Duplicate Question check Using NLP

In [1]:
```
#importing the Libery
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

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

In [2]:
```
#loading the dataset in pandas dataframe
df = pd.read_csv('train.csv')
```

In [3]:
```
#check shape of the dataset
df.shape
```

Out[3]: (404290, 6)

In [4]:
```
#check first five rows of the dataset
df.head()
```

Out[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

In [5]:
```
#take randam 30000 dataset
new_df = df.sample(30000,random_state=2)
```

In [6]:
```
#check missing values of the dataset
new_df.isnull().sum()
```

Out[6]:
```
id              0
qid1            0
qid2            0
question1       0
question2       0
is_duplicate    0
dtype: int64
```

In [7]:
```
#check new dataset
new_df.head()
```

Out[7]:

	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
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
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
151235	151235	237843	50930	Consequences of Bhopal gas tragedy?	What was the reason behind the Bhopal gas trag...	0

In [8]:
```
#check newdataset missing value in dataset
new_df.isnull().sum()
```

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

In [9]:
```
#check duplicate value in dataset
new_df.duplicated().sum()
```

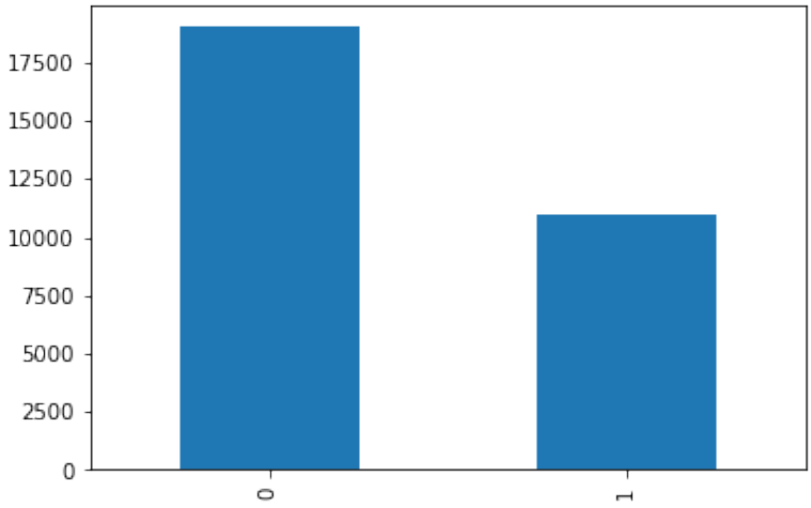
Out[9]:
```
0
```

In [10]:
```
# 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']
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
```

Out[10]: <AxesSubplot:>



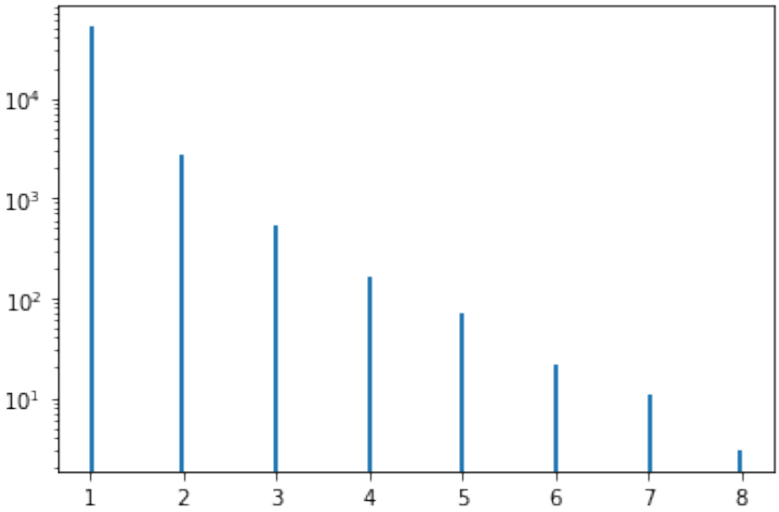
In [11]:
```
# Repeated questions

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
```

In [12]:
```
# Repeated questions histogram

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



In [13]:
```
# Feature Engineering

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

In [14]:
```
new_df.head()
```

Out[14]:

	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

In [15]:
```
new_df['q1_num_words'] = new_df['question1'].apply(lambda row: len(
new_df['q2_num_words'] = new_df['question2'].apply(lambda row: len(
new_df.head()
```

Out[15]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q
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	

In [16]:
```
#check coommon words in dataset
def common_words(row):
    w1 = set(map(lambda word: word.lower().strip(), row['question1'
    w2 = set(map(lambda word: word.lower().strip(), row['question2'
    return len(w1 & w2)
```

In [17]:
```
new_df['word_common'] = new_df.apply(common_words, axis=1)
new_df.head()
```

Out[17]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q
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	

In [18]:
```
def total_words(row):
    w1 = set(map(lambda word: word.lower().strip(), row['question1'
    w2 = set(map(lambda word: word.lower().strip(), row['question2'
    return (len(w1) + len(w2))
```

In [19]:
```
new_df['word_total'] = new_df.apply(total_words, axis=1)
new_df.head()
```

Out[19]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q
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	

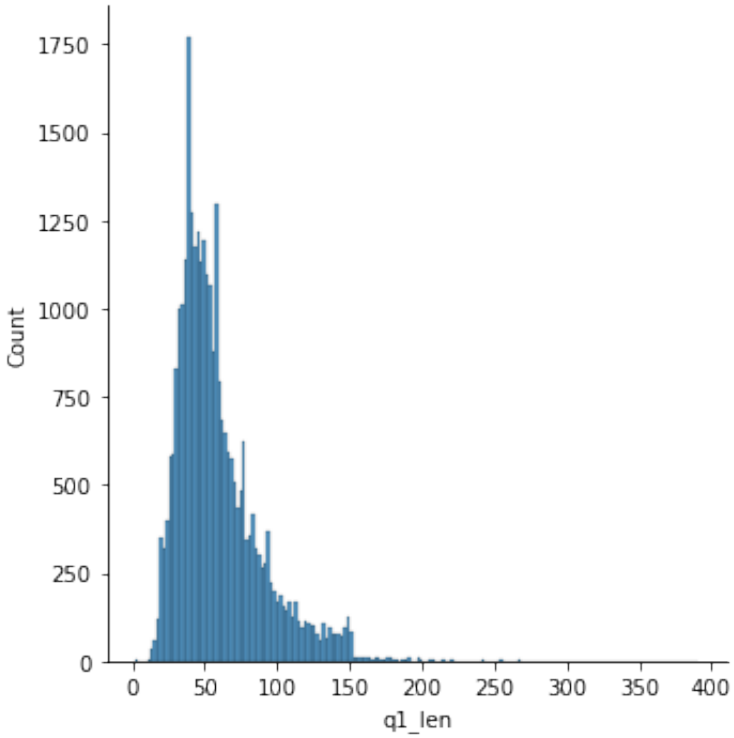
In [20]:
```
new_df['word_share'] = round(new_df['word_common']/new_df['word_tot
new_df.head()
```

Out[20]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q
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	

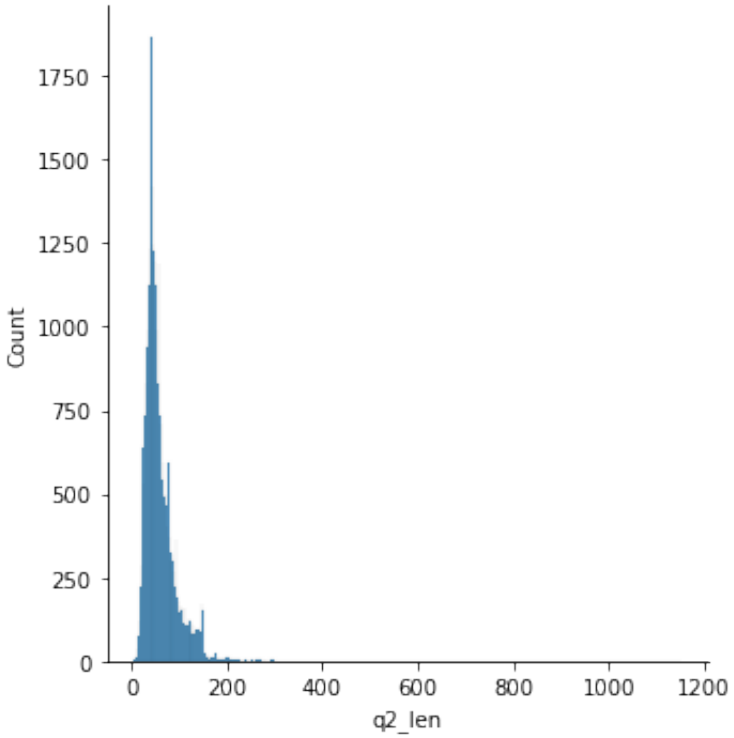
In [21]:
```
# Analysis of features
sns.displot(new_df['q1_len'])
print('minimum characters',new_df['q1_len'].min())
print('maximum characters',new_df['q1_len'].max())
print('average num of characters',int(new_df['q1_len'].mean()))
```

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



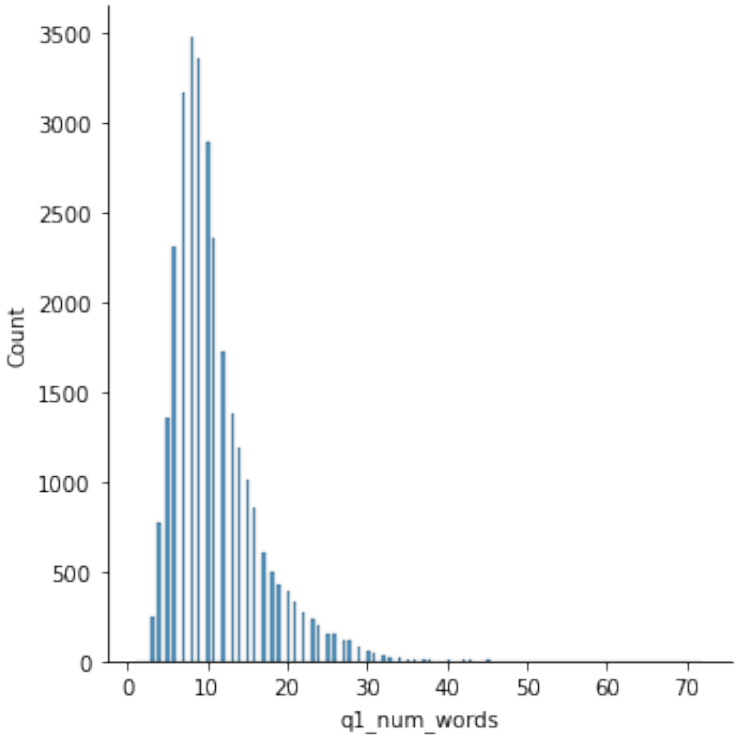
In [22]:
```
sns.displot(new_df['q2_len'])
print('minimum characters',new_df['q2_len'].min())
print('maximum characters',new_df['q2_len'].max())
print('average num of characters',int(new_df['q2_len'].mean()))
```

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



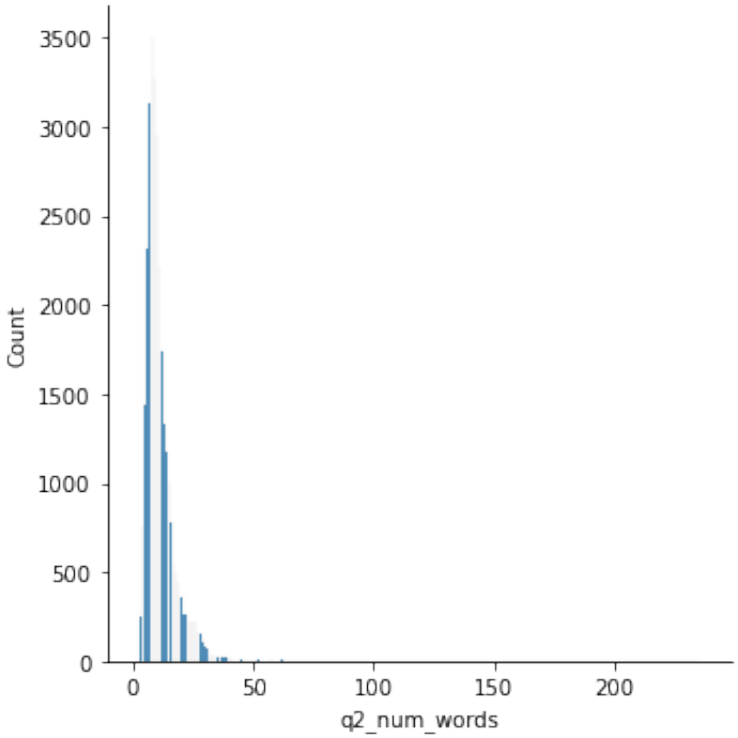
In [23]:
```
sns.displot(new_df['q1_num_words'])
print('minimum words',new_df['q1_num_words'].min())
print('maximum words',new_df['q1_num_words'].max())
print('average num of words',int(new_df['q1_num_words'].mean()))
```

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

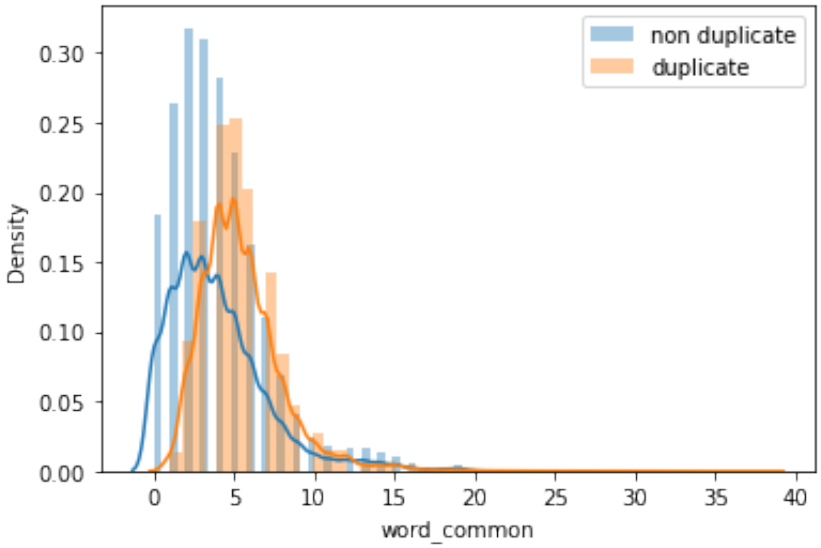


In [24]:
```
sns.displot(new_df['q2_num_words'])
print('minimum words',new_df['q2_num_words'].min())
print('maximum words',new_df['q2_num_words'].max())
print('average num of words',int(new_df['q2_num_words'].mean()))
```

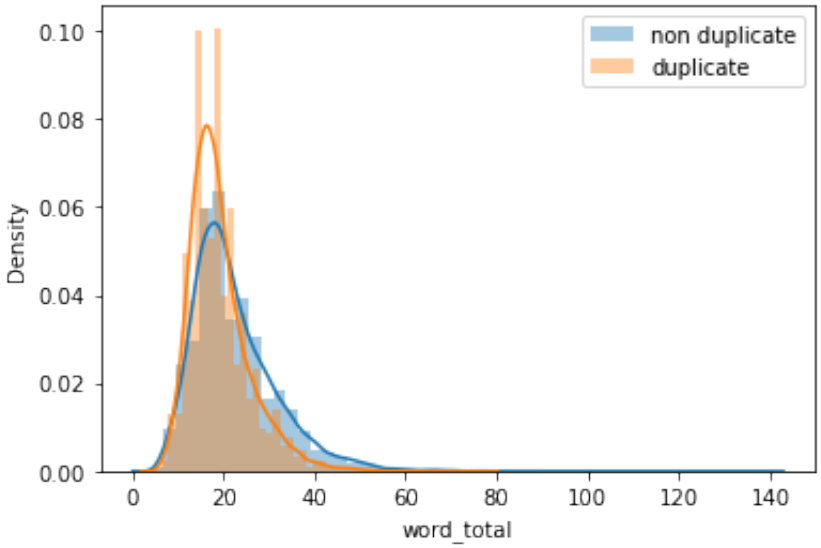
```
minimum words 1
maximum words 237
average num of words 11
```



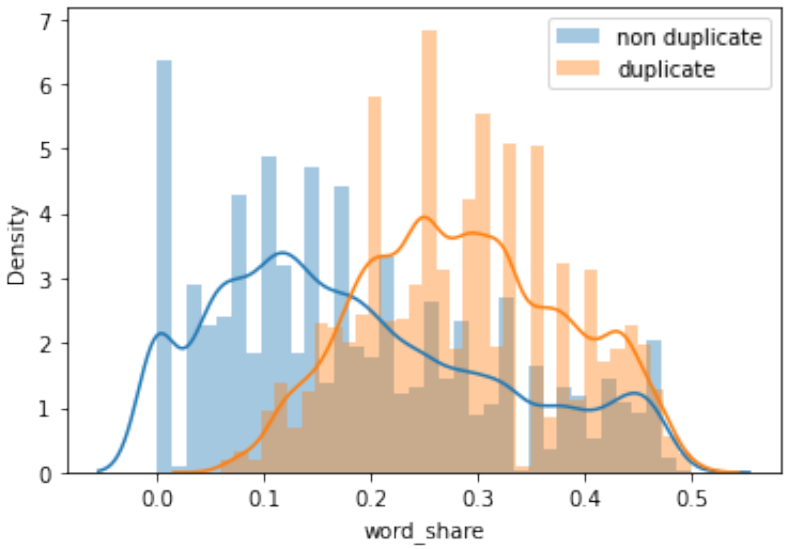
In [25]:
```
# common words
sns.distplot(new_df[new_df['is_duplicate'] == 0]['word_common'],lab
sns.distplot(new_df[new_df['is_duplicate'] == 1]['word_common'],lab
plt.legend()
plt.show()
```



In [26]:
```
# total words
sns.distplot(new_df[new_df['is_duplicate'] == 0]['word_total'],labe
sns.distplot(new_df[new_df['is_duplicate'] == 1]['word_total'],labe
plt.legend()
plt.show()
```



In [27]:
```
# word share
sns.distplot(new_df[new_df['is_duplicate'] == 0]['word_share'],labe
sns.distplot(new_df[new_df['is_duplicate'] == 1]['word_share'],labe
plt.legend()
plt.show()
```



In [28]:
```
ques_df = new_df[['question1','question2']]
ques_df.head()
```

Out[28]:

	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...
367788	Why do so many people in the U.S. hate the sou...	My boyfriend doesnt feel guilty when he hurts ...
151235	Consequences of Bhopal gas tragedy?	What was the reason behind the Bhopal gas trag...

In [29]:
```
final_df = new_df.drop(columns=['id','qid1','qid2','question1','que
print(final_df.shape)
final_df.head()
```

```
(30000, 8)
```

Out[29]:

	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_
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	

In [30]:
```
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)
```

In [31]:
```
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
```

Out[31]: (30000, 6000)

In [33]:
```
final_df = pd.concat([final_df, temp_df], axis=1)
print(final_df.shape)
final_df.head()
```

```
(30000, 6008)
```

Out[33]:

	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_
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	

5 rows × 6008 columns

In [35]:
```
#spliting the dataset in train and test
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(final_df.iloc[:,1:
```

In [36]:
```
#using RandomForestClassifier
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)
```

Out[36]: 0.7683333333333333

In [37]:
```
#using XGBClassifier
from xgboost import XGBClassifier
xgb = XGBClassifier()
xgb.fit(X_train,y_train)
y_pred = xgb.predict(X_test)
accuracy_score(y_test,y_pred)
```

```
[08:57:05] WARNING: C:/Users/Administrator/workspace/xgboost-win64
_release_1.5.0/src/learner.cc:1115: Starting in XGBoost 1.3.0, the
default evaluation metric used with the objective 'binary:logistic
' was changed from 'error' to 'logloss'. Explicitly set eval_metri
c if you'd like to restore the old behavior.
```

Out[37]: 0.7645

In [ ]: