1500, 'mind': 908, 'nude': 970, 'see': 1247, 'lobster': 831, 'bisqu': 151, 'bussel': 208, 'sprout': 1339, 'risotto': 1200, 'filet': 536, 'need': 948, 'pepperand': 1051, 'cours': 327, 'none': 964, 'bode': 169, 'someon': 1309, 'either': 463, 'cold': 287, 'date': 361, 'unbeliev': 1490, 'bargain': 117, 'folk': 562, 'otto': 1000, 'make': 859, 'welcom': 1569, 'special': 1327, 'main': 857, 'uninspir': 1498, 'muststop': 938, 'whenev': 1575, 'isnt': 762, 'world': 1597, 'worstannoy': 1601, 'drunk': 443, 'fun': 592, 'chef': 259, 'doubl': 423, 'cheeseburg': 257, 'singl': 1288, 'patti': 1041, 'fall': 515, 'apart': 62, 'pictur': 1065, 'upload': 1509, 'yeah': 1614, 'coupl': 325, 'sport': 1335, 'event': 486, 'tv': 1486, 'possibl': 1098, 'theyd': 1427, 'descript': 387, 'yum': 1625, 'eel': 457, 'yet': 1619, 'mayowel': 876, 'hardest': 669, 'decis': 368, 'honestli': 705, 'ms': 929, 'suppos': 1393, 'eye': 508, 'stay': 1347, 'money': 919, 'flavour': 554, 'almost': 42, 'build': 201, 'freez': 578, 'couldnt': 323, 'close': 280, 'point': 1088, 'ayc': 100, 'light': 821, 'dark': 360, 'set': 1261, 'mood': 922, 'base': 119, 'effort': 460, 'gratitud': 630, 'owner': 1015, 'privileg': 1112, 'working': 1596, 'creami': 340, 'parent': 1030, 'similar': 1283, 'complaint': 299, 'silent': 1282, 'pizza': 1073, 'peanut': 1045, 'fast': 523, 'wouldv': 1605, 'godfath': 617, 'tough': 1466, 'short': 1273, 'stick': 1353, 'recal': 1158, 'charg': 250, 'tap': 1405, 'exquisit': 503, 'plu': 1086, 'buck': 198, 'par': 1028, 'thu': 1442, 'far': 521, 'twice': 1487, 'self': 1251, 'proclaim': 1116, 'coffe': 286, 'wildli': 1582, 'veggitarian': 1525, 'platter': 1080, 'wrong': 1610, 'madison': 854, 'ironman': 761, 'job': 776, 'dedic': 370, 'boba': 168, 'tea': 1412, 'jenni': 773, 'patio': 1039, 'outstand': 1005, 'goat': 616, 'skimp': 1292, 'mac': 850, 'bachi': 104, 'stink': 1355, 'burn': 205, 'saganaki': 1216, '100': 1, 'hate': 672, 'disagre': 404, 'fellow': 532, 'yelper': 1618, 'later': 803, 'neighborhood': 951, 'conveni': 313, 'locat': 832, 'pull': 1130, 'soooo': 1315, 'gave': 600, 'rate': 1147, 'pleas': 1082, 'third': 1434, 'write': 1609, 'stir': 1356, 'noodl': 966, 'count': 324, 'box': 180, '12': 3, 'bore': 174, 'servicecheck': 1260, 'greedi': 636, 'corpor': 317, 'dime': 398, 'atroci': 85, 'summer': 1388, 'charm': 251, 'outdoor': 1001, 'toast': 1449, 'english': 472, 'muffin': 931, 'untoast': 1504, 'high': 691, 'hous': 716, 'bu': 197, 'boy': 181, 'basic': 121, 'figur': 535, 'joke': 780, 'publicli': 1128, 'loudli': 839, 'bbq': 126, 'lighter': 822, 'fare': 522, 'public': 1127, 'two': 1488, 'happi': 666, 'inhous': 751, 'downsid': 429, 'without': 1589, 'doubt': 424, 'except': 495, 'month': 921, 'favorit': 527, 'shawarrrrrrma': 1267, 'black': 155, 'pea': 1043, 'unreal': 1502, 'vinaigrett': 1534, 'seen': 1249, 'especi': 480, '400': 15, 'mom': 918, 'pleasant': 1083, 'honor': 706, 'hut': 727, 'coupon': 326, 'truli': 1480, 'dirti': 403, 'replenish': 1180, 'plain': 1075, 'yucki': 1623, 'standard': 1342, '17': 6, 'omg': 987, 'delicioso': 380, 'authent': 91, 'spaghetti': 1326, 'whatsoev': 1574, 'veget': 1523, 'tucson': 1481, 'vegasther': 1522, 'chipotl': 266, 'classywarm': 277, 'succul': 13

79, 'basebal': 120, 'brick': 188, 'oven': 1007, 'app': 64, 'multipl': 933, 'ten': 1417, 'terribl': 1
420, 'equal': 479, 'shouldnt': 1275, 'pancak': 1025, 'genuin': 604, 'enthusiast': 476, 'sadli': 121
4, 'gordon': 624, 'ramsey': 1141, 'shall': 1265, 'sharpli': 1266, 'life': 820, 'door': 422, 'offer':
978, 'cool': 315, 'turn': 1485, 'els': 466, 'buy': 210, 'handl': 661, 'rowdi': 1207, 'find': 540, 'd
espic': 390, 'soup': 1322, 'lukewarm': 848, 'crave': 336, 'deserv': 388, 'stomach': 1357, 'ach': 27,
'rest': 1183, 'drop': 442, 'ball': 111, 'space': 1325, 'tini': 1446, 'elegantli': 464, 'comfort': 29
4, 'usual': 1513, 'eggplant': 462, 'green': 638, 'bean': 127, 'outta': 1006, 'part': 1031, 'inconsid
er': 741, 'hi': 690, 'dinner': 400, 'outshin': 1003, 'halibut': 657, 'told': 1453, 'happen': 665, 'c
ar': 228, 'front': 585, 'starv': 1345, '90': 20, 'disgrac': 410, 'def': 373, 'ethic': 483, 'contin
u': 312, 'andddd': 53, 'anyon': 58, 'past': 1035, 'stuf': 1371, 'crystal': 352, 'shop': 1272, 'mal
l': 860, 'aria': 75, 'summar': 1386, 'nay': 945, 'transcend': 1471, 'bring': 189, 'joy': 781, 'memor
i': 892, 'pneumat': 1087, 'condiment': 306, 'dispens': 413, 'ian': 728, 'kid': 788, 'option': 995,
'kiddo': 789, 'perfect': 1052, 'famili': 516, 'impecc': 737, 'simpli': 1286, 'bouchon': 177, 'accoun
t': 26, 'screw': 1240, 'remind': 1179, 'pop': 1094, 'san': 1225, 'francisco': 575, 'buldogi': 202,
'gourmet': 627, 'frustrat': 588, 'petti': 1059, 'hungri': 724, 'assur': 81, 'teeth': 1414, 'sore': 1
318, 'complet': 300, 'becom': 131, 'regular': 1171, 'profession': 1117, 'companion': 297, 'meeveryt
h': 887, 'ground': 645, 'smear': 1299, 'beensteppedinandtrackedeverywher': 133, 'pile': 1067, 'bir
d': 149, 'poop': 1091, 'furthermor': 594, 'oper': 991, 'websit': 1562, 'weve': 1573, 'mistak': 912,
'expertconnisseur': 502, 'topic': 1460, 'jerk': 774, 'strike': 1366, 'rush': 1211, 'nicest': 957, 'a
cross': 29, 'biscuit': 150, '40': 14, 'absolutley': 22, 'awkward': 98, '15lb': 5, 'cow': 331, '34t
h': 12, 'gristl': 642, 'fat': 524, 'steiner': 1350, 'dollar': 418, 'anyway': 61, 'fs': 589, 'breakfa
stlunch': 186, 'week': 1565, 'mention': 893, 'combin': 291, 'pear': 1046, 'almond': 41, 'big': 144,
'winner': 1585, 'spicier': 1333, 'prefer': 1104, 'ribey': 1191, 'mesquit': 895, 'anytim': 60, 'goood
d': 623, 'connoisseur': 307, 'certainli': 244, 'contain': 311, 'driest': 438, 'relax': 1175, 'venu':
1529, 'group': 646, 'etc': 482, 'nargil': 943, 'tater': 1411, 'tot': 1463, 'southwest': 1324, 'pai
d': 1020, 'vanilla': 1518, 'smooth': 1302, 'profiterol': 1118, 'choux': 269, 'pastri': 1037, 'az': 1
01, 'new': 954, 'carli': 231, 'due': 446, 'acknowledg': 28, '35': 13, 'foodand': 565, 'forget': 569,
'margarita': 866, 'ventil': 1527, 'upgrad': 1508, 'letdown': 817, 'rather': 1148, 'camelback': 222,
'flower': 558, 'cartel': 234, 'trim': 1476, '70': 18, 'claim': 274, 'bill': 147, 'jewel': 775, 'la':
795, 'exactli': 491, 'nearli': 946, 'limit': 826, 'boil': 170, 'crab': 333, 'leg': 812, 'toro': 146
2, 'cavier': 243, 'extraordinari': 506, 'thinli': 1433, 'slice': 1294, 'wagyu': 1543, 'truffl': 147
9, 'long': 833, 'attach': 86, 'ga': 596, 'awesom': 97, 'wors': 1599, 'humili': 721, 'worker': 1595,
'mebunch': 882, 'name': 941, 'call': 219, 'conclus': 305, 'fill': 537, 'daili': 358, 'tragedi': 147
0, 'struck': 1369, 'crawfish': 337, 'monster': 920, 'funni': 593, 'multigrain': 932, 'pumpkin': 113
1, 'pecan': 1047, 'fluffi': 559, 'airlin': 36, 'noca': 962, 'lettuc': 818, 'thoroughli': 1436, 'home
mad': 702, 'thin': 1430, 'cheesecurd': 258, 'typic': 1489, 'glanc': 611, 'finger': 542, 'item': 765,
'havent': 674, 'gone': 620, 'greasi': 633, 'unhealthi': 1497, 'might': 902, 'similarli': 1284, 'deli
veri': 384, 'man': 861, 'apolog': 63, '45': 17, 'expens': 499, 'pack': 1019, 'togo': 1452, 'tiramis
u': 1448, 'cannoli': 223, 'upway': 1510, 'sun': 1389, 'whole': 1578, 'bunch': 203, 'choos': 268, 'fr
enchman': 579, 'martini': 871, 'opinion': 992, 'entre': 478, 'gc': 601, 'sampl': 1224, 'thirti': 143
5, 'vacant': 1514, 'yellowtail': 1617, 'carpaccio': 232, 'stranger': 1362, 'hello': 686, 'strang': 1
361, 'boyfriend': 182, 'recent': 1160, 'donut': 421, 'save': 1234, 'room': 1204, 'mayb': 874, 'howe
v': 717, 'particular': 1033, 'suffer': 1382, 'tapa': 1406, 'vinegrett': 1535, 'babi': 103, 'palm': 1

023, 'believ': 137, 'hanker': 664, 'forth': 570, 'theft': 1426, 'eew': 458, 'overhaul': 1010, 'wit': 1588, 'guest': 650, 'regularli': 1172, 'super': 1392, 'swung': 1398, 'deepli': 372, 'effici': 459, 'fan': 519, 'sucker': 1381, 'dri': 437, '15': 4, 'cheap': 252, 'oliv': 985, 'perpar': 1056, 'present': 1106, 'giant': 606, 'lightli': 823, 'dust': 448, 'powder': 1102, 'sugar': 1383, 'fo': 560, 'accomod': 25, 'veganveggi': 1521, 'crumbi': 349, 'pale': 1022, 'color': 290, 'instead': 756, 'crouton': 346, 'itll': 767, 'crema': 341, 'caf': 214, 'expand': 497, 'wish': 1587, 'philadelphia': 1061, 'sit': 1289, 'fairli': 513, 'crisp': 343, 'north': 967, 'scottsdal': 1238, 'soooooo': 1317, 'freak': 576, 'paper': 1027, 'reheat': 1173, 'ok': 982, 'wedg': 1564, 'sorri': 1319, 'tongu': 1456, 'cheek': 255, 'bloodi': 163, 'despit': 391, 'yellow': 1616, 'saffron': 1215, 'thru': 1441, 'mean': 878, 'half': 656, 'somehow': 1308, 'luck': 846, 'noncustom': 963, 'focus': 561, 'grandmoth': 629, 'hostess': 712, 'four': 574, 'blue': 167, 'shirt': 1268, 'vibe': 1533, 'drastic': 432, 'highqual': 694, 'caesar': 213, 'promptli': 1122, 'madhous': 853, 'proven': 1124, 'dead': 364, 'greatest': 635, 'macaron': 851, 'insan': 752, 'inform': 749, 'werent': 1572, 'somewhat': 1311, 'edibl': 454, 'promis': 1120, 'fail': 511, 'deliv': 383, 'averag': 92, 'plater': 1079, 'sitdown': 1290, 'togeth': 1451, 'poorli': 1093, 'construct': 310, 'italian': 763, 'scream': 1239, 'legit': 813, 'booksomethat': 172, 'duo': 447, 'violinist': 1536, 'song': 1313, 'request': 1181, 'baklava': 110, 'falafel': 514, 'baba': 102, 'ganoush': 597, 'mgm': 899, 'courteou': 329, 'eclect': 453, 'onion': 989, 'ring': 1197, 'work': 1594, 'nobu': 961, 'googl': 622, 'smashburg': 1298, 'gem': 602, 'plantain': 1076, 'spend': 1330, 'panna': 1026, 'cotta': 321, 'atmosphere1': 84, 'slaw': 1293, 'drench': 435, 'mayo': 875, 'piano': 1064, 'soundtrack': 1321, 'amazingrg': 48, 'fillet': 538, 'relleno': 1176, 'plate': 1078, 'sergeant': 1254, 'pepper': 1050, 'auju': 90, 'hawaiian': 675, 'breez': 187, 'mango': 864, 'magic': 856, 'pineappl': 1069, 'smoothi': 1303, 'theyr': 1428, 'mortifi': 923, 'needless': 949, 'drip': 440, 'mostli': 924, '2007': 9, 'brought': 192, 'hospit': 710, 'industri': 746, 'paradis': 1029, 'refrain': 1166, 'cibo': 272, 'longer': 834, 'famou': 518, 'read': 1152, 'pro': 1113, 'simpl': 1285, 'dough': 426, 'tonight': 1457, 'elk': 465, 'specialand': 1328, 'hook': 707, 'classic': 276, 'quaint': 1134, 'compliment': 301, 'thank': 1424, 'dylan': 449, 'tummi': 1482, 'gratuiti': 631, 'larger': 799, 'fli': 555, 'appl': 70, 'juic': 783, 'han': 659, 'nan': 942, 'bare': 116, 'ryan': 1212, 'edinburgh': 455, 'revisit': 1188, 'chines': 263, 'naan': 939, 'pine': 1068, 'nut': 972, 'airport': 37, 'speedi': 1329, 'calligraphi': 220, 'anyth': 59, 'complain': 298, 'stood': 1358, 'begin': 135, 'awkwardli': 99, 'open': 990, 'extens': 504, 'wide': 1579, 'array': 77, 'inflat': 748, 'smaller': 1297, 'grow': 647, 'rapidli': 1144, 'lil': 825, 'fuzzi': 595, 'fabul': 509, 'wonton': 1592, 'thick': 1429, 'level': 819, 'spice': 1331, 'crowd': 347, 'older': 984, 'mid': 900, 'arepa': 74, 'jalapeno': 769, 'shoe': 1270, 'leather': 808, 'defin': 374, 'block': 161, 'lowkey': 843, 'nonfanc': 965, 'afford': 32, 'sour': 1323, 'sunday': 1390, 'tradit': 1469, 'hunan': 723, 'style': 1374, 'bother': 175, 'flair': 549, 'nutshel': 973, 'restaraunt': 1184, 'market': 869, 'sewer': 1263, 'girlfriend': 607, 'veal': 1519, 'satifi': 1230, 'join': 778, 'club': 281, 'via': 1532, 'email': 468, 'case': 235, 'colder': 288, 'flavorless': 553, 'describ': 386, 'tepid': 1419, 'chain': 246, 'easili': 450, 'nacho': 940, 'crazi': 338, 'juri': 784, 'lawyer': 806, 'court': 328, '785': 19, 'wienerschnitzel': 1580, 'idea': 731, 'brother': 191, 'law': 805, 'hereа': 688, 'tribut': 1475, 'held': 683, 'salsa': 1221, 'youll': 1621, 'pissd': 1071, 'surpris': 1395, 'goldencrispi': 619, 'fell': 530, 'flat': 550, 'bruschetta': 195, 'devin': 394, 'employe': 469, 'lastli': 801, 'mozzarella': 928, 'neglig': 950, 'unwelcom': 1505, 'consist': 309, 'fruit': 587, 'peach': 1044, 'offici': 979, 'blown': 166, 'put': 1133, 'plastic': 1077, 'oppos': 994, 'cram': 334, 'takeout': 1403, 'crpe': 348, 'delic': 378, 'aw': 95, 'fair': 512, 'kabuki': 785, 'overhip': 1011, 'unde

```
rservic': 1492, 'maria': 868, 'articl': 79, 'fuck': 590, 'caballero': 212, 'head': 677, 'oyster': 10
17, 'round': 1206, 'disbelief': 408, 'qualifi': 1135, 'version': 1531, 'low': 842, 'toler': 1454, 'p
olit': 1090, 'wash': 1553, 'otherwis': 999, 'heat': 681, 'coconut': 284, 'fella': 531, 'huevo': 718,
'ranchero': 1143, 'appeal': 67, 'pricey': 1110, 'foodservic': 566, 'tempi': 1416, 'gloveseveryth': 6
13, 'deep': 371, 'oil': 981, 'pleasur': 1084, 'plethora': 1085, 'seal': 1242, 'approv': 71, 'colle
g': 289, 'class': 275, 'start': 1344, 'edit': 456, 'besid': 140, 'costco': 320, 'uniqu': 1499, 'weir
d': 1568, 'hardli': 670, 'groceri': 643, 'store': 1360, 'ownerchef': 1016, 'japanes': 771, 'dude': 4
45, 'arent': 73, 'doughi': 427, 'inch': 739, 'wire': 1586, 'albondiga': 39, 'tomato': 1455, 'meatba
l': 880, 'three': 1439, 'occas': 976, 'medium': 885, 'bloodiest': 164, 'refus': 1169, 'anymor': 57,
'killer': 790, 'chai': 245, 'latt': 804, 'allergi': 40, 'warn': 1552, 'clue': 282, 'mediterranean':
884, 'rotat': 1205, 'concern': 304, 'mellow': 890, 'strawberri': 1363, 'unprofession': 1501, 'loya
l': 845, 'patron': 1040, 'occasion': 977, 'pat': 1038, 'mmmm': 914, 'bellagio': 138, 'anticip': 56,
'weak': 1561, 'bought': 178, 'sal': 1218, 'fav': 525, 'unexperienc': 1495, 'steakhous': 1349, 'prope
rli': 1123, 'understand': 1493, 'concept': 303, 'guacamol': 648, 'postino': 1099, 'poison': 1089, 'b
atch': 122, 'yay': 1613, 'hilari': 695, 'christma': 271, 'eve': 484, 'rememb': 1178, 'biggest': 146,
'entir': 477, 'teamwork': 1413, 'degre': 376, 'ri': 1189, 'calamari': 218, 'fondu': 563, 'lost': 83
7, 'forev': 568, 'scene': 1237, 'itdefinit': 764, 'denni': 385, 'downright': 428, 'waaaaaayyyyyyyyy
i': 1542, 'sangria': 1227, 'glass': 612, 'ridicul': 1194, 'brisket': 190, 'neat': 947, 'trippi': 147
8, 'hurri': 725, 'reserv': 1182, 'stretch': 1365, 'cashew': 237, 'undercook': 1491, 'chipolt': 265,
'ranch': 1142, 'dip': 401, 'saus': 1233, 'douchey': 425, 'indoor': 745, 'garden': 598, 'con': 302,
'spotti': 1337, 'neither': 952, 'ensu': 475, 'bing': 148, 'carb': 229, 'profound': 1119, 'deuchebagg
eri': 393, 'smoke': 1301, 'solidifi': 1307, 'ala': 38, 'cart': 233, 'blame': 157, 'herewhat': 689,
'del': 377, 'hamburg': 658, 'hell': 684, 'gotten': 626, 'yaall': 1611, 'shot': 1274, 'firebal': 544,
'disapppoint': 406, 'correct': 318, 'heimer': 682, 'prettyoff': 1108, 'caus': 242, 'own': 1014, 'vom
it': 1539, 'circumst': 273, 'brownish': 193, 'obvious': 975, 'movi': 926, 'ha': 653, 'flop': 557, 'p
roblem': 1115, '1199': 2, 'bigger': 145, 'sub': 1376, 'unwrap': 1506, 'mile': 903, 'brushfir': 196,
'hasnt': 671, 'mirag': 910, 'refri': 1168, 'crusti': 351, 'caterpillar': 240, 'appetit': 69, 'instan
tli': 755, 'ninja': 960, 'hadnt': 654, 'pour': 1101, 'wound': 1606, 'draw': 433}
```


Question 03 (of 05): How many different words are there in the dataset? List some examples.

Answer: There are 1628 different words in the dataset. Examples of this are yellowtail, vote, wagyu, and unprofession.

Task 08 (of 10): Print the first row and the first column of the word count matrix.

In [127]:
```
print("first row:\n",count_matrix[0,])
print("first column:\n",count_matrix[:,0])
```

```
first row:
   (0, 1074)    1
  (0, 840)      1
  (0, 1607)     1
first column:
   (59, 0)      1
  (209, 0)      1
  (376, 0)      1
  (420, 0)      1
  (430, 0)      1
```

Question 04 (of 05): How many words are there in the first review? How many reviews contain the word '10'?

Answer: There are 3 words in the first review. There are 5 reviews with the word '10.'

Task 09 (of 10): Convert the text into a term frequency-inverse document frequency matrix. *Hint:* Use `sklearn`'s `TfidfVectorizer` object.

In [129]:
```
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(data['text'])
```

Task 10 (of 10): Print the first row and the first column of the term frequency-inverse document frequency matrix.

In [130]:
```
print("first row:\n",tfidf_matrix[0,])
print("first column:\n",tfidf_matrix[:,0])
```

```
first row:
   (0, 1607)    0.7682465003477824
  (0, 840)      0.5160632498654986
  (0, 1074)     0.37878230798394597
first column:
   (59, 0)      0.3428916833970035
  (209, 0)      0.3826230726526059
  (376, 0)      0.3235489794962004
  (420, 0)      0.4205898038666185
  (430, 0)      0.33512598551846573
```

Question 05 (of 05): What is the term frequency-inverse document frequency of the word '10' in the 60th review?

Answer: The value is 0.343 rounded to the 3rd decimal.