preprocess_dataset

June 2, 2020

In [53]:
```
import csv
import pandas as pd
import numpy as np
from sklearn.utils import shuffle
# filename1 = "X_test.txt"
# filename2 = "X_train.txt"

# df_data_test = pd.read_csv(filename1,header=None, error_bad_lines=False)
# df_data_train = pd.read_csv(filename1,header=None, error_bad_lines=False)
# df_data.to_csv("X_test_2.csv",index=False)
# print (df_data.head(1).shape)

filename3 = "X_train_cleaned.csv"
filename4 = "y_train.csv"

filename5 = "X_test_cleaned.csv"
filename6 = "y_test.csv"

df_train_data = pd.read_csv(filename3)
df_train_labels = pd.read_csv(filename4)
df_test_data = pd.read_csv(filename5)
df_test_labels = pd.read_csv(filename6)

df_train_total = pd.concat([df_train_data,df_train_labels],axis=1)
df_test_total = pd.concat([df_test_data,df_test_labels],axis=1)
```

In [54]:
```
np_train = np.array(df_train_total)
np_test = np.array(df_test_total)
```

In [55]:
```
np_train = shuffle(np_train)
np_test = shuffle(np_test)
```

In [56]:
```
np_full = np.concatenate((np_train,np_test))
```

In [57]:
```
np_full.shape
```

Out[57]:
```
(10299, 562)
```

```
In [58]: new_data_train_dict = [5,5,5,5,5,5,5]
         new_data_train = []
         for i in np_full:
             if(new_data_train_dict[int(i[-1])] > 0):
                 new_data_train.append(i)
                 new_data_train_dict[int(i[-1])] -= 1
             continue

In [59]: new_data_train = np.array(new_data_train)

In [60]: new_data_train.shape

Out[60]: (30, 562)

In [61]: list_data = new_data_train.tolist()

In [62]: # list_data[0]

In [63]: # list_data[0][-1]

In [64]: new_format_data = []
         for i in list_data:
             data = map(str,i[:20])
             data = ",".join(data)
             data = '"' + data + ','
             data += str(i[-1]) +'"\n'
             new_format_data.append(data)

In [65]: # new_format_data[3]

In [66]: # new_format_data[0]

In [67]: f = open("preprocessed_data_to_blockchain.txt","w")

In [68]: for i in new_format_data:
             f.write(i)

In [69]: f.close()
```