EDA

September 28, 2023

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:
```
df=pd.read_csv("train.csv")
```

[3]:
```
df.shape
```

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

[4]:
```
df.sample(10)
```

[4]:
```
            id    qid1    qid2  \
205145  205145  141679  132751
257915  257915  206460   85144
34545    34545   63272   63273
115895  115895  102576   85988
45624    45624   81729   81730
361524  361524   56331    6313
310963  310963  435169  435170
118642  118642  192787  136779
326896  326896  453269  453270
67601    67601  116964  116965

                                                question1  \
205145                 Which instrument is easy to learn?
257915  What is a dominant allele? How does it differ …
34545                  Why did Soviets hate Jews so much?
115895  How do I recover deleted messages in whatsapp …
45624   What is difference between living for work and…
361524  Why Indian government abruptly announced the d…
310963       What are the best novels to read on Wattpad?
118642        What are some examples of simple sentences?
326896  Do people who take pain in the name of God fin…
67601   What is the difference between netbook and lap…

                                                question2  is_duplicate
```

```
205145  Which is the most easiest music instrument to …             1
257915  What is the difference between dominant and re…             1
34545   What was Adolf Hitler’s rationale for hating J…             0
115895  Can you delete erased WhatsApp chat messages s…             0
45624   What is the difference between living to work …             0
361524  What are the possible implications of Demoneti…             1
310963      What are some good novels to read on Wattpad?             1
118642  What are some examples of simple sentences wit…             1
326896  How do I stop feeling nervous when I am in a w…             0
67601   What is the difference between a Chromebook an…             0
```

[5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 404290 entries, 0 to 404289
Data columns (total 6 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   id            404290 non-null  int64
 1   qid1          404290 non-null  int64
 2   qid2          404290 non-null  int64
 3   question1     404289 non-null  object
 4   question2     404288 non-null  object
 5   is_duplicate  404290 non-null  int64
dtypes: int64(4), object(2)
memory usage: 18.5+ MB
```

[6]:
```
# missing values
df.isnull().sum()
```

[6]:
```
id              0
qid1            0
qid2            0
question1       1
question2       2
is_duplicate    0
dtype: int64
```

[14]:
```
# Duplicate Values
df.duplicated().sum()
```

[14]:
```
0
```

[15]:
```
df.columns
```

[15]:
```
Index(['id', 'qid1', 'qid2', 'question1', 'question2', 'is_duplicate'],
dtype='object')
```

[16]:
```
df
```

[16]:
```
            id    qid1    qid2  \
0            0       1       2
1            1       3       4
2            2       5       6
3            3       7       8
4            4       9      10
…        …     …     …
404285  404285  433578  379845
404286  404286   18840  155606
404287  404287  537928  537929
404288  404288  537930  537931
404289  404289  537932  537933

                                                question1  \
0       What is the step by step guide to invest in sh…
1       What is the story of Kohinoor (Koh-i-Noor) Dia…
2       How can I increase the speed of my internet co…
3       Why am I mentally very lonely? How can I solve…
4       Which one dissolve in water quikly sugar, salt…
…                                                   …
404285  How many keywords are there in the Racket prog…
404286          Do you believe there is life after death?
404287                                  What is one coin?
404288  What is the approx annual cost of living while…
404289              What is like to have sex with cousin?

                                                question2  is_duplicate
0       What is the step by step guide to invest in sh…              0
1       What would happen if the Indian government sto…              0
2       How can Internet speed be increased by hacking…              0
3       Find the remainder when [math]23^{24}[/math] i…              0
4                 Which fish would survive in salt water?              0
…                                                   …           …
404285  How many keywords are there in PERL Programmin…              0
404286         Is it true that there is life after death?              1
404287                                  What's this coin?              0
404288  I am having little hairfall problem but I want…              0
404289      What is it like to have sex with your cousin?              0

[404290 rows x 6 columns]
```

[18]:
```
df.head()
```

[18]:
```
   id  qid1  qid2                                          question1  \
0   0     1     2  What is the step by step guide to invest in sh…
```

```
1   1     3     4  What is the story of Kohinoor (Koh-i-Noor) Dia…
2   2     5     6  How can I increase the speed of my internet co…
3   3     7     8  Why am I mentally very lonely? How can I solve…
4   4     9    10  Which one dissolve in water quikly sugar, salt…

                                           question2  is_duplicate
0  What is the step by step guide to invest in sh…             0
1  What would happen if the Indian government sto…             0
2  How can Internet speed be increased by hacking…             0
3  Find the remainder when [math]23^{24}[/math] i…             0
4            Which fish would survive in salt water?             0
```

[19]:

```python
df.tail()
```

[19]:

```
            id    qid1    qid2  \
404285  404285  433578  379845
404286  404286   18840  155606
404287  404287  537928  537929
404288  404288  537930  537931
404289  404289  537932  537933

                                                question1  \
404285  How many keywords are there in the Racket prog…
404286          Do you believe there is life after death?
404287                                  What is one coin?
404288  What is the approx annual cost of living while…
404289              What is like to have sex with cousin?

                                                question2  is_duplicate
404285  How many keywords are there in PERL Programmin…             0
404286         Is it true that there is life after death?             1
404287                                  What's this coin?             0
404288  I am having little hairfall problem but I want…             0
404289      What is it like to have sex with your cousin?             0
```

[ ]:

```

```

[ ]:

```

```

[9]:

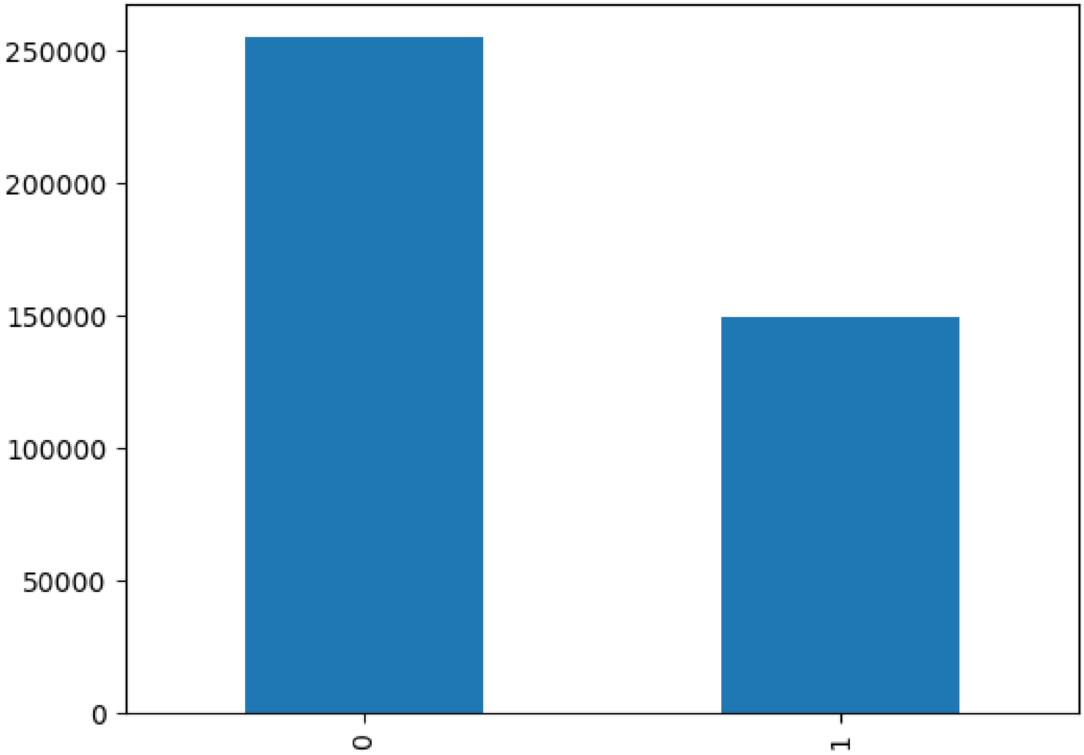
```python
# Distribution of duplicate and non-duplicate questions
print(df['is_duplicate'].value_counts())
print((df['is_duplicate'].value_counts()/df['is_duplicate'].count())*100)
df['is_duplicate'].value_counts().plot(kind='bar')
```

```
0    255027
1    149263
Name: is_duplicate, dtype: int64
0    63.080215
```

```
1    36.919785
Name: is_duplicate, dtype: float64
```

[9]:

```
<Axes: >
```



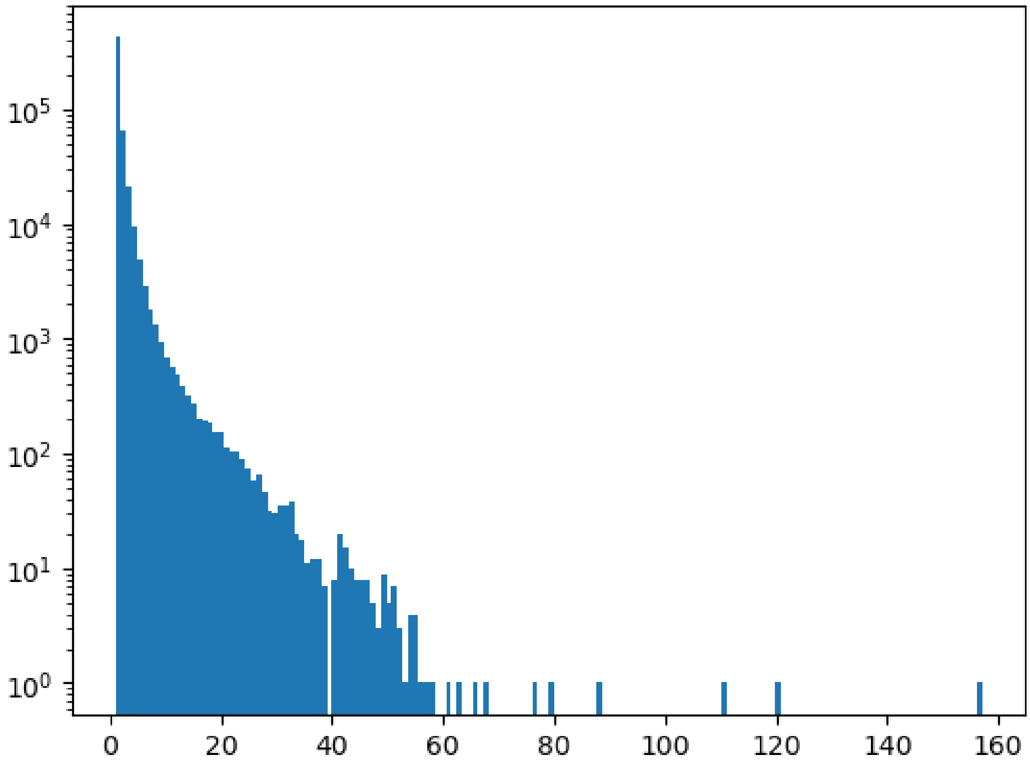
[10]:

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

```
Number of unique questions 537933
Number of quenstions getting repeated 111780
```

[13]:

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


10^5
10^4
10^3
10^2
10^1
10^0
0
20
40
60
80
100
120
140
160

```
[ ]:
```