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

```
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
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

```
new_df = df.sample(30000)
```

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

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

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

```
    0
```

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

	question1	question2
181533	What are social media marketing strategies?	What social media sites are the most effective...
199964	What are some things new employees should know...	What are some things new employees should know...
185210	Was Turkey justified in shooting down the Russ...	Did Putin want a Russian jet shot down to esca...
298306	What are the differences between tint and shade?	What is the difference between tint and shade ...
244386	What does my birth chart say about me ?	What does my birth chart say about me?

Next steps: Generate code with ques_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)
```

```
temp_df
```

	0	1	2	3	4	5	6	7	8	9	...	2990	2991	2992	2993	2994	2995	2996	2997	2998	2999
181533	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
199964	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
185210	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
298306	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
244386	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
178532	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
332383	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
54903	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
367982	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
296652	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

30000 rows × 6000 columns

```
temp_df['is_duplicate'] = new_df['is_duplicate']
```

```
temp_df.head()
```

	0	1	2	3	4	5	6	7	8	9	...	2991	2992	2993	2994	2995	2996	2997	2998	2999	is_duplicate
181533	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
199964	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
185210	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
298306	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
244386	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 6001 columns

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(temp_df.iloc[:,0:-1].values,temp_df.iloc[:,-1].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)
```

```
0.7355
```

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

```
0.7211666666666666
```