bow-with-basic-features

September 30, 2023

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
data = pd.read_csv('train.csv')
```

[3]:
```
data.shape
```

[3]:
```
(404290, 6)
```

[4]:
```
data.head()
```

[4]:

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh…	What is the step by step guide to invest in sh…	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia…	What would happen if the Indian government sto…	0
2	2	5	6	How can I increase the speed of my internet co…	How can Internet speed be increased by hacking…	0
3	3	7	8	Why am I mentally very lonely? How can I solve…	Find the remainder when [math]23^{24}[/math] i…	0
4	4	9	10	Which one dissolve in water quikly sugar, salt…	Which fish would survive in salt water?	0

[5]:
```
new_data = data.sample(30000,random_state=2)
```

[6]:
```
new_data.isnull().sum()
```

[6]:
```
id              0
qid1            0
qid2            0
question1       0
question2       0
```

```
is_duplicate    0
dtype: int64
```

[7]:
```
new_data.head()
```

[7]:
```
            id    qid1    qid2  \
398782  398782  496695  532029
115086  115086  187729  187730
327711  327711  454161  454162
367788  367788  498109  491396
151235  151235  237843   50930

                                                 question1  \
398782  What is the best marketing automation tool for…
115086  I am poor but I want to invest. What should I do?
327711  I am from India and live abroad. I met a guy f…
367788  Why do so many people in the U.S. hate the sou…
151235                Consequences of Bhopal gas tragedy?

                                                 question2  is_duplicate
398782  What is the best marketing automation tool for…              1
115086  I am quite poor and I want to be very rich. Wh…              0
327711  T.I.E.T to Thapar University to Thapar Univers…              0
367788  My boyfriend doesnt feel guilty when he hurts …              0
151235  What was the reason behind the Bhopal gas trag…              0
```

[8]:
```
new_data.isnull().sum()
```

[8]:
```
id              0
qid1            0
qid2            0
question1       0
question2       0
is_duplicate    0
dtype: int64
```

[9]:
```
new_data.duplicated().sum()
```

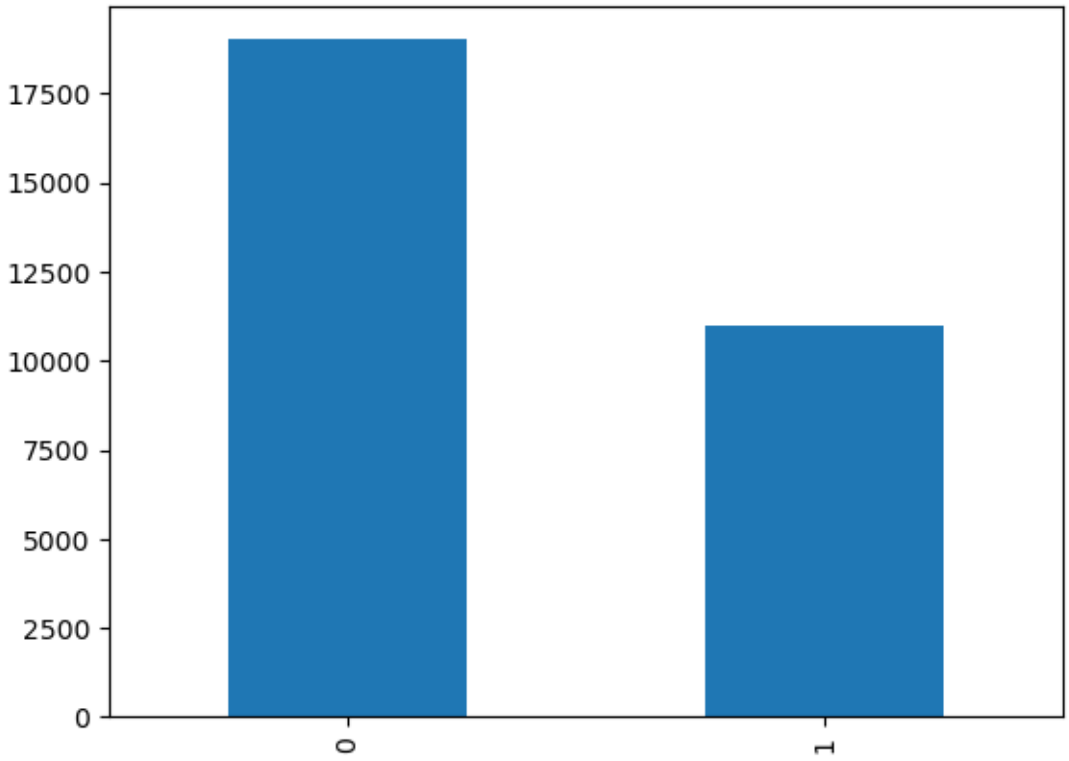
[9]:
```
0
```

[10]:
```
# Distribution of duplicate and non-duplicate questions
print(new_data['is_duplicate'].value_counts())
print((new_data['is_duplicate'].value_counts()/new_data['is_duplicate'].
  ↪count())*100)
new_data['is_duplicate'].value_counts().plot(kind = 'bar')
```

```
0    19013
1    10987
```

```
Name: is_duplicate, dtype: int64
0    63.376667
1    36.623333
Name: is_duplicate, dtype: float64
```

[10]: <Axes: >

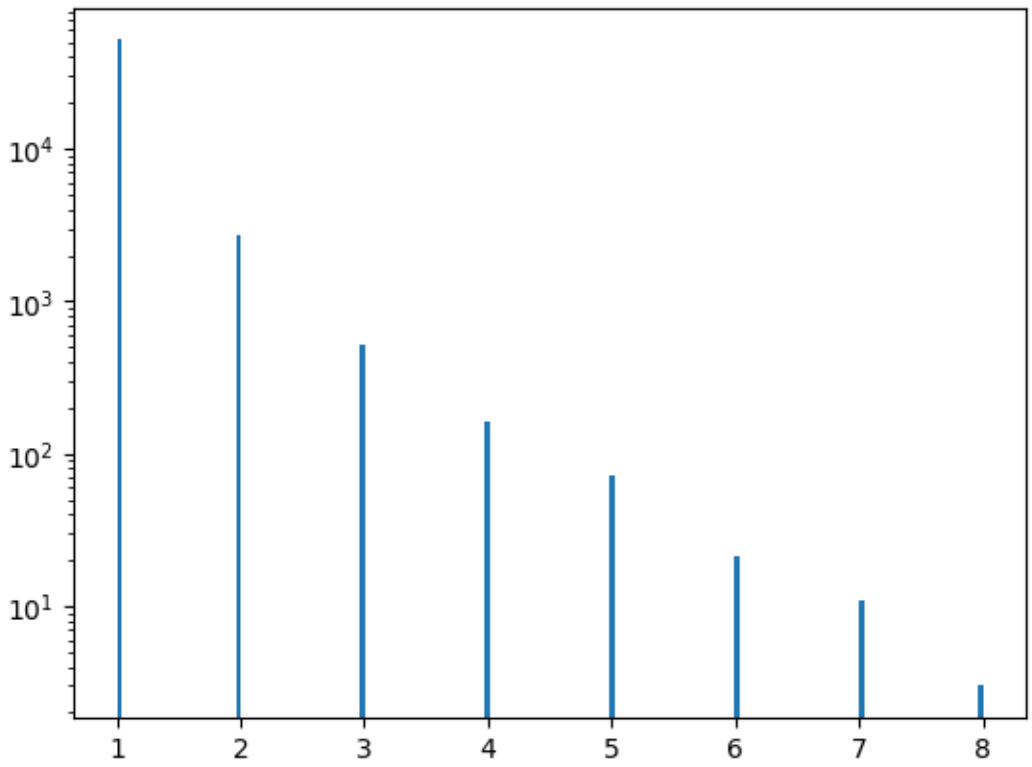


[11]:
```
# Repated quetions
qid = pd.Series(new_data['qid1'].tolist() + new_data['qid2'].tolist())
print("Number odf unique quesions",np.unique(qid).shape[0])
x =qid.value_counts()>1
print('Number of quetions getting repeated',x[x].shape[0])
```

```
Number odf unique quesions 55299
Number of quetions getting repeated 3480
```

[ ]:

[12]:
```
# Repeated questions histogram
plt.hist(qid.value_counts().values,bins=160)
plt.yscale('log')
plt.show()
```



[13]:
```
# F-eature Engineering
new_data['q1_len'] = new_data['question1'].str.len()
new_data['q2_len'] = new_data['question1'].str.len()
```

[14]:
```
new_data.head()
```

[14]:

	id	qid1	qid2	question1	question2	is_duplicate
398782	398782	496695	532029	What is the best marketing automation tool for…	What is the best marketing automation tool for…	1
115086	115086	187729	187730	I am poor but I want to invest. What should I do?		
327711	327711	454161	454162	I am from India and live abroad. I met a guy f…		
367788	367788	498109	491396	Why do so many people in the U.S. hate the sou…		
151235	151235	237843	50930	Consequences of Bhopal gas tragedy?		

```
115086  I am quite poor and I want to be very rich. Wh…             0
327711  T.I.E.T to Thapar University to Thapar Univers…             0
367788  My boyfriend doesnt feel guilty when he hurts …             0
151235  What was the reason behind the Bhopal gas trag…             0

        q1_len  q2_len
398782      76      76
115086      49      49
327711     105     105
367788      59      59
151235      35      35
```

[15]:
```
new_data['q1_num_words'] = new_data['question1'].apply(lambda row: len(row.
 ↪split(" ")))
new_data['q2_num_words'] = new_data['question2'].apply(lambda row: len(row.
 ↪split(" ")))
```

[16]:
```
new_data
```

[16]:
```
            id    qid1    qid2  \
398782  398782  496695  532029
115086  115086  187729  187730
327711  327711  454161  454162
367788  367788  498109  491396
151235  151235  237843   50930
…          …       …       …
243932  243932   26193  356455
91980    91980  154063  154064
266955  266955  133017  384210
71112    71112  122427  122428
312470  312470  436915  436916

                                                 question1  \
398782  What is the best marketing automation tool for…
115086  I am poor but I want to invest. What should I do?
327711  I am from India and live abroad. I met a guy f…
367788  Why do so many people in the U.S. hate the sou…
151235                Consequences of Bhopal gas tragedy?
…                                                      …
243932         What are some good web scraping tutorials?
91980   Can I apply for internet banking in SBI withou…
266955  How much HE laundry detergent do you use in a …
71112   What is the best way to understand and learn m…
312470  What would the Modi-led government do in case …

                                                 question2  is_duplicate  \
398782  What is the best marketing automation tool for…              1
```

```
115086  I am quite poor and I want to be very rich. Wh…             0
327711  T.I.E.T to Thapar University to Thapar Univers…             0
367788  My boyfriend doesnt feel guilty when he hurts …             0
151235  What was the reason behind the Bhopal gas trag…             0
…                                                    …           …
243932          What are some good web scraping programs?             1
91980   I have internet banking kit of SBI but it's no…             0
266955  Can I use regular Dawn dishsoap in my dishwash…             0
71112       What are some of the best ways to learn math?             1
312470  If Pakistan mounts a 26/11 type attack again, …             1

        q1_len  q2_len  q1_num_words  q2_num_words
398782      76      76            12            12
115086      49      49            12            15
327711     105     105            25            17
367788      59      59            12            30
151235      35      35             5             9
…          …       …             …             …
243932      42      42             7             7
91980       68      68            12            12
266955      73      73            14            17
71112       51      51            10            10
312470      87      87            15            14

[30000 rows x 10 columns]
```

[17]:
```
new_data.head()
```

[17]:
```
            id    qid1    qid2  \
398782  398782  496695  532029
115086  115086  187729  187730
327711  327711  454161  454162
367788  367788  498109  491396
151235  151235  237843   50930

                                                question1  \
398782  What is the best marketing automation tool for…
115086  I am poor but I want to invest. What should I do?
327711  I am from India and live abroad. I met a guy f…
367788  Why do so many people in the U.S. hate the sou…
151235                Consequences of Bhopal gas tragedy?

                                                question2  is_duplicate  \
398782  What is the best marketing automation tool for…             1
115086  I am quite poor and I want to be very rich. Wh…             0
327711  T.I.E.T to Thapar University to Thapar Univers…             0
367788  My boyfriend doesnt feel guilty when he hurts …             0
```

```
151235  What was the reason behind the Bhopal gas trag…             0

        q1_len  q2_len  q1_num_words  q2_num_words
398782      76      76            12            12
115086      49      49            12            15
327711     105     105            25            17
367788      59      59            12            30
151235      35      35             5             9
```

[ ]:

[18]:
```
def common_words(row):
    w1=set(map(lambda word: word.lower().strip(),row['question1'].split(" ")))
    w2=set(map(lambda word: word.lower().strip(),row['question2'].split(" ")))
    return len(w1 & w2)
```

[19]:
```
new_data['word_common'] = new_data.apply(common_words, axis=1)
new_data.head()
```

[19]:
```
            id    qid1    qid2  \
398782  398782  496695  532029
115086  115086  187729  187730
327711  327711  454161  454162
367788  367788  498109  491396
151235  151235  237843   50930

                                                question1  \
398782  What is the best marketing automation tool for…
115086  I am poor but I want to invest. What should I do?
327711  I am from India and live abroad. I met a guy f…
367788  Why do so many people in the U.S. hate the sou…
151235                Consequences of Bhopal gas tragedy?

                                                question2  is_duplicate  \
398782  What is the best marketing automation tool for…              1
115086  I am quite poor and I want to be very rich. Wh…              0
327711  T.I.E.T to Thapar University to Thapar Univers…              0
367788  My boyfriend doesnt feel guilty when he hurts …              0
151235  What was the reason behind the Bhopal gas trag…              0

        q1_len  q2_len  q1_num_words  q2_num_words  word_common
398782      76      76            12            12           11
115086      49      49            12            15            7
327711     105     105            25            17            2
367788      59      59            12            30            0
151235      35      35             5             9            3
```

[20]:
```
def total_words(row):
    w1=set(map(lambda word: word.lower().strip(),row['question1'].split(" ")))
    w2=set(map(lambda word: word.lower().strip(),row['question2'].split(" ")))
    return (len(w1) + len(w2))
```

[21]:
```
new_data['word_total'] = new_data.apply(total_words, axis=1)
new_data.head()
```

[21]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_total
398782	398782	496695	532029	What is the best marketing automation tool for…	What is the best marketing automation tool for…	1	76	76	12	12	11	24
115086	115086	187729	187730	I am poor but I want to invest. What should I do?	I am quite poor and I want to be very rich. Wh…	0	49	49	12	15	7	23
327711	327711	454161	454162	I am from India and live abroad. I met a guy f…	T.I.E.T to Thapar University to Thapar Univers…	0	105	105	25	17	2	34
367788	367788	498109	491396	Why do so many people in the U.S. hate the sou…	My boyfriend doesnt feel guilty when he hurts …	0	59	59	12	30	0	32
151235	151235	237843	50930	Consequences of Bhopal gas tragedy?	What was the reason behind the Bhopal gas trag…	0	35	35	5	9	3	13

[22]:
```
new_data['word_share']= round(new_data['word_common']/new_data['word_total'],2)
new_data.head()
```

[22]:

	id	qid1	qid2	question1	\
398782	398782	496695	532029		
115086	115086	187729	187730		
327711	327711	454161	454162		
367788	367788	498109	491396		
151235	151235	237843	50930		

```
398782  What is the best marketing automation tool for…
115086  I am poor but I want to invest. What should I do?
327711  I am from India and live abroad. I met a guy f…
367788  Why do so many people in the U.S. hate the sou…
151235                Consequences of Bhopal gas tragedy?

                                                 question2  is_duplicate  \
398782  What is the best marketing automation tool for…                1
115086  I am quite poor and I want to be very rich. Wh…                0
327711  T.I.E.T to Thapar University to Thapar Univers…                0
367788  My boyfriend doesnt feel guilty when he hurts …                0
151235  What was the reason behind the Bhopal gas trag…                0

        q1_len  q2_len  q1_num_words  q2_num_words  word_common  word_total  \
398782      76      76            12            12           11          24
115086      49      49            12            15            7          23
327711     105     105            25            17            2          34
367788      59      59            12            30            0          32
151235      35      35             5             9            3          13

        word_share
398782        0.46
115086        0.30
327711        0.06
367788        0.00
151235        0.23
```

[23]:
```
new_data.shape
```

[23]:
```
(30000, 13)
```

[ ]:
```

```

[24]:
```
# Analysis of features
#Calculate the minimum and maximum values from the "q1_len" column
min_q1_len = new_data['q1_len'].min()
max_q1_len = new_data['q1_len'].max()
```

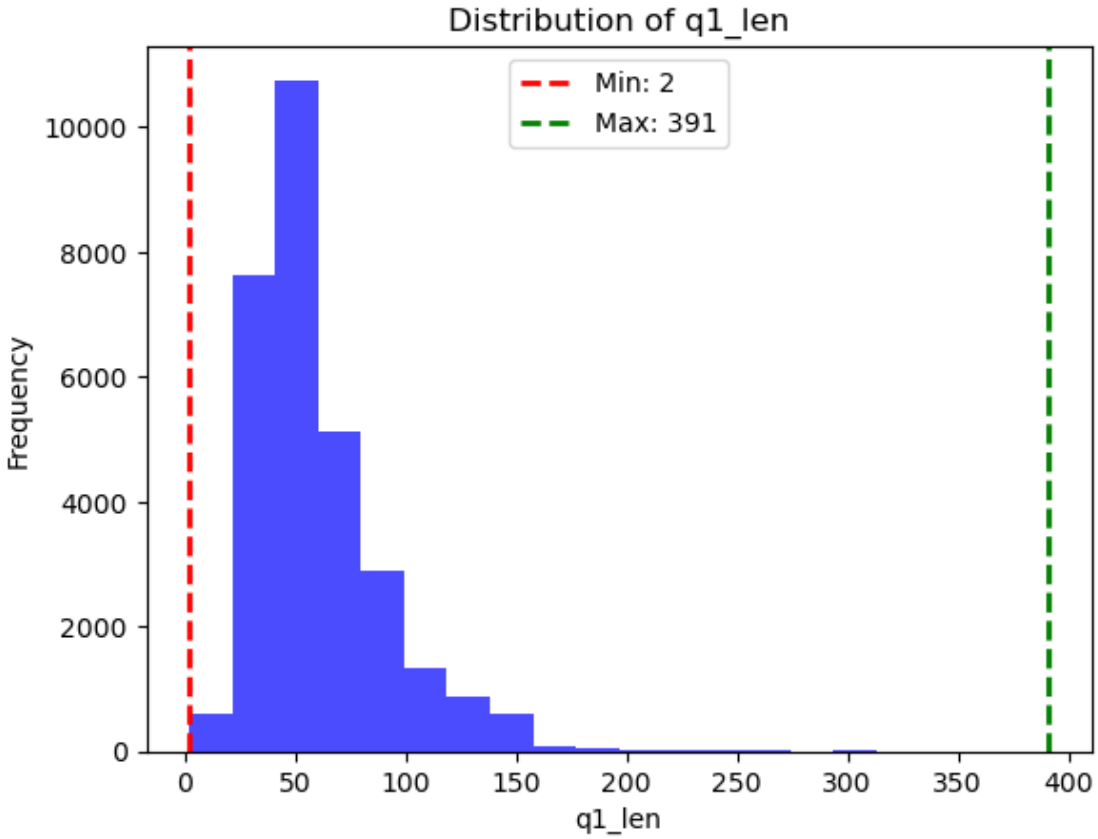
[25]:
```
print(f"Minimum q1_len: {min_q1_len}")
print(f"Maximum q1_len: {max_q1_len}")
```

```
Minimum q1_len: 2
Maximum q1_len: 391
```

[26]:
```
#calculate average
print('average num of characters',int(new_data['q1_len'].mean()))
```

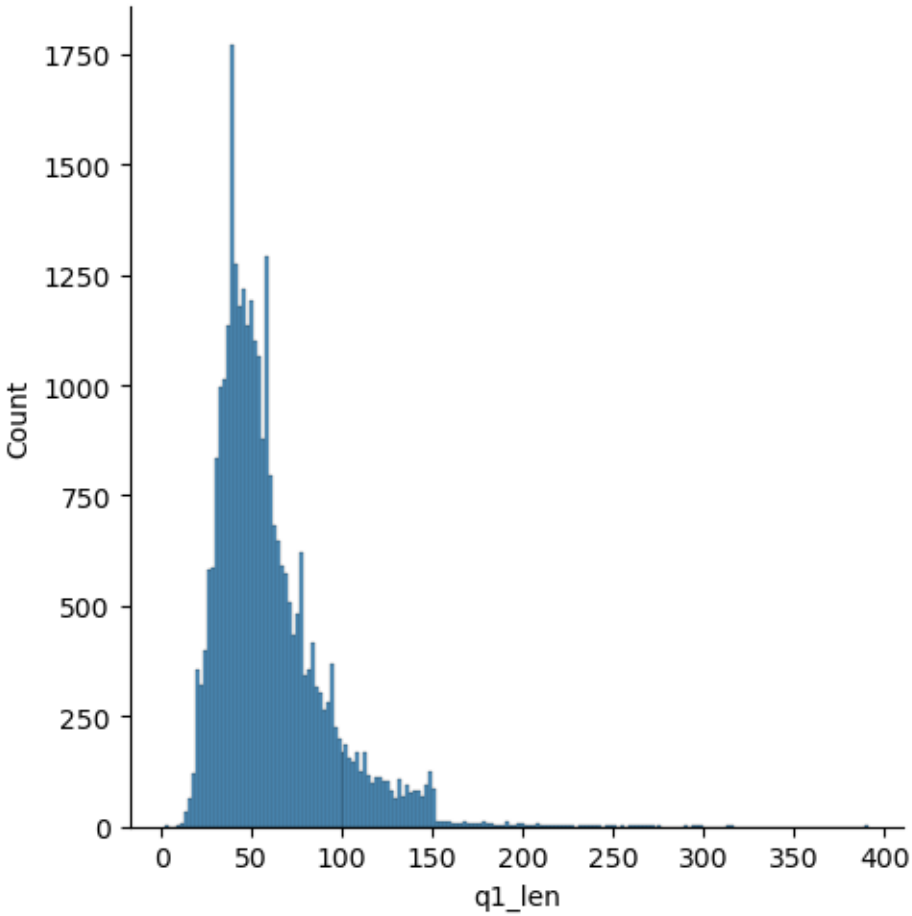
```
average num of characters 59
```

[27]:
```
# Plotting a histogram of q1_len column
plt.hist(new_data['q1_len'], bins=20, color='blue', alpha=0.7)
plt.xlabel('q1_len')
plt.ylabel('Frequency')
plt.title('Distribution of q1_len')
plt.axvline(min_q1_len, color='red', linestyle='dashed', linewidth=2,␣
 ↪label=f'Min: {min_q1_len}')
plt.axvline(max_q1_len, color='green', linestyle='dashed', linewidth=2,␣
 ↪label=f'Max: {max_q1_len}')
plt.legend()
plt.show()
```



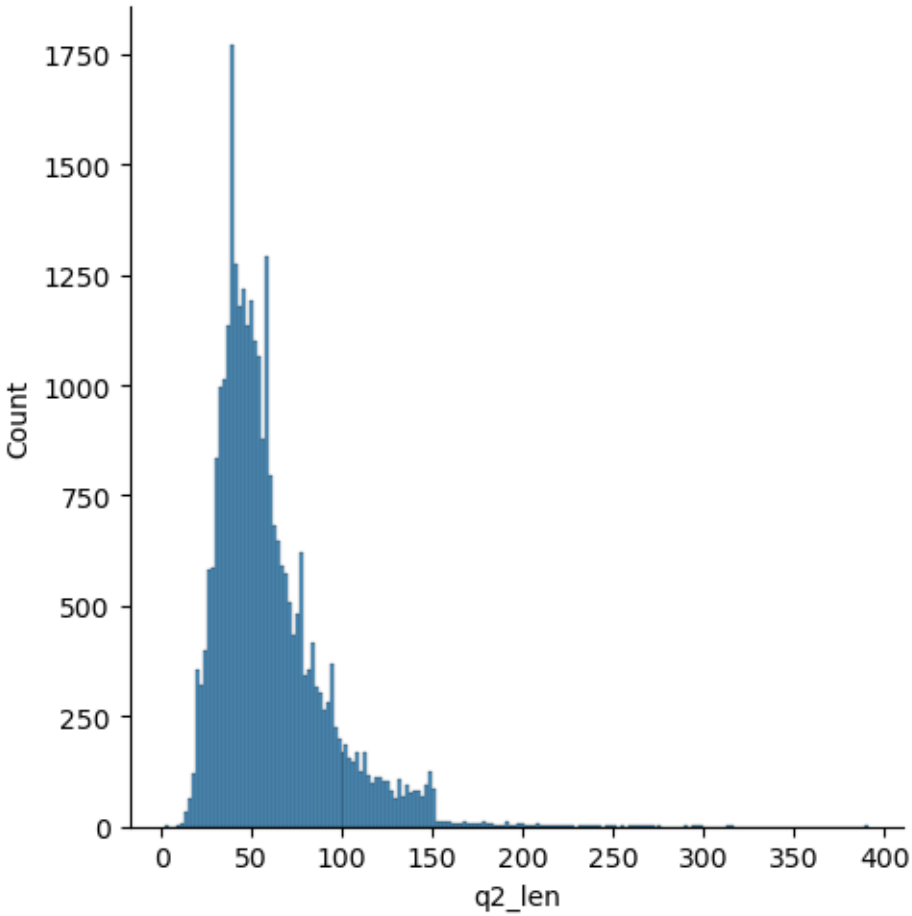
[28]:
```
sns.displot(new_data['q1_len'])
```

[28]:
```
<seaborn.axisgrid.FacetGrid at 0x205ab673c90>
```



[29]:
```
sns.displot(new_data['q2_len'])
print('minimum characters',new_data['q2_len'].min())
print('maximum characters',new_data['q2_len'].max())
print('average num of characters',int(new_data['q2_len'].mean()))
```

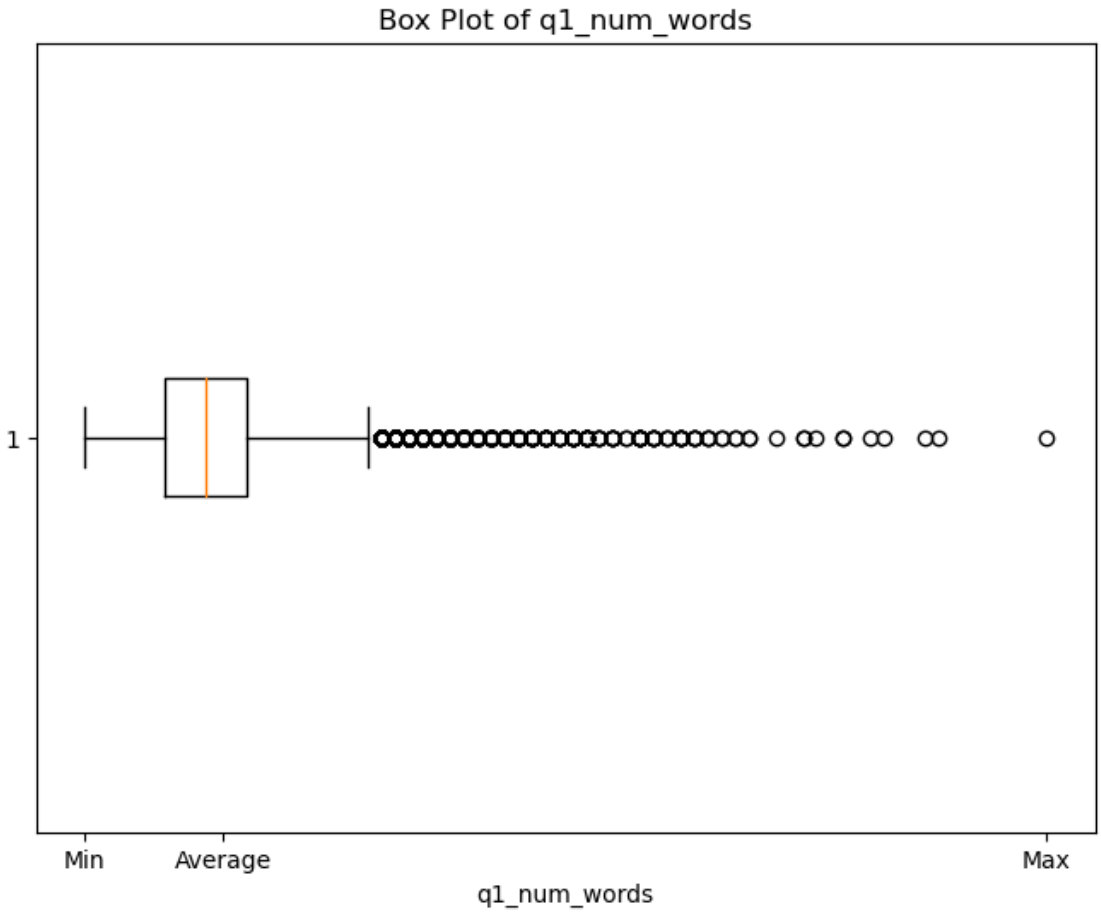
```
minimum characters 2
maximum characters 391
average num of characters 59
```



[30]:
```
min_q1_num_words = new_data['q1_num_words'].min()
max_q1_num_words = new_data['q1_num_words'].max()
average_q1_num_words = new_data['q2_num_words'].mean()
print(f"Minimum q1_num_words: {min_q1_num_words}")
print(f"Maximum q1_num_words: {max_q1_num_words}")
print(f"Average q1_num_words: {average_q1_num_words}")
```

```
Minimum q1_num_words: 1
Maximum q1_num_words: 72
Average q1_num_words: 11.232133333333334
```

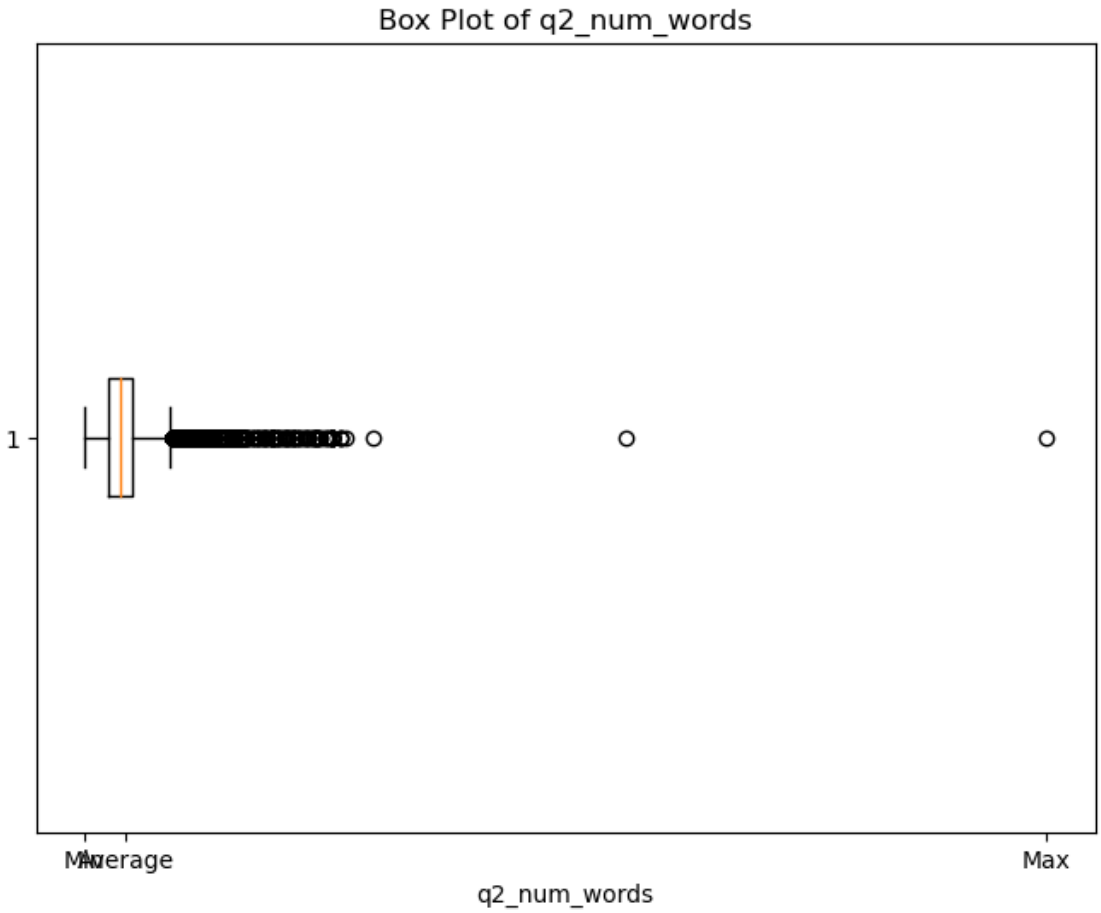
[31]:
```
plt.figure(figsize=(8, 6))
plt.boxplot(new_data['q1_num_words'], vert=False)
plt.xlabel('q1_num_words')
plt.title('Box Plot of q1_num_words')
plt.xticks([min_q1_num_words, max_q1_num_words, average_q1_num_words], ['Min',␣
↪'Max', 'Average'])
plt.show()
```



[32]:
```
min_q2_num_words = new_data['q2_num_words'].min()
max_q2_num_words = new_data['q2_num_words'].max()
average_q2_num_words = new_data['q2_num_words'].mean()
print(f"Minimum q2_num_words: {min_q2_num_words}")
print(f"Maximum q2_num_words: {max_q2_num_words}")
print(f"Average q2_num_words: {average_q2_num_words}")
```

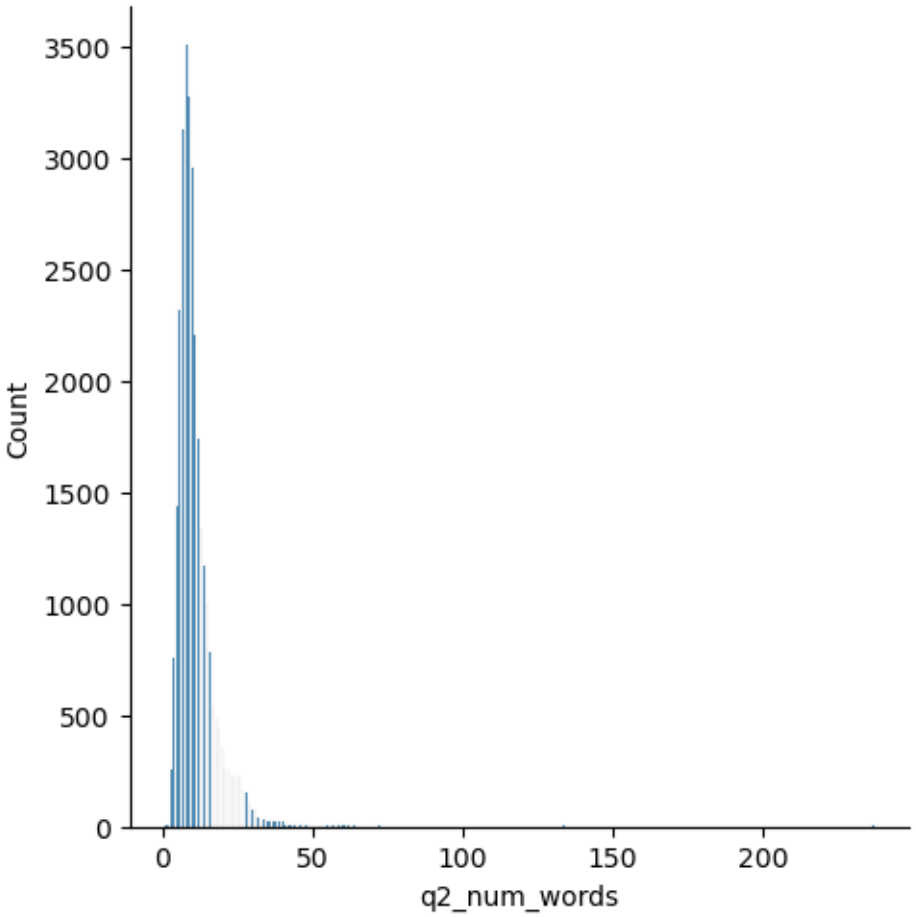
```
Minimum q2_num_words: 1
Maximum q2_num_words: 237
Average q2_num_words: 11.232133333333334
```

[33]:
```
plt.figure(figsize=(8, 6))
plt.boxplot(new_data['q2_num_words'], vert=False)
plt.xlabel('q2_num_words')
plt.title('Box Plot of q2_num_words')
plt.xticks([min_q2_num_words, max_q2_num_words, average_q2_num_words], ['Min',␣
↪'Max', 'Average'])
plt.show()
```



[34]:
```
sns.displot(new_data['q2_num_words'])
```

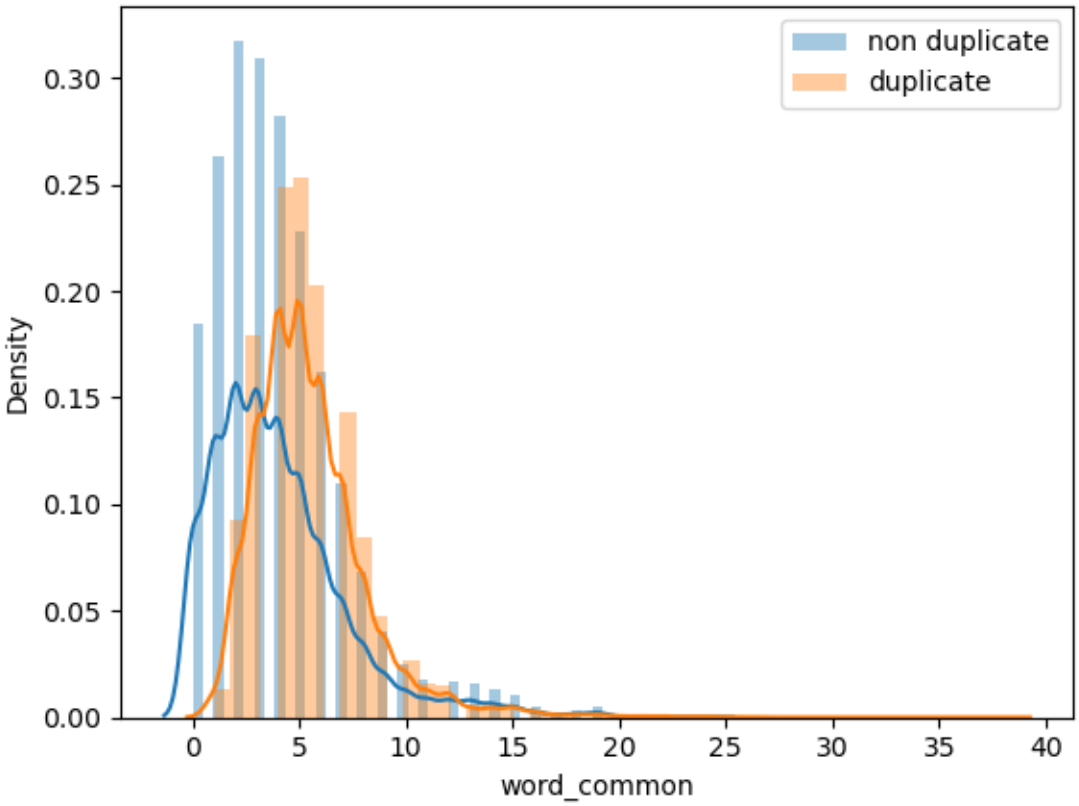
[34]: <seaborn.axisgrid.FacetGrid at 0x205ae5b9810>



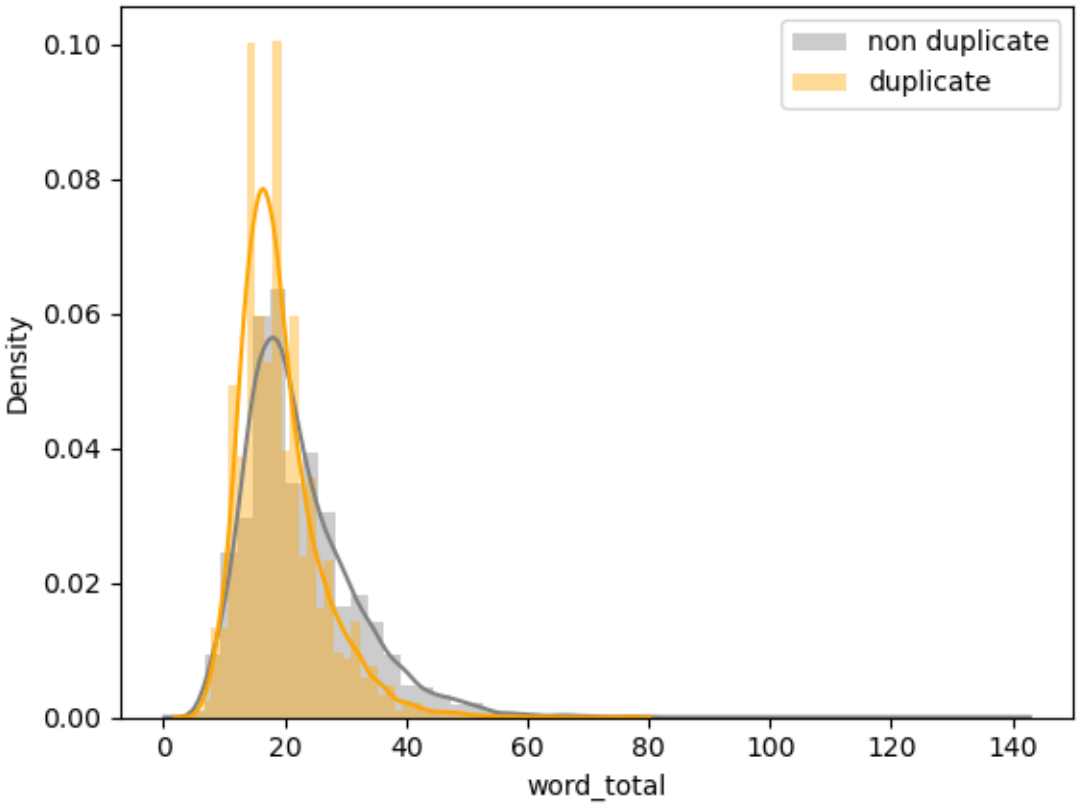
[35]:
```
print(new_data.columns)
```

```
Index(['id', 'qid1', 'qid2', 'question1', 'question2', 'is_duplicate',
       'q1_len', 'q2_len', 'q1_num_words', 'q2_num_words', 'word_common',
       'word_total', 'word_share'],
      dtype='object')
```

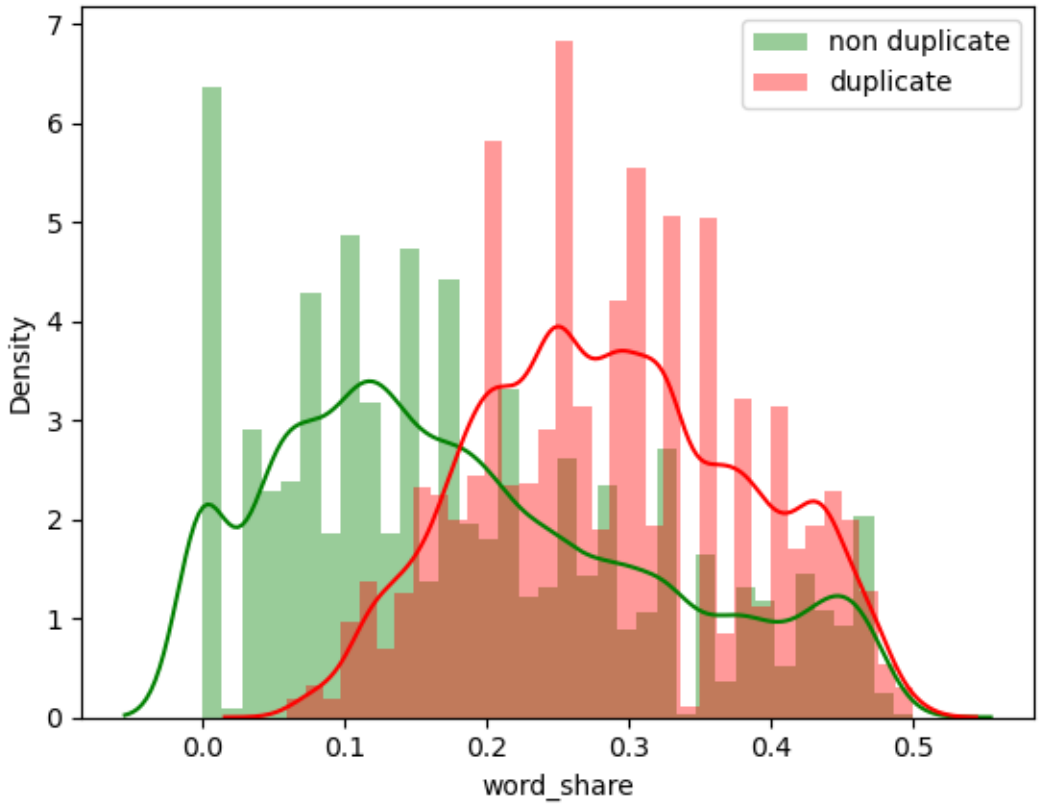
[36]:
```
sns.distplot(new_data[new_data['is_duplicate'] == 0]['word_common'],label='non␣
 ↪duplicate')
sns.distplot(new_data[new_data['is_duplicate'] ==␣
 ↪1]['word_common'],label='duplicate')
plt.legend()
plt.show()
```



[37]:
```
sns.distplot(new_data[new_data['is_duplicate'] == 0]['word_total'],label='non␣
 ↪duplicate',color='gray')
sns.distplot(new_data[new_data['is_duplicate'] ==␣
 ↪1]['word_total'],label='duplicate',color='orange')
plt.legend()
plt.show()
```



[38]:
```
# word share
sns.distplot(new_data[new_data['is_duplicate'] == 0]['word_share'],label='non␣
 ↪duplicate',color='green')
sns.distplot(new_data[new_data['is_duplicate'] ==␣
 ↪1]['word_share'],label='duplicate',color='red')
plt.legend()
plt.show()
```



[39]:
```
question_data = new_data[['question1','question2']]
```

[40]:
```
question_data
```

[40]:
```
                                                question1  \
398782  What is the best marketing automation tool for…
115086  I am poor but I want to invest. What should I do?
327711  I am from India and live abroad. I met a guy f…
367788  Why do so many people in the U.S. hate the sou…
151235                Consequences of Bhopal gas tragedy?
…                                                     …
243932         What are some good web scraping tutorials?
91980   Can I apply for internet banking in SBI withou…
266955  How much HE laundry detergent do you use in a …
71112   What is the best way to understand and learn m…
312470  What would the Modi-led government do in case …

                                                question2
398782  What is the best marketing automation tool for…
115086  I am quite poor and I want to be very rich. Wh…
327711  T.I.E.T to Thapar University to Thapar Univers…
```

```
367788  My boyfriend doesnt feel guilty when he hurts …
151235  What was the reason behind the Bhopal gas trag…
…                                                    …
243932          What are some good web scraping programs?
91980   I have internet banking kit of SBI but it's no…
266955  Can I use regular Dawn dishsoap in my dishwash…
71112       What are some of the best ways to learn math?
312470  If Pakistan mounts a 26/11 type attack again, …

[30000 rows x 2 columns]
```

[41]:
```
final_data=new_data.drop(columns=['id','qid1','qid2','question1','question2'])
print(final_data.shape)
final_data.head()
```

```
(30000, 8)
```

[41]:

	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_total	word_share
398782	1	76	76	12	12	11	24	0.46
115086	0	49	49	12	15	7	23	0.30
327711	0	105	105	25	17	2	34	0.06
367788	0	59	59	12	30	0	32	0.00
151235	0	35	35	5	9	3	13	0.23

[42]:
```
from sklearn.feature_extraction.text import CountVectorizer
# merge texts
questions = list(question_data['question1']) + list(question_data['question2'])

cv = CountVectorizer(max_features=3000)
q1_arr, q2_arr = np.vsplit(cv.fit_transform(questions).toarray(),2)
```

[43]:
```
temp_df1 = pd.DataFrame(q1_arr, index= question_data.index)
temp_df2 = pd.DataFrame(q2_arr, index= question_data.index)
temp_data = pd.concat([temp_df1, temp_df2], axis=1)
temp_data.shape
```

[43]:
```
(30000, 6000)
```

[44]:
```
final_data = pd.concat([final_data, temp_data], axis=1)
print(final_data.shape)
final_data.head()
```

```
(30000, 6008)
```

[44]:
```
        is_duplicate  q1_len  q2_len  q1_num_words  q2_num_words  word_common  \
398782             1      76      76            12            12           11
115086             0      49      49            12            15            7
327711             0     105     105            25            17            2
367788             0      59      59            12            30            0
151235             0      35      35             5             9            3

        word_total  word_share  0  1  …  2990  2991  2992  2993  2994  2995  \
398782          24        0.46  0  0  …     0     0     0     0     0     0
115086          23        0.30  0  0  …     0     0     0     0     0     0
327711          34        0.06  0  0  …     0     0     0     0     0     0
367788          32        0.00  0  0  …     0     0     0     1     0     0
151235          13        0.23  0  0  …     0     0     0     0     0     0

        2996  2997  2998  2999
398782     0     0     0     0
115086     0     0     0     0
327711     0     0     0     0
367788     0     0     0     0
151235     0     0     0     0

[5 rows x 6008 columns]
```

[45]:
```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(final_data.iloc[:,-1:].
  ↪values,final_data.iloc[:,0].values,test_size=0.2,random_state=1)
```

[46]:
```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
rf = RandomForestClassifier()
rf.fit(X_train,y_train)
y_pred = rf.predict(X_test)
accuracy_score(y_test,y_pred)
```

[46]:
```
0.6353333333333333
```

[47]:
```
from xgboost import XGBClassifier
xgb = XGBClassifier()
xgb.fit(X_train,y_train)
y_pred = xgb.predict(X_test)
accuracy_score(y_test,y_pred)
```

[47]:
```
0.6353333333333333
```

1 Advanced Features

1. Token Features cwc_min: This is the ratio of the number of common words to the length of the smaller question

cwc_max: This is the ratio of the number of common words to the length of the larger question

csc_min: This is the ratio of the number of common stop words to the smaller stop word count among the two questions

csc_max: This is the ratio of the number of common stop words to the larger stop word count among the two questions

ctc_min: This is the ratio of the number of common tokens to the smaller token count among the two questions

ctc_max: This is the ratio of the number of common tokens to the larger token count among the two questions

last_word_eq: 1 if the last word in the two questions is same, 0 otherwise

first_word_eq: 1 if the first word in the two questions is same, 0 otherwise

2. Length Based Features mean_len: Mean of the length of the two questions (number of words)

abs_len_diff: Absolute difference between the length of the two questions (number of words)

longest_substr_ratio: Ratio of the length of the longest substring among the two questions to the length of the smaller question

3. Fuzzy Features fuzz_ratio: fuzz_ratio score from fuzzywuzzy

fuzz_partial_ratio: fuzz_partial_ratio from fuzzywuzzy

token_sort_ratio: token_sort_ratio from fuzzywuzzy

token_set_ratio: token_set_ratio from fuzzywuzzy

[ ]: