In [1]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [2]:
```
folder_path = "/content/drive/MyDrive/NLP-Project/yelp_dataset"
```

- Set the path of dataset

Pre-Processing

In [3]:
```
import os
import json
import pandas as pd
filename = 'yelp_academic_dataset_business.json'
data = []

with open(os.path.join(folder_path, filename), 'r') as f:
    for line in f:
        try:
            json_data = json.loads(line)
            if 'review_count' in json_data and json_data['review_count'] > 50:
                data.append(json_data)
                if len(data) >= 10:
                    break
        except json.JSONDecodeError as e:
            print(f"Error decoding JSON in {filename}: {e}")

business_df = pd.DataFrame(data)
business_df.head()
```

Out[3]:

	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	
0	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	80	1	{'R
1	0bPLkL0QhhPO5kt1_EXmNQ	Zio's Italian Market	2575 E Bay Dr	Largo	FL	33771	27.916116	-82.760461	4.5	100	0	
2	MUTTqe8uqyMdBl186RmNeA	Tuna Bar	205 Race St	Philadelphia	PA	19106	39.953949	-75.143226	4.0	245	1	
3	ROeacJQwBeh05Rqg7F6TCg	BAP	1224 South St	Philadelphia	PA	19147	39.943223	-75.162568	4.5	205	1	
4	9OG5YkX1g2GReZM0AskizA	Romano's Macaroni Grill	5505 S Virginia St	Reno	NV	89502	39.476117	-119.789339	2.5	339	1	{'Re

- Select the business that contains more than 50 reviews.
- If there are more number of reviews it will be effiecient when we summarize all the review of each individual business.
- Print first 5 business to verify the correctness.

In [4]:
```
print(business_df.columns)
```

```
Index(['business_id', 'name', 'address', 'city', 'state', 'postal_code',
       'latitude', 'longitude', 'stars', 'review_count', 'is_open',
       'attributes', 'categories', 'hours'],
      dtype='object')
```

- Above are the features of the business data frame.

In [5]:
```
business_df['business_id']
```

Out[5]:
```
0    MTSW4McQd7CbVtyjqoe9mw
1    0bPLkL0QhhPO5kt1_EXmNQ
2    MUTTqe8uqyMdBl186RmNeA
3    ROeacJQwBeh05Rqg7F6TCg
4    9OG5YkX1g2GReZM0AskizA
5    tMkwHmWFUEXrC9ZduonpTg
6    QdN72BWoyFypdGJhhI5r7g
7    kV_Q1oqis8Qli8dUoGpTyQ
8    aPNXGTDkf-4bjhyMBQxqpQ
9    ljxNT9p0y7YMPx0fcNBGig
Name: business_id, dtype: object
```

- Print the business_id's

In [6]:
```
business_df['name']
```

Out[6]:
```
0                    St Honore Pastries
1                  Zio's Italian Market
2                             Tuna Bar
3                                 BAP
4               Romano's Macaroni Grill
5                    The Green Pheasant
6                           Bar One
7                        Ardmore Pizza
8                           Craft Hall
9    Tony's Restaurant & 3rd Street Cafe
Name: name, dtype: object
```

In [7]:
```
filename = 'yelp_academic_dataset_review.json'
data = []

with open(os.path.join(folder_path, filename), 'r') as f:
    for line in f:
        try:
            json_data = json.loads(line)
            if 'business_id' in json_data and json_data['business_id'] in business_df['business_id'].values:
                data.append(json_data)
        except json.JSONDecodeError as e:
            print(f"Error decoding JSON in {filename}: {e}")

review_df = pd.DataFrame(data)
review_df.head()
```

Out[7]:

	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	4.0	0	0	0	Great bar Happy Hour 4-7 every day. Wine & Dra...	2016-01-30 03:16:46
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	5.0	0	0	0	This place is top notch, with phenomenal servi...	2016-11-10 16:52:33
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	1.0	0	0	0	Please, this place makes a semi-new menu and r...	2013-04-11 02:40:03
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	5.0	0	0	0	Stopped in to check out this new spot around t...	2017-12-16 00:13:06
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	1.0	0	0	0	I live in the neighborhood and used to order a...	2018-04-28 00:46:05

- Fetch all the reviews from the review dataset that are given to the businesses that we have chosen in the above step.
- This basically means we are removing out all the unwanted reviews and storing the required reviwes in a dataframe.
- Print the first 5 reviews to verify the correctness.

In [8]:
```
print(review_df.columns)
```

```
Index(['review_id', 'user_id', 'business_id', 'stars', 'useful', 'funny',
       'cool', 'text', 'date'],
      dtype='object')
```

Feature Extraction

In [9]:
```
columns = ['stars', 'useful', 'funny', 'cool']
clean_review = review_df.drop(columns=columns)
```

- Drop all the columns such as 'stars', 'useful', 'funny', and 'cool' which are not required and are unnecessary

- Drop all the columns such as 'stars', 'useful', 'funny', and 'cool' which are not required and are unnecessary.

In [10]:
```
list(clean_review.columns)
```

Out[10]:
```
['review_id', 'user_id', 'business_id', 'text', 'date']
```

- List all the columns that are left after removing unwanted columns that are important.

Case Conversion

In [11]:
```
clean_review['text'] = clean_review['text'].str.lower()
```

- The above code converts all the characters to lower case, making summarizing convenient.

In [12]:
```
clean_review.head()
```

Out[12]:

	review_id	user_id	business_id	text	date
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour 4-7 every day. wine & dra...	2016-01-30 03:16:46
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch, with phenomenal servi...	2016-11-10 16:52:33
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please, this place makes a semi-new menu and r...	2013-04-11 02:40:03
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05

Sentence Tokenization

In [13]:
```
import nltk
nltk.download('punkt')
from nltk.tokenize import sent_tokenize
tokens = []
for i in clean_review['text']:
  tokens.append(nltk.sent_tokenize(i))

clean_review['sentence'] = tokens
clean_review
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

Out[13]:

	review_id	user_id	business_id	text	date	sentence
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour 4-7 every day. wine & dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch, with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please, this place makes a semi-new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...
...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance, to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	5 stars all around! the sushi here is amazing....	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...

1522 rows × 6 columns

- Import NLTK and punkt used for tokenization.
- We tokenize each text into a sentence using the sent_tokenize() method.
- Then create a new column in the data frame and store all the sentence tokens in it.
- Display the data frame to view the changes.

Sentence Count

In [14]:

```
sent_count = []
for i in clean_review['sentence']:
  count = 0
  for c in i:
    count+=1
  sent_count.append(count)

clean_review['sent_count'] = sent_count
clean_review
```

Out[14]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour 4-7 every day. wine & dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch, with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please, this place makes a semi-new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance, to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	5 stars all around! the sushi here is amazing....	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 7 columns

- In the above code block we create a new column named 'sent_count' which stores the number of sentences'
- In the previous step we created a column 'sentence' which stores all the sentences in an array format, so 'sent_count' is the length of the 'sentence' array.

Eliminate non-alphabetical characters

In [15]:
```
import re
raw_text = []
for i in clean_review['text']:
  raw_text.append(re.sub(r"[^a-zA-Z]", " ", i))

clean_review['text'] = raw_text
clean_review
```

Out[15]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvlNjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 7 columns

- In the above step we remove all the characters that are not alphabets and replace them with a space.
- As you can see in the first row '4-7' is removed in the text column.
- This step is mostly done to remove punctuations, special characters, numbers, etc.
- It is important because of simplicity, consistency, focus on content, reduced dimensionality, and improved model performance.

Character Count

In [16]:

```
char_count = []
for i in clean_review['text']:
  count = 0
  for c in i:
    if c != ' ':
      count+=1
  char_count.append(count)

clean_review['char_count'] = char_count
clean_review
```

Out[16]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvlNjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 8 columns

- The above code block creates a new column named 'char_count' which stores the number of characters in the text for each review.
- It is important because it gives us insights about the text length which can be used in summarization. It ensures consistency in text normalization and also ensures quality control.

Word Tokenization

In [17]:

```
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize
tokens = []
for i in clean_review['text']:
  tokens.append(nltk.word_tokenize(i))

clean_review['tokens'] = tokens
clean_review
```

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

Out[17]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 9 columns

- In the above code we create a new column named 'tokens' which stores the tokens that are extracted from the text using nltk. tokenize library.
- It is important for text analysis. Using this we can perform feature extraction and semantic analysis, enhancing the accuracy of this application.
- Then we printed the reviews to ensure the proper update.

Word count before removing stop words

In [18]:
```
word_count_b = []
for i in clean_review['tokens']:
  count = 0
  for c in i:
    count+=1
  word_count_b.append(count)

clean_review['word_count_b'] = word_count_b
clean_review
```

Out[18]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvlNjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 10 columns

- The above code calculates and stores the word count of tokens that were generated in the previous step.

Removing Stop Words

In [19]:
```
import nltk
from nltk.corpus import stopwords

nltk.download("stopwords")
sr = stopwords.words('english')
for i, tokens in enumerate(clean_review['tokens']):
  for token in tokens:
    if token in stopwords.words('english'):
      clean_review['tokens'][i].remove(token)
clean_review
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

Out[19]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvlNjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOlth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 10 columns

- The above code removes all the stop words like 'in', 'the', etc in English.
- It is important because it removes all the unimportant words which improves the accuracy of the text summarization and sentiment analysis.
- In the displayed table you can see that all the stop words are removed from the 'tokens' column.

Character count after removing stop words

In [20]:

```
char_count = []
for i in clean_review['tokens']:
  count = 0
  for c in i:
    if c != ' ':
      count+=1
  char_count.append(count)

clean_review['char_count_a'] = char_count
clean_review
```

Out[20]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvlNjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOlth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 11 columns

- The above code stores the character count after removing the stop words.

Number of stop words

In [21]:

```
word_count_a = []
stopword_count = []
for i in clean_review['tokens']:
  count = 0
  for c in i:
    count+=1
  word_count_a.append(count)

clean_review['word_count_a'] = word_count_a
#calculating stopwords count
for index, row in clean_review.iterrows():
  stopword_count.append(row['word_count_b'] - row['word_count_a'])

clean_review['stopword_count'] = stopword_count
clean_review
```

Out[21]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvlNjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOlth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 13 columns

- The above code stores the word count after removing the stop words.

Sentence Density

In [22]:

```
sent_density = []

for index, row in clean_review.iterrows():
  sent_density.append(row['sent_count']/(row['word_count_a']+1))

clean_review['sent_density'] = sent_density
clean_review
```

Out[22]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOlth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 14 columns

- The above code calculates the sentence density of every text.
- Create a new column in the dataframe to store the sentence density.
- This helps us understand the relationship between number of sentences and number of words, which in turn helps in text readability and text complexity.

Word Density

In [23]:
```
word_density = []

for index, row in clean_review.iterrows():
  word_density.append(row['word_count_a']/(row['char_count_a']+1))

clean_review['word_density'] = word_density
clean_review
```

Out[23]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 15 columns

- The above code calculates the word density of every text.
- Create a new column in the data frame to store the word density.

Stop word density

In [24]:

```
stopword_density = []

for index, row in clean_review.iterrows():
  stopword_density.append(row['stopword_count']/(row['word_count_b']+1))

clean_review['stopword_density'] = stopword_density
clean_review
```

Out[24]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvlNjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOlth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 16 columns

- The above code cerates a new column and stores the density of stop words.

Spelling correction

In [25]:
```
!pip install pyspellchecker
```

```
Collecting pyspellchecker
  Downloading pyspellchecker-0.8.1-py3-none-any.whl (6.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.8/6.8 MB 16.2 MB/s eta 0:00:00
Installing collected packages: pyspellchecker
Successfully installed pyspellchecker-0.8.1
```

- Install a library to check the spellings.

In [26]:
```
#dont run taking too long
from spellchecker import SpellChecker
spell = SpellChecker()
def correct_spelling(tokens):


    for i, token in enumerate(tokens):
        correction = spell.correction(token)
        if correction:
            tokens[i] = correction

    return tokens

for i, tokens in enumerate(clean_review['tokens']):
```

```
  if i == 3:
    break
  for token in tokens:
    corrected_tokens = correct_spelling(tokens)
    clean_review['tokens'][i] = corrected_tokens

clean_review
```

```
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
<ipython-input-26-1a689424e541>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = corrected_tokens
```

Out[26]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_IprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 16 columns

- The above code corrects the spelling in the text using the SpellChecker module
- This helps to improve the model's accuracy and performance.

Stemming

In [27]:

```
from nltk.stem import PorterStemmer
porter = PorterStemmer()
```

```
stem = []
for i, tokens in enumerate(clean_review['tokens']):
  for token in tokens:
    stem.append(porter.stem(token))
  clean_review['tokens'][i] = stem
  stem = []
clean_review
```

```
Streaming output truncated to the last 5000 lines.
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  clean_review['tokens'][i] = stem
<ipython-input-27-f6fca1472b34>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```

Out[27]:

	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvlNjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOlth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

1522 rows × 16 columns

- The above code uses PorterStemmer on tokens to reduce the words to their base word or the root word.
- This is done to improve the text summarization and improve the correctness of the result.

Visulalization

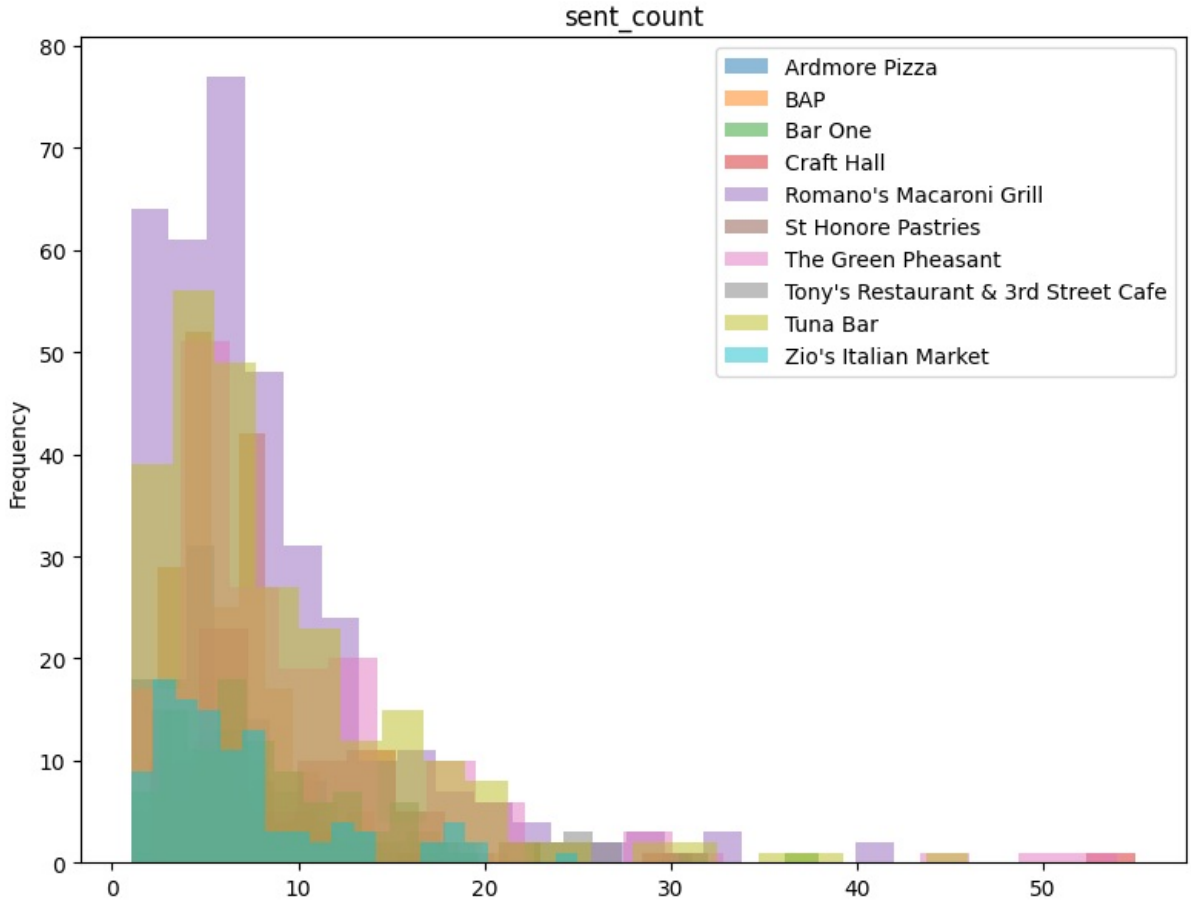
In [28]:
```
business_dict = {}
for i, row in business_df.iterrows():
  business_dict[row['business_id']] = row['name']
business_dict
```

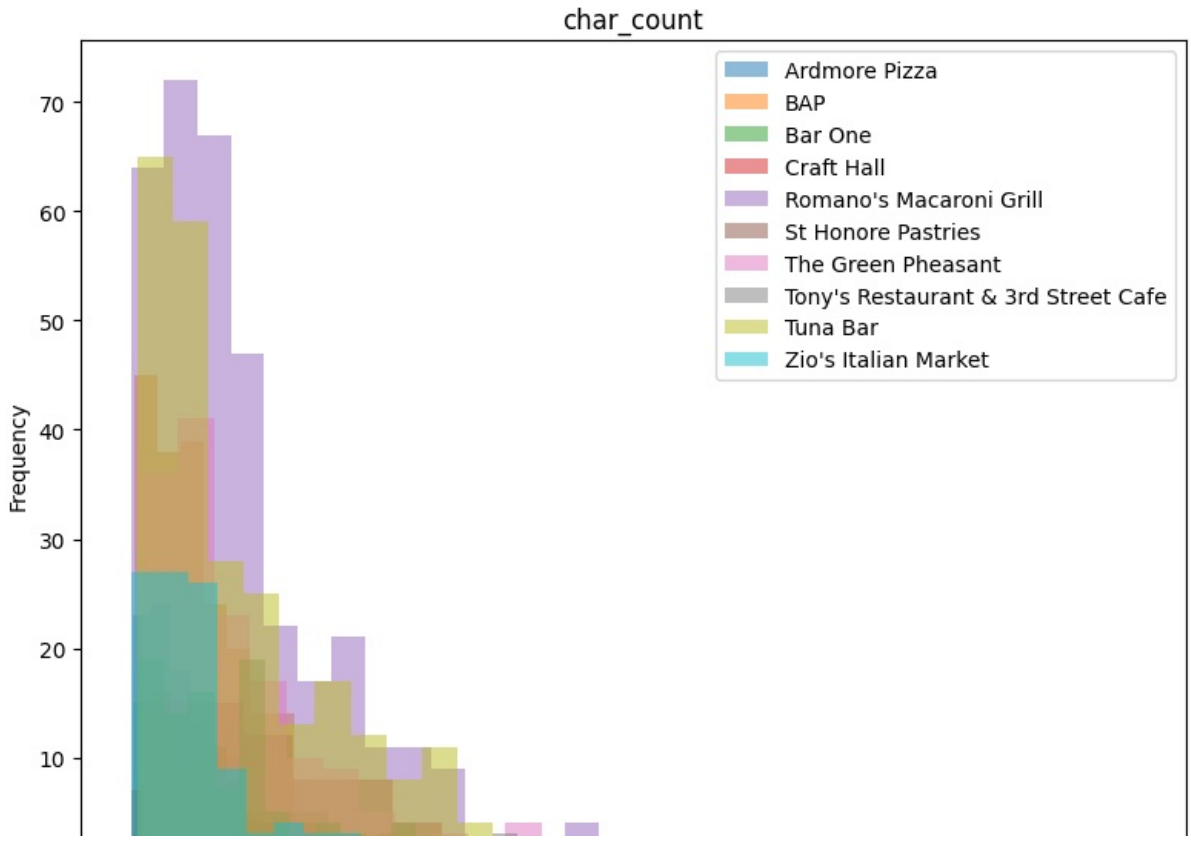
Out[28]:
```
{'MTSW4McQd7CbVtyjqoe9mw': 'St Honore Pastries',
 '0bPLkL0QhhPO5kt1_EXmNQ': "Zio's Italian Market",
 'MUTTqe8uqyMdBl186RmNeA': 'Tuna Bar',
 'ROeacJQwBeh05Rqg7F6TCg': 'BAP',
 '9OG5YkX1g2GReZM0AskizA': "Romano's Macaroni Grill",
 'tMkwHmWFUEXrC9ZduonpTg': 'The Green Pheasant',
 'QdN72BWoyFypdGJhhI5r7g': 'Bar One',
 'kV_Q1oqis8Qli8dUoGpTyQ': 'Ardmore Pizza',
 'aPNXGTDkf-4bjhyMBQxqpQ': 'Craft Hall',
 'ljxNT9p0y7YMPx0fcNBGig': "Tony's Restaurant & 3rd Street Cafe"}
```

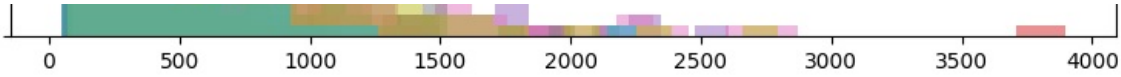
In [29]:
```
import pandas as pd
import matplotlib.pyplot as plt

clean_review['business_name'] = clean_review['business_id'].map(business_dict)
grouped_df = clean_review.groupby('business_name')

columns_to_plot = ['sent_count', 'char_count', 'word_count_b', 'char_count_a', 'word_count_a', 'stopword_count'
```

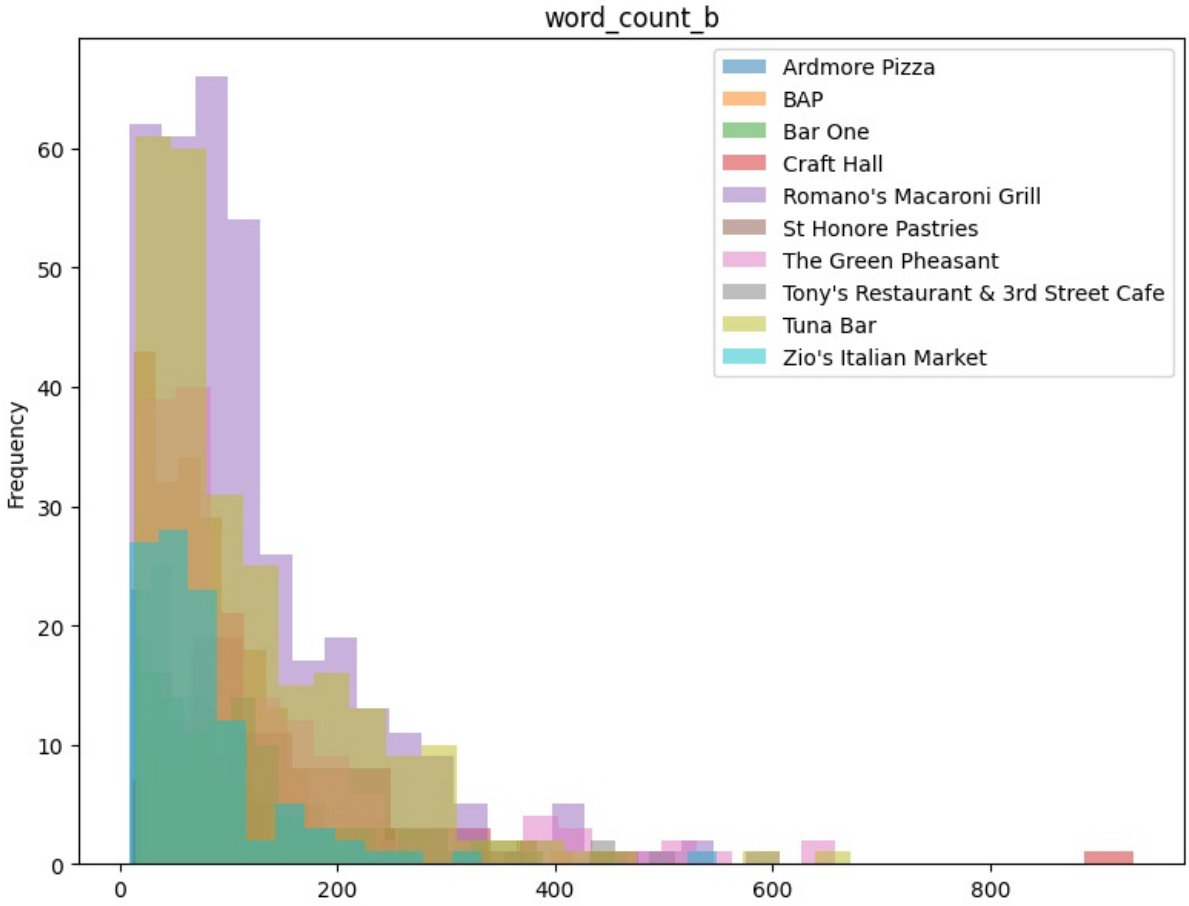
- Group the data frame of different columns.

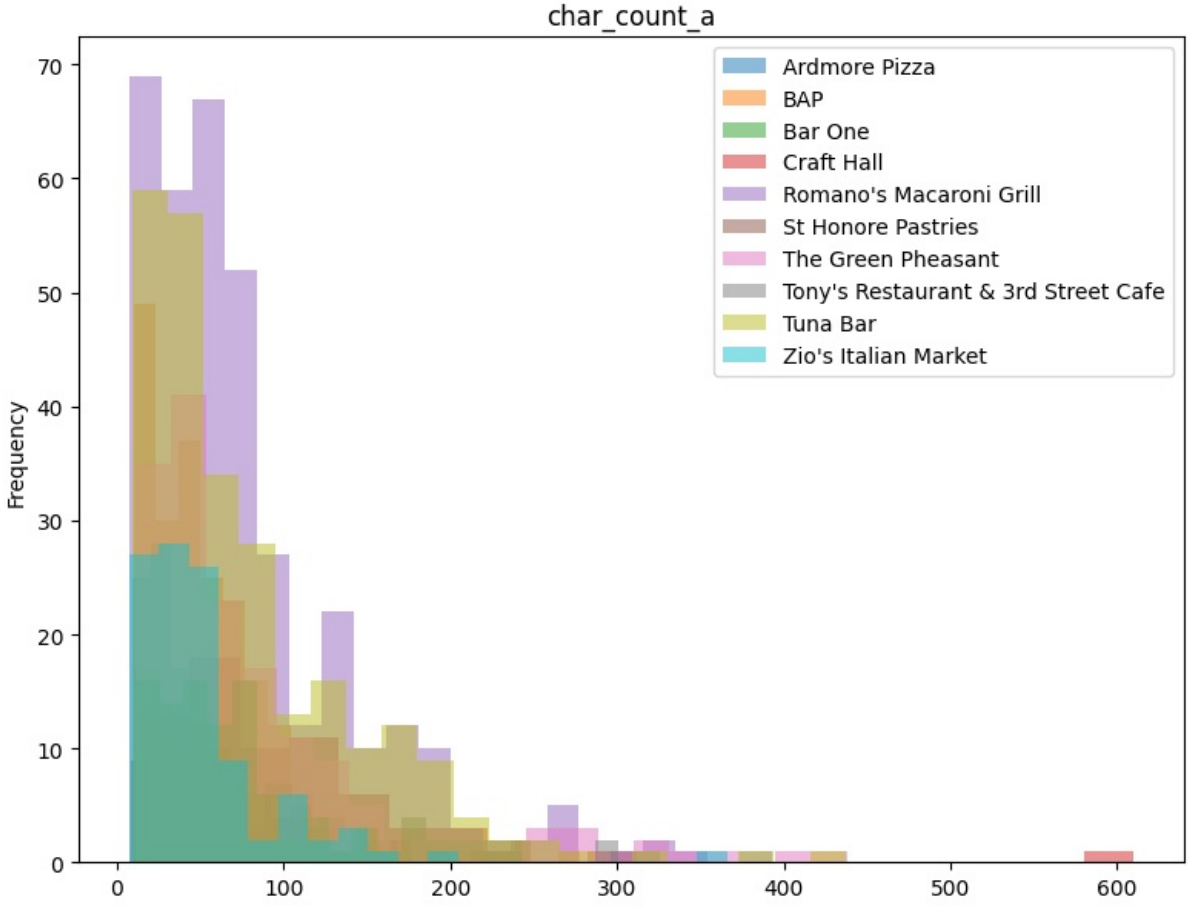
In [30]:
```
columns_to_plot = ['sent_count', 'char_count', 'word_count_b', 'char_count_a', 'word_count_a', 'stopword_count'

fig, axes = plt.subplots(nrows=len(columns_to_plot), figsize=(8, 6 * len(columns_to_plot)))

for i, col in enumerate(columns_to_plot):
    ax = axes[i]
    for label, df in grouped_df:
        df[col].plot(kind="hist", bins=20, alpha=0.5, ax=ax, label=label)

    ax.set_title(col)
    ax.legend()

plt.tight_layout()
plt.show()
```

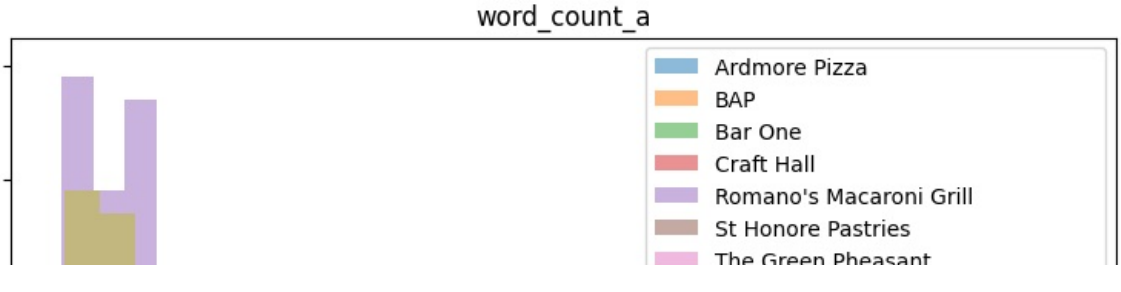


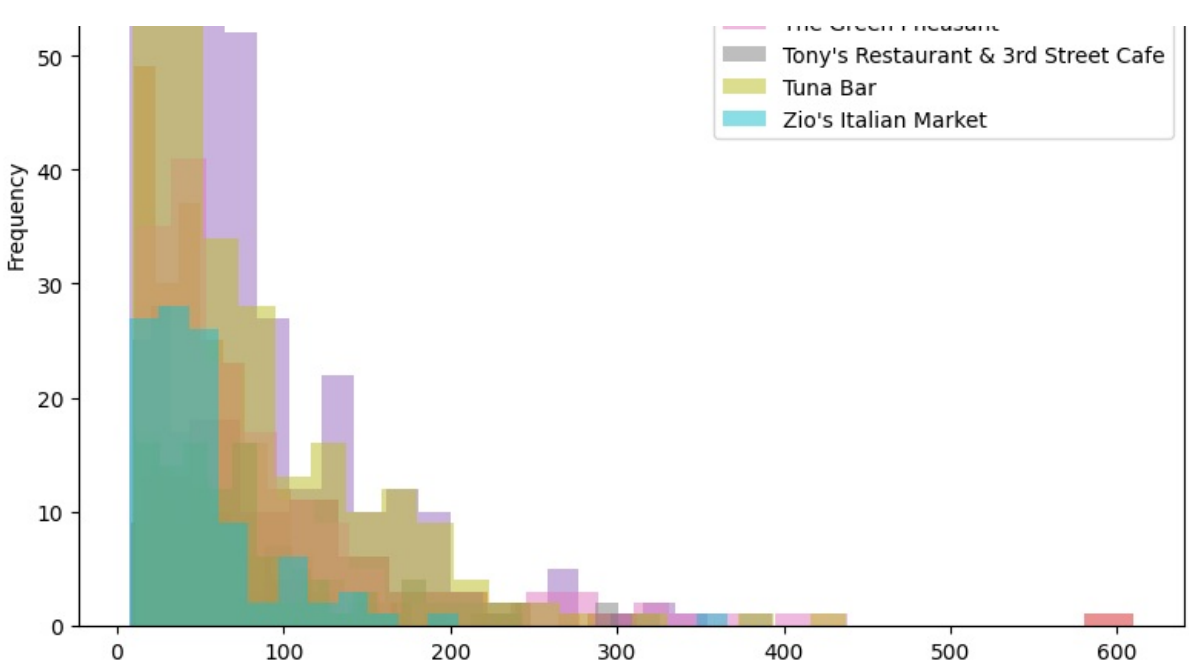


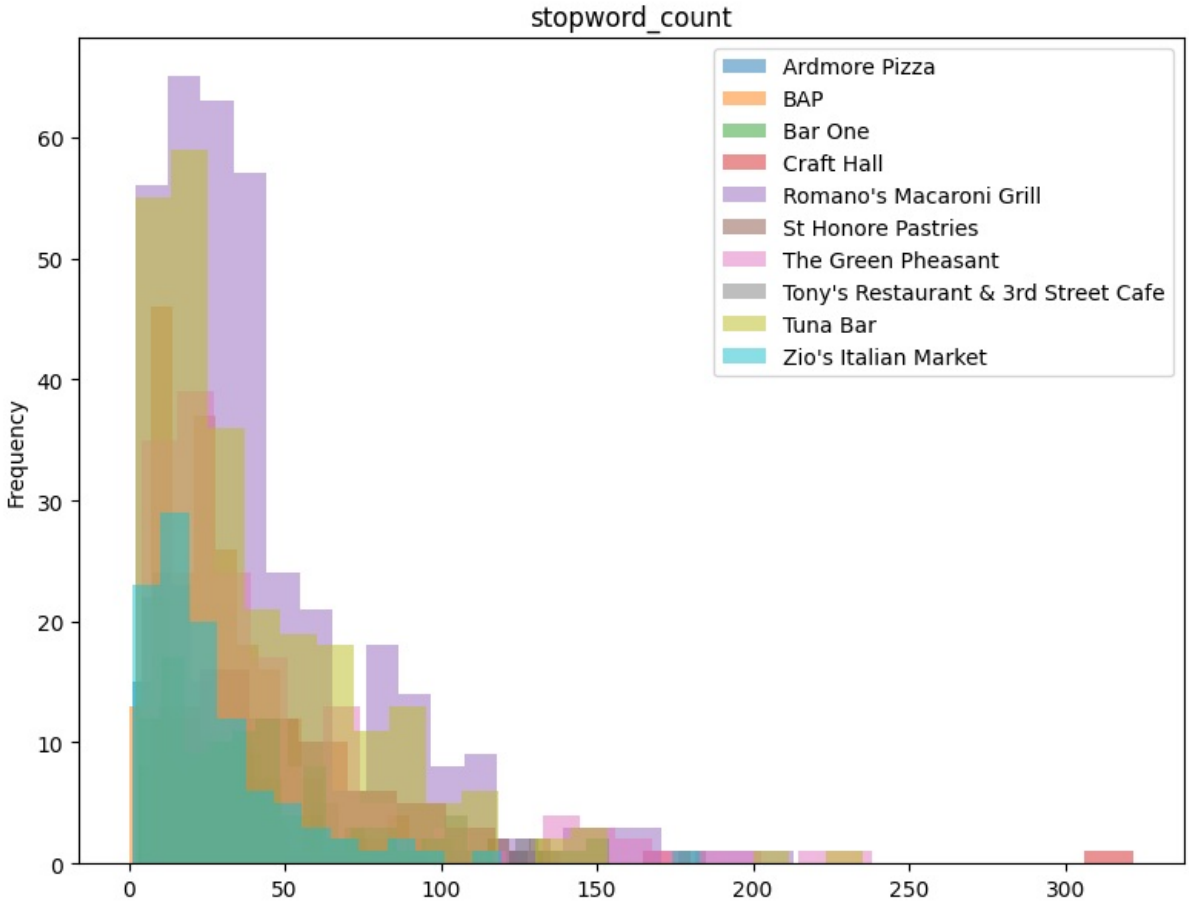

0
0
500
1000
1500
2000
2500
3000
3500
4000

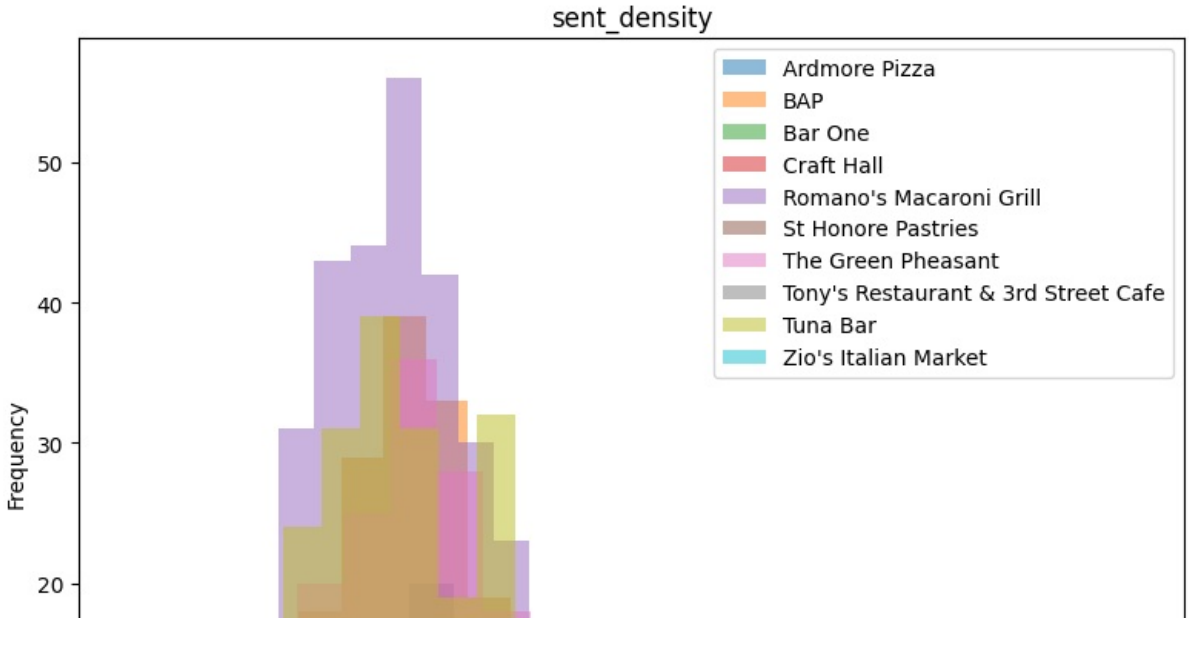

word_count_b
Frequency
0
10
20
30
40
50
60
0
200
400
600
800
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market

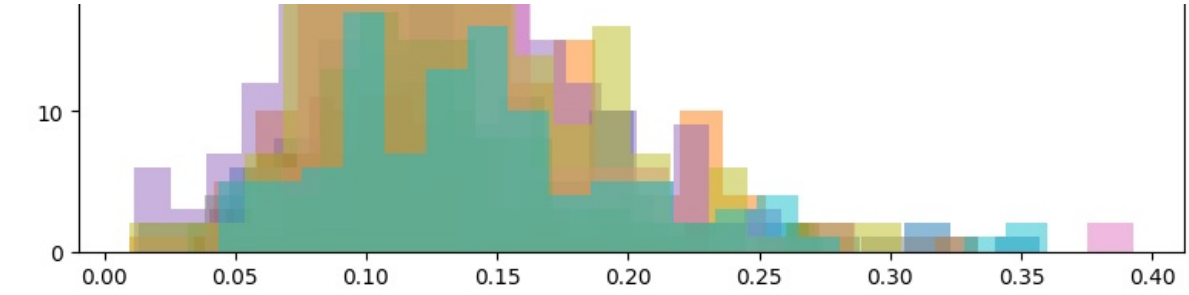

char_count_a
Frequency
0
10
20
30
40
50
60
70
0
100
200
300
400
500
600
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market

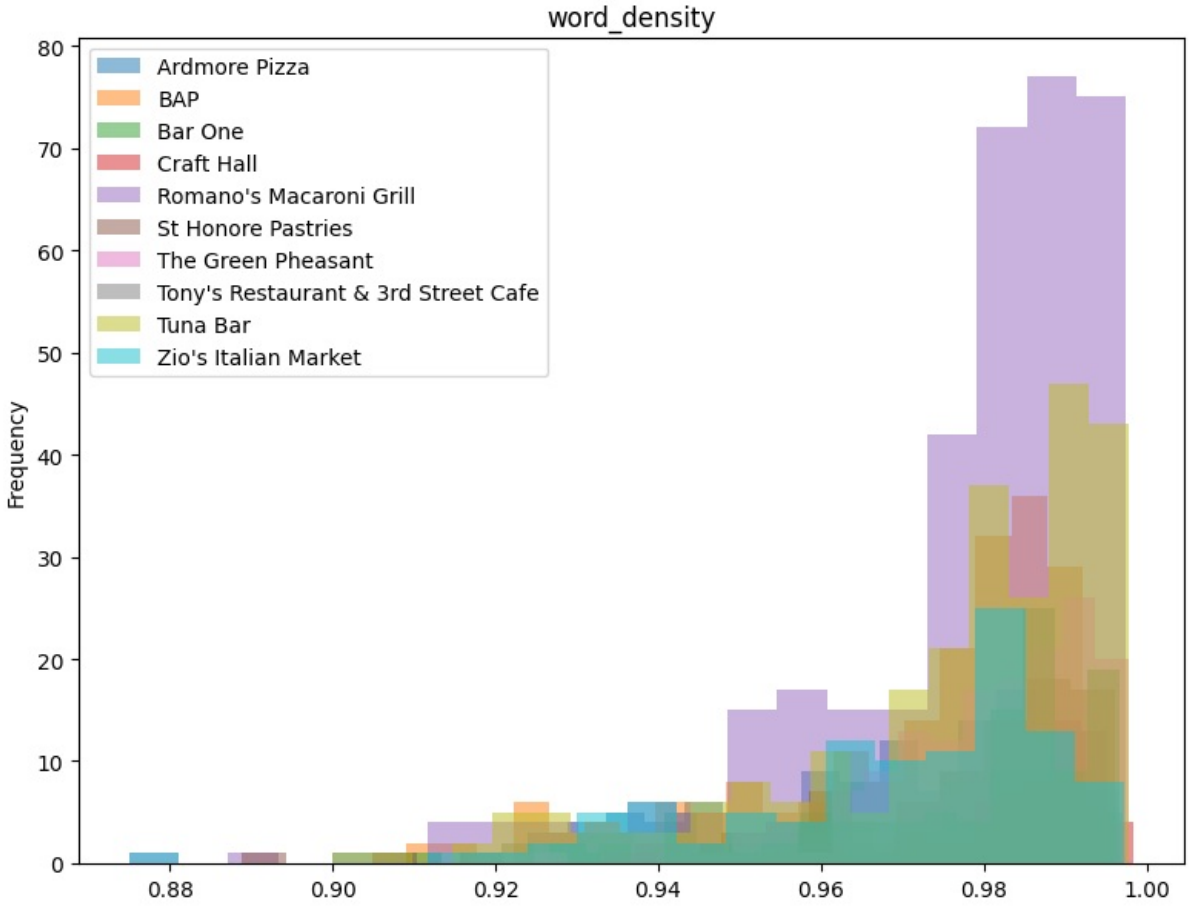

word_count_a
60
70
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant

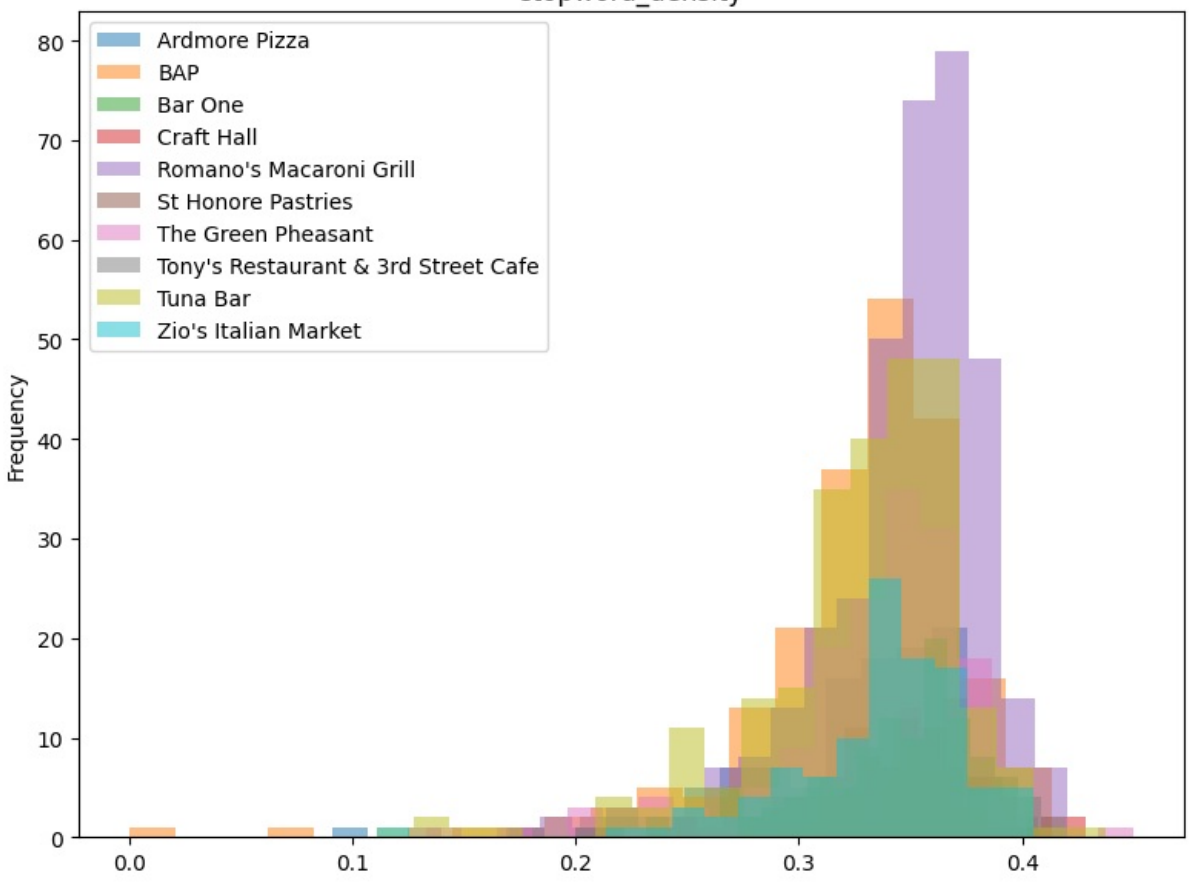

Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market
Frequency
0
10
20
30
40
50
0
100
200
300
400
500
600


stopword_count
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market
Frequency
0
10
20
30
40
50
60
0
50
100
150
200
250
300


sent_density
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market
Frequency
20
30
40
50

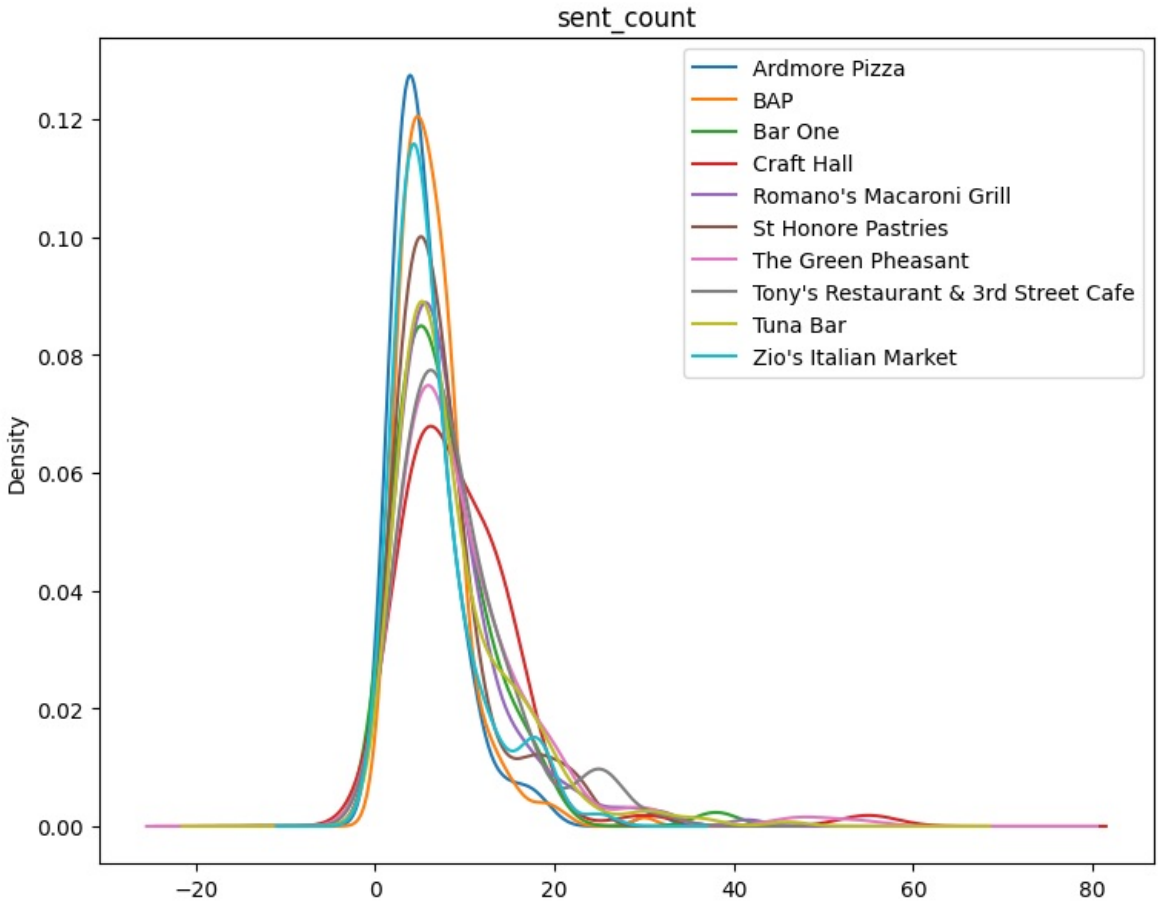

0
10
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

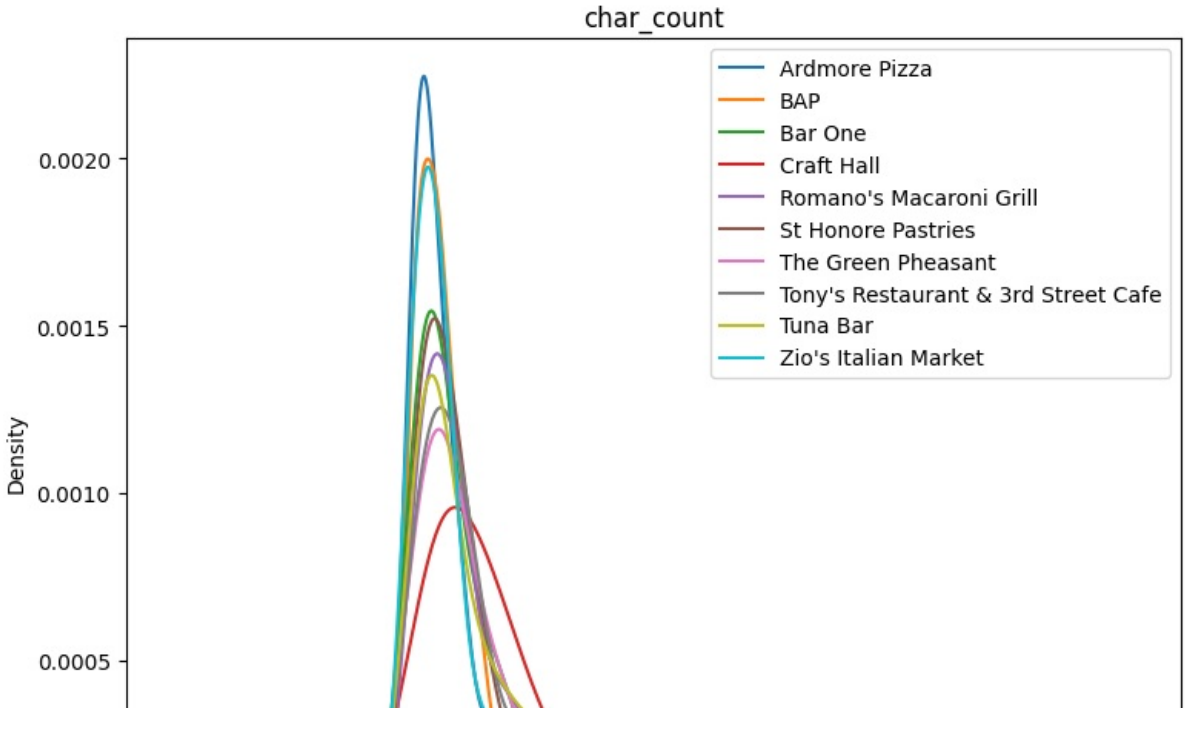

word_density
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market
Frequency
0
10
20
30
40
50
60
70
80
0.88
0.90
0.92
0.94
0.96
0.98
1.00

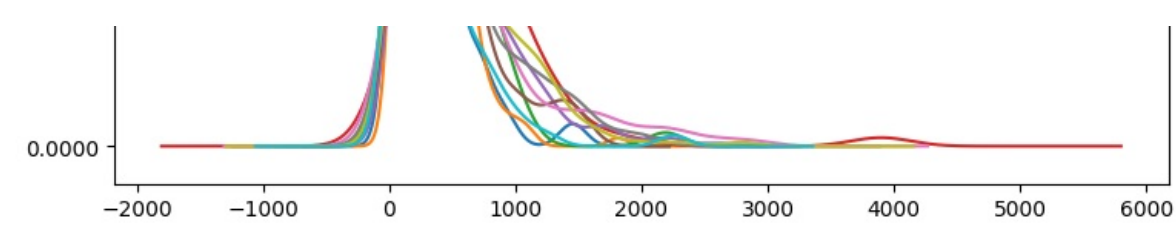

stopword_density
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market
Frequency
0
10
20
30
40
50
60
70
80
0.0
0.1
0.2
0.3
0.4

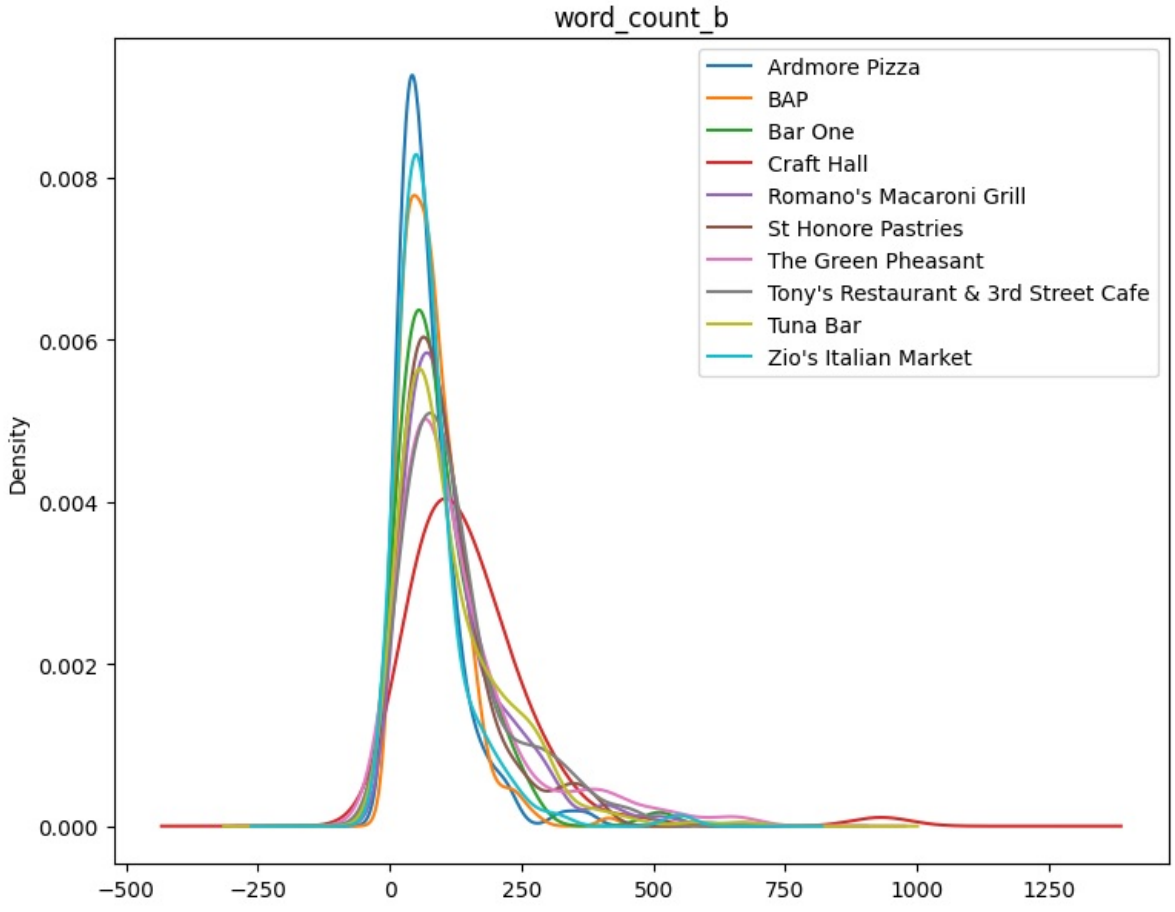
- The above are the histograms whose x-axis is the count and the y-axis is the frequency.
- The above histograms show the frequencies of sentence count, character count, word count after eliminating stop words, character count after eliminating stop words, stop word count, sentence density, word density, density of stop words.
- From the visulaization we can identify the patterns of the reviews and distribution of text related to different metric accross differnt businesses. It gives a comparition accross different businesses.

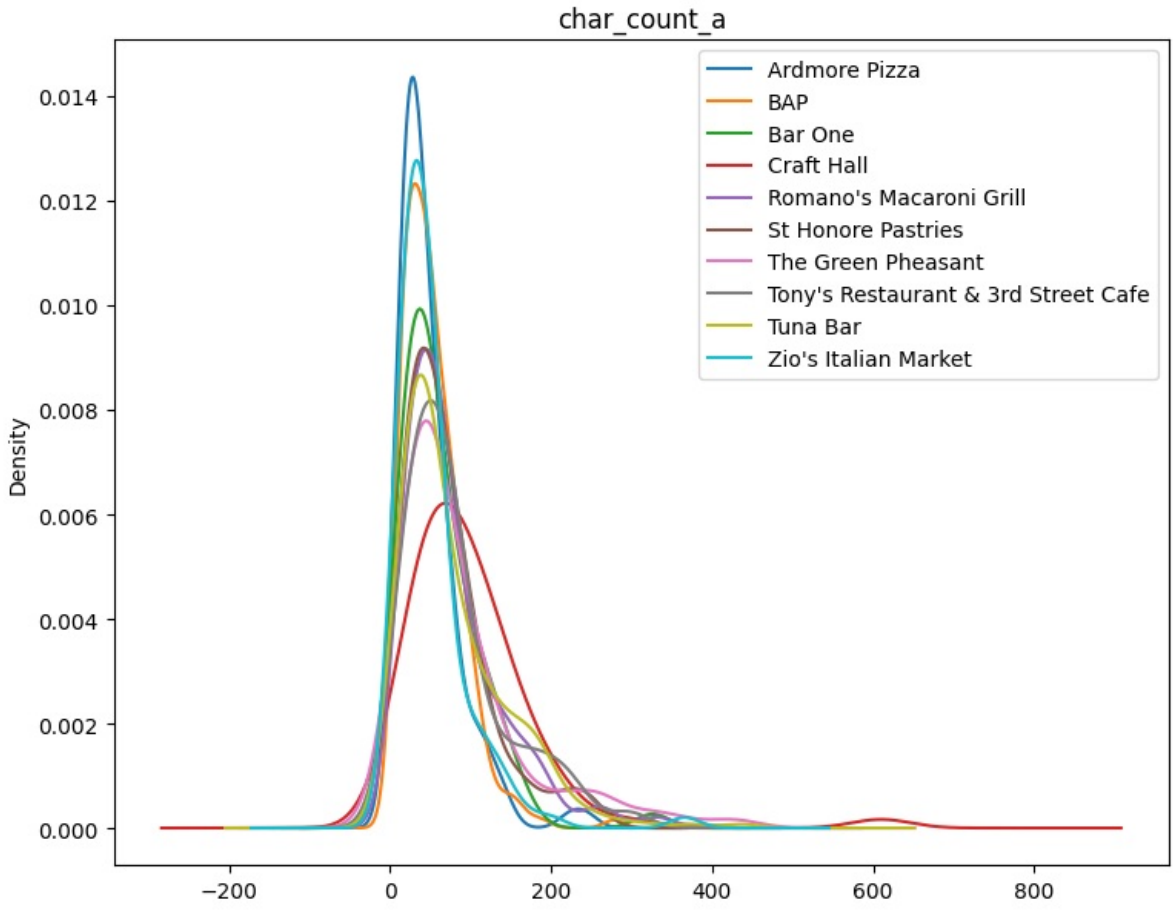
In [31]:
```
columns_to_plot = ['sent_count', 'char_count', 'word_count_b', 'char_count_a', 'word_count_a', 'stopword_count'

fig, axes = plt.subplots(nrows=len(columns_to_plot), figsize=(8, 6 * len(columns_to_plot)))

for i, col in enumerate(columns_to_plot):
    ax = axes[i]
    for label, df in grouped_df:
        df[col].plot(kind="kde", ax=ax, label=label)

    ax.set_title(col)
    ax.legend()

plt.tight_layout()
plt.show()
```

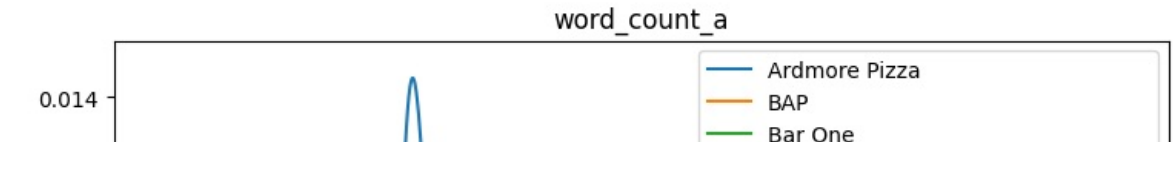


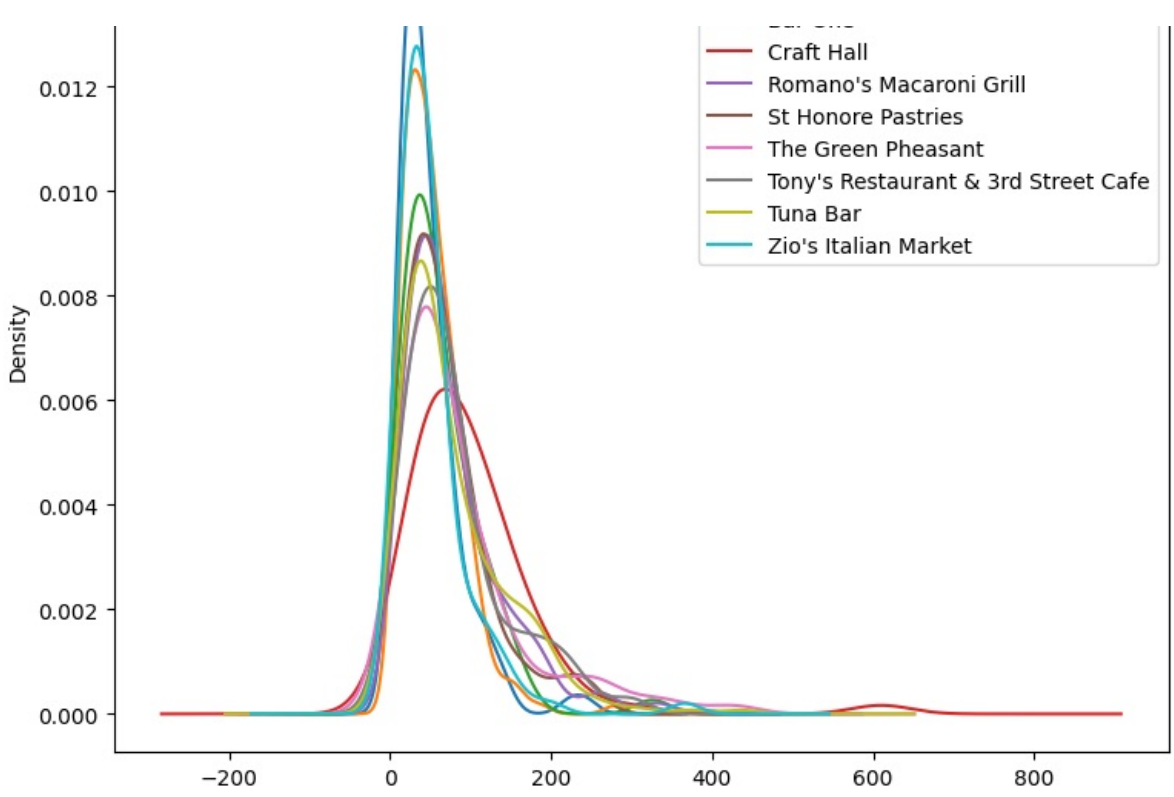



0.0000
−2000
−1000
0
1000
2000
3000
4000
5000
6000

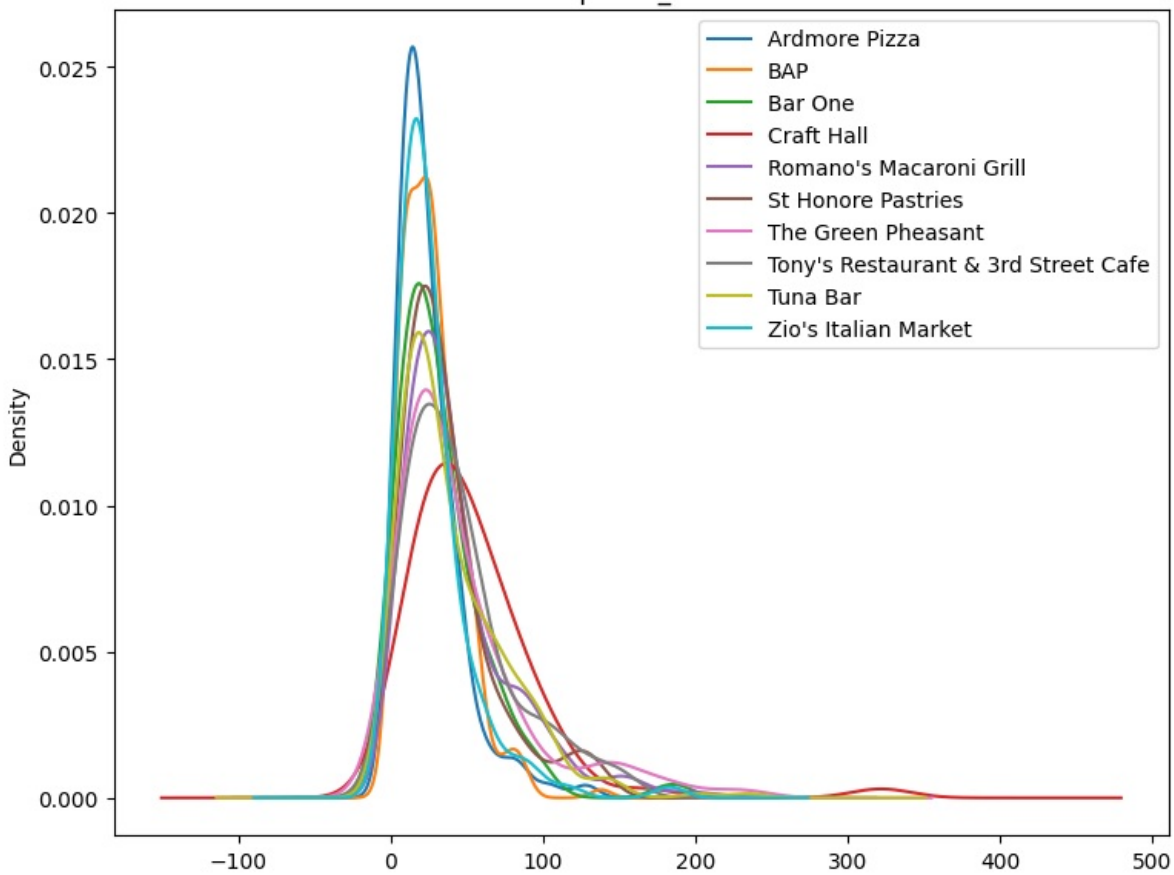

word_count_b
Density
0.008
0.006
0.004
0.002
0.000
−500
−250
0
250
500
750
1000
1250
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market


char_count_a
Density
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−200
0
200
400
600
800
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market

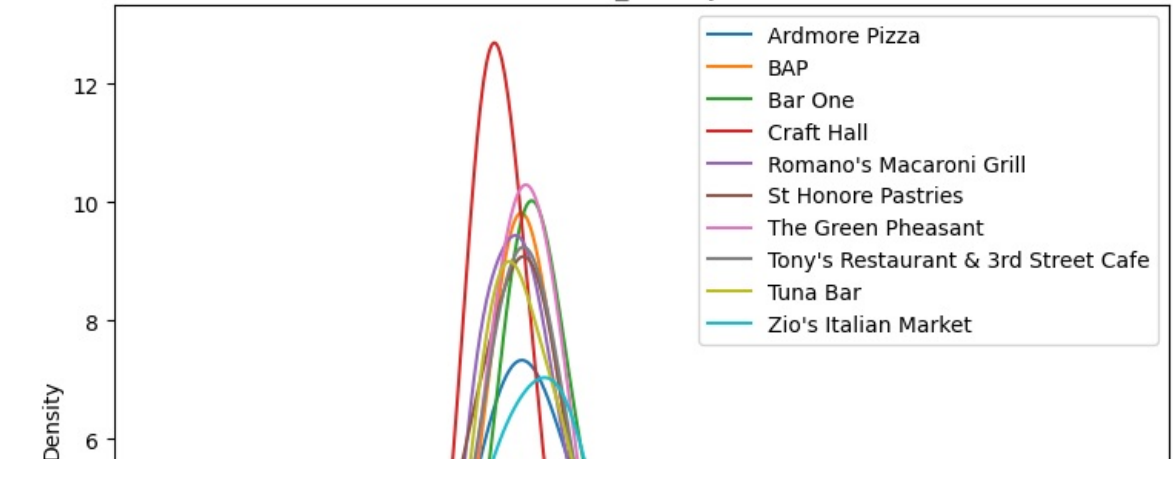

word_count_a
0.014
Ardmore Pizza
BAP
Bar One

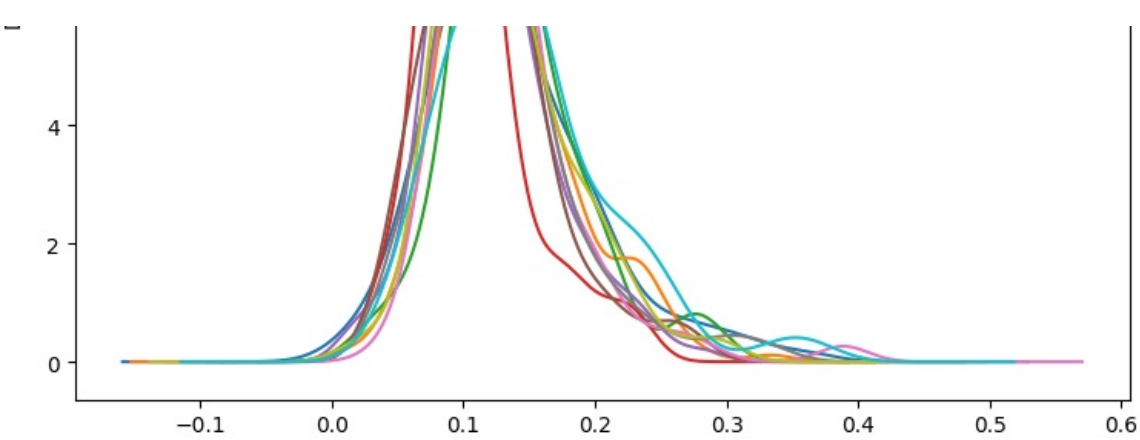


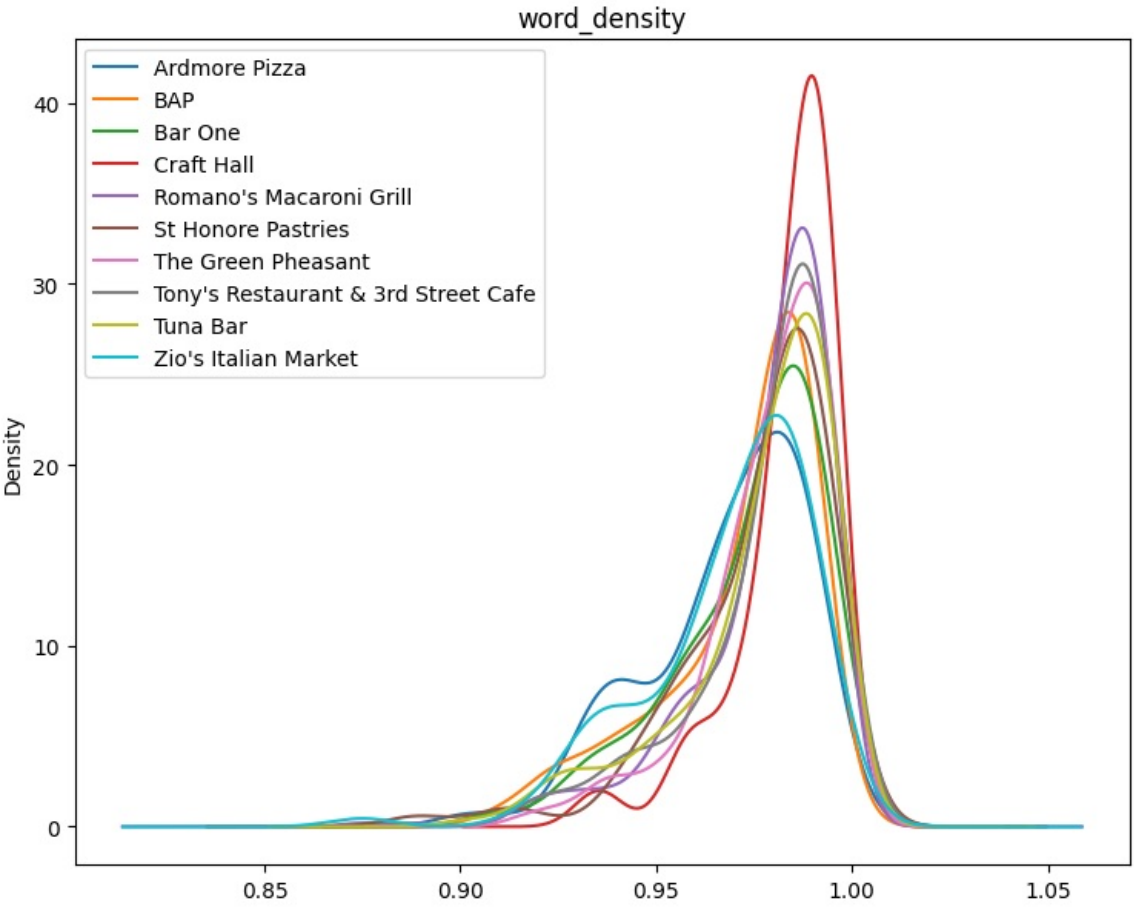
stopword_count

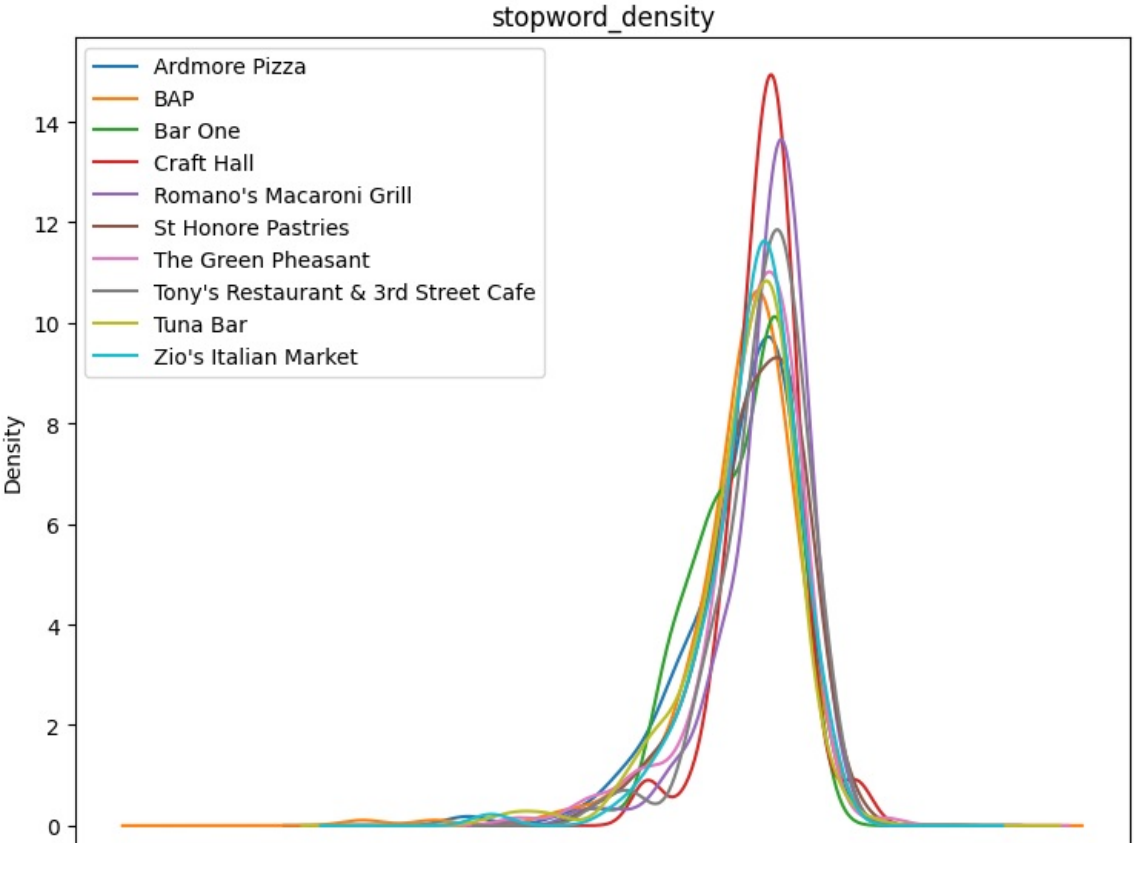


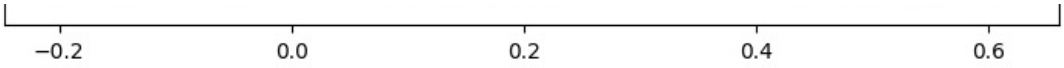
sent_density




4
2
0
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6


word_density
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market
Density
40
30
20
10
0
0.85
0.90
0.95
1.00
1.05


stopword_density
Ardmore Pizza
BAP
Bar One
Craft Hall
Romano's Macaroni Grill
St Honore Pastries
The Green Pheasant
Tony's Restaurant & 3rd Street Cafe
Tuna Bar
Zio's Italian Market
Density
14
12
10
8
6
4
2
0



- The above are the Kernel Density Estimation plots for different columns in the data frame.
- We can observe the probability density distribution of text metrics related to different businesses.

In [32]:
```
import nltk
from nltk.corpus import stopwords

nltk.download("stopwords")
sr = stopwords.words('english')
for i, tokens in enumerate(clean_review['tokens']):
  for token in tokens:
    if token in stopwords.words('english'):
      clean_review['tokens'][i].remove(token)
clean_review
```

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

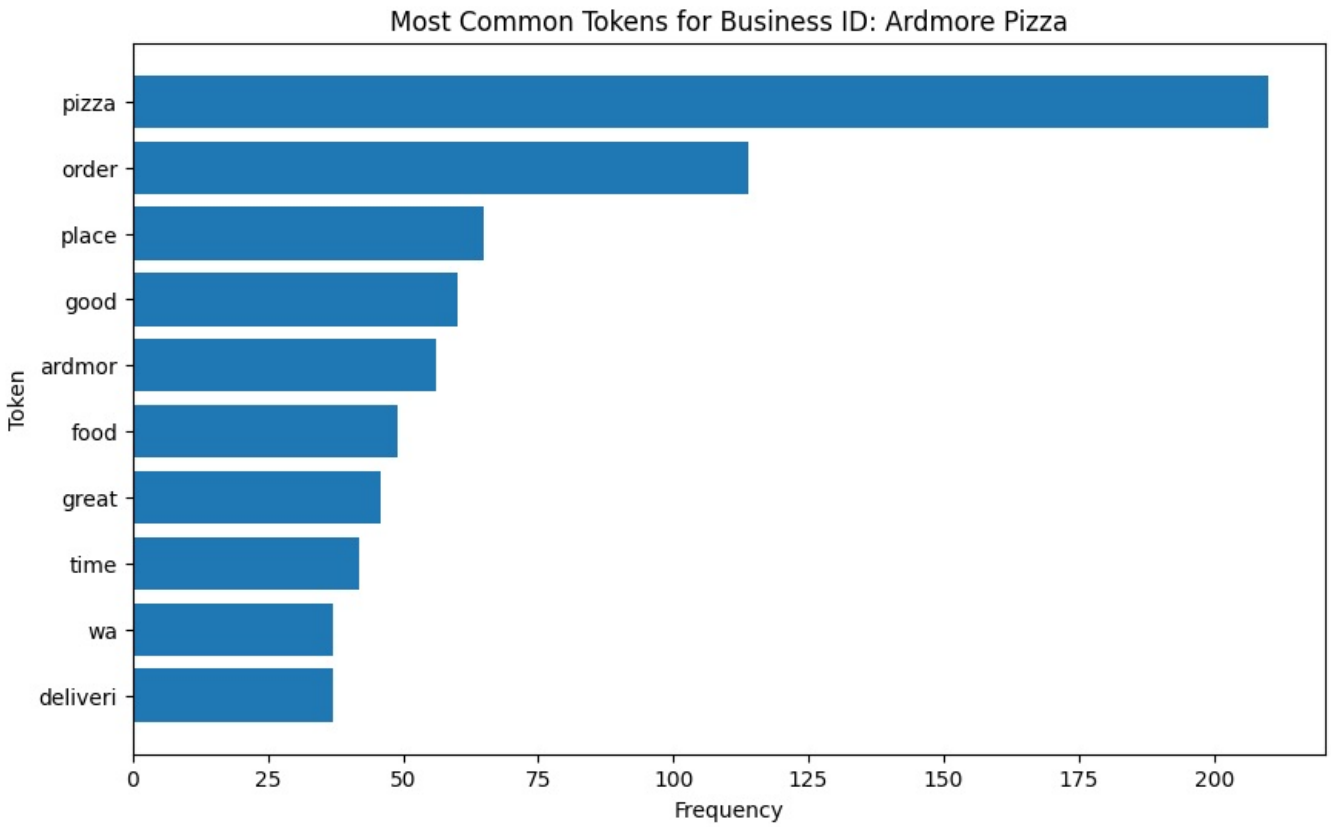
Out[32]:

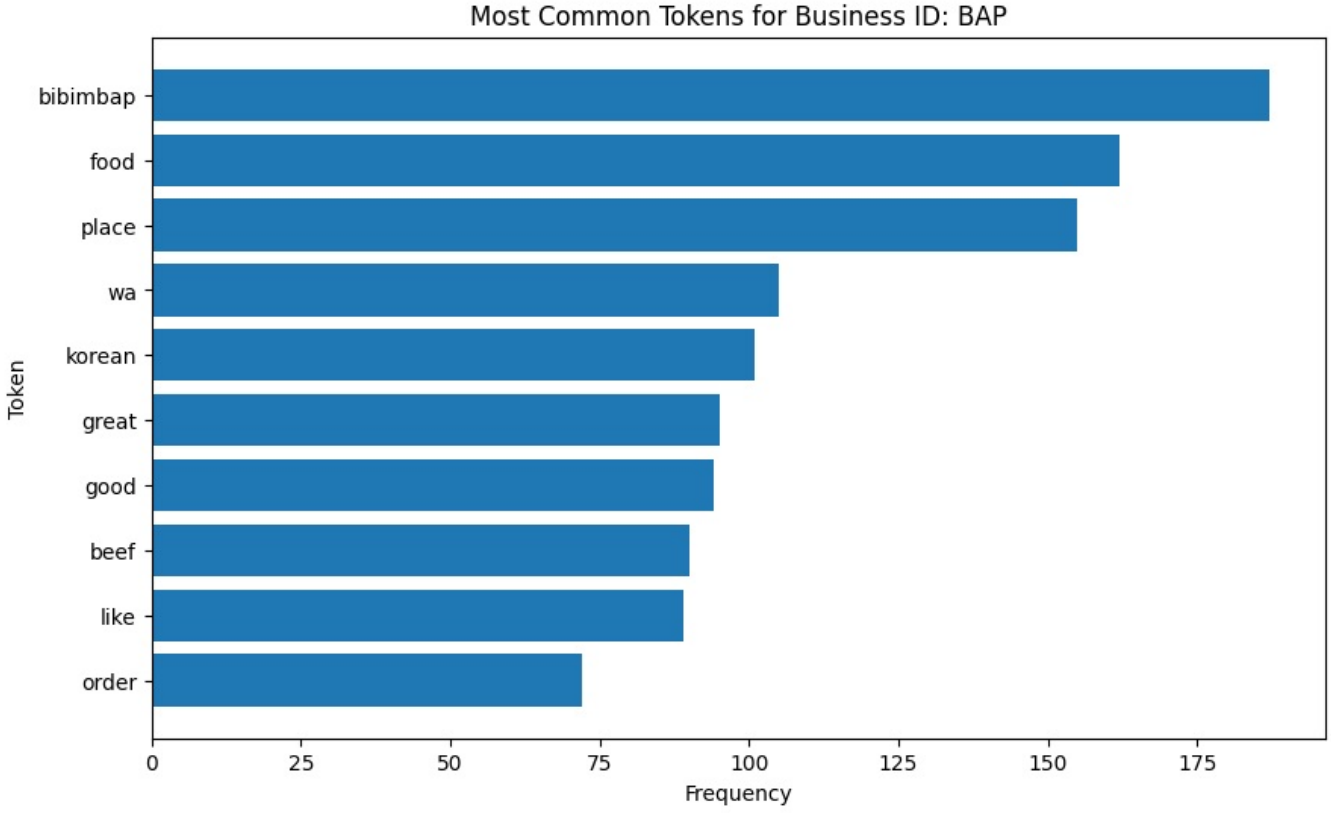
	review_id	user_id	business_id	text	date	sentence	sent_count
0	mMwnX1vc3tQUeDNS2wiKFw	f10WH1fXhy-68r4AEEhAWA	9OG5YkX1g2GReZM0AskizA	great bar happy hour every day wine dra...	2016-01-30 03:16:46	[great bar happy hour 4-7 every day., wine & d...	4
1	z_fgvINjKZCw5RgByaTxxw	dldfg-X_QbBkhR2DOsQFWg	QdN72BWoyFypdGJhhI5r7g	this place is top notch with phenomenal servi...	2016-11-10 16:52:33	[this place is top notch, with phenomenal serv...	13
2	rkDzWtbZ2_en8HZDCUbF1Q	-TbX3AYOIEyo6-b67MT8eA	9OG5YkX1g2GReZM0AskizA	please this place makes a semi new menu and r...	2013-04-11 02:40:03	[please, this place makes a semi-new menu and ...	15
3	XYaDbPKyJAu4k2aUOIth5g	Qsk0aTclam9W_DIK6bx42A	MUTTqe8uqyMdBl186RmNeA	stopped in to check out this new spot around t...	2017-12-16 00:13:06	[stopped in to check out this new spot around ...	9
4	tpLolBuBTx_Ncx3RSf7WBw	TJW1aEzjhaxbD10fjhokfQ	MUTTqe8uqyMdBl186RmNeA	i live in the neighborhood and used to order a...	2018-04-28 00:46:05	[i live in the neighborhood and used to order ...	8
...	...	...	...	...	...	...	...
1517	qNrqLFzUotJXqhO_8k2fEw	BgbMh5k8Gd3YoQOfX915Xw	tMkwHmWFUEXrC9ZduonpTg	everything on the menu is absolutely incredibl...	2020-02-11 14:27:39	[everything on the menu is absolutely incredib...	3
1518	9r-TVMhfk5ncJ_Cc_lprUQ	cC9flaguB3JXdQSghVT03Q	MUTTqe8uqyMdBl186RmNeA	from the ambiance to the service and the food...	2021-11-29 01:42:04	[from the ambiance, to the service and the foo...	11
1519	AJtRqi_xQJs5YoTnuUlaTw	u9kFHr0ZyuvXYejCaxz7ew	aPNXGTDkf-4bjhyMBQxqpQ	went to a frontier event with a friend today a...	2019-04-29 23:48:18	[went to a frontier event with a friend today ...	14
1520	Me2ixr2UWaqnMWTGFwkyRQ	Y_CWjc7mz6jaebDxyyVViw	MUTTqe8uqyMdBl186RmNeA	tuna bar has quickly become a favorite of phil...	2021-12-04 01:21:49	[tuna bar has quickly become a favorite of phi...	7
1521	yP6qECsUSGs4Jsnr2Pu5KQ	BpfStJAeH3-8mnvZKwh1qg	MUTTqe8uqyMdBl186RmNeA	stars all around the sushi here is amazing ...	2017-12-31 13:51:49	[5 stars all around!, the sushi here is amazin...	18

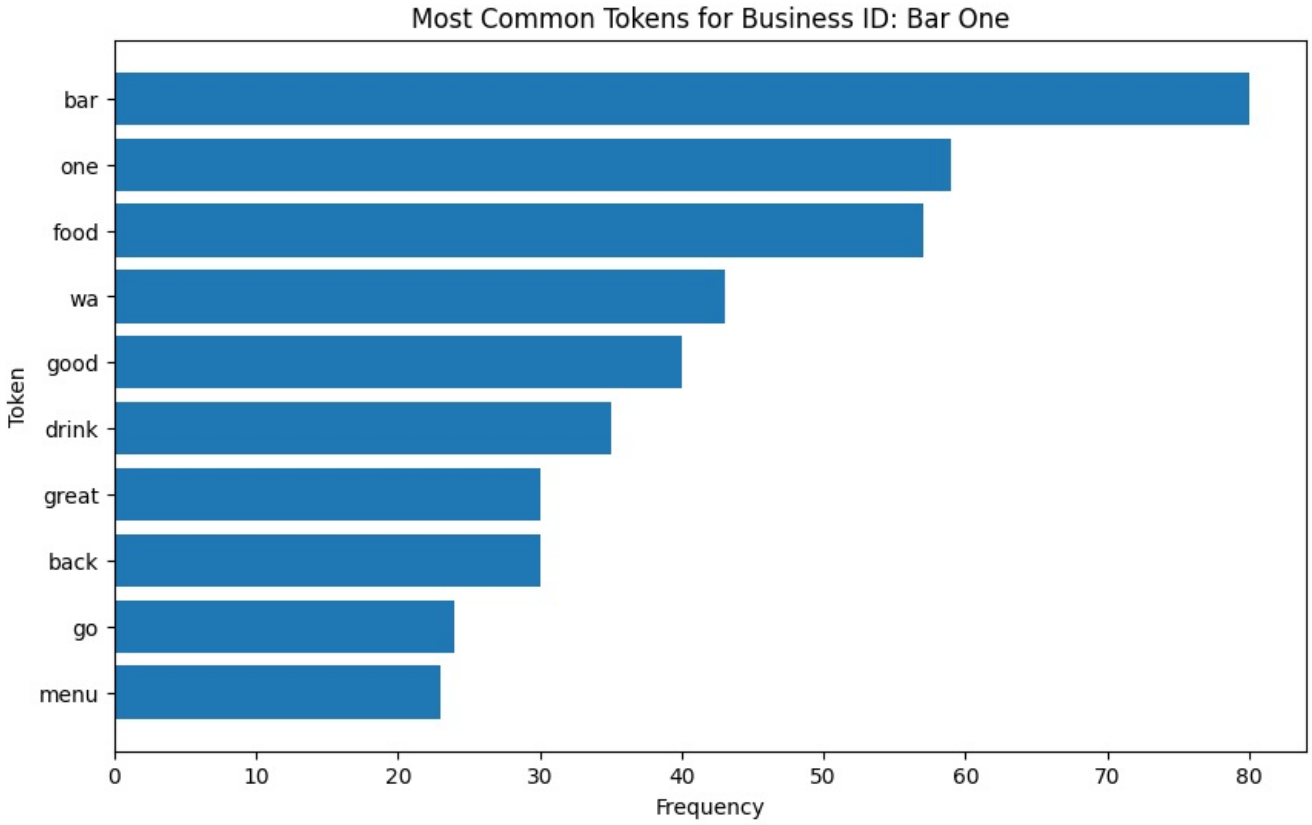
1522 rows × 17 columns

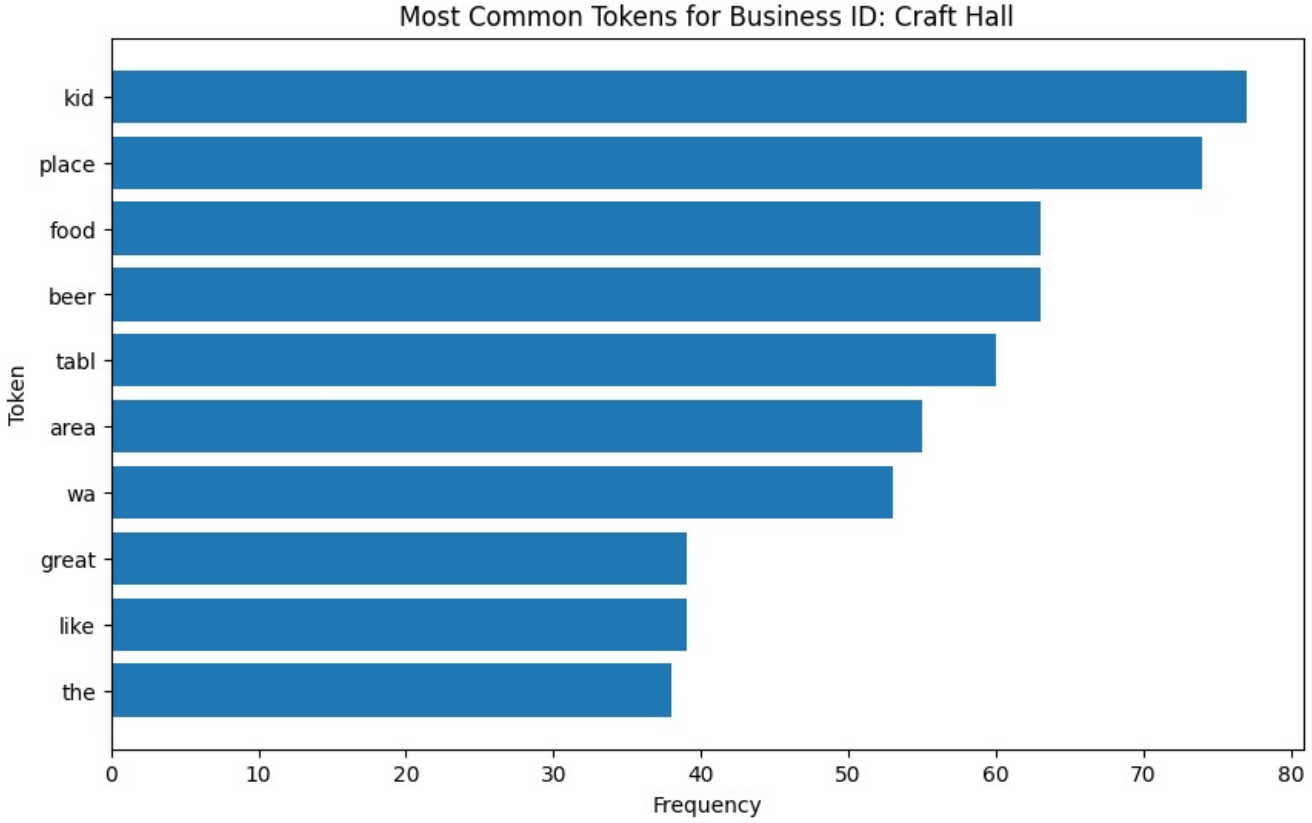
In [33]:
```
from collections import Counter
import matplotlib.pyplot as plt
top_n = 10

for business_name, group in grouped_df:


    all_tokens = [token for sublist in group['tokens'] for token in sublist]

    token_counts = Counter(all_tokens)
```

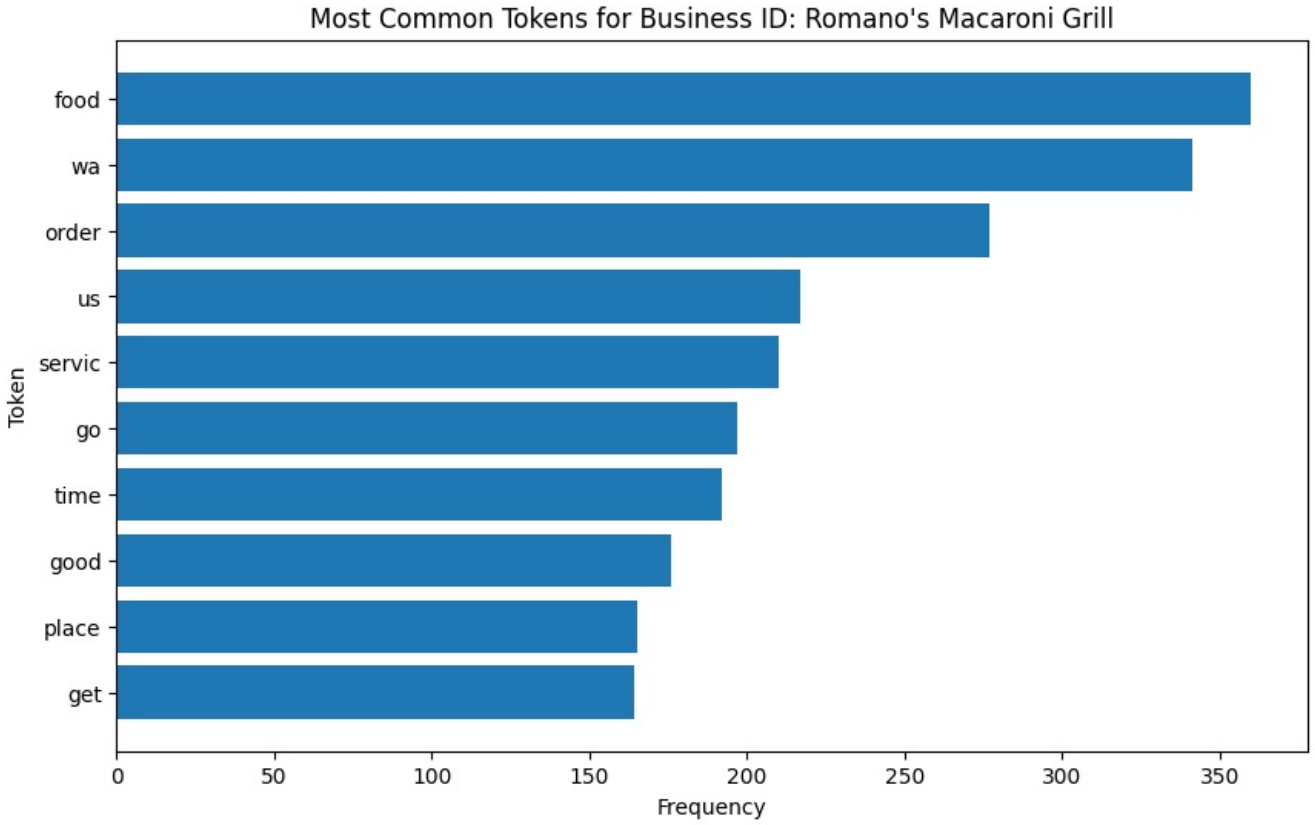
```
most_common_tokens = token_counts.most_common(top_n)

most_common_df = pd.DataFrame(most_common_tokens, columns=['Token', 'Frequency'])

plt.figure(figsize=(10, 6))
plt.barh(most_common_df['Token'], most_common_df['Frequency'])
plt.xlabel('Frequency')
plt.ylabel('Token')
plt.title(f'Most Common Tokens for Business ID: {business_name}')
plt.gca().invert_yaxis()
plt.show()
```

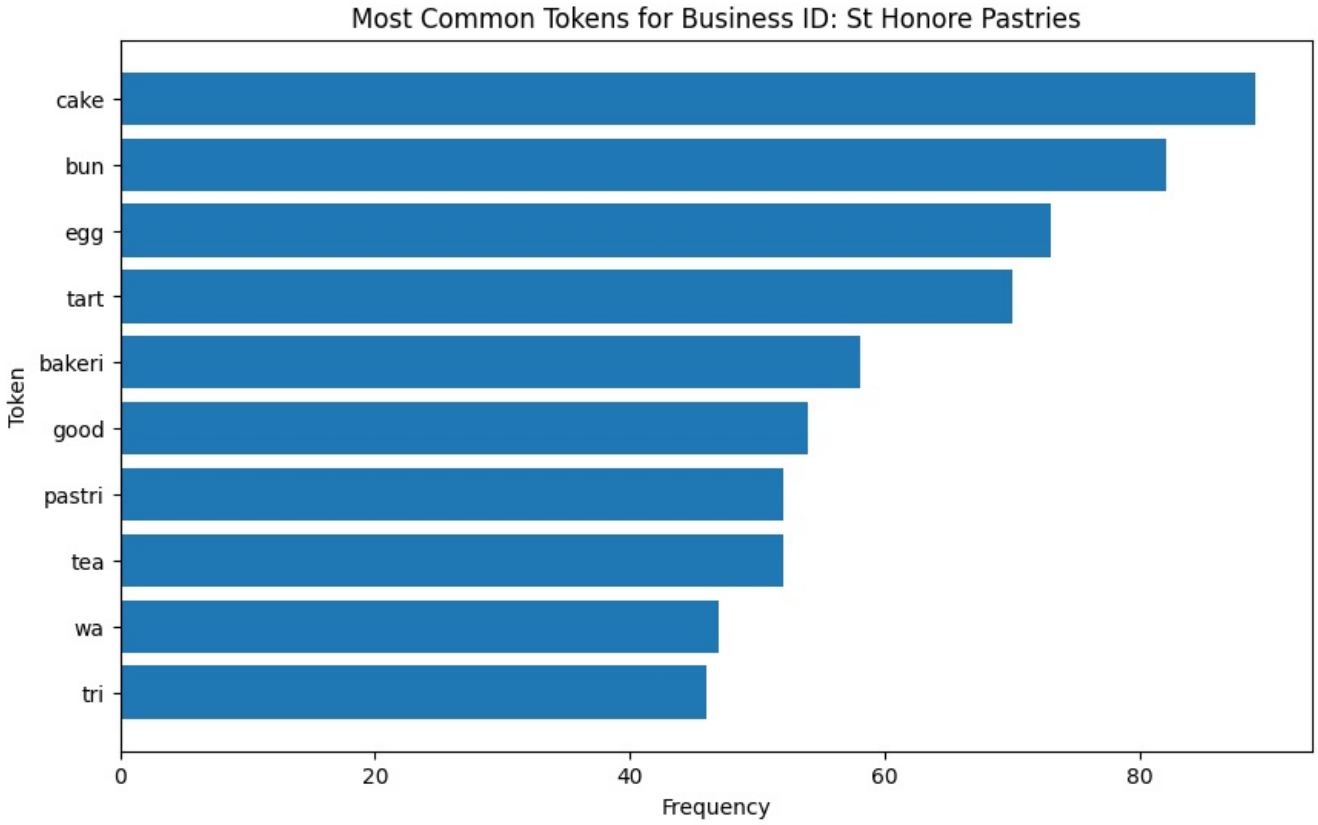


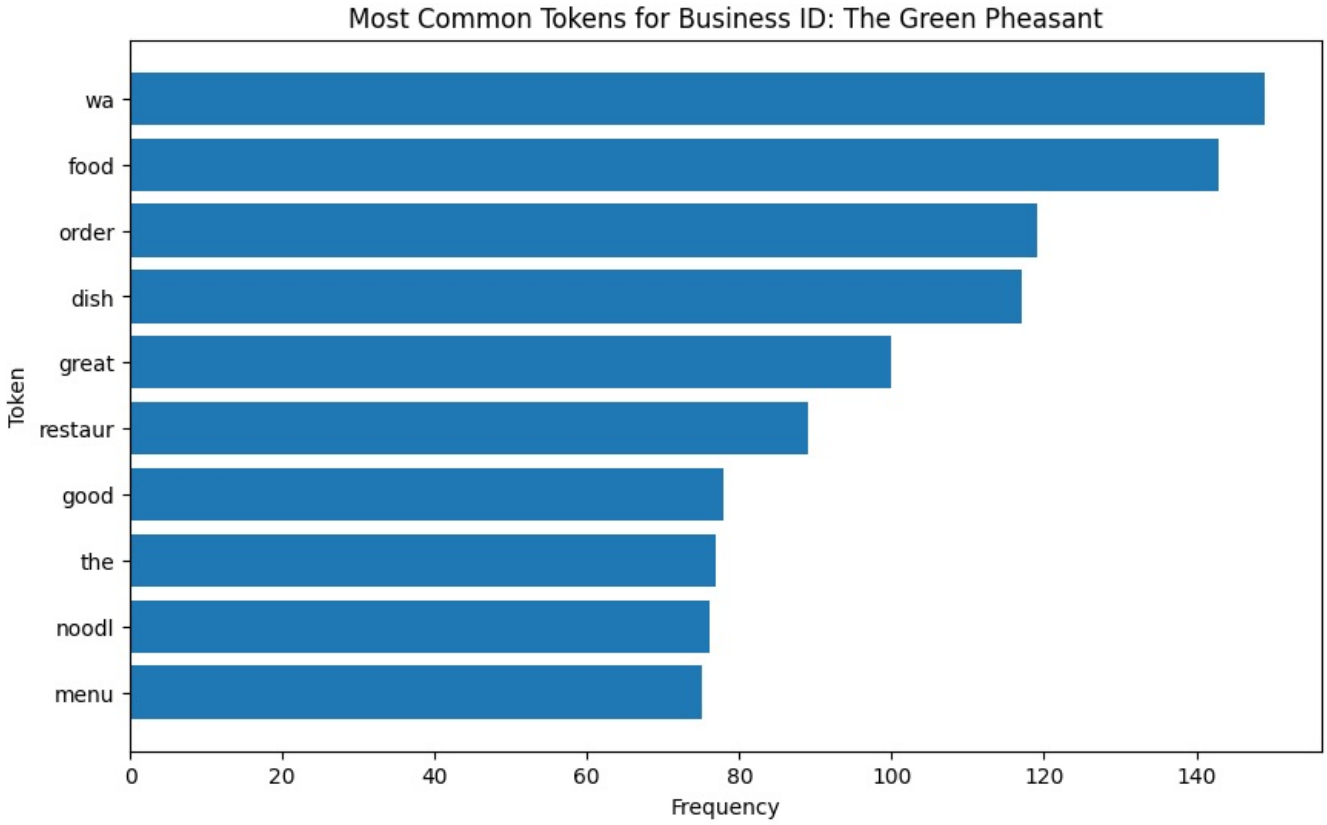


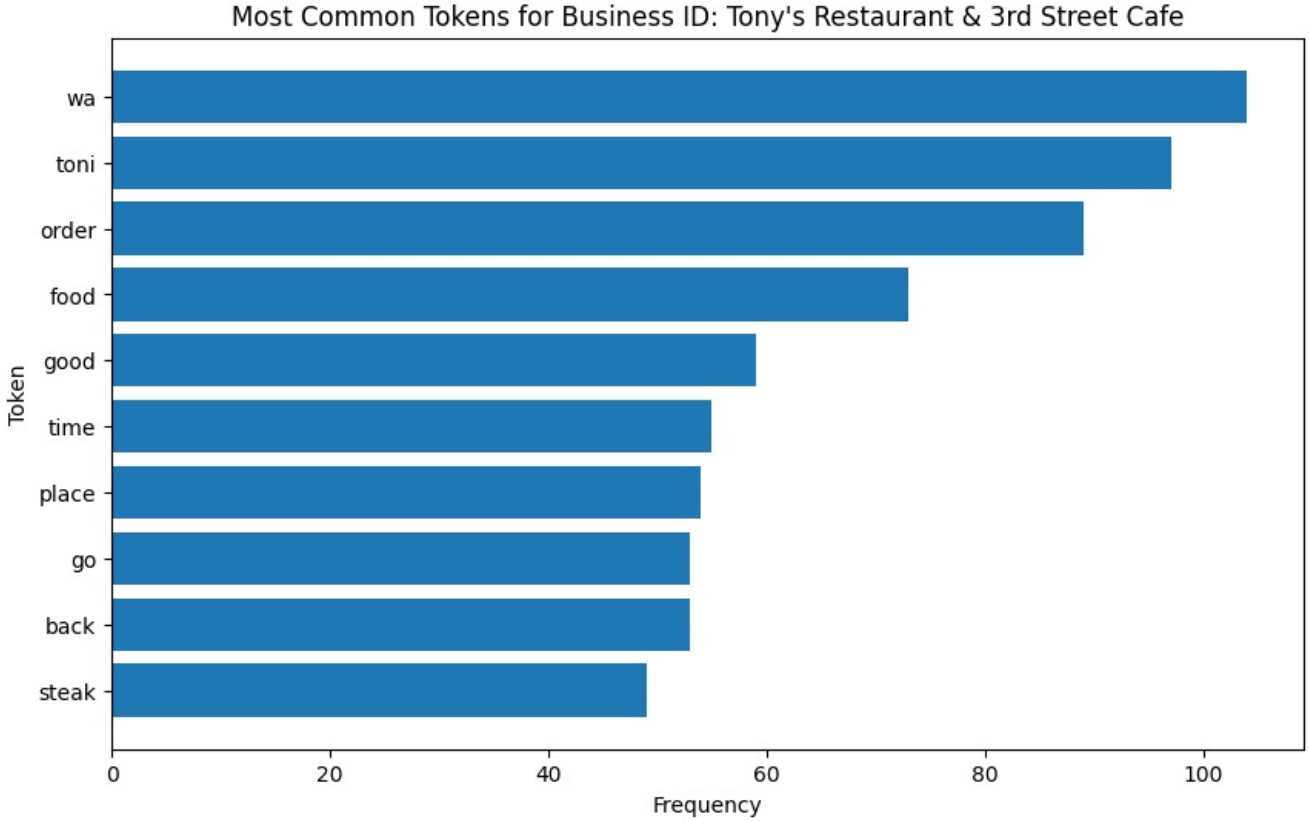

Most Common Tokens for Business ID: Bar One
bar
one
food
wa
good
drink
great
back
go
menu
Token
0
10
20
30
40
50
60
70
80
Frequency

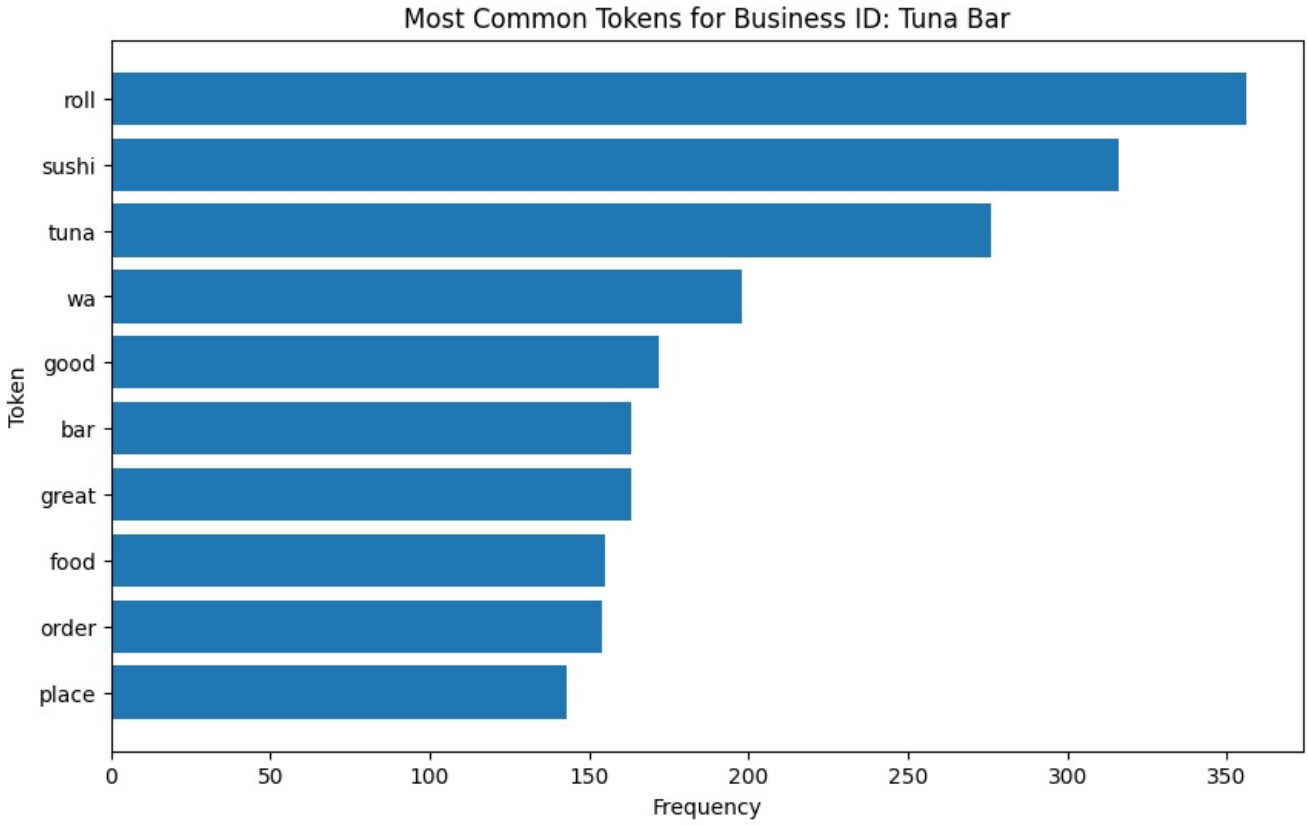

Most Common Tokens for Business ID: Craft Hall
kid
place
food
beer
tabl
area
wa
great
like
the
Token
0
10
20
30
40
50
60
70
80
Frequency

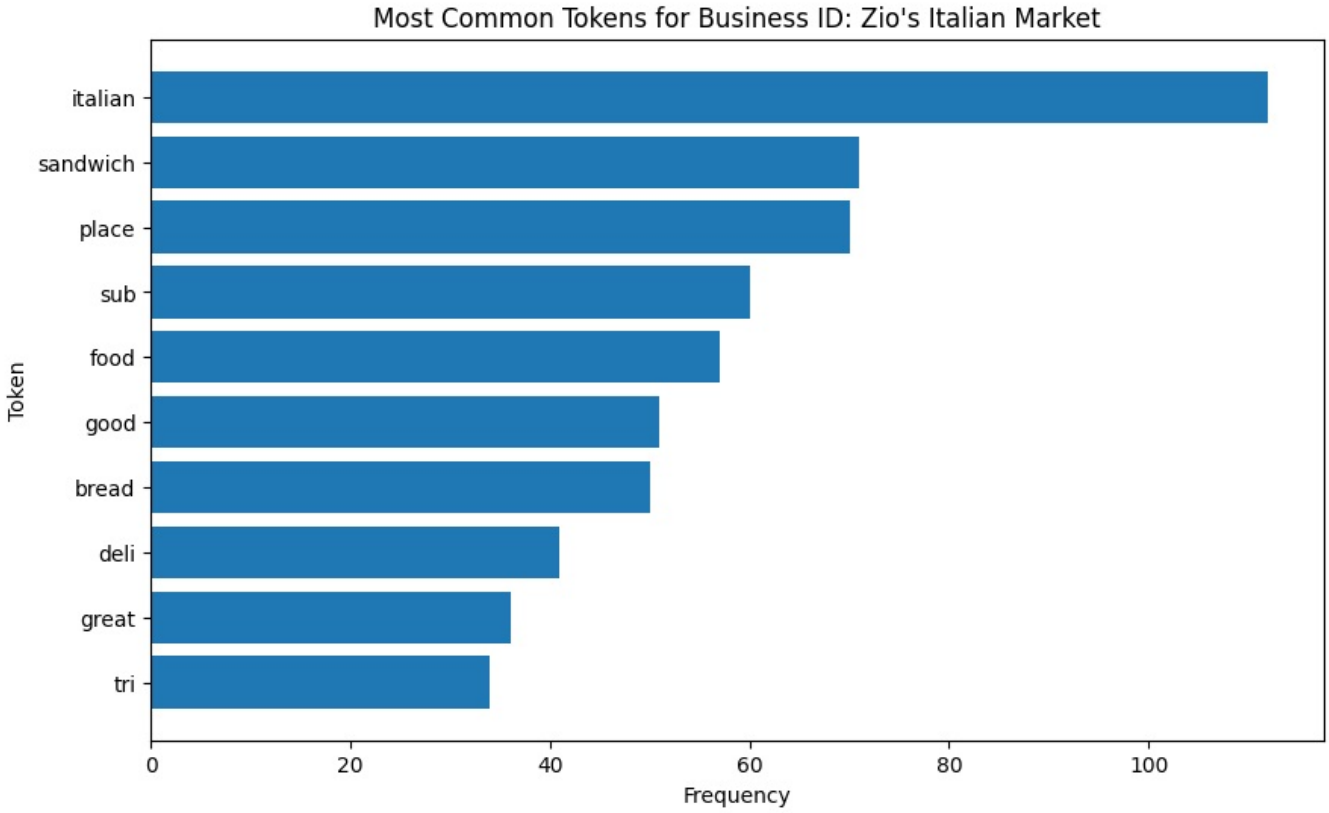

Most Common Tokens for Business ID: Romano's Macaroni Grill
Token
food
wa
order
us
servic
go
time
good
place
get
0
50
100
150
200
250
300
350
Frequency


Most Common Tokens for Business ID: St Honore Pastries
Token
cake
bun
egg
tart
bakeri
good
pastri
tea
wa
tri
0
20
40
60
80
Frequency


Most Common Tokens for Business ID: The Green Pheasant
wa
food
order
dish
great
restaur
good
the
noodl
menu
Token
0
20
40
60
80
100
120
140
Frequency


Most Common Tokens for Business ID: Tony's Restaurant & 3rd Street Cafe
wa
toni
order
food
good
time
place
go
back
steak
Token
0
20
40
60
80
100
Frequency

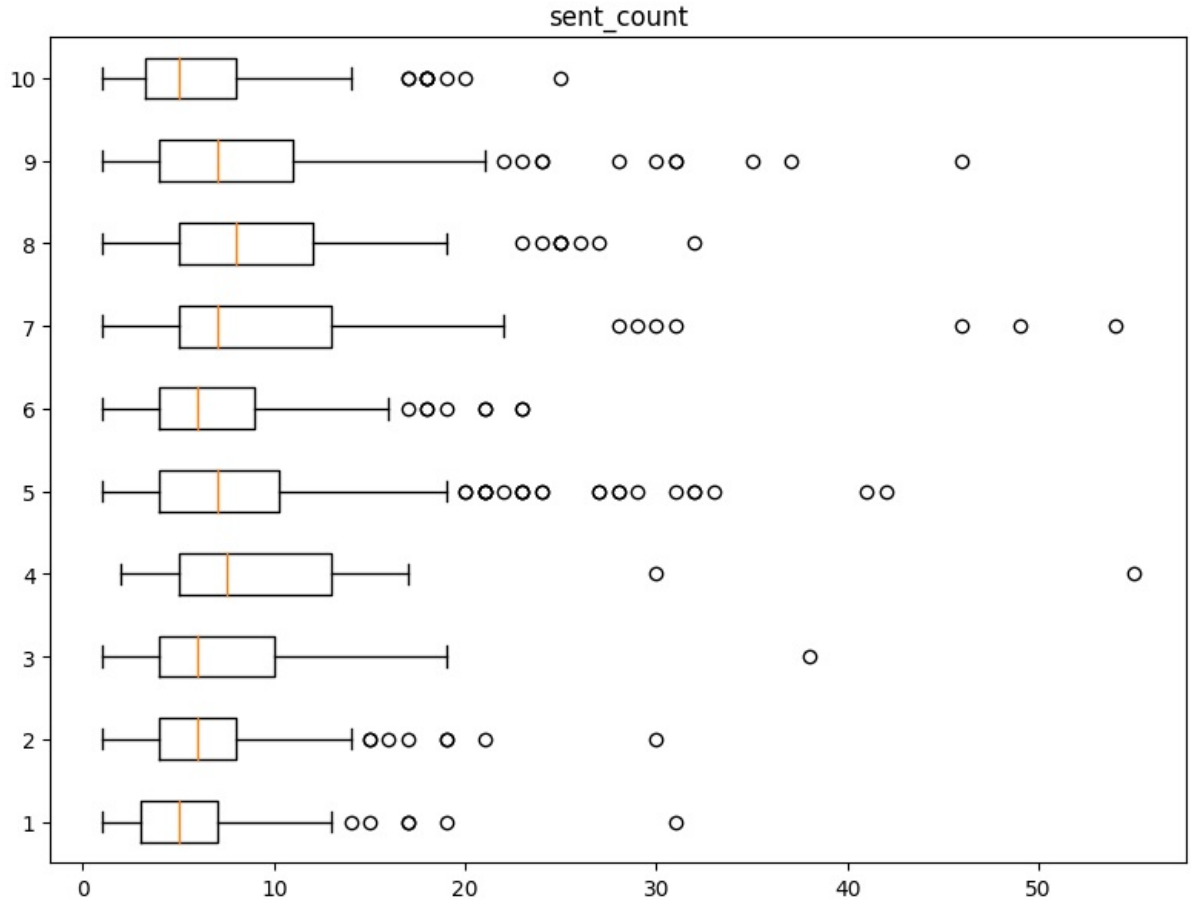


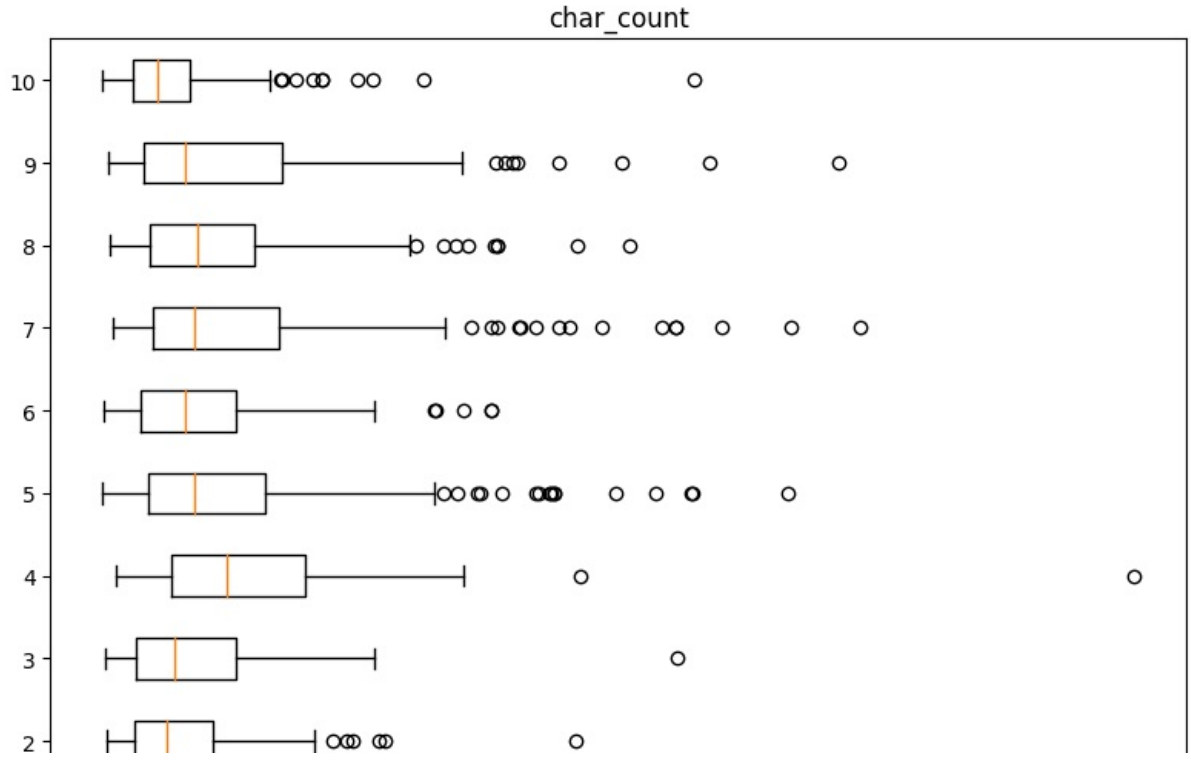


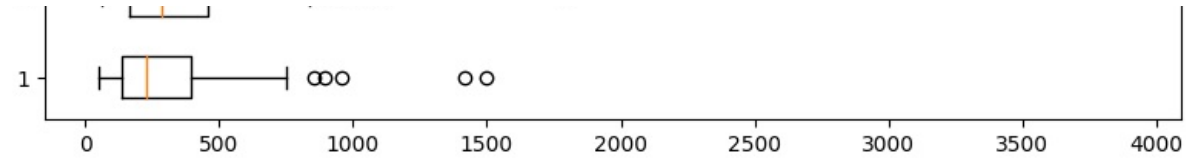
- The above graphs represent the most common tokens for different businesses.
- This can be used to understand customer semantics.
- For example in the first graph the most common token used in the reviews is Italian. So if this business has an overall positive review that means people are liking Italian food.

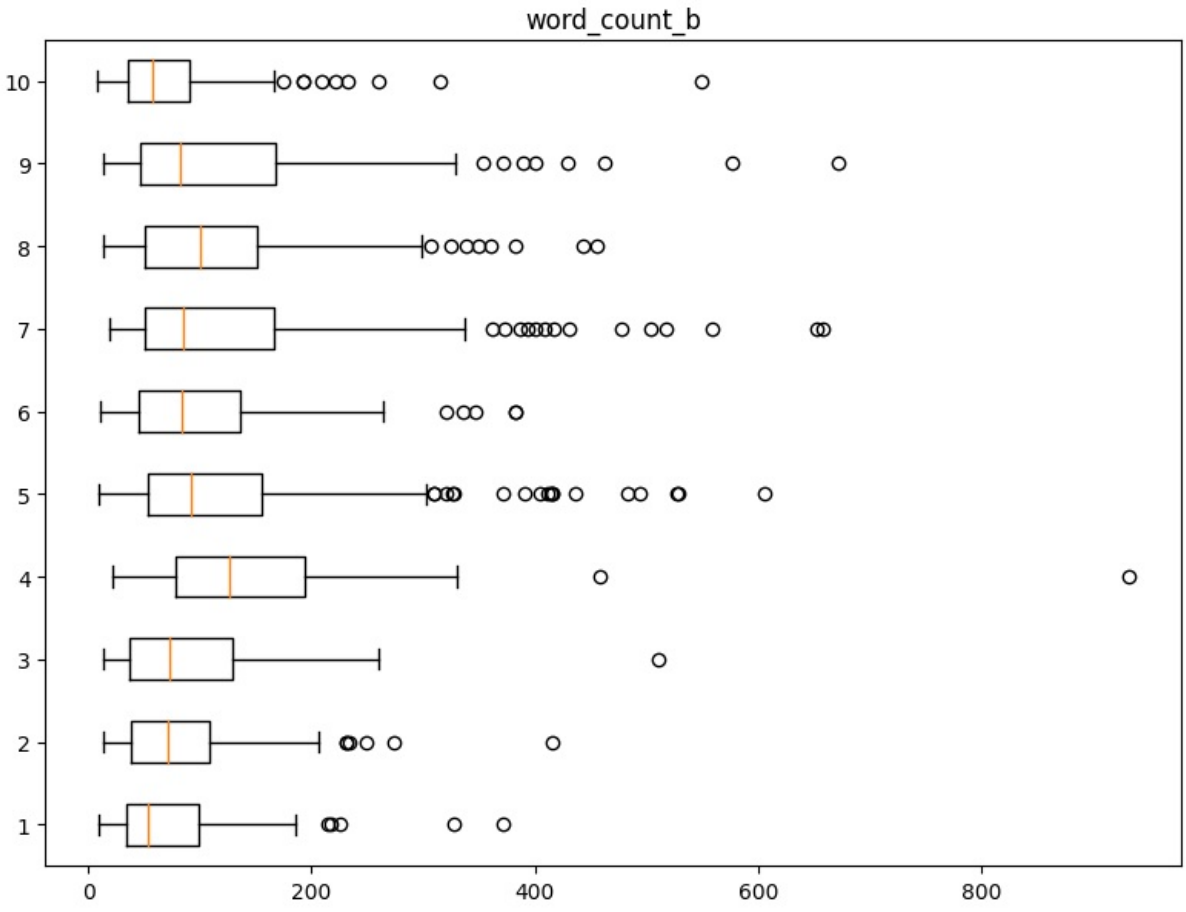
Box Plot

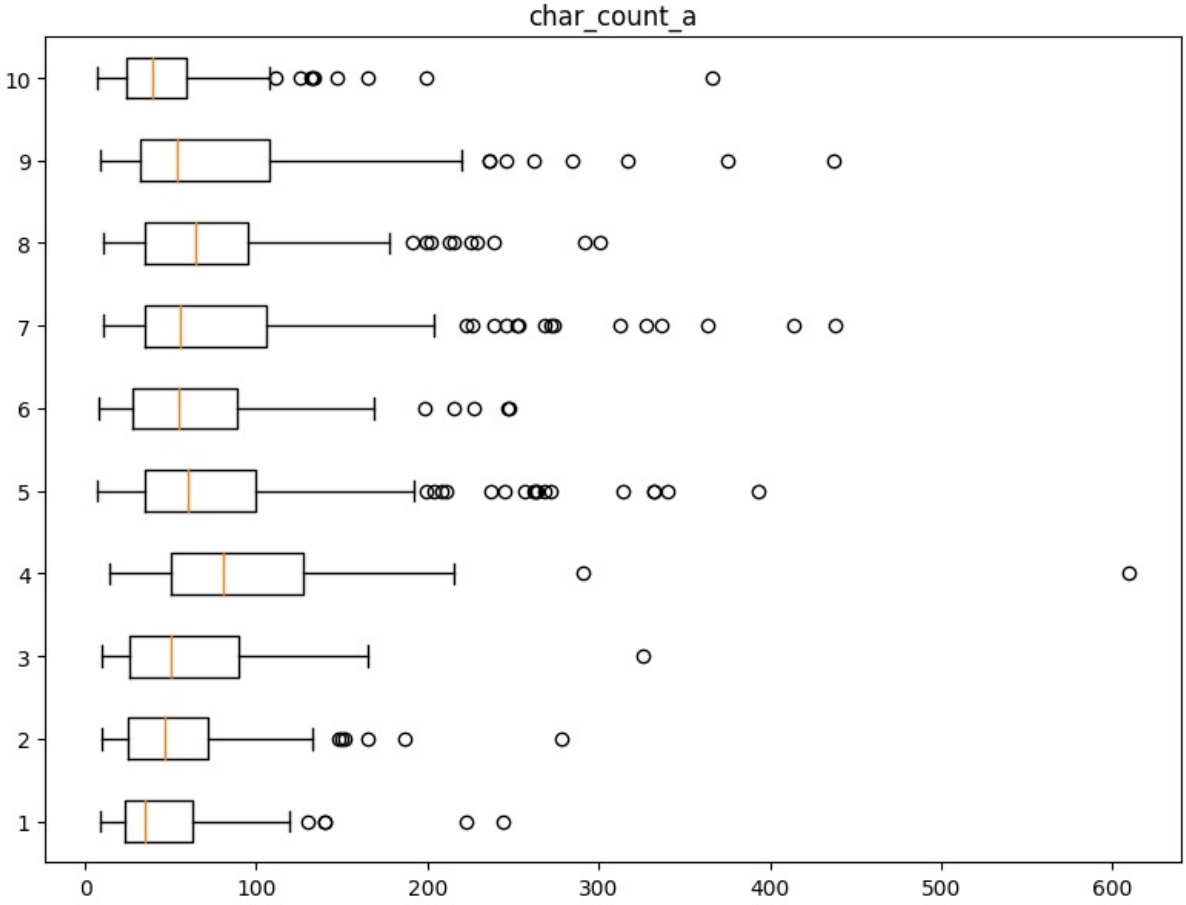
In [34]:
```
columns_to_plot = ['sent_count', 'char_count', 'word_count_b', 'char_count_a', 'word_count_a', 'stopword_count']

# Create subplots
fig, axes = plt.subplots(nrows=len(columns_to_plot), figsize=(8, 6 * len(columns_to_plot)))

# Plot each column
for i, col in enumerate(columns_to_plot):
    ax = axes[i]
    # Extract values for the current column from each group and convert them into a list of lists
    data = [group[col].values for _, group in grouped_df]
    # Plot boxplot for the current column
    ax.boxplot(data, vert=False)
    ax.set_title(col)  # Set title for the subplot
    ax.label = []

plt.tight_layout()
plt.show()
```

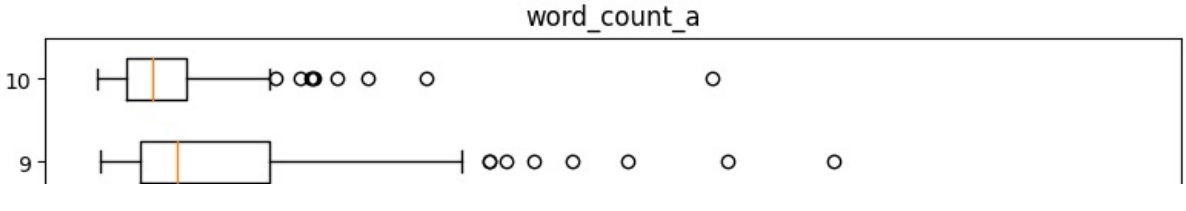


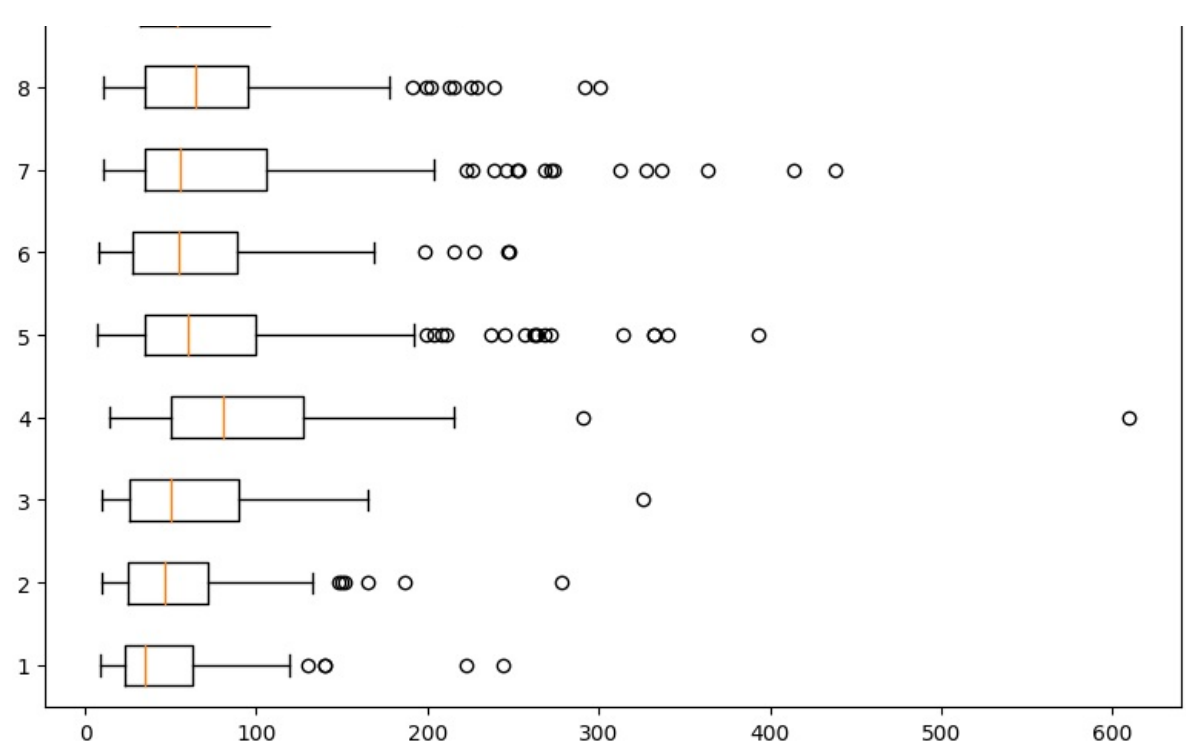



1
0
500
1000
1500
2000
2500
3000
3500
4000

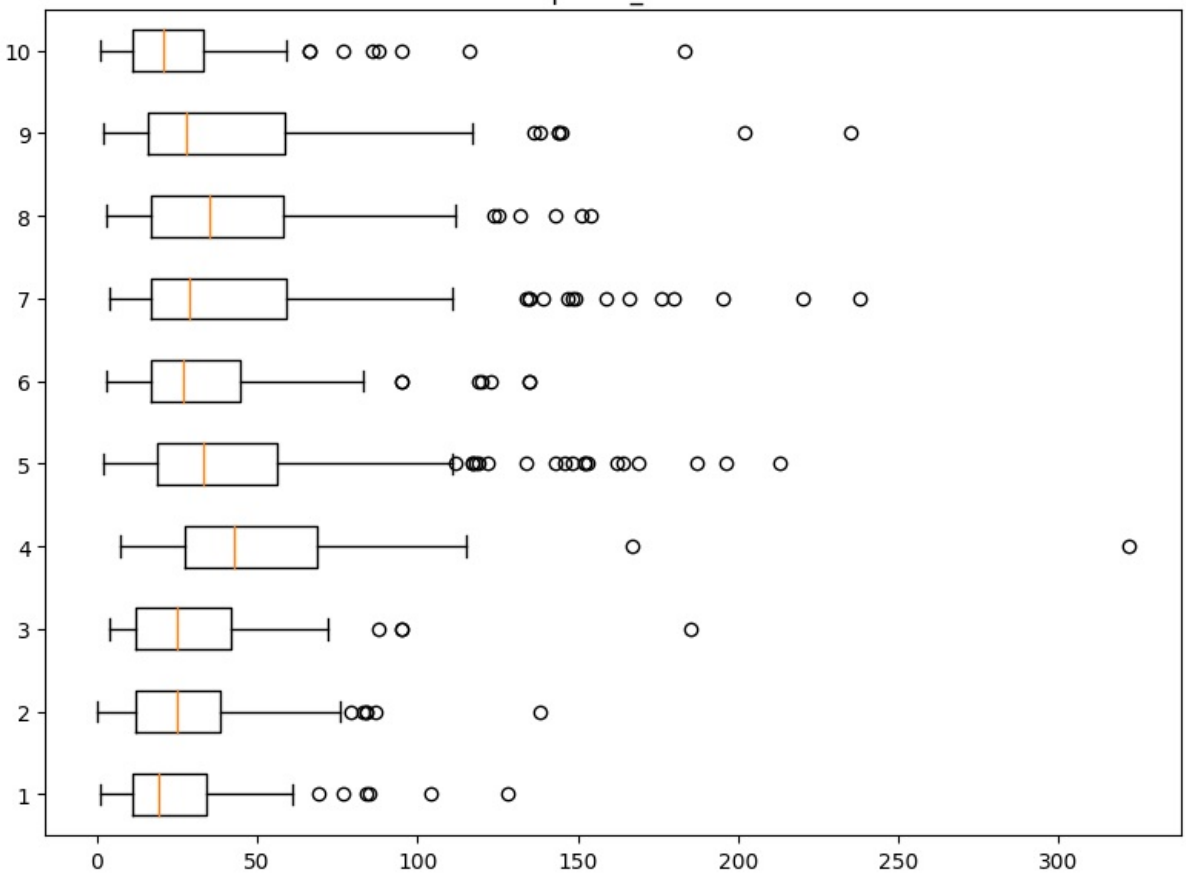

word_count_b
10
9
8
7
6
5
4
3
2
1
0
200
400
600
800


char_count_a
10
9
8
7
6
5
4
3
2
1
0
100
200
300
400
500
600

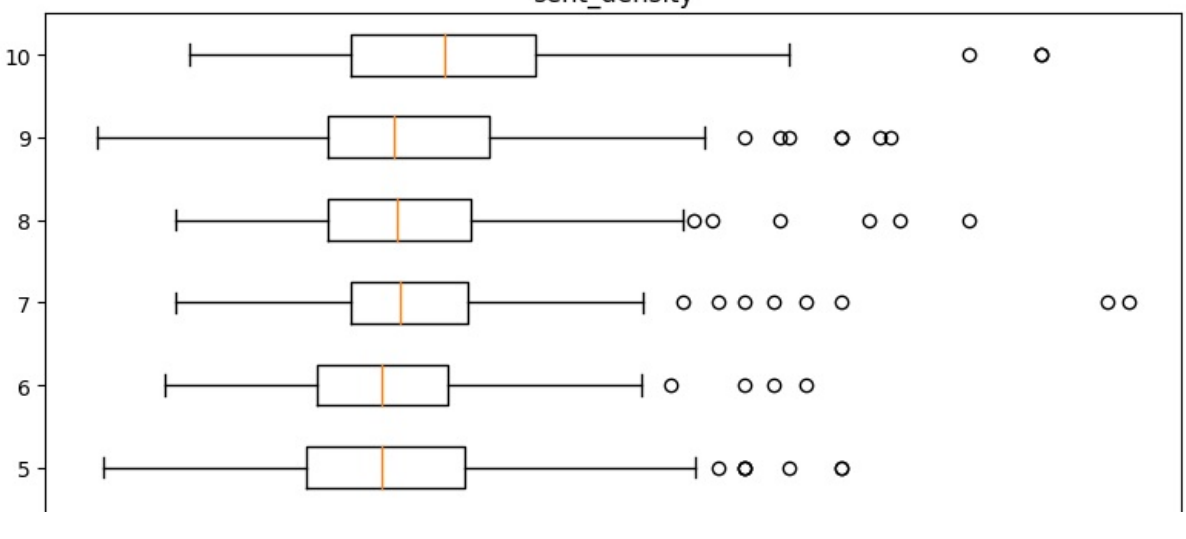

word_count_a
10
9

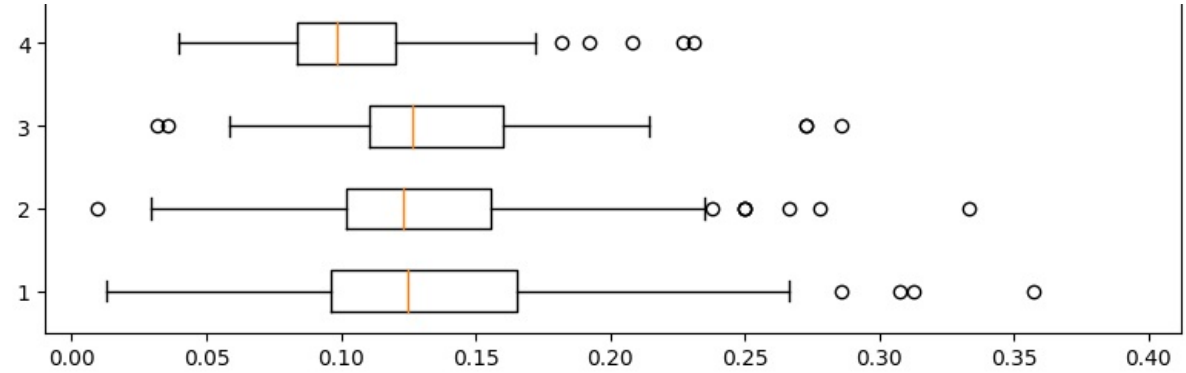


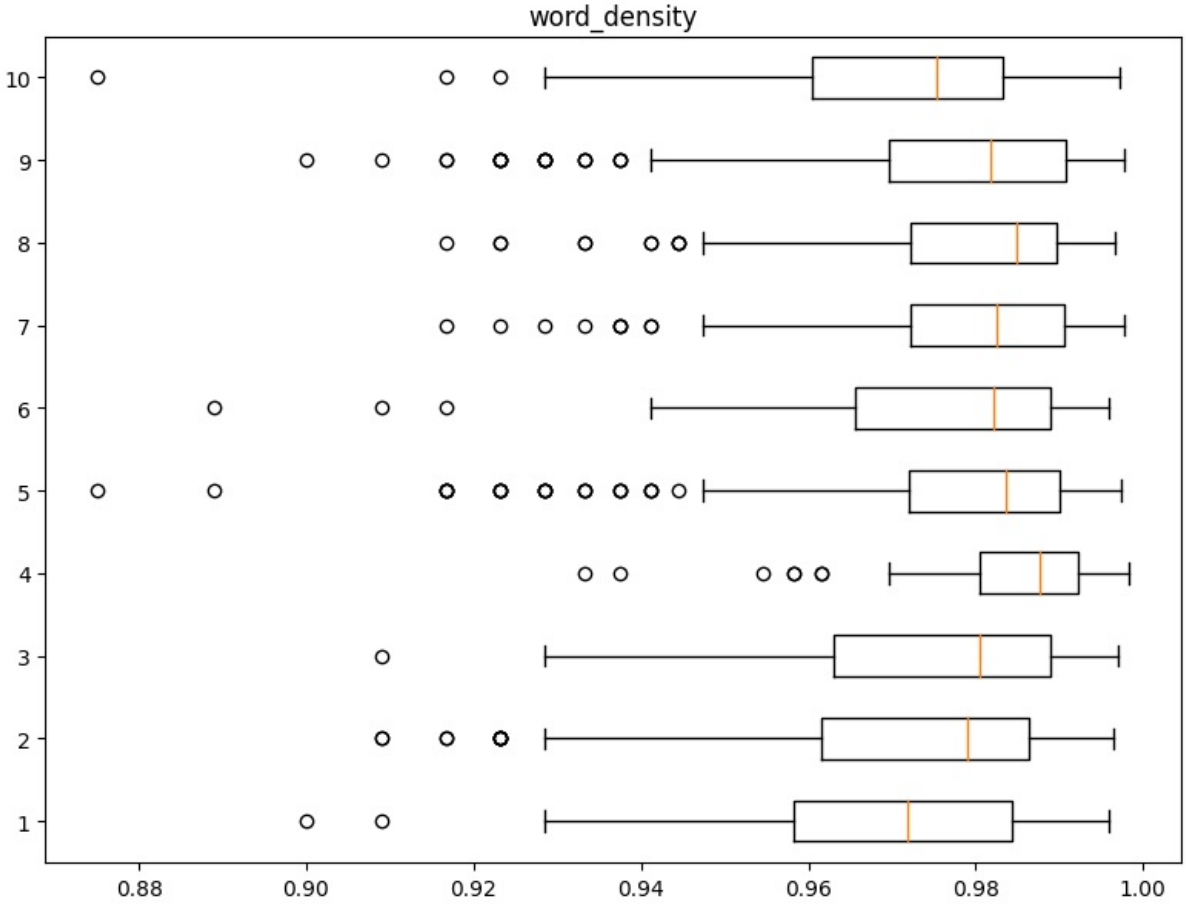
stopword_count

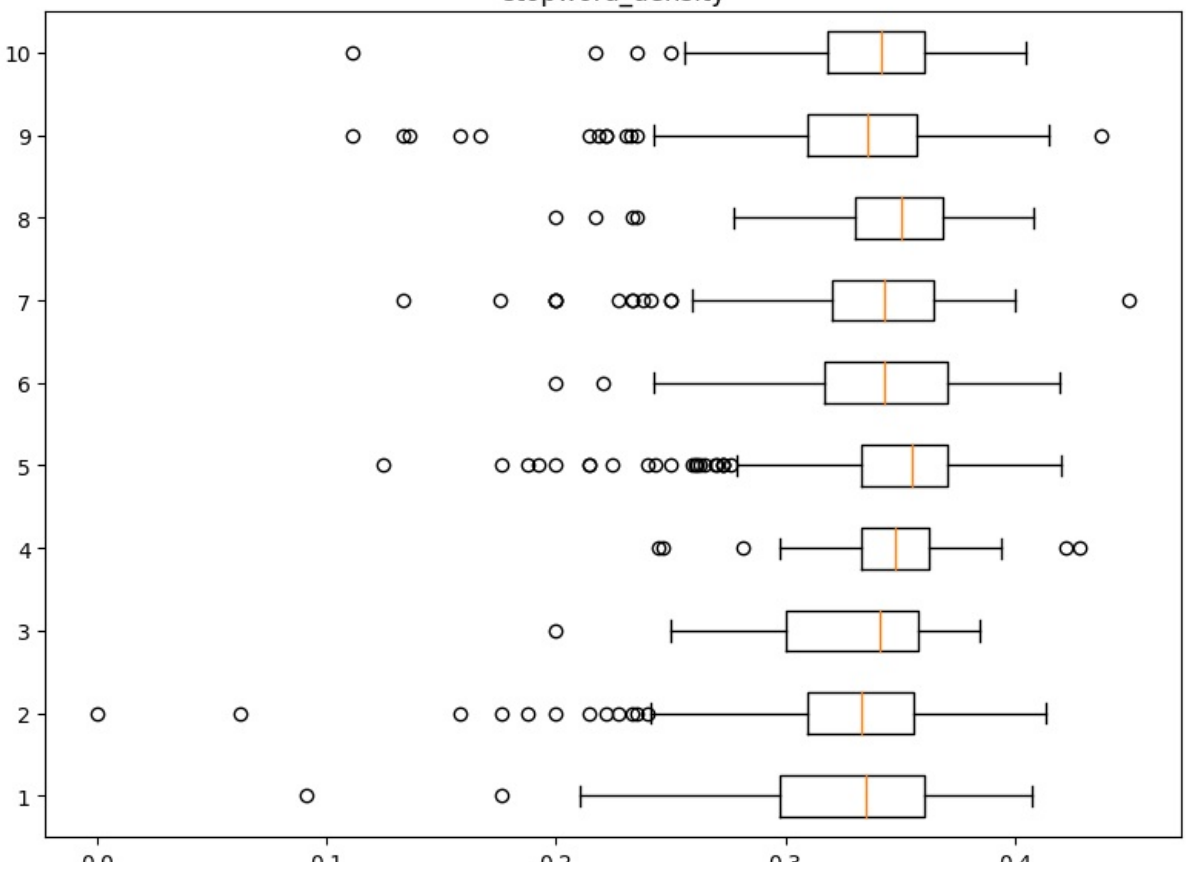


sent_density




4
3
2
1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40


word_density
10
9
8
7
6
5
4
3
2
1
0.88
0.90
0.92
0.94
0.96
0.98
1.00

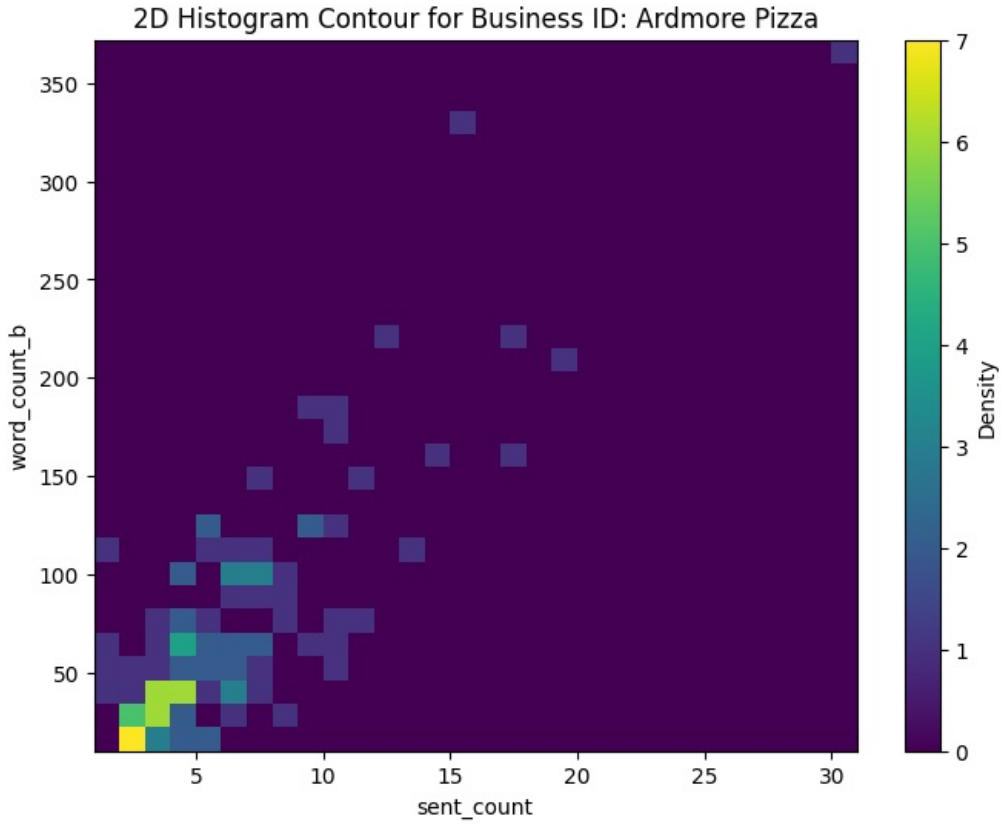

stopword_density
10
9
8
7
6
5
4
3
2
1

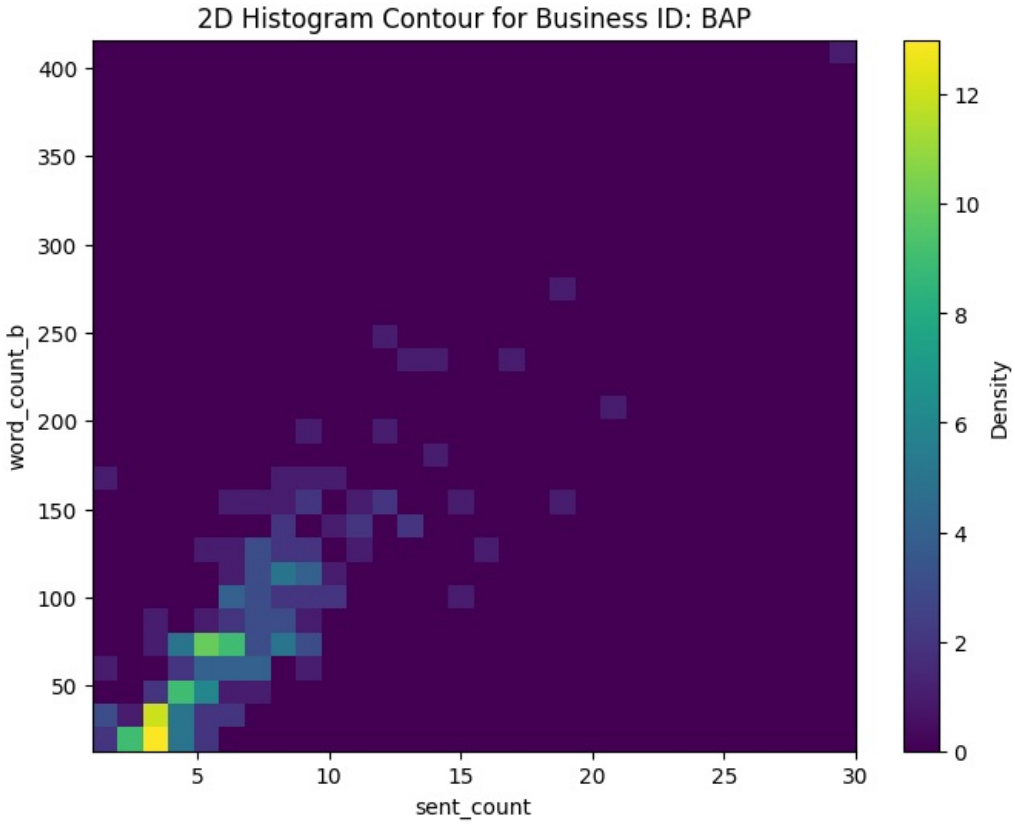
0.0 0.1 0.2 0.3 0.4

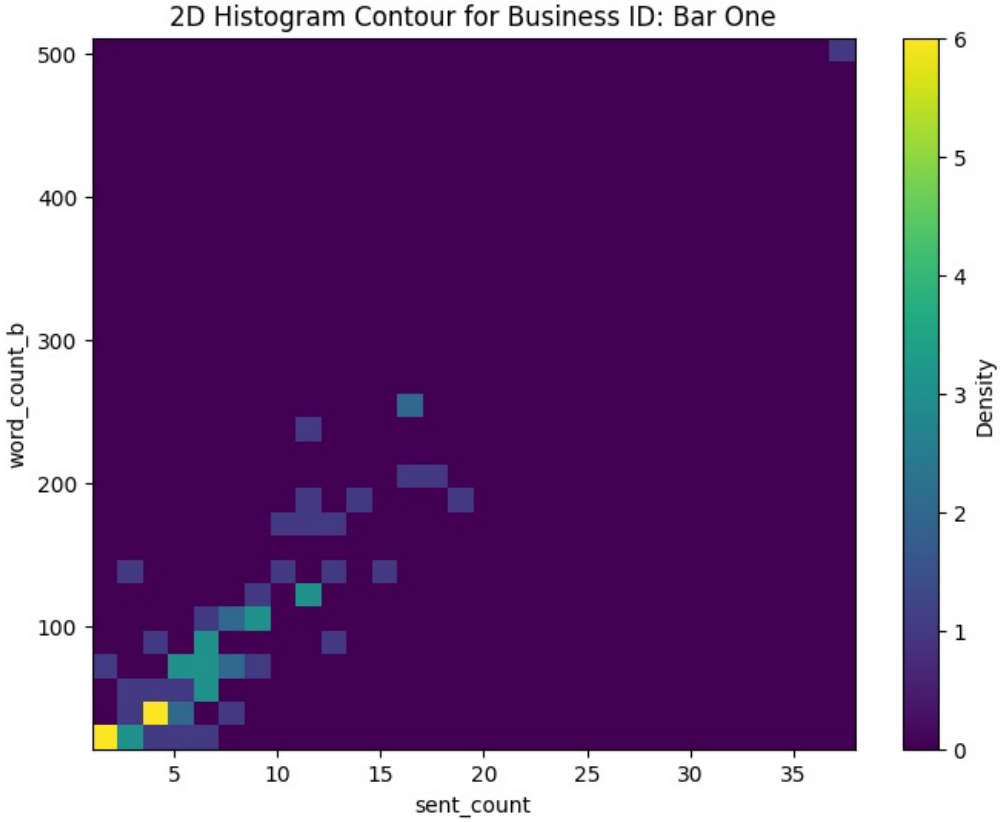
- The above are the box plots for different columns like sentence count, character count, word count after removing the stop words, character count after removing the stop words, stop word count, sentence density, word density, and stop word density.

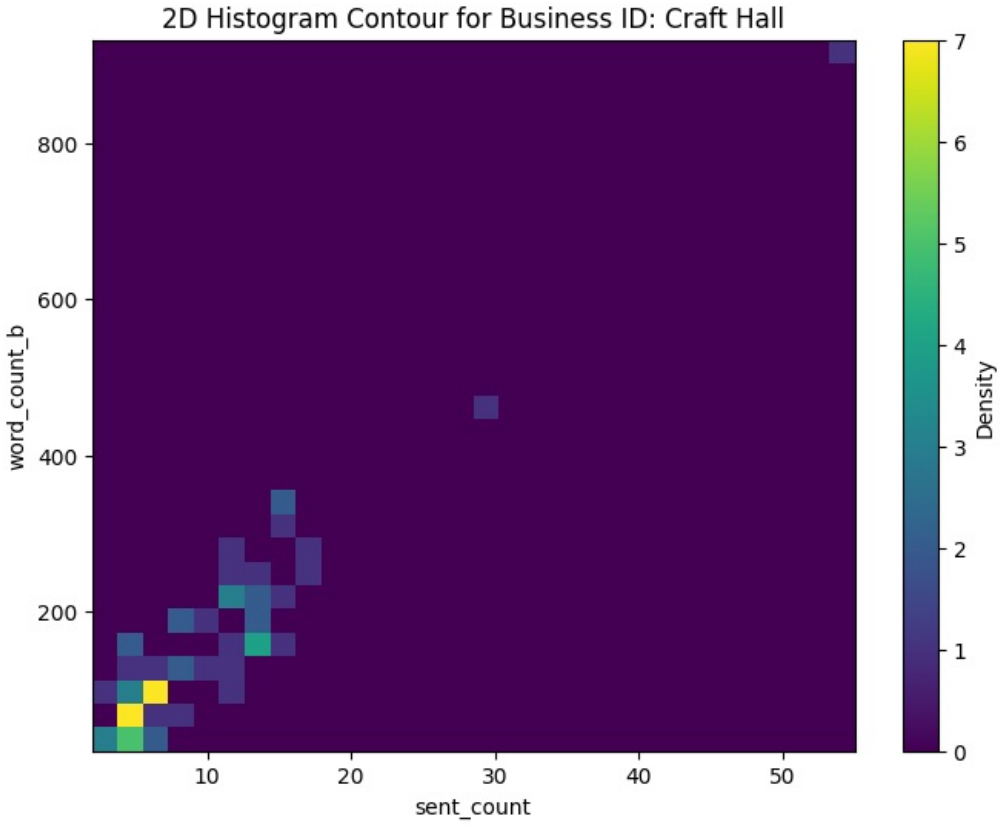
2D Histogram Contour

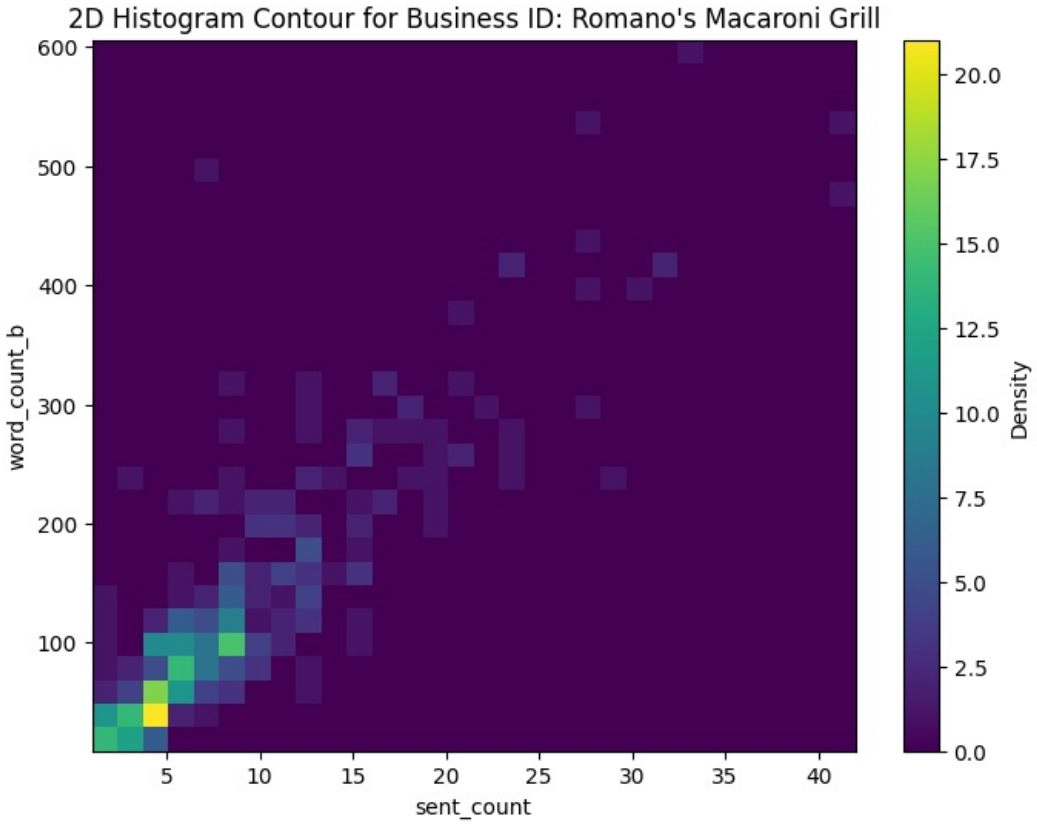
In [35]:
```
x_column = 'sent_count'
y_column = 'word_count_b'

for business_name, group in grouped_df:
    x_data = group[x_column]
    y_data = group[y_column]


    plt.figure(figsize=(8, 6))
    plt.hist2d(x_data, y_data, bins=30, cmap='viridis')
    plt.colorbar(label='Density')
    plt.xlabel(x_column)
    plt.ylabel(y_column)
    plt.title(f'2D Histogram Contour for Business ID: {business_name}')
    plt.show()
```

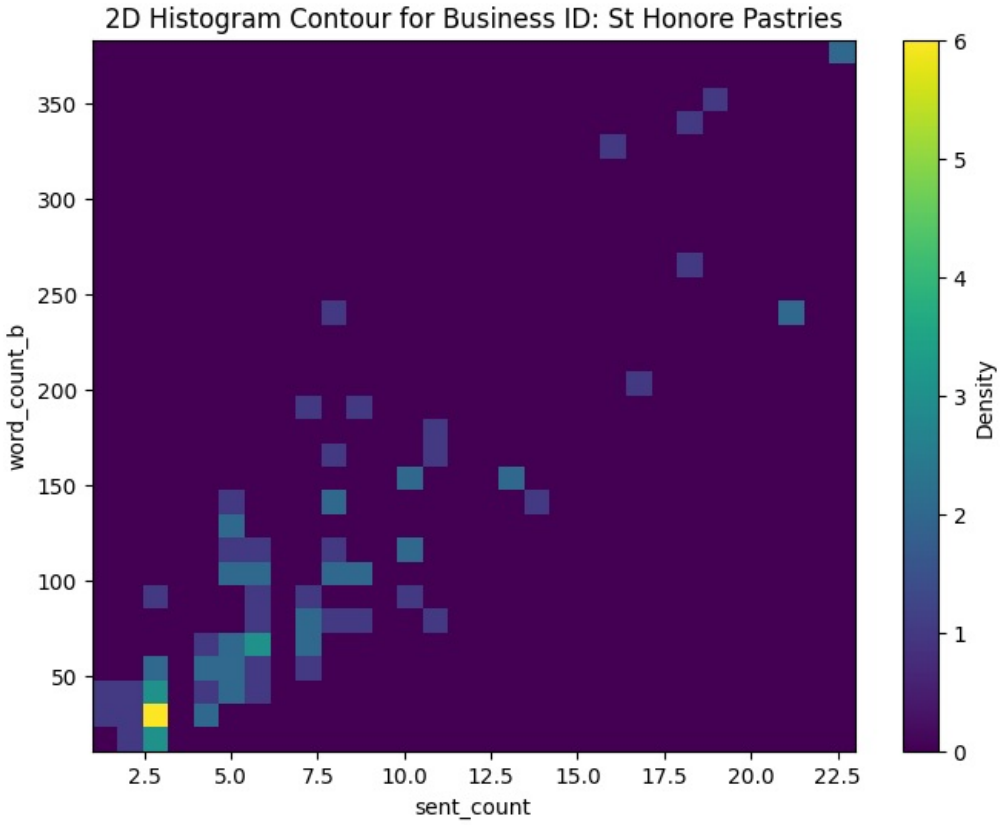


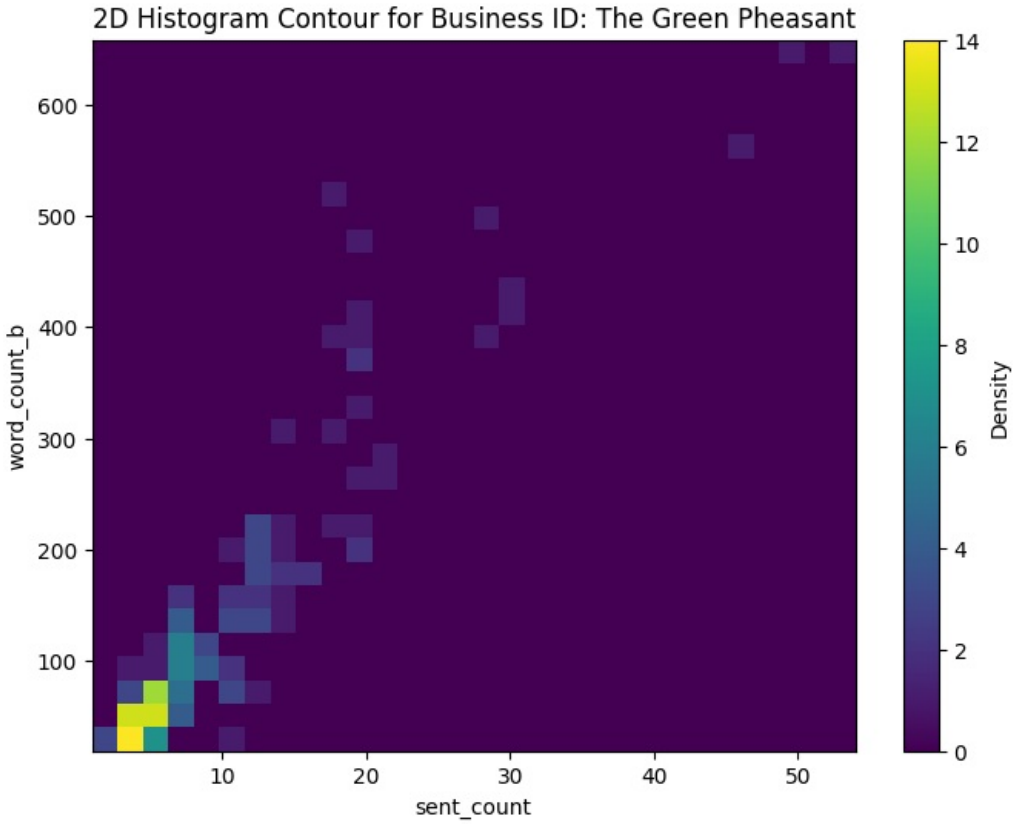

2D Histogram Contour for Business ID: BAP
word_count_b
sent_count
Density
400
350
300
250
200
150
100
50
5
10
15
20
25
30
12
10
8
6
4
2
0

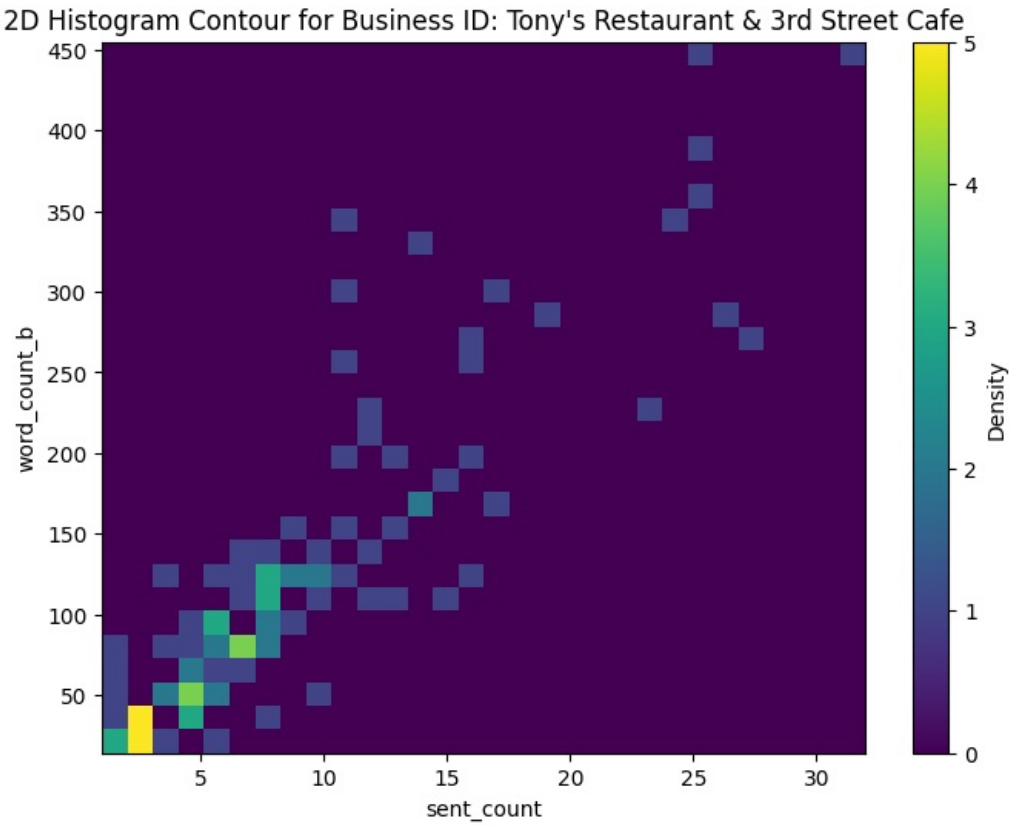

2D Histogram Contour for Business ID: Bar One
word_count_b
sent_count
Density
500
400
300
200
100
5
10
15
20
25
30
35
6
5
4
3
2
1
0

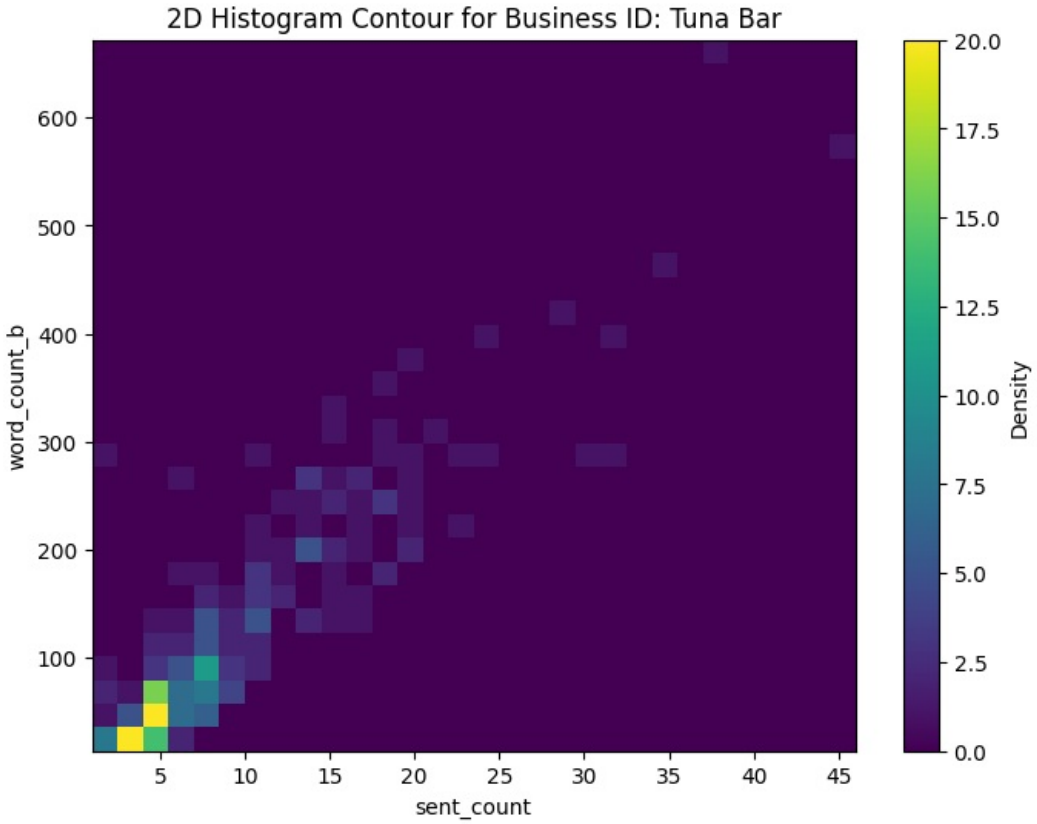

2D Histogram Contour for Business ID: Craft Hall
word_count_b
800
600
400
200
10
20
30
40
50
sent_count
Density
7
6
5
4
3
2
1
0

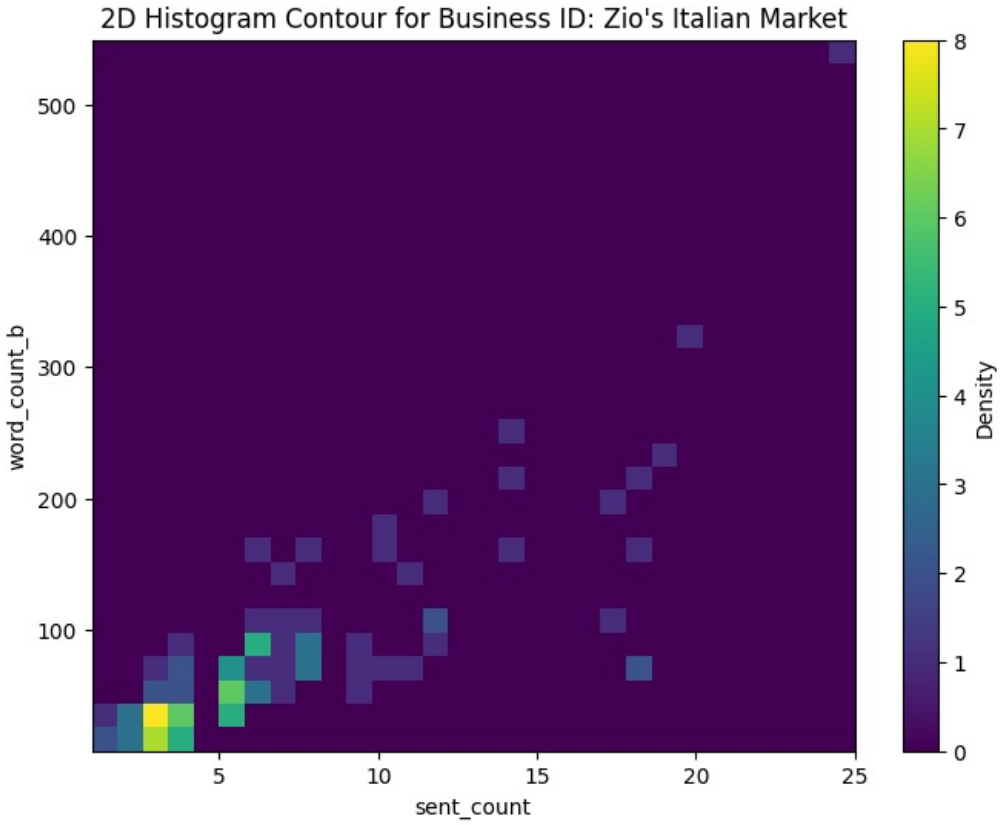

2D Histogram Contour for Business ID: Romano's Macaroni Grill
word_count_b
sent_count
Density
600
500
400
300
200
100
5
10
15
20
25
30
35
40
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


2D Histogram Contour for Business ID: St Honore Pastries
word_count_b
sent_count
Density
350
300
250
200
150
100
50
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
6
5
4
3
2
1
0


2D Histogram Contour for Business ID: The Green Pheasant
word_count_b
sent_count
Density
600
500
400
300
200
100
10
20
30
40
50
14
12
10
8
6
4
2
0


2D Histogram Contour for Business ID: Tony's Restaurant & 3rd Street Cafe
word_count_b
sent_count
Density
450
400
350
300
250
200
150
100
50
5
10
15
20
25
30
5
4
3
2
1
0


2D Histogram Contour for Business ID: Tuna Bar
word_count_b
sent_count
Density
600
500
400
300
200
100
5
10
15
20
25
30
35
40
45
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0



- The above 2D histogram contour plot represents between sentence count and word count before eliminating the stop words for each business.
- This basically over laps the sentence count and word count.
- For example, as you can see in the first graph, it says that the density is high where is 2 sentences which contains word count between 20 to 30.

Implementation

Merge Data

In [36]:
```
data = []

for business_name, group in grouped_df:
    all_reviews = ' '.join(group['text'])
    data.append({'business_name': business_name, 'text': all_reviews})

business_reviews_df = pd.DataFrame(data)
```

```
print(business_reviews_df)
```

```
                        business_name  \
0                        Ardmore Pizza
1                                  BAP
2                             Bar One
3                          Craft Hall
4              Romano's Macaroni Grill
5                   St Honore Pastries
6                   The Green Pheasant
7  Tony's Restaurant & 3rd Street Cafe
8                              Tuna Bar
9                 Zio's Italian Market

                                              text
0  bunch of high school college kids running the ...
1  this place is fantastic   delicious  simple  h...
2  this place is top notch  with phenomenal servi...
3  this is a great place to take guests visiting ...
4  great bar happy hour     every day  wine   dra...
5  this is nice little chinese bakery in the hear...
6  cute ambience that would be great for date nig...
7  we had been driving around for some time  on a...
8  stopped in to check out this new spot around t...
9  the worst chicken parm  sandwich i ve ever eat...
```

- The above code combines all the reviews with the business id in one data frame.

In [37]:
```
tokens = []
for i in business_reviews_df['text']:
  tokens.append(nltk.word_tokenize(i))

business_reviews_df['tokens'] = tokens
business_reviews_df.head()
```

Out[37]:

	business_name	text	tokens
0	Ardmore Pizza	bunch of high school college kids running the ...	[bunch, of, high, school, college, kids, runni...
1	BAP	this place is fantastic delicious simple h...	[this, place, is, fantastic, delicious, simple...
2	Bar One	this place is top notch with phenomenal servi...	[this, place, is, top, notch, with, phenomenal...
3	Craft Hall	this is a great place to take guests visiting ...	[this, is, a, great, place, to, take, guests, ...
4	Romano's Macaroni Grill	great bar happy hour every day wine dra...	[great, bar, happy, hour, every, day, wine, dr...

- The above code generates tokens for the text.
- Each word is considered as a token.

In [38]:
```
for i, tokens in enumerate(business_reviews_df['tokens']):
  for token in tokens:
    if token in stopwords.words('english'):
      business_reviews_df['tokens'][i].remove(token)
business_reviews_df
```

Out[38]:

	business_name	text	tokens
0	Ardmore Pizza	bunch of high school college kids running the ...	[bunch, high, school, college, kids, running, ...
1	BAP	this place is fantastic delicious simple h...	[place, fantastic, delicious, simple, healthy,...
2	Bar One	this place is top notch with phenomenal servi...	[place, top, notch, phenomenal, service, fanta...
3	Craft Hall	this is a great place to take guests visiting ...	[great, place, take, guests, visiting, childre...
4	Romano's Macaroni Grill	great bar happy hour every day wine dra...	[great, bar, happy, hour, every, day, wine, dr...
5	St Honore Pastries	this is nice little chinese bakery in the hear...	[nice, little, chinese, bakery, heart, philade...
6	The Green Pheasant	cute ambience that would be great for date nig...	[cute, ambience, would, great, date, night, ev...
7	Tony's Restaurant & 3rd Street Cafe	we had been driving around for some time on a...	[driving, around, time, weekday, last, week, l...
8	Tuna Bar	stopped in to check out this new spot around t...	[stopped, check, new, spot, around, corner, us...
9	Zio's Italian Market	the worst chicken parm sandwich i ve ever eat...	[worst, chicken, parm, sandwich, ever, eaten, ...

Removing stopwords from the list tokens before fidning the token frequency and other step for text summarization

In [39]:
```
from collections import Counter
token_freq_list = []
for i in business_reviews_df['tokens']:

  token_freq_list.append(Counter(i))

business_reviews_df['token_freq'] = token_freq_list
business_reviews_df.head()
```

Out[39]:

	business_name	text	tokens	token_freq
0	Ardmore Pizza	bunch of high school college kids running the ...	[bunch, high, school, college, kids, running, ...	{'bunch': 2, 'high': 4, 'school': 2, 'college'...
1	BAP	this place is fantastic delicious simple h...	[place, fantastic, delicious, simple, healthy,...	{'place': 146, 'fantastic': 8, 'delicious': 65...
2	Bar One	this place is top notch with phenomenal servi...	[place, top, notch, phenomenal, service, fanta...	{'place': 18, 'top': 6, 'notch': 2, 'phenomena...
3	Craft Hall	this is a great place to take guests visiting ...	[great, place, take, guests, visiting, childre...	{'great': 39, 'place': 65, 'take': 12, 'guests...
4	Romano's Macaroni Grill	great bar happy hour every day wine dra...	[great, bar, happy, hour, every, day, wine, dr...	{'great': 119, 'bar': 22, 'happy': 24, 'hour':...

- The above code creates a new column to store the frequencies of each token.
- As you can see in the above table there is a column named as 'token_freq' which stores the frequency of each token.

In [40]:
```
review_df_grouped = review_df.groupby('business_id')
business_reviews_df['og_text'] = review_df_grouped['text']
for i, row in business_reviews_df.iterrows():
    max_frequency = max(row['token_freq'].values())
    for word in row['token_freq'].keys():
        row['token_freq'][word] = row['token_freq'][word] / max_frequency


tokens = []
for i in business_reviews_df['og_text']:
  all_reviews = ' '.join(i[1])

  tokens.append(nltk.sent_tokenize(all_reviews))


business_reviews_df['sent_token'] = tokens
business_reviews_df = business_reviews_df.drop(columns='og_text')
business_reviews_df.head()
```

Out[40]:

	business_name	text	tokens	token_freq	sent_token
0	Ardmore Pizza	bunch of high school college kids running the ...	[bunch, high, school, college, kids, running, ...	{'bunch': 0.00980392156862745, 'high': 0.01960...	[The worst Chicken Parm., Sandwich I've ever e...
1	BAP	this place is fantastic delicious simple h...	[place, fantastic, delicious, simple, healthy,...	{'place': 0.38320209973753283, 'fantastic': 0....	[Great bar Happy Hour 4-7 every day., Wine & D...
2	Bar One	this place is top notch with phenomenal servi...	[place, top, notch, phenomenal, service, fanta...	{'place': 0.1125, 'top': 0.0375, 'notch': 0.01...	[This is nice little Chinese bakery in the hea...
3	Craft Hall	this is a great place to take guests visiting ...	[great, place, take, guests, visiting, childre...	{'great': 0.15725806451612903, 'place': 0.2620...	[Stopped in to check out this new spot around ...
4	Romano's Macaroni Grill	great bar happy hour every day wine dra...	[great, bar, happy, hour, every, day, wine, dr...	{'great': 0.12395833333333334, 'bar': 0.022916...	[This place is top notch, with phenomenal serv...

- The above code normalizes the token frequencies of each business to their maximum frequency.

In [41]:
```
sentence_scores = {}
sentence_scores_list = []
for i, row in business_reviews_df.iterrows():
  for sent in row['sent_token']:
    for word in sent.split():
      if word.lower() in row['token_freq'].keys():
        if sent not in sentence_scores.keys():
          sentence_scores[sent] = row['token_freq'][word.lower()]
        else:
          sentence_scores[sent] += row['token_freq'][word.lower()]
  sentence_scores_list.append(sentence_scores)
  sentence_scores = {}
business_reviews_df['sent_score'] = sentence_scores_list
business_reviews_df.head()
```

Out[41]:

	business_name	text	tokens	token_freq	sent_token	sent_score
0	Ardmore Pizza	bunch of high school college kids running the ...	[bunch, high, school, college, kids, running, ...	{'bunch': 0.00980392156862745, 'high': 0.01960...	[The worst Chicken Parm., Sandwich I've ever e...	{'The worst Chicken Parm.': 1.053921568627451,...
1	BAP	this place is fantastic delicious simple h...	[place, fantastic, delicious, simple, healthy,...	{'place': 0.38320209973753283, 'fantastic': 0....	[Great bar Happy Hour 4-7 every day., Wine & D...	{'Great bar Happy Hour 4-7 every day.': 0.6089...
2	Bar One	this place is top notch with phenomenal servi...	[place, top, notch, phenomenal, service, fanta...	{'place': 0.1125, 'top': 0.0375, 'notch': 0.01...	[This is nice little Chinese bakery in the hea...	{'This is nice little Chinese bakery in the he...
3	Craft Hall	this is a great place to take guests visiting ...	[great, place, take, guests, visiting, childre...	{'great': 0.15725806451612903, 'place': 0.2620...	[Stopped in to check out this new spot around ...	{'Stopped in to check out this new spot around...
4	Romano's Macaroni Grill	great bar happy hour every day wine dra...	[great, bar, happy, hour, every, day, wine, dr...	{'great': 0.12395833333333334, 'bar': 0.022916...	[This place is top notch, with phenomenal serv...	{'This place is top notch, with phenomenal ser...

- The above code calculates a score for each sentence in the text based on the frequency occured in each business.

In [42]:
```
from heapq import nlargest
summary_list = []
for i, row in business_reviews_df.iterrows():
    max_frequency = max(row['token_freq'].values())
    select_length = int(len(row['sent_token']) * 0.3)
    summary = nlargest(select_length, row['sent_score'], key=row['sent_score'].get)
    final_summary = [re.sub(r'\n', '', word) for word in summary]
    summary = ' '.join(final_summary)
    summary_list.append(summary)

business_reviews_df['summary'] = summary_list
business_reviews_df.head()
```

Out[42]:

	business_name	text	tokens	token_freq	sent_token	sent_score	summary
0	Ardmore Pizza	bunch of high school college kids running the ...	[bunch, high, school, college, kids, running, ...	{'bunch': 0.00980392156862745, 'high': 0.01960...	[The worst Chicken Parm., Sandwich I've ever e...	{'The worst Chicken Parm.': 1.053921568627451,...	And it seems for me three times I have to wait...
1	BAP	this place is fantastic delicious simple h...	[place, fantastic, delicious, simple, healthy,...	{'place': 0.38320209973753283, 'fantastic': 0....	[Great bar Happy Hour 4-7 every day., Wine & D...	{'Great bar Happy Hour 4-7 every day.': 0.6089...	I and my wife's first reno attempt at a restau...
2	Bar One	this place is top notch with phenomenal servi...	[place, top, notch, phenomenal, service, fanta...	{'place': 0.1125, 'top': 0.0375, 'notch': 0.01...	[This is nice little Chinese bakery in the hea...	{'This is nice little Chinese bakery in the he...	And usually, there will be one or two middle a...
3	Craft Hall	this is a great place to take guests visiting ...	[great, place, take, guests, visiting, childre...	{'great': 0.15725806451612903, 'place': 0.2620...	[Stopped in to check out this new spot around ...	{'Stopped in to check out this new spot around...	Vibes: simple, casual but refined, dark wooden...
4	Romano's Macaroni Grill	great bar happy hour every day wine dra...	[great, bar, happy, hour, every, day, wine, dr...	{'great': 0.12395833333333334, 'bar': 0.022916...	[This place is top notch, with phenomenal serv...	{'This place is top notch, with phenomenal ser...	We ordered four things:cheesesteak pretzels: w...

- The above code is used to summarize all the reviews of a particular business.
- The sentences are selected based on the score that is generated in the above step. The top 30 sentences are chosen.
- Then it cleans the extra lines and combines them and stores it in a new column,

In [43]:
```
!pip install vaderSentiment
```

```
Collecting vaderSentiment
  Downloading vaderSentiment-3.3.2-py2.py3-none-any.whl (125 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 126.0/126.0 kB 2.2 MB/s eta 0:00:00
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from vaderSentiment) (2.31.
0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from reques
ts->vaderSentiment) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->vaderSen
timent) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->va
derSentiment) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->va
derSentiment) (2024.2.2)
Installing collected packages: vaderSentiment
Successfully installed vaderSentiment-3.3.2
```

- Install vaderSentiment for sentiment analysis.

In [44]:
```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

def sentiment_scores(sentence):
```

```
    sid_obj = SentimentIntensityAnalyzer()

    sentiment_dict = sid_obj.polarity_scores(sentence)
    return sentiment_dict
```

- The above code calculates the score using VADER sentiment tool of the summarised data.

In [45]:
```
sentiment = []
for i, row in business_reviews_df.iterrows():
    sentence = row['summary']
    score = sentiment_scores(sentence)
    sentiment.append(score['compound'])



business_reviews_df['sentiment'] = sentiment
business_reviews_df.head()
```

Out[45]:

	business_name	text	tokens	token_freq	sent_token	sent_score	summary	sentiment
0	Ardmore Pizza	bunch of high school college kids running the ...	[bunch, high, school, college, kids, running, ...	{'bunch': 0.00980392156862745, 'high': 0.01960...	[The worst Chicken Parm., Sandwich I've ever e...	{'The worst Chicken Parm.': 1.053921568627451,...	And it seems for me three times I have to wait...	1.0
1	BAP	this place is fantastic delicious simple h...	[place, fantastic, delicious, simple, healthy,...	{'place': 0.38320209973753283, 'fantastic': 0....	[Great bar Happy Hour 4-7 every day., Wine & D...	{'Great bar Happy Hour 4-7 every day.': 0.6089...	I and my wife's first reno attempt at a restau...	1.0
2	Bar One	this place is top notch with phenomenal servi...	[place, top, notch, phenomenal, service, fanta...	{'place': 0.1125, 'top': 0.0375, 'notch': 0.01...	[This is nice little Chinese bakery in the hea...	{'This is nice little Chinese bakery in the he...	And usually, there will be one or two middle a...	1.0
3	Craft Hall	this is a great place to take guests visiting ...	[great, place, take, guests, visiting, childre...	{'great': 0.15725806451612903, 'place': 0.2620...	[Stopped in to check out this new spot around ...	{'Stopped in to check out this new spot around...	Vibes: simple, casual but refined, dark wooden...	1.0
4	Romano's Macaroni Grill	great bar happy hour every day wine dra...	[great, bar, happy, hour, every, day, wine, dr...	{'great': 0.12395833333333334, 'bar': 0.022916...	[This place is top notch, with phenomenal serv...	{'This place is top notch, with phenomenal ser...	We ordered four things:cheesesteak pretzels: w...	1.0

- The sentiment score is calculated and stored in the new column.
- In the above table you can see the sentiment column contains the score.

Overall text sentiment

Performance Test

In [46]:
```
sentiment = []
for i, row in business_reviews_df.iterrows():
    sentence = row['text']
    score = sentiment_scores(sentence)
    sentiment.append(score['compound'])


sentiment
#business_reviews_df['sentiment'] = sentiment
#business_reviews_df.head()
```

Out[46]:
```
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]
```

- The above is the performance check, a sentiment of the review before summarization.
- All the reviews are positive and after summarization also all the reviews are positive which means the model is working fine.