Assignment 2 - Parallel Programming!

Imports

In [28]:

```
import utils
import images
import images_MP
```

Setup the Project ¶

In [4]:

```
utils.create_config_file()
```

```
Config file setup properly.
```

In [3]:

```
images.download_data()
```

```
Downloading [########################################################
####] 50/50
```

Exploratory Data Analysis (EDA)

In [13]:

```
%%time
df = images.get_df()
```

```
CPU times: user 102 ms, sys: 23.1 ms, total: 125 ms
Wall time: 212 ms
```

In [11]:

```
df.shape
```

Out[11]:

```
(1500, 20)
```

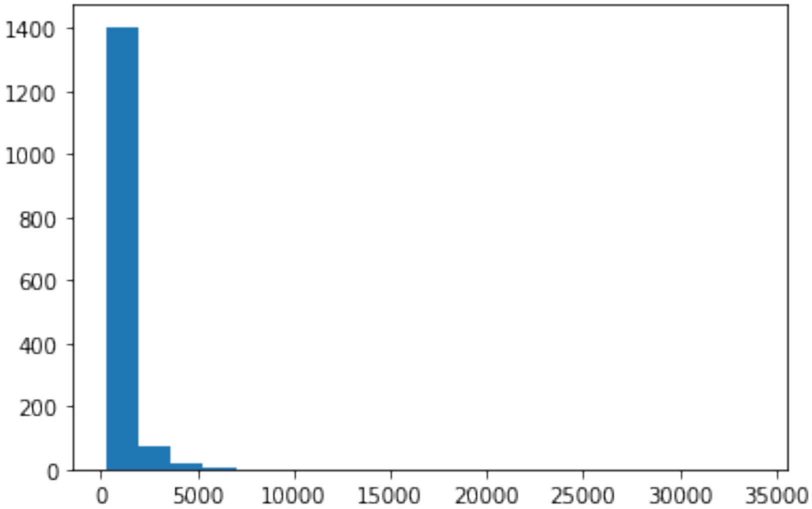
In [16]:

```
# todo: more eda analysis here.
df.sample(5)
```

444	IbXuuGhiO48	2020-03-11T12:18:57-04:00	2020-03-28T01:05:00-04:00	2020-03-12T10:27:02-04:00	4240	2832	#E3E7E2	No
722	nrC2TA0CK8w	2020-03-29T02:07:23-04:00	2020-03-29T15:55:24-04:00	2020-03-29T05:36:27-04:00	3601	2401	#0A1016	A sere day duri quaranti tir
690	61L3f70h5Nc	2020-03-31T02:36:08-04:00	2020-04-07T01:01:54-04:00	2020-03-31T02:42:01-04:00	3541	5312	#161719	No

In [22]:

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
x=df.downloads
plt.hist(x, bins=20)
plt.show()
```



In [25]:
```
df.likes.describe()
```

Out[25]:
```
count    1500.000000
mean       52.029333
std        34.835329
min         7.000000
25%        29.000000
50%        43.000000
75%        65.000000
max       250.000000
Name: likes, dtype: float64
```

In [34]:
```
avg_width=df.width.mean()
print(avg_width)
avg_height=df.height.mean()
print(avg_height)
```

```
4341.398666666667
4749.532
```

In [37]:
```
df.color.unique()
```

Out[37]:
```
array(['#C80115', '#18130F', '#F2E8E2', ..., '#FDAC56', '#2C5E7A',
       '#E7945A'], dtype=object)
```

In [43]:
```
df.description.dtypes
df['description'] = df['description'].replace({None: 'Not provided'})
df.description
```

Out[43]:
```
0                         Painted red brick wall texture
1                                           Not provided
2            Grand Central during Coronavirus Pandemic.
3                                         Remote working
4                                           Not provided
                             ...
1495                                        Not provided
1496                             reflective water ripples
1497                                        Not provided
1498    Textured blue cement wall background wallpaper.
1499                                        Not provided
Name: description, Length: 1500, dtype: object
```

Downloading Images

Serial Way

In [41]:

```
%%time
images.download_images(quality='regular')
```

```
Found 1500 images in 1 files. Starting to download...
This may take a while.
Downloading [########################################################
####] 1500/1500
Done!
CPU times: user 2.98 s, sys: 959 ms, total: 3.94 s
Wall time: 8.27 s
```

Parallel Way

In [42]:

```
%%time
images.download_images(quality='regular')
```

```
Found 1500 images in 1 files. Starting to download...
This may take a while.
Downloading [########################################################
####] 1500/1500
Done!
CPU times: user 3.09 s, sys: 1.01 s, total: 4.1 s
Wall time: 9.26 s
```

Resizing Images

Serial Way

In [72]:

```
%%time
images.create_thumbnail(size=(128, 128))
```

```
Found 1500 images in 1 files. Starting for processing...
This may take a while.
Processing [#########################################################
###] 1500/1500
Done!
CPU times: user 1min 57s, sys: 10.7 s, total: 2min 8s
Wall time: 2min 41s
```

Conclusion

You have completed your assignment! Now, it is time to share your results and conclusions!

You may need to comment about three things.

1. Your dataset. Explain your EDA findings.
2. Serial and Parallel way differences. What is the difference btw downloading and resizing?
3. Your timing results of both operations in both serial and parallel way.

1. I printed a sample from the dataset and created a histogram with at least 20 bins from the downloads.After words I describe the likes field and did an outlier analysis using a 5 number summary.Then found the average size of the image ratio of the whole dataset using the width and height, then did a unique number of colors of the dataset. I replaces all the None fields in the description field with Not provided text expanded the url field into multiple columns.

2. Serial processing allows only one object at a time to be processed, whereas parallel processing assumes that various objects are processed simultaneously.

3. Serial wall time: 8.27 s parallel way wall time: 9.26 s

```
In [ ]:
```