simpleHashing

March 5, 2023

[11]:
```
import hashlib
import scipy
import matplotlib.pyplot as plt
%matplotlib inline
import time
import numpy as np
```

[2]:
```
def file_hash(filepath):
    with open(filepath, 'rb') as f:
        return md5(f.read()).hexdigest()
```

[3]:
```
import os
```

[4]:
```
os.getcwd()
```

[4]:
```
'/Users/sridhararunachalam/Desktop/MiniProject'
```

[5]:
```
files_list = os.listdir()
print(len(files_list))
```

```
13
```

[6]:
```
import hashlib, os
duplicates = []
hash_keys = dict()
for index, filename in  enumerate(os.listdir('.')):  #listdir('.') = current␣
 ↪directory
    if os.path.isfile(filename):
        with open(filename, 'rb') as f:
            filehash = hashlib.md5(f.read()).hexdigest()
        if filehash not in hash_keys:
            hash_keys[filehash] = index
        else:
            duplicates.append((index,hash_keys[filehash]))
```

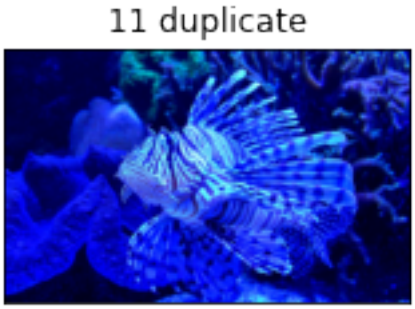
[7]:
```
duplicates
```

[7]:
```
[(10, 1), (11, 7)]
```

[16]:
```
for file_indexes in duplicates[:30]:
    try:
        a = plt.imread(files_list[file_indexes[1]])
        b = plt.imread(files_list[file_indexes[0]])
        plt.subplot(121),plt.imshow(a)
        plt.title(file_indexes[1]), plt.xticks([]), plt.yticks([])

        plt.subplot(122),plt.imshow(b)
        plt.title(str(file_indexes[0]) + ' duplicate'), plt.xticks([]), plt.
 ↪yticks([])
        plt.show()

    except OSError as e:
        continue
```



[19]:
```
for index in duplicates:
    os.remove(files_list[index[0]])
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
```

```
/var/folders/hb/fxb4rq1x5rb303y7zmkhprl00000gn/T/ipykernel_47470/295358519.py in
 ↪<module>
      1 for index in duplicates:
----> 2     os.remove(files_list[index[0]])

FileNotFoundError: [Errno 2] No such file or directory: 'cherryBlossom copy'
```

[ ]:

[ ]: