Task 3 Group 6 README

Jack Cook

Govardhan Digumurthi

Thomas Okonkwo

Description of dataset

The dataset used is the output from task 1. The name of the csv file is `Group6_Task_1_Output.csv`. The file contains corrected data by cosine similarity.

Work done

The following is a list of work done in order:

The following is a list in order of what happens in this program:

- The spark configuration is setup
- A file path to `Group6_Task_1_Output.csv` is read in from the command line
- Useful re-usable functions for accessing dataframes are defined
- The dataframe is randomly split into 80/20 train/test sets
- Columns of vectors are created that will contain features for the networks:
  - a) all features in the dataset
  - b) Only the feature with the highest correlatin as determined in task 2 (specifically #Bedroom)
- Two random forest regression models are created using the training data
- The two models and the datasets (training and testing) are exported

Instructions to run program

The cluster is logged into using `ssh cookjc@hadoop-nn001.cs.okstate.edu`

The output from `Task 1` is needed, so it is copied onto the cluster

```
scp Group6_Task_1_Output.csv cookjc@hadoop-nn001.cs.okstate.edu:/home/cookjc
```

The file needs to be stored in the hadoop file system:

```
hdfs dfs -copyFromLocal Group6_Task_1_Output.csv /user/cookjc
```

This file must be executed using `spark-submit`:

```
bin/spark-submit Group6_Task_3_Code.py Group6_Task_1_Output.csv
```

Discussion of results

There are 4 total outputs for this task:

- Group6_Task_3_Output_RF_A - a folder containing the random forest model using all the features
- Group6_Task_3_Output_RF_B - a folder containing the random forest model only using the #Bathroom feature
- Group6_Task_3_Output_Test.csv - The testing data (20% of original dataset)
- Group6_Task_3_Output_Train.csv - The training data (80% of the original dataset)