README file for PA2 by Pranav Subramanian

Running the file on EC2 instances involves

- setting up on EC2 instances,
- Running the script
- running with Docker file

Please follow the steps below to setup and run the Python script

Setup on EC2

1. Install spark – Can run the command sudo wget apache http://archive.apache.org/dist/spark/spark-2.0.0/spark-2.0.0-bin-hadoop2.7.tgz

2. Do sudo tar –zxvf spark-2.0.0-bin-hadoop2.7.tgz

3. Install Hadoop –

   run the command sudo tar -zxvf hadoop to unzip tar file if needed

4. Install Python

   Run python 3 get-pip.py

5. Edit bashrc

   Do ls –a and set HADOOP_HOME to be the location of the Hadoop folder

   Modify PATH to include Hadoop Home

   ```
   # User specific environment
   PATH="$HOME/.local/bin:$HOME/bin:$PATH"
   export PATH=$PATH:$HADOOP_HOME/bin/
   export HADOOP_HOME=/home/ec2-user/hadoop-2.7.3
   ```

6. Uploading to cloud (if needed)

   For example, I uploaded the file using the command:

Sudo scip –I "instanconnect.pem"
mnt/c/pranav/CS643_Cloud_Computing/TraininDataset.csv EC2-user@EC2-3-89-57-96.compute-1.amazonaws.com:~/.

7. Run the python script – you may need to edit lines 3 & 4 for os.environ and sys.path.append to your location for spark folder

   Can run python3 PredictionModel.py pathtocsv

8. While running program may see output like this:

9. Run with Docker

   Using the dockerfile can create image by running a command like docker build –t image name .

   Can run the docker file by running a command like docker run imagename

Link to Github

https://github.com/GitDeveloperLite/Models

Link to DockerHUb

https://hub.docker.com/repository/docker/ps789/models

Thank you for taking the time to read the README file for PA2!