In [1]:
```
import pandas as pd
import sklearn as sk
import numpy as np
import sys
```

In [2]:
```
%%time
#importing the csv file That contains all the bigram features
df=pd.read_csv("new_bigram_of_byte.csv");
```

```
CPU times: total: 15min 41s
Wall time: 16min 9s
```

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

Out[3]:
```
(10869, 66052)
```

In [4]:
```
df.columns
```

Out[4]:
```
Index(['Unnamed: 0.1', 'Unnamed: 0', 'FileName', '00 01', '01 00', '00 02',
       '02 00', '00 03', '03 00', '00 04',
       ...
       'f7 f7', 'f8 f8', 'f9 f9', 'fa fa', 'fb fb', 'fc fc', 'fd fd', 'fe fe',
       'ff ff', '?? ??'],
      dtype='object', length=66052)
```

In [ ]:
```

```

In [5]:
```
class_labels=pd.read_csv('trainLabels.csv')
class_labels=class_labels["Class"]
class_labels.shape
df=df[1:]
```

In [33]:
```
from sklearn.feature_selection import SelectKBest, chi2, f_regression
select_kbest_object = SelectKBest(score_func=chi2, k=2000)
```

In [34]:
```
%%time
top_features=select_kbest_object.fit(df.drop('FileName',axis=1),class_labels)
```

```
CPU times: total: 13.6 s
Wall time: 13.6 s
```

In [35]:
```
top_features_df=pd.DataFrame(top_features.scores_)
```

In [36]:
```
all_features_columns_df=pd.DataFrame(df.columns[1:])
```

In [37]:
```
bigram_df_imp_feature_score=pd.concat([top_features_df,all_features_columns_df],axis=1)
```

In [38]:
```
bigram_df_imp_feature_score
```

Out[38]:

	0	0
0	1.895041e+07	Unnamed: 0
1	1.895041e+07	FileName
2	9.187146e+03	00 01
3	9.740081e+03	01 00
4	1.808368e+04	00 02
...	...	...
66046	3.437062e+04	fc fc
66047	9.429014e+04	fd fd
66048	5.329319e+04	fe fe
66049	8.349287e+04	ff ff
66050	3.083854e+07	?? ??

66051 rows × 2 columns

In [39]:
```
bigram_df_imp_feature_score.columns=["Byte Bigram Top 2000 Feature scores","Byte Bigram Top 2000 Feature Names"]
```

In [40]:
```
bigram_df_imp_feature_score=bigram_df_imp_feature_score.nlargest(2000,"Byte Bigram Top 2000 Feature scores")
```

In [41]:
```
top_2000_features_Names=list(bigram_df_imp_feature_score["Byte Bigram Top 2000 Feature Names"])
```

In [42]:
```
top_2000_byte_bigram_features=pd.concat([df["FileName"],df[top_2000_features_Names]],axis=1)
```

In [43]:
```
top_2000_byte_bigram_features.to_csv("Bytes_final_bigram.csv")
```

In [44]:
```
top_2000_byte_bigram_features.shape
```

Out[44]:
```
(10868, 2001)
```