DVC Management with S3

New Dataset

1. Checkout new branch and add new remote url

```
git checkout -b babystroller
dvc remote add -d babystroller s3://vama-sceneuds-images/babystroller
```

1. Add a folder `babystroller` and within it, two folders `images` & `labels`. Add the data into the folders. Create a `babystroller.dvc` & cache the data and hashes with the below command.

```
dvc add babystroller
```

1. Push data to S3. **VERY IMPT!**: do not just use `dvc push` as other data will be uploaded to this folder in S3. Be specific as shown below.

```
dvc remote default babystroller
dvc push babystroller.dvc
```

1. Upload changes to repository.

```
git add *
git add .dvc/config
git commit -m "your commit msg"
git push --set-upstream origin babystroller
```

1. Send a merge request to master

Adding New Data to Existing Dataset

1. Checkout to the branch with the dataset you need, and update changes from master branch.

```
git checkout wheelchair
git merge master
```

1. Pull your dataset from S3 after ensuring the remote url is correct.

```
dvc remote default wheelchair
dvc pull wheelchair.dvc

# add your new data into the downloaded folders, then update the cache
and hashes
dvc add wheelchair
```

1. Push data to S3. **VERY IMPT!**: do not just use `dvc push` as other data will be uploaded to this folder in S3. Be specific as shown below.

```
dvc remote default wheelchair # not required if already set previously
dvc push wheelchair.dvc
```

1. Upload changes to repository.

```
git add *
git add .dvc/config
git commit -m "your commit msg"
git push
```

1. Send a merge request to master

Load from Existing Dataset

1. First, remember to export the AWS access key and secret key
2. Suppose the object of interest you want to download is babystroller
3. Check that the number of images matches what is written in `information.txt`

```
dvc remote default babystroller
dvc pull babystroller.dvc
```