

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

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



Mounted at /content/drive

```
df = pd.read_csv('/content/drive/MyDrive/Dataset ML LAB 205122021/train.csv')
```

```
df.shape
```

(404290, 6)

```
df.head()
```



	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	

```
new_df = df.sample(30000,random_state=2)
```

```
new_df.isnull().sum()
```

```
id          0
qid1        0
qid2        0
question1   0
question2   0
is_duplicate 0
dtype: int64
```

```
new_df.head()
```

	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?	I am quite poor and I want to be very rich. Wh...	0	
				I am from India and	T I F T to Thanar		

Next steps:

[Generate code with new_df](#)

 [View recommended plots](#)

```
new_df.isnull().sum()
```

```
id          0
qid1        0
qid2        0
question1   0
question2   0
is_duplicate 0
dtype: int64
```

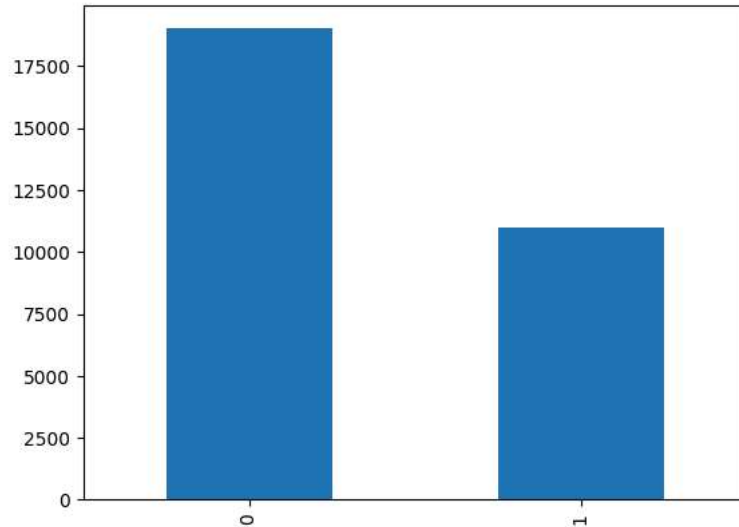
```
new_df.duplicated().sum()
```

0

Distribution of duplicate and non-duplicate questions

```
print(new_df['is_duplicate'].value_counts())
print((new_df['is_duplicate'].value_counts()/new_df['is_duplicate'].count()*100)
new_df['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
<Axes: >
```



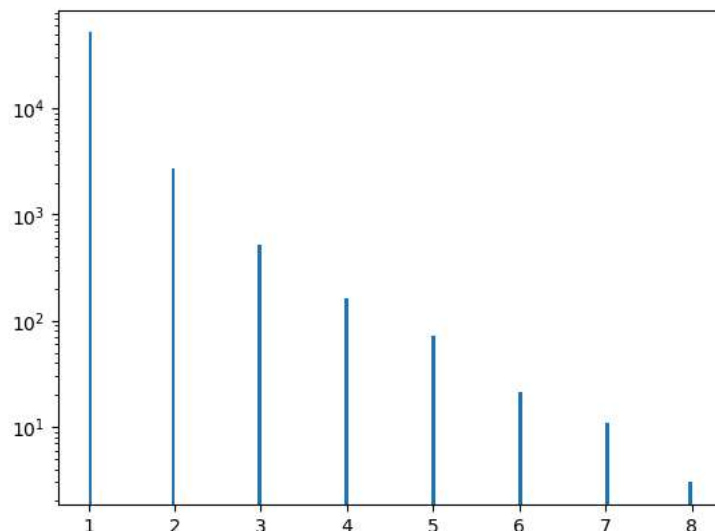
Repeated questions

```
qid = pd.Series(new_df['qid1'].tolist() + new_df['qid2'].tolist())
print('Number of unique questions', np.unique(qid).shape[0])
x = qid.value_counts()>1
print('Number of questions getting repeated', x[x].shape[0])
```

```
Number of unique questions 55299
Number of questions getting repeated 3480
```

Repeated questions histogram

```
plt.hist(qid.value_counts().values, bins=160)
plt.yscale('log')
plt.show()
```



```
# Feature Engineering
```

```
new_df['q1_len'] = new_df['question1'].str.len()
new_df['q2_len'] = new_df['question2'].str.len()
```

```
new_df.head()
```

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len
398782	398782	496695	532029	What is the best marketing automation tool for...	What is the best marketing automation tool for...	1	76	77
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	57
				I am from India	T.I.E.T to Thapar			

Next steps:

[Generate code with new_df](#)
[View recommended plots](#)

```
new_df['q1_num_words'] = new_df['question1'].apply(lambda row: len(row.split(" ")))
new_df['q2_num_words'] = new_df['question2'].apply(lambda row: len(row.split(" ")))
new_df.head()
```

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len
398782	398782	496695	532029	What is the best marketing automation tool for...	What is the best marketing automation tool for...	1	76	77
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	57
				I am from India and live	T.I.E.T to Thapar			

Next steps:

[Generate code with new_df](#)
[View recommended plots](#)

```
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)
```

```
new_df['word_common'] = new_df.apply(common_words, axis=1)
new_df.head()
```

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len
398782	398782	496695	532029	What is the best marketing automation tool for...	What is the best marketing automation tool for...	1	76	77
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	57
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	120
				Why do so many people in the U.S. hate the sou...	My boyfriend doesnt feel guilty when he hurts ...			

Next steps:

[Generate code with new_df](#)[View recommended plots](#)

```
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))
```

```
new_df['word_total'] = new_df.apply(total_words, axis=1)
new_df.head()
```

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len
398782	398782	496695	532029	What is the best marketing automation tool for...	What is the best marketing automation tool for...	1	76	77
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	57
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	120
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	146
151235	151235	237843	50930	Consequences of Bhopal gas tragedv?	What was the reason behind the Bhopal	0	35	50

Next steps:

[Generate code with new_df](#)[View recommended plots](#)

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

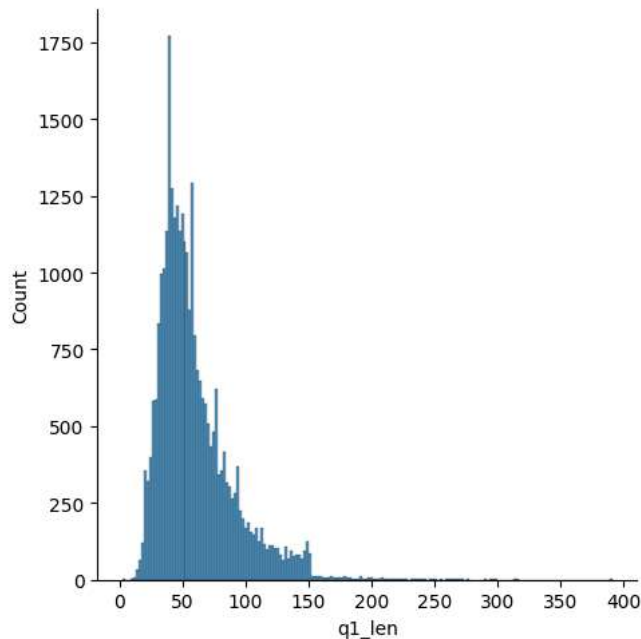
	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len
398782	398782	496695	532029	What is the best marketing automation tool for...	What is the best marketing automation tool for...	1	76	77
115086	115086	187729	187730	I am poor but I want to invest. What should I do?	I am quite poor and I want to be very rich. Wh...	0	49	57
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	120
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	146
151235	151235	237843	50930	Consequences of Bhopal gas tragedy?	What was the reason behind the Bhopal gas trag...	0	35	50

Next steps:

[Generate code with new_df](#)[View recommended plots](#)

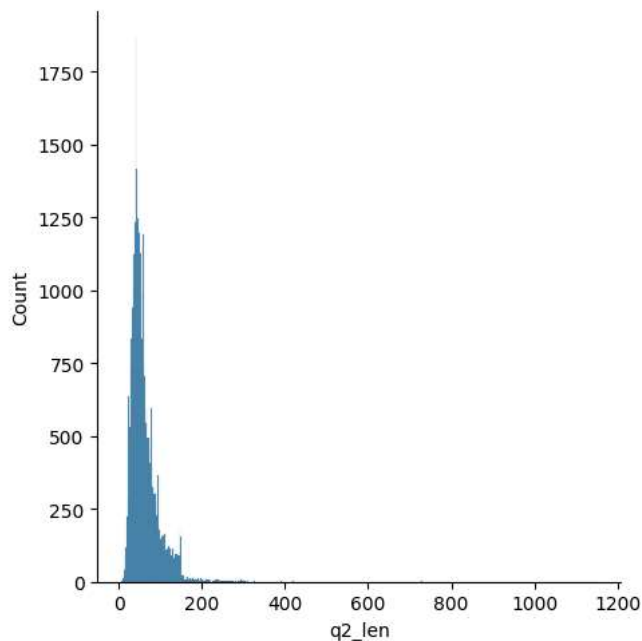
```
# Analysis of features
sns.displot(new_df['q1_len'])
print('minimum characters',new_df['q1_len'].min())
print('maximum characters',new_df['q1_len'].max())
print('average num of characters',int(new_df['q1_len'].mean()))
```

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



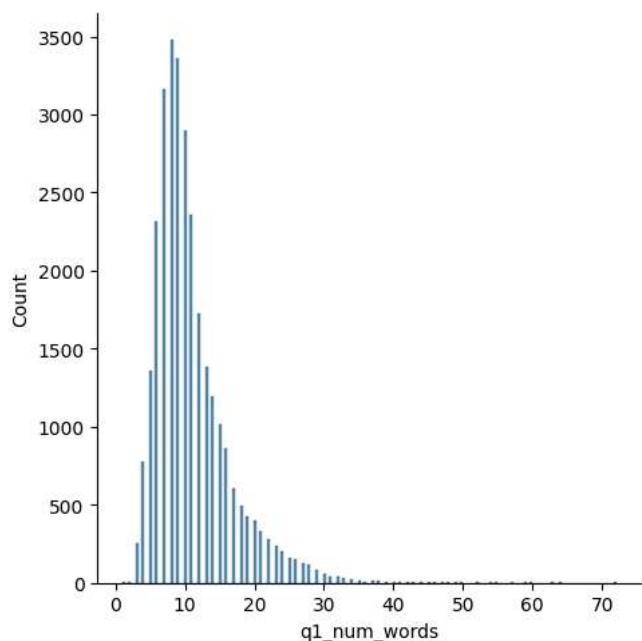
```
sns.displot(new_df['q2_len'])
print('minimum characters',new_df['q2_len'].min())
print('maximum characters',new_df['q2_len'].max())
print('average num of characters',int(new_df['q2_len'].mean()))
```

```
minimum characters 6  
maximum characters 1151  
average num of characters 60
```



```
sns.displot(new_df['q1_num_words'])  
print('minimum words', new_df['q1_num_words'].min())  
print('maximum words', new_df['q1_num_words'].max())  
print('average num of words', int(new_df['q1_num_words'].mean()))
```

```
minimum words 1  
maximum words 72  
average num of words 10
```

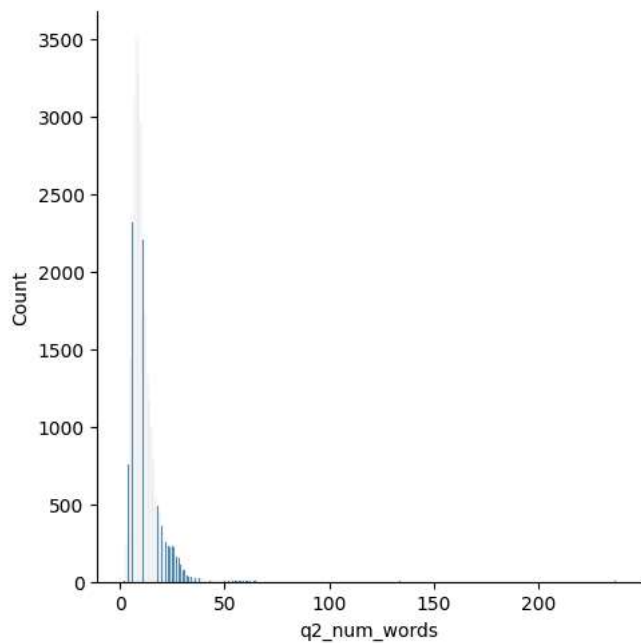


```
sns.displot(new_df['q2_num_words'])  
print('minimum words', new_df['q2_num_words'].min())  
print('maximum words', new_df['q2_num_words'].max())  
print('average num of words', int(new_df['q2_num_words'].mean()))
```

```

minimum words 1
maximum words 237
average num of words 11

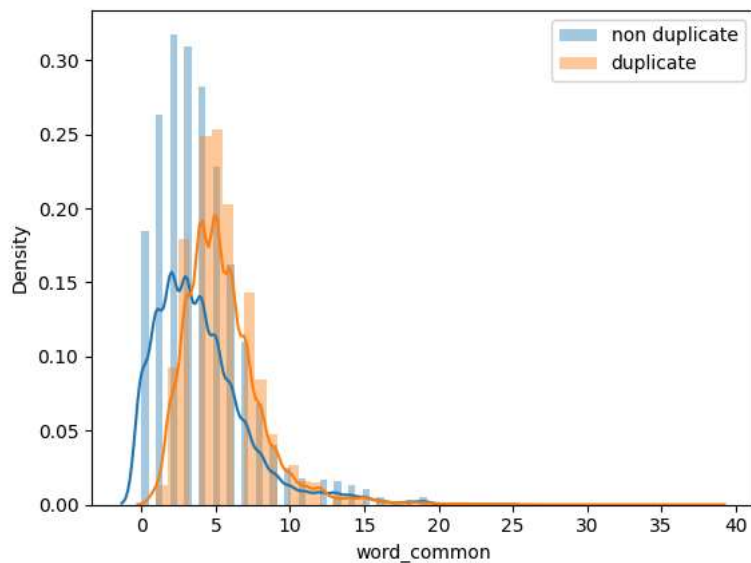
```



```

# common words
sns.distplot(new_df[new_df['is_duplicate'] == 0]['word_common'],label='non duplicate')
sns.distplot(new_df[new_df['is_duplicate'] == 1]['word_common'],label='duplicate')
plt.legend()
plt.show()

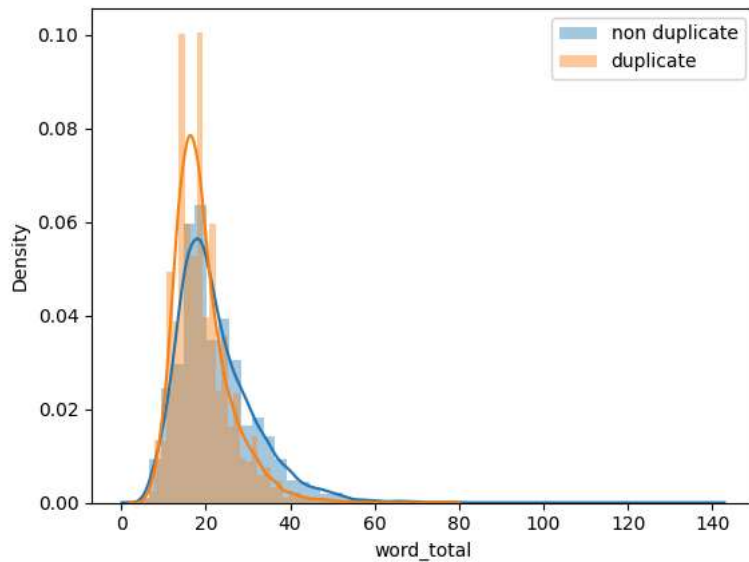
```



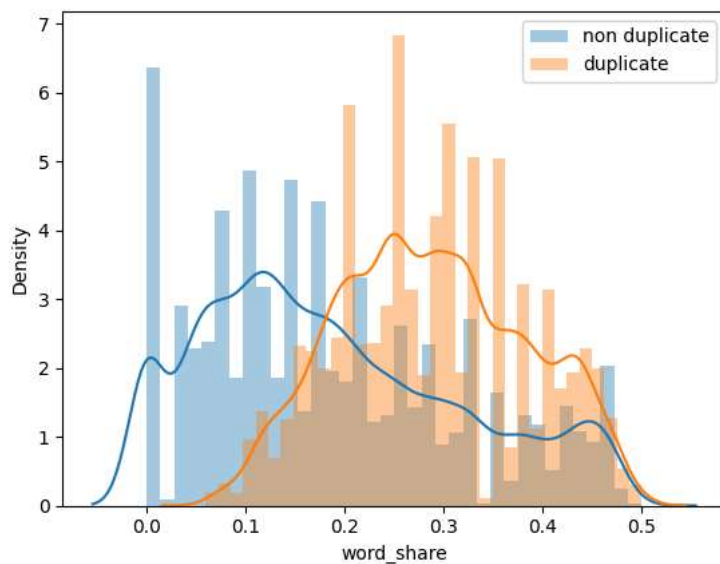
```

# total words
sns.distplot(new_df[new_df['is_duplicate'] == 0]['word_total'],label='non duplicate')
sns.distplot(new_df[new_df['is_duplicate'] == 1]['word_total'],label='duplicate')
plt.legend()
plt.show()

```



```
# word share
sns.distplot(new_df[new_df['is_duplicate'] == 0]['word_share'],label='non duplicate')
sns.distplot(new_df[new_df['is_duplicate'] == 1]['word_share'],label='duplicate')
plt.legend()
plt.show()
```



```
ques_df = new_df[['question1','question2']]
ques_df.head()
```

	question1	question2	
398782	What is the best marketing automation tool for...	What is the best marketing automation tool for...	
115086	I am poor but I want to invest. What should I do?	I am quite poor and I want to be very rich. Wh...	
327711	I am from India and live abroad. I met a guy f...	T.I.E.T to Thapar University to Thapar Univers...	
607700	Why do so many people in the U.S. hate	My boyfriend doesnt feel guilty when he	

Next steps: [Generate code with ques_df](#) [View recommended plots](#)

```
final_df = new_df.drop(columns=['id','qid1','qid2','question1','question2'])
print(final_df.shape)
final_df.head()
```


(30000, 8)

	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_tot
398782	1	76	77	12	12	11	:
115086	0	49	57	12	15	7	:
327711	0	105	120	25	17	2	:
367788	0	59	146	12	30	0	:
151235	0	35	50	5	9	3	:

Nex

[Generate code with final_df](#)[View recommended plots](#)

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

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

```
temp_df1 = pd.DataFrame(q1_arr, index= ques_df.index)
temp_df2 = pd.DataFrame(q2_arr, index= ques_df.index)
temp_df = pd.concat([temp_df1, temp_df2], axis=1)
temp_df.shape
```

(30000, 6000)

```
final_df = pd.concat([final_df, temp_df], axis=1)
print(final_df.shape)
final_df.head()
```

(1000, 6008)

	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_total	word_share	0	1	...	2990	2991	2992	2993	2994
18782	1	76	77	12	12	11	24	0.46	0	0	...	0	0	0	0	0
5086	0	49	57	12	15	7	23	0.30	0	0	...	0	0	0	0	0
17711	0	105	120	25	17	2	34	0.06	0	0	...	0	0	0	0	0
17788	0	59	146	12	30	0	32	0.00	0	0	...	0	0	0	0	1
11235	0	35	50	5	9	3	13	0.23	0	0	...	0	0	0	0	0

ows × 6008 columns

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

```
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)
```

0.7738333333333334

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

0.7661666666666667

✓ 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