Only Back of word

September 28, 2023

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

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

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

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

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

[4]:
```
   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
```

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

[5]:
```
            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
```

[6]:
```
new_df = df.sample(30000)
```

[7]:
```
new_df.isnull().sum()
```

[7]:
```
id              0
qid1            0
qid2            0
question1       0
question2       0
is_duplicate    0
dtype: int64
```

[8]:
```
new_df.duplicated().sum()
```

[8]:
```
0
```

[9]:
```
ques_df = new_df[['question1','question2']]
ques_df
```

[9]:
```
                                                question1  \
364622  What are the main factors that shape a country…
117949  What are the biggest questions that are yet in…
114373  How do I memorize various articles and section…
218214  How do I redeem debit card cashback points of …
399235                 How do you know if you're in love?
…                                                     …
384242          How much does youtube pay per 1000 views?
195552            Who are the enemies and allies of Iran?
242576   How does the I-card of an IAS officer look like?
126675  Who will win the 2016 U.S. presidential electi…
12053   What are the most convenient flights between L…

                                                question2
364622                    What does foreign policy mean?
117949  What are the biggest failures of theory of evo…
```

```
114373  What is meant by article 18 of Indian constitu…
218214  How can I redeem my SBI credit card points onl…
399235              How do you know when it is true love?
…                                                     …
384242  About how much is often gained from a monetize…
195552        What are each country's allies and enemies?
242576  What does an IAS officer's ID card look like? …
126675  Who will win the us 2016 presidential election…
12053   Are there any non-stop flights between Leeds a…

[30000 rows x 2 columns]
```

[10]:
```
ques_df = new_df[['question1','question2']]
ques_df.head()
```

[10]:
```
                                                 question1  \
364622  What are the main factors that shape a country…
117949  What are the biggest questions that are yet in…
114373  How do I memorize various articles and section…
218214  How do I redeem debit card cashback points of …
399235                 How do you know if you're in love?

                                                 question2
364622                      What does foreign policy mean?
117949  What are the biggest failures of theory of evo…
114373  What is meant by article 18 of Indian constitu…
218214  How can I redeem my SBI credit card points onl…
399235              How do you know when it is true love?
```

[11]:
```
ques_df = new_df[['question1','question2']]
ques_df.tail()
```

[11]:
```
                                                 question1  \
384242          How much does youtube pay per 1000 views?
195552            Who are the enemies and allies of Iran?
242576   How does the I-card of an IAS officer look like?
126675  Who will win the 2016 U.S. presidential electi…
12053   What are the most convenient flights between L…

                                                 question2
384242  About how much is often gained from a monetize…
195552        What are each country's allies and enemies?
242576  What does an IAS officer's ID card look like? …
126675  Who will win the us 2016 presidential election…
12053   Are there any non-stop flights between Leeds a…
```

[12]:
```
# Replace NaN values with an empty string
ques_df.fillna('', inplace=True)
```

```
C:\Users\mohsi\AppData\Local\Temp\ipykernel_5536\2570563329.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  ques_df.fillna('', inplace=True)
```

[13]:
```
question1=[30000]
question2=[30000]
```

[14]:
```
from sklearn.feature_extraction.text import CountVectorizer

# Assuming you have already processed and cleaned your 'questions' data

# Create a CountVectorizer
cv = CountVectorizer(max_features=30000, lowercase=True, stop_words='english')
cv = CountVectorizer(max_features=1000, lowercase=True, stop_words='english')
# Fit and transform your questions data

questions = list(ques_df['question1']) + list(ques_df['question2'])
q_matrix = cv.fit_transform(questions)
# Split the matrix into two equal parts
#split_idx = q_matrix.shape[0] // 2  # Use shape[0] to get the number of rows
#q1_arr = q_matrix[:split_idx].toarray()
#q2_arr = q_matrix[split_idx:].toarray()
q1_arr, q2_arr = np.vsplit(cv.fit_transform(questions).toarray(),2)
# Create DataFrames with the correct shapes
#temp_df1 = pd.DataFrame(q1_arr, index=ques_df.index[:split_idx])
#temp_df2 = pd.DataFrame(q2_arr, index=ques_df.index[split_idx:])
temp_df1 = pd.DataFrame(q1_arr, index= ques_df.index)
temp_df2 = pd.DataFrame(q2_arr, index= ques_df.index)

# Concatenate the DataFrames
temp_df = pd.concat([temp_df1, temp_df2], axis=1)
```

[15]:
```
temp_df.shape
```

[15]:
```
(30000, 2000)
```

[16]:
```
temp_df
```

[16]:

	0	1	2	3	4	5	6	7	8	9	…	990	991	992	\
364622	0	0	0	0	0	0	0	0	0	0	…	0	0	0	
117949	0	0	0	0	0	0	0	0	0	0	…	0	0	0	

```
114373    0    0    0    0    0    0    0    0    0    0  …    0    0    0
218214    0    0    0    0    0    0    0    0    0    0  …    0    0    0
399235    0    0    0    0    0    0    0    0    0    0  …    0    0    0
…       …    …    …    …    …    …    …    …    …    …    …    …    …    …
384242    0    0    0    1    0    0    0    0    0    0  …    0    0    0
195552    0    0    0    0    0    0    0    0    0    0  …    0    0    0
242576    0    0    0    0    0    0    0    0    0    0  …    0    0    0
126675    0    0    0    0    0    0    0    0    0    0  …    0    0    0
12053     0    0    0    0    0    0    0    0    0    0  …    0    0    0

         993  994  995  996  997  998  999
364622     0    0    0    0    0    0    0
117949     0    0    0    0    0    0    0
114373     0    0    0    0    0    0    0
218214     0    0    0    0    0    0    0
399235     0    0    0    0    0    0    0
…        …    …    …    …    …    …    …
384242     0    0    0    0    0    1    0
195552     0    0    0    0    0    0    0
242576     0    0    0    0    0    0    0
126675     0    0    0    0    0    0    0
12053      0    0    0    0    0    0    0

[30000 rows x 2000 columns]
```

[17]:
```
temp_df['is_duplicate']=new_df['is_duplicate']
```

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

[18]:
```
        0  1  2  3  4  5  6  7  8  9  …  991  992  993  994  995  996  997  \
364622  0  0  0  0  0  0  0  0  0  0  …    0    0    0    0    0    0    0
117949  0  0  0  0  0  0  0  0  0  0  …    0    0    0    0    0    0    0
114373  0  0  0  0  0  0  0  0  0  0  …    0    0    0    0    0    0    0
218214  0  0  0  0  0  0  0  0  0  0  …    0    0    0    0    0    0    0
399235  0  0  0  0  0  0  0  0  0  0  …    0    0    0    0    0    0    0

         998  999  is_duplicate
364622     0    0             0
117949     0    0             0
114373     0    0             0
218214     0    0             0
399235     0    0             1

[5 rows x 2001 columns]
```

[19]:
```
q1_arr = q1_arr.astype(np.float32)
q2_arr = q2_arr.astype(np.float32)
```

[20]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(temp_df.iloc[:,:-1].values,␣
↪temp_df.iloc[:, -1].values, test_size=0.1, random_state=1)
```

[21]:
```
pip install xgboost
```

```
Defaulting to user installation because normal site-packages is not
writeableNote: you may need to restart the kernel to use updated packages.

Requirement already satisfied: xgboost in
c:\users\mohsi\appdata\roaming\python\python311\site-packages (2.0.0)
Requirement already satisfied: numpy in d:\pkg\anaconda\lib\site-packages (from
xgboost) (1.24.3)
Requirement already satisfied: scipy in d:\pkg\anaconda\lib\site-packages (from
xgboost) (1.10.1)
```

[22]:
```
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score
xgb = XGBClassifier()
xgb.fit(X_train,y_train)
y_pred = xgb.predict(X_test)
accuracy_score(y_test,y_pred)
```

[22]:
```
0.7126666666666667
```

[23]:
```
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)
```

[23]:
```
0.7093333333333334
```

[ ]:
```

```