In [1]:
```
#Import the required libraries
# basics
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Others
import os
from PIL import Image
```

In [2]:
```
#Conduct EDA
def listdir_nohidden(path):
    return [file for file in os.listdir(path) if not file.startswith('.')]

def num_files_in_directory(path):
    return len([file for file in os.listdir(path) if not file.startswith('.')])
```

In [3]:
```
#project path
project_path = "/Users/freazx/Documents/ONE TAB/CODING"
train_path = project_path + "/chest_xray/setB/train/"
val_path = project_path + "/chest_xray/setB/val/"
test_path = project_path + "/chest_xray/setB/test/"
```

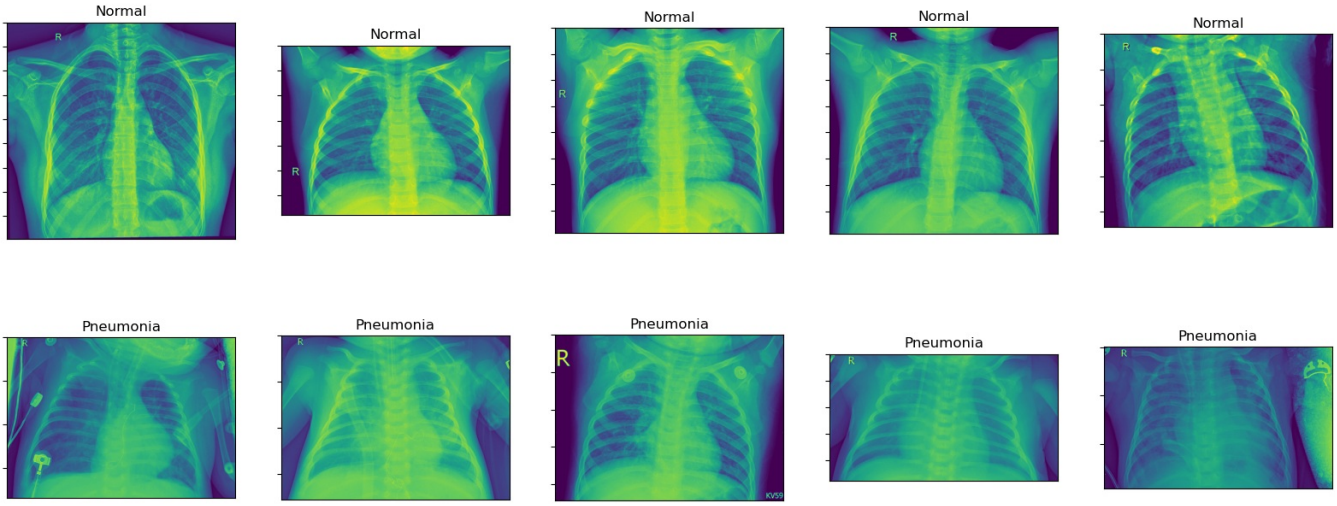
In [4]:
```
train_normal_path = train_path + 'NORMAL/'
train_pneumonia_path = train_path + 'PNEUMONIA/'
train_normal_len = num_files_in_directory(train_normal_path)
train_pneumonia_len = num_files_in_directory(train_pneumonia_path)
train_sum_len = train_normal_len + train_pneumonia_len

print("[Train] Number of NORMAL Images: ", train_normal_len)
print("[Train] Number of PNEUMONIA Images: ", train_pneumonia_len)
print("[Train] Number of TOTAL Images: ", train_sum_len)
```

```
[Train] Number of NORMAL Images:  1106
[Train] Number of PNEUMONIA Images:  2998
[Train] Number of TOTAL Images:  4104
```

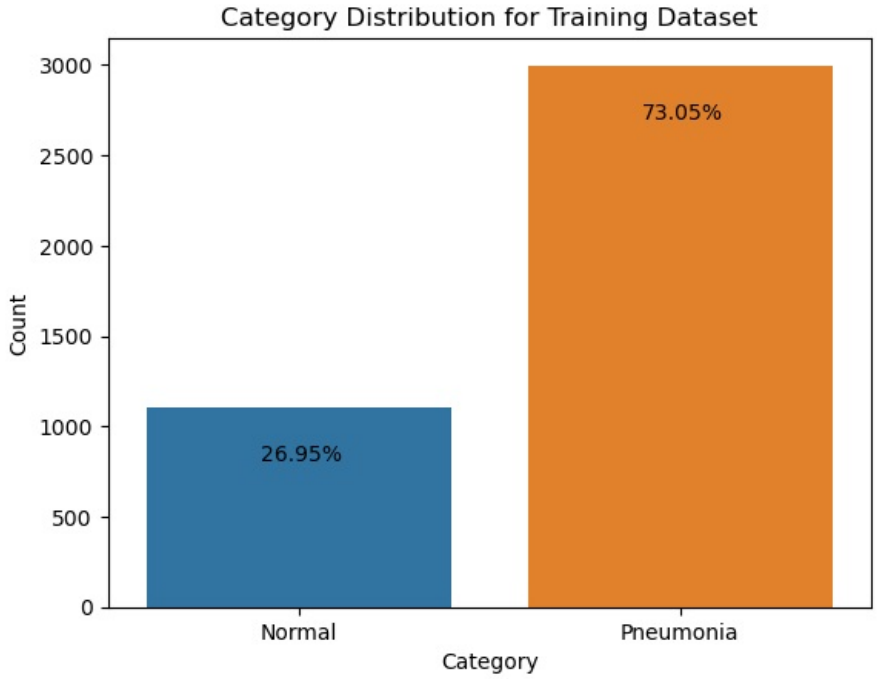
In [5]:
```
val_normal_path = val_path + 'NORMAL/'
val_pneumonia_path = val_path + 'PNEUMONIA/'

print("[Validation] Number of NORMAL Images: ", num_files_in_directory(val_normal_path))
print("[Validation] Number of PNEUMONIA Images: ", num_files_in_directory(val_pneumonia_path))
print("[Validation] Number of TOTAL Images: ", num_files_in_directory(val_normal_path) + num_files_in_directory
```

```
[Validation] Number of NORMAL Images:  236
[Validation] Number of PNEUMONIA Images:  635
[Validation] Number of TOTAL Images:  871
```

In [6]:
```
plt.figure(figsize=(20, 8))
num = 5
for index in range(num):
    n_img_title = os.listdir(train_normal_path)[index]
    n_img_path = train_normal_path + n_img_title

    plt.subplot(2, num, index+1)
    plt.imshow(Image.open(n_img_path))
    plt.tick_params(axis='both', which='both', bottom=False, top=False, labelbottom=False, labelleft=False)
    plt.title('Normal')
    ;

    p_img_title = os.listdir(train_pneumonia_path)[index]
    p_img_path = train_pneumonia_path + p_img_title

    plt.subplot(2, num, index+num+1)
    plt.imshow(Image.open(p_img_path))
    plt.tick_params(axis='both', which='both', bottom=False, top=False, labelbottom=False, labelleft=False)
    plt.title('Pneumonia')
    ;
```



In [8]:
```
sum_len = train_normal_len + train_pneumonia_len
count_list = [['Normal', train_normal_len, train_normal_len/sum_len], \
              ['Pneumonia', train_pneumonia_len, train_pneumonia_len/sum_len]]
count_df = pd.DataFrame(count_list, columns=['Category', 'Count', 'Percentage'])
values = [train_normal_len, train_pneumonia_len]

ax = sns.barplot(x='Category', y='Count', data=count_df)
ax.set_title("Category Distribution for Training Dataset")

for index, value in enumerate(values):
    plt.text(index-0.1, value-300, str(round(value/sum(values)*100, 2)) + "%");
```



In [9]:
```
#all images are in different sizes
for index in range(20):
    sample_image = train_normal_path + listdir_nohidden(train_normal_path)[index]
    image = Image.open(sample_image)
    width, height = image.size
    print('image', index+1, ':', width, 'x', height)
```

```
image 1 : 2359 x 2234
image 2 : 1828 x 1357
image 3 : 2194 x 1966
image 4 : 2172 x 1963
image 5 : 1284 x 1086
image 6 : 1432 x 1184
image 7 : 1442 x 1152
image 8 : 1410 x 1106
image 9 : 2373 x 2663
image 10 : 1426 x 1049
image 11 : 1692 x 1302
image 12 : 2172 x 1615
image 13 : 1680 x 1353
image 14 : 1654 x 1364
image 15 : 1372 x 1109
image 16 : 1638 x 1236
image 17 : 2218 x 2183
image 18 : 2309 x 2601
image 19 : 1114 x 815
image 20 : 1928 x 1303
```