```
# -*- coding: utf-8 -*-
"""Data_Preprocessing.ipynb

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1ALcO_L0DWWHcrm6Qcxv-KH4TzpSpAlak

<center><h1>Mini Project 3 - Convolutional Neural Network</h1>
<h3>Data Preprocessing</h3>
<h4>This file performs some of the operations on Data Preprocessing and Analysis.</h4></center>

<h3>Team Members:</h3>
<center>
Yi Zhu, 260716006<br>
Fei Peng, 260712440<br>
Yukai Zhang, 260710915
</center>

# Importations
"""

from google.colab import drive
drive.mount('/content/drive')

# make path = './' in-case you are running this locally
path = '/content/drive/My Drive/ECSE_551_F_2020/Mini_Project_03/'

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from google.colab import files
from sklearn.preprocessing import LabelEncoder
from scipy.stats import entropy

"""# Data Preprocessing"""

dataset = pd.read_csv(path+"TrainLabels.csv")
# reddit_test = pd.read_csv(path+"test.csv")

y = dataset['class']

class Data_Processing:
    def __init__(self, data, name='New Data'):
        self.data = data
        self.name = name

    def show_y_dist(self, ydata):
        plt.figure(figsize=(8,4))
        plt.subplot(111), sns.countplot(x='class', data=ydata)
        plt.title('Distribution of Class')
        plt.savefig("Distribution of Class.png", dpi=1200)
        files.download("Distribution of Class.png")
        plt.show()

data_analysis = Data_Processing(dataset.values, 'TrainLabels.csv')
data_analysis.show_y_dist(dataset)

# calculate the data entropy
le = LabelEncoder()  # encoder for classes
```

```
63 le.fit(y)
64 y_label = le.transform(y)
65 n_k = len(le.classes_)
66 N = len(y)
67 theta_k = np.zeros(n_k)  # probability of class k
68 # compute theta values
69 for k in range(n_k):
70     count_k = (y_label==k).sum()
71     theta_k[k] = count_k / N
72
73 print("Data entropy is", entropy(theta_k, base=2))
```