Neeraj Asthana
nasthan2
3/7/2016
Stat 480 Homework 5

This report includes a summary of all the scripts I have written for each exercise, the command line code to run those scripts, and the results. All commands are also summarized in a file called “commands.txt”.

Exercise 1:
Map Script: min_temperature_map.py
Reduce Script: min_temperature_reduce.py
Command to run:

```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
 -files /home/host-data/STAT480DataScience/HW5/min_temperature_map.py,\
/home/host-data/STAT480DataScience/HW5/min_temperature_reduce.py \
-input input/ncdc/all \
-output outputs/HW5Exercise1 \
-mapper "/home/host-data/STAT480DataScience/HW5/min_temperature_map.py" \
-reducer "/home/host-data/STAT480DataScience/HW5/min_temperature_reduce.py"
```

Results File: part-00000Exercise1
Results:

```
1901    -333
1902    -328
1903    -306
1904    -294
1905    -328
1906    -250
1907    -350
1908    -378
1909    -378
1910    -372
```

Exercise 2:
Map Script: num_trusted_map.py
Reduce Script: num_trusted_reduce.py
Command to run:

```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
 -files /home/host-data/STAT480DataScience/HW5/num_trusted_map.py,\
/home/host-data/STAT480DataScience/HW5/num_trusted_reduce.py \
-input input/ncdc/all \
-output outputs/HW5Exercise2 \
-mapper "/home/host-data/STAT480DataScience/HW5/num_trusted_map.py" \
-reducer "/home/host-data/STAT480DataScience/HW5/num_trusted_reduce.py"
```

Results File: part-00000Exercise2
Results:

```
1901    6564
1902    6565
1903    6511
1904    6582
1905    6561
1906    5474
1907    5461
```

```
    1908   6584
    1909   7534
    1910   7645
```

Exercise 3:

Map Script: min_max_trusted_map.py

Reduce Script: min_max_trusted_reduce.py

Command to run:

```
    hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
     -files /home/host-data/STAT480DataScience/HW5/min_max_trusted_map.py,\
    /home/host-data/STAT480DataScience/HW5/min_max_trusted_reduce.py \
    -input input/ncdc/all \
    -output outputs/HW5Exercise3 \
    -mapper "/home/host-data/STAT480DataScience/HW5/min_max_trusted_map.py" \
    -reducer "/home/host-data/STAT480DataScience/HW5/min_max_trusted_reduce.py"
```

Results File: part-00000Exercise3

Results (year, number trusted, minimum temperature, maximum temperature):

```
    1901   6564   -333   317
    1902   6565   -328   244
    1903   6511   -306   289
    1904   6582   -294   256
    1905   6561   -328   283
    1906   5474   -250   294
    1907   5461   -350   283
    1908   6584   -378   289
    1909   7534   -378   278
    1910   7645   -372   294
```

Exercise 4:

Map Script: mean_temperature_map.py

Reduce Script: mean_temperature_reduce.py

Command to run:

```
    hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
     -files /home/host-data/STAT480DataScience/HW5/mean_temperature_map.py,\
    /home/host-data/STAT480DataScience/HW5/mean_temperature_reduce.py \
    -input input/ncdc/all \
    -output outputs/HW5Exercise4 \
    -mapper "/home/host-data/STAT480DataScience/HW5/mean_temperature_map.py" \
    -reducer "/home/host-data/STAT480DataScience/HW5/mean_temperature_reduce.py"
```

Results File: part-00000Exercise4

Results (year, number of observations, mean):

```
    1901   6564   46.6985070079
    1902   6565   21.6595582635
    1903   6511   48.2417447397
    1904   6582   33.3222424795
    1905   6561   43.3322664228
    1906   5474   47.0834855681
    1907   5461   31.7641457608
    1908   6584   28.8365735115
    1909   7534   26.5653039554
    1910   7645   35.5586657946
```