# google-automl-prep

March 29, 2020

[2]:
```
import os
import pandas as pd
```

[3]:
```
# hardcoded
data_folders = ["dog", "cat"]
data_folders
```

[3]:
```
['dog', 'cat']
```

[4]:
```
# array of arrays, containing the list files, grouped by folder
filenames = [os.listdir(f) for f in data_folders]
[print(f[1]) for f in filenames]
[len(f) for f in filenames]
```

```
dog.1.jpg
cat.1.jpg
```

[4]:
```
[12500, 12500]
```

[5]:
```
files_dict = dict(zip(data_folders, filenames))
```

[6]:
```
base_gcs_path = 'gs://rdc-automl-catsvsdogs/'
```

[7]:
```
# What we want:
# gs://rdc-automl-catsvsdogs/dog/dog.0.jpg, 'dog'
# base_gcs_path + dict_key + '/' + filename

data_array = []

for (dict_key, files_list) in files_dict.items():
    for filename in files_list:
        # print(base_gcs_path + dict_key + '/' + filename)
        if '.jpg' not in filename:
            continue # don't include non-photos

        label = dict_key

        data_array.append((base_gcs_path + dict_key + '/' + filename , label))
```

[9]:
```
dataframe = pd.DataFrame(data_array)
```

[10]:
```
dataframe
```

[10]:

| | 0 | 1 |
|---|---|---|
| 0 | gs://rdc-automl-catsvsdogs/dog/dog.0.jpg | dog |
| 1 | gs://rdc-automl-catsvsdogs/dog/dog.1.jpg | dog |
| 2 | gs://rdc-automl-catsvsdogs/dog/dog.10.jpg | dog |
| 3 | gs://rdc-automl-catsvsdogs/dog/dog.100.jpg | dog |
| 4 | gs://rdc-automl-catsvsdogs/dog/dog.1000.jpg | dog |
| … | … | … |
| 24995 | gs://rdc-automl-catsvsdogs/cat/cat.9995.jpg | cat |
| 24996 | gs://rdc-automl-catsvsdogs/cat/cat.9996.jpg | cat |
| 24997 | gs://rdc-automl-catsvsdogs/cat/cat.9997.jpg | cat |
| 24998 | gs://rdc-automl-catsvsdogs/cat/cat.9998.jpg | cat |
| 24999 | gs://rdc-automl-catsvsdogs/cat/cat.9999.jpg | cat |

[25000 rows x 2 columns]

[11]:
```
dataframe.to_csv('all_data.csv', index=False, header=False)
```